Linear Models

Lecture Notes

Thorsten Dickhaus
University of Bremen
Summer term 2026
Version: 7. April 2026

Preliminary Remarks

The material for this script has partly been compiled during my stand-in professorship at Clausthal University of Technology in the summer term 2011. Further important sources were the script about inferential likelihood theory by Prof. Guido Giani (German Diabetes Center Düsseldorf) and the script about probability theory and statistics by Dr. Wolfgang Meyer, Forschungszentrum Jülich. I am indebted to all my academic teachers for their material and guidance.

Mareile Große Ruse and Konstantin Schildknecht have helped with manuscript preparation.

Exercises and R programs for this course are available from me upon request. At some occasions, I will refer to these in the text.

List of Abbreviations and Symbols

$\nu << \mu$	The measure ν is absolutely continuous with respect to the measure μ.
$B(p,q)$	Beta function, $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$
$\mathcal{B}(\Omega)$	Some σ-field over Ω, often: the system of Borel sets of Ω
$\lceil x \rceil$	Smallest integer larger than or equal to x
χ^2_ν	Chi-square distribution with ν degrees of freedom
$\complement M$	Complement of the set M
δ_a	Dirac measure in the point a
$\stackrel{\mathcal{D}}{=}$	Equality in distribution
F_X	Cumulative distribution function of the real-valued random variable X
$\lfloor x \rfloor$	Largest integer smaller than or equal to x
$\Gamma(\cdot)$	Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt,\ x > 0$
$\mathrm{im}(X)$	Image of X
I_p	Identity matrix in $\mathbb{R}^{p\times p}$
iid.	independent and identically distributed
$\mathbf{1}_M$	Indicator function of the set M
$\inf M$	Infimum of the set M
$L_1(\mu)$	Space of functions that are integrable with respect to the measure μ
$L_2(\mu)$	Space of functions that are square-integrable with respect to the measure μ
λ	Lebesgue measure on $\mathbb{R}$
λ^n	Lebesgue measure on $\mathbb{R}^n$

$\mathcal{L}(X)$	Distribution (law) of the random variable X
LFC	Least favorable configuration
$p(y_i, \theta)$	Likelihood of observational unit i under θ
$Z(y, \theta)$	(Joint) likelihood of the entire sample under θ
$\ell(y_i, \theta)$	Log-Likelihood of observational unit i under θ
$L(y, \theta)$	Log-Likelihood of the entire sample under θ
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with parameters μ and σ^2
Φ	Cumulative distribution function of the $\mathcal{N}(0, 1)$ distribution
$\varphi(\cdot)$	Lebesgue density of the $\mathcal{N}(0, 1)$ distribution
$\dot{\ell}(y_i, \theta)$	Score function at observational unit i under θ
$\dot{L}(y, \theta)$	Score function at the entire sample under θ
$\langle \cdot, \cdot \rangle_{\mathbb{R}^k}$	Standard scalar product in $\mathbb{R}^k$
$\sup M$	Supremum of the set M
supp(F)	Support of the cumulative distribution function F
tr(A)	Trace of the matrix A
$A^\top$	Transpose of the matrix A (analogous for vectors)
UNI$[a, b]$	Uniform distribution on the interval $[a, b]$
$\xrightarrow{w}$	weak convergence

Table of Contents

Chapter 0

Introduction and Examples

Regression analysis deals with the analysis of (systematic) relationships between a (univariate) *response variable* Y and a set of k *explanatory variables (covariates, regressors)* $X_1, \ldots, X_k$. In contrast to the situation in, e. g., classical physics, which deals with deterministic (quantitative) "laws of nature" of the form $y = f(x_1, \ldots, x_k)$ with $\vec{x} = (x_1, \ldots, x_k)$ as "input" and y as "output", statistics assumes random "disturbances" (measurement errors, imprecise measurements, etc). Thus, the output / response Y is treated as a random variable, whose distribution depends on the covariates.

The goal of regression analysis is the investigation of the influence of the explanatory variables on the (conditional) expectation of the response. Hence, we are modelling

$$\mathbb{E}[Y|X_1 = x_1, \ldots, X_k = x_k] = f(x_1, \ldots, x_k).$$

The observables $Y_1, \ldots, Y_n$, which describe a sample of the response, can then be decomposed into a systematic component and a random component:

$$Y_i = \mathbb{E}[Y_i|X_{i,1} = x_{i,1}, \ldots, X_{i,k} = x_{i,k}] + \varepsilon_i = f(x_{i,1}, \ldots, x_{i,k}) + \varepsilon_i$$

with $\mathbb{E}[\varepsilon_i] = 0$, for all $1 \leq i \leq n$. The vector $\vec{x}_i = (x_{i,1}, \ldots, x_{i,k})$ is called "profile of covariates" of the i-th measurement (observational unit) and ε_i is called error term of the i-th measurement (observational unit). (Generalized) *linear* regression models consider the special case where f is a linear function of the values (realizations) of the covariates.

Definition 0.1 ((Generalized) linear model)
Let $\vec{X} := (X_1, \ldots, X_k)$, $\vec{x} := (x_1, \ldots, x_k)$, *and* $\eta := g(\mathbb{E}[Y|\vec{X} = \vec{x}])$.
Model assumption: $\eta = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$. *We call* g *link function,* β_0 *intercept, and* $(\beta_1, \ldots, \beta_k)^\top$ *vector of regression coefficients (the parameters of the model). The covariates* $X_1, \ldots, X_k$ *are also referred to as* independent variables *and* Y *as the* dependent variable.

Scheme 0.2 (Overview of GLMs)

GLM stands for "generalized linear model".

Model	Scale level of Y	Link function g
ANCOVA (MLR)	continuous	id. (components of $\vec{X}$ have continuous scale)
ANOVA	continuous	id. ($\vec{X}$ has categorical components)
log-linear	$Y \in (0, \infty)$	$\eta := \mathbb{E}[\log(Y)\|\vec{X} = \vec{x}]$
Poisson	$Y \in \mathbb{N}_0$	$\log$
logistic	dichotomous	$\text{logit}(x) = \ln(x/(1-x))$
Cox (proportional hazards)	time interval	$\eta = \ln(h(t))$, $h(t)$ hazard function

Table 0.1: Overview of generalized linear regression models

Example 0.3 (Real data)

Rent index data, cf. `R` *program.*

Chapter 1

Basic Notions

1.1 Decision making under uncertainty, statistical models

Two decisive changes when going from *probability theory* to *mathematical statistics* are:

(1) The modelling is typically done on the support (range) of random variables, and not on their domain (which is the underlying "universe" or "population").

(2) Instead of deducing the "correct" distribution of some random variable Y defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we consider a *family of indexed probability measures* $(\mathbb{P}_\theta)_{\theta\in\Theta}$. We try to assess for which values of θ the probability measure $\mathbb{P}_\theta$ describes the (unknown or only partially known) distribution of Y best or good enough, according to certain criteria. In other words, we assess for which values of θ the distribution $\mathbb{P}_\theta$ is "compatible" with realizations y of Y. These realizations are called *observations* or *samples*. The samples are our sources of information, and they constitute the basis for *statistical inference* about θ.

More concretely: In *probability theory*, the fundamental object is the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Random variables are measurable mappings $Y : \Omega \to \mathcal{Y}$. Typically, the task is to compute $\mathcal{L}(Y) \equiv \mathbb{P}^Y = \mathbb{P} \circ Y^{-1}$. This is a probability measure on $\mathcal{Y}$, called the "distribution of Y".

Example 1.1
Consider rolling two fair dice independently from each other. Here, $\Omega = \{1, \dots, 6\}^2$, $\mathcal{A} = 2^\Omega$, *and* $\mathbb{P} = (UNI\{1, \dots, 6\})^{\otimes 2}$. *We call* $(\Omega, \mathcal{A}, \mathbb{P})$ *a Laplacian probability space. Now, let* $Y : \Omega \to \{2, \dots, 12\} = \mathcal{Y}$ *denote the sum of the results of the two dice. Then, for* $j \in \mathcal{Y}$ *we have that*

$$\mathbb{P}^Y(\{j\}) = \mathbb{P}(Y = j) = \mathbb{P}(\{\omega \in \Omega : Y(\omega) = j\}),$$

e. g., $\mathbb{P}^Y(\{7\}) = \mathbb{P}(Y = 7) = \mathbb{P}(\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}) = 6/36 = 1/6.$

In *(inferential) statistics*, the task is different. Namely, we want to draw conclusions (make inference) about $\mathbb{P}$ or $\mathbb{P}^Y$, respectively, only on the basis of observations $Y = y$. For instance, under

the scope of Example 1.1 one could ask the question whether the two dice are indeed "fair" or not. To address this question, it is near at hand to repeat the aforementioned experiment many times and record the results in a tally chart.

From now on, we denote by Y a random variable which describes the possible outcomes of a given experiment.[1] Since we can draw our statistical conclusions about $\mathcal{L}(Y)$ only on the basis of the sample $Y = y$, it is near at hand to consider the support or range of Y now as the fundamental object. Thus, let from now on $\mathcal{Y}$ denote the *sample space* corresponding to Y, i. e., the set of all possible realizations of Y. Furthermore, let $\mathcal{B}(\mathcal{Y}) \subseteq 2^{\mathcal{Y}}$ denote a σ-field over $\mathcal{Y}$. The elements of $\mathcal{B}(\mathcal{Y})$ are called measurable subsets of $\mathcal{Y}$ or *events.*

Let $\mathbb{P}^Y$ denote the distribution of Y. We assume that we have uncertainty about $\mathbb{P}^Y$, but that we are sure that $\mathbb{P}^Y \in \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. We may interpret the value of θ as the unknown and unobservable state of nature. This leads to the following fundamental definition.

Definition 1.2 (Statistical model)

A triple $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathcal{P})$ with $\mathcal{Y} \neq \emptyset$ some non-empty set, $\mathcal{B}(\mathcal{Y}) \subseteq 2^{\mathcal{Y}}$ a σ-field over $\mathcal{Y}$, and $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ a family of probability measures on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is called a statistical model (or: statistical experiment).

If $\Theta \subseteq \mathbb{R}^k$, $k \in \mathbb{N}$, then we call $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathcal{P})$ a parametric statistical model, $\theta \in \Theta$ the parameter, and Θ the parameter space.

Remark 1.3

Although the underlying "universe" (the domain of Y, the "target population" Ω) does not explicitly appear in Definition 1.2, one should nevertheless always be clear about the target population for which our data are representative!

Example 1.4

a) *In a big industrial production process, one is interested in the proportion of defectively produced pieces. To assess this proportion, a sample of size n is randomly drawn from the produced pieces. The number $n \in \mathbb{N}$ is fixed in advance by the management of the company. After the termination of this quality check, it is reported how many of the n checked pieces have been defective.*

$$\mathcal{Y} = \{0, \ldots, n\}, \mathcal{B}(\mathcal{Y}) = 2^{\mathcal{Y}} \text{ (power set)}, (\mathbb{P}_\theta)_{\theta \in \Theta} = (\text{Bin}(n, \theta))_{0 \leq \theta \leq 1}, \Theta = [0, 1] \ni \theta.$$

b) *Assume that the feature "intelligence quotient (IQ)" is normally distributed in a given target population (e. g., among the inhabitants of Bremen). Assume further that some researcher is, for demographic reasons, interested in the expectation and the variance of this normal*

[1] Witting (1985): "We think of all the data material summarized as one "observation" [...]." (translation by the author, the observation will be denoted as $Y = y$).

distribution. To assess these population characteristics, n randomly drawn inhabitants of Bremen perform IQ tests independently from each other under the same, standardized conditions, resulting in n IQ values.

$$\mathcal{Y} = \mathbb{R}^n, \mathcal{B}(\mathcal{Y}) = \mathcal{B}(\mathbb{R}^n), \Theta = \mathbb{R} \times \mathbb{R}_{\geq 0}, \theta = (\mu, \sigma^2), (\mathbb{P}_\theta)_{\theta \in \Theta} = ((\mathcal{N}(\mu, \sigma^2))^{\otimes n})_{(\mu, \sigma^2) \in \Theta}.$$

Critical points:

The IQ can neither become negative nor infinitely large. Moreover, the IQ cannot take every value in an interval, because the underlying computing formula only involves rational numbers.

Hence, our statistical model is here only an approximate description of the true data-generating process. In general, every mathematical model is (only) an abstraction of reality.

c) *In an agricultural research institute, k different varieties of wheat are grown on n plots of land each. One is interested in the average (expected) yield per variety. One is willing to assume that all $(k \times n)$ yield measurements are stochastically independent and that each measurement is normally distributed with a variety-specific expectation μ_i, $1 \leq i \leq k$. One assumes that the measurement variability is only due to the technical setup used, and that it is thus known and the same for all $(k \times n)$ mesurements. In particular, it is assumed that no effect of the plot on the yield exists, or at least that it is of negligible magnitude (as compared to the variety effect).*

$$\mathcal{Y} = \mathbb{R}^{n \cdot k}, \quad \mathcal{B}(\mathcal{Y}) = \mathcal{B}(\mathbb{R}^{n \cdot k}), \quad \Theta = \mathbb{R}^k, \quad \theta = (\mu_1, \ldots, \mu_k)^\top =: \vec{\mu},$$

$$\begin{aligned} (\mathbb{P}_\theta)_{\theta \in \Theta} &= \bigotimes_{i=1}^{n} \mathcal{N}_k(\vec{\mu}, \sigma^2 \cdot I_k), \quad \sigma^2 > 0 \text{ known} \\ &\widehat{=} \quad \mathcal{N}_{n \cdot k} \left[\begin{pmatrix} \vec{\mu} \\ \vdots \\ \vec{\mu} \end{pmatrix}, \sigma^2 I_{n \cdot k} \right]. \end{aligned}$$

In such a situation, the measurements are typically arranged in matrix form.

Statistical inference is concerned with making assertions about the true, but unknown distribution $\mathbb{P}^Y$ or about the true, but unknown value of the parameter θ, respectively. We formalize this as a statistical decision problem.

Definition 1.5

Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta \in \Theta})$ denote a statistical model. A decision rule is a measurable map $\delta : \mathcal{Y} \to (A, \mathcal{A})$. The measurable space $(A, \mathcal{A})$ is called action space. Any function loss $: \Theta \times A \to \mathbb{R}_{\geq 0}$, which is measurable in its second argument, is called a loss function. The tuple $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta \in \Theta}, A, \mathcal{A}, loss)$ is called a statistical decision problem.

The risk of a decision rule δ under θ is the (under θ) expected loss of δ, i. e.,

$$R(\theta, \delta) := \mathbb{E}_\theta\big[loss(\theta, \delta)\big] = \int_{\mathcal{Y}} loss(\theta, \delta(y))\mathbb{P}_\theta(dy).$$

Example 1.6

(a) Point estimation:

Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\theta, 1)^{\otimes n})_{\theta\in\Theta=\mathbb{R}})$.

Assume that our task is to report a real number $\hat{\theta} = \hat{\theta}(y) \in \Theta$ *(on the basis of the data* $Y = y = (y_1, \ldots, y_n)^\top$*), which is "close to the true value of* θ*".*

We formalize this task as a statistical decision problem. To this end, we add to the model $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ *the action space* $(A, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *and the quadratic loss* $loss(\theta, a) = (\theta - a)^2, a \in A = \mathbb{R} = \Theta$*. Consider the special choice* $\hat{\theta}(y) = \bar{y}_n = n^{-1}\sum_{i=1}^n y_i$*. We calculate that*

$$\begin{aligned} R(\theta, \hat{\theta}) &= \mathbb{E}_\theta\big[(\theta - \bar{Y}_n)^2\big] \\ &= \mathbb{E}_\theta\big[\theta^2 - 2\theta\bar{Y}_n + \bar{Y}_n^2\big] \\ &= \theta^2 - 2\theta^2 + (\theta^2 + \frac{1}{n}) = \frac{1}{n}, \end{aligned}$$

because $\mathbb{E}_\theta\big[\bar{Y}_n^2\big] = \big(\mathbb{E}_\theta\big[\bar{Y}_n\big]\big)^2 + Var_\theta\left(\bar{Y}_n\right)$ *and* $Var_\theta\left(\bar{Y}_n\right) = n^{-2}\sum_{i=1}^n Var_\theta\left(Y_i\right) = 1/n$*. With growing sample size* n*, our estimate becomes more and more precise (its quadratic risk decreases with* n*).*

(b) Hypothesis test:

Assume that, under the model from Part (a)*, we want to decide whether the true value of* θ *lies in a given subset* $\Theta_0 \subset \Theta$ *or in* $\Theta_1 := \Theta \setminus \Theta_0$ *(where both* Θ_0 *and* Θ_1 *are non-empty).*

The corresponding decision space is binary, meaning that is consists of exactly two elements: $A = \{a_0, a_1\}$*. W.l.o.g., we can choose* $(A, \mathcal{A}) = (\{0, 1\}, 2^{\{0,1\}})$*. A reasonable loss function is given by*

$$loss(\theta, a) = \ell_1\,\mathbf{1}_{\{a=1,\theta\in\Theta_0\}} + \ell_2\,\mathbf{1}_{\{a=0,\theta\in\Theta_1\}}$$

for non-negative real constants ℓ_1 *und* ℓ_2*. Thus,*

$$R(\theta, \delta) = \begin{cases} \ell_1\mathbb{P}_\theta(\delta(Y) = 1), & \text{if } \theta \in \Theta_0, \\ \ell_2\mathbb{P}_\theta(\delta(Y) = 0), & \text{if } \theta \in \Theta_1. \end{cases}$$

The so-called "type I error probability" is weighted with ℓ_1 *and the so-called "type II error probability" is weighted with* ℓ_2*. It is also possible to choose* $\ell_1 = \ell_1(\theta)$ *and* $\ell_2 = \ell_2(\theta)$ *as functions of the value of the parameter, in order to "punish" severely wrong decisions stronger than less severe ones.*

In order to choose between concurring decision rules for the same problem, (optimality) criteria for decision rules are needed. As the (pointwise) risk depends on the value of the parameter, it can happen that a decision rules which "performs well" locally (i. e., on a certain $\Theta^* \subset \Theta$) exhibits a "bad performance" outside of Θ^*. Two commonly considered global criteria are given by the minimax and the Bayes approach.

Definition 1.7
Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta}, A, \mathcal{A}, loss)$ denote a statistical decision problem. Furthermore, let $\mathcal{M}$ denote a set of (concurring) decision rules, i. e., a set of maps from $\mathcal{Y}$ to $(A, \mathcal{A})$.

a) The decision rule δ_1 is called (uniformly) better than the decision rule δ_2, if $\forall\theta \in \Theta$: $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ holds true and if there exists a $\theta_0 \in \Theta$ with $R(\theta_0, \delta_1) < R(\theta_0, \delta_2)$. A decision rule $\delta^* \in \mathcal{M}$ is called admissible in $\mathcal{M}$, if there exists no better decision rule in $\mathcal{M}$.

b) The rule $\delta^* \in \mathcal{M}$ is called uniformly best decision rule in $\mathcal{M}$, if

$$\forall\theta \in \Theta : \forall\delta \in \mathcal{M} : R(\theta, \delta) \geq R(\theta, \delta^*).$$

c) A decision rule δ^* is called minimax in $\mathcal{M}$, if

$$\sup_{\theta\in\Theta} R(\theta, \delta^*) = \inf_{\delta\in\mathcal{M}} \sup_{\theta\in\Theta} R(\theta, \delta).$$

d) Assume that the parameter space Θ is equipped with a σ-field $\mathcal{F}_\Theta$, that the loss function loss is measurable wrt. both of its arguments, and that $\theta \mapsto \mathbb{P}_\theta(B)$ is measurable for all $B \in \mathcal{B}(\mathcal{Y})$.

Let π be a probability measure on $(\Theta, \mathcal{F}_\Theta)$, which expresses the data analyst's uncertainty about the parameter value before the start of the experiment (prior distribution of ϑ, where ϑ denotes a random variable whose realization equals the parameter value θ). The Bayes risk associated with π of a decision rule $\delta \in \mathcal{M}$ is given by

$$\begin{aligned} R_\pi(\delta) &:= \mathbb{E}_\pi\big[R(\vartheta, \delta)\big] \\ &:= \int_\Theta R(\theta, \delta)\pi(d\theta) \\ &= \int_\Theta \int_{\mathcal{Y}} loss(\theta, \delta(y))\mathbb{P}_\theta(dy)\pi(d\theta). \end{aligned}$$

The decision rule $\delta_\pi \in \mathcal{M}$ is called Bayes rule or Bayes-optimal in $\mathcal{M}$ with respect to π, if

$$R_\pi(\delta_\pi) = \inf_{\delta\in\mathcal{M}} R_\pi(\delta).$$

e) If $\mathbb{P}_\theta$ is absolutely continuous wrt. some probability measure μ for all $\theta \in \Theta$, and if π is absolutely continuous wrt. the probability measure ν with densities $f_{Y|\vartheta=\theta}$ and f_ϑ, respectively,

then we define the posterior distribution of the parameter (denoted by $\mathbb{P}^{\vartheta|Y=y}$) by means of the following ν-density:

$$f_{\vartheta|Y=y}(\theta) = \frac{f_\vartheta(\theta) \cdot f_{Y|\vartheta=\theta}(y)}{\int_\Theta f_\vartheta(t) f_{Y|\vartheta=t}(y)\nu(dt)}$$

(Bayes formula for densities).

f) Assume that we choose a prior distribution from a parametric family. If the posterior distribution necessarily belongs to the same parametric family (with "updated" (hyper-)parameters, depending on the data y), then we call the prior and the statistical model conjugate.

Bayesian inference becomes easy in the presence of conjugacy. For complex models without conjugate priors, the posterior distribution can typically only be evaluated numerically. In the latter case, one often employs so-called Markov Chain Monte Carlo (MCMC) algorithms. Bayesian methods are very popular in practice. We will discuss some examples in Chapter 5.

Theorem 1.8 (Criterion for Bayes optimality)
Let the expectation operator $\mathbb{E}$ refer to the joint distribution of parameter and data. A decision rule δ^* is Bayes-optimal with respect to the chose prior π, if $\delta^*(Y) = \operatorname{argmin}_{a\in A} \mathbb{E}\big[loss(\vartheta, a)|Y\big]$ almost surely, meaning that

$$\mathbb{E}\big[loss(\vartheta, \delta^*(y))|Y = y\big] \leq \mathbb{E}\big[loss(\vartheta, a)|Y = y\big]$$

for all $a \in A$ and for almost all $y \in \mathcal{Y}$.

Proof: The proof is an application of the tower equation for conditional expectations. Namely, let δ be any decision rule. Then,

$$R_\pi(\delta) = \mathbb{E}\big[\mathbb{E}\big[\text{loss}(\vartheta, \delta(Y))|Y\big]\big] \geq \mathbb{E}\big[\mathbb{E}\big[\text{loss}(\vartheta, \delta^*(Y))|Y\big]\big] = R_\pi(\delta^*).$$

■

Corollary 1.9
Consider the point estimation problem $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta\subseteq\mathbb{R}}, \mathbb{R}, \mathcal{B}(\mathbb{R}), loss)$ for a real-valued parameter. Assume that a prior distribution π is chosen, and let the expectation operator $\mathbb{E}$ refer to the resulting joint distribution of parameter and data.

(a) In the case of $loss(\theta, a) = (\theta - a)^2$ (L_2-loss), the conditional expectation $\mathbb{E}\big[\vartheta|Y\big]$ (i. e., the posterior mean) is Bayes-optimal estimator of $\vartheta = \theta$ wrt. π.

(b) In the case of $loss(\theta, a) = |\theta - a|$ (L_1-loss), every posterior median, i. e., every $\hat{\theta}_\pi$ with (almost surely) $\mathbb{P}(\vartheta \leq \hat{\theta}_\pi|Y) \geq \frac{1}{2}$ and $\mathbb{P}(\vartheta \geq \hat{\theta}_\pi|Y) \geq \frac{1}{2}$, is Bayes-optimal estimator of $\vartheta = \theta$ wrt. π.

Proof: The proof is a straightforward application of the L_2-projection property of the (conditional) expectation and the L_1-minimization property of a median, respectively. ■

1.2 Basics of estimation theory

Definition 1.10

Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ denote a statistical model. Let $p \in \mathbb{N}$, let $\varrho(\theta)$ with $\varrho : \Theta \to \mathbb{R}^p$ denote a (derived) parameter, and let loss be a loss function.
The statistical decision problem $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta}, \mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), loss)$ is called (point) estimation problem for $\varrho(\theta)$.
A decision rule $\hat{\varrho} : \mathcal{Y} \to \mathbb{R}^p$ is called estimation rule, the random variable $\hat{\varrho}(Y)$ is called an estimator for $\varrho(\theta)$ and the value $\hat{\varrho}(y) \in \mathbb{R}^p$ is called estimate for $\varrho(\theta)$ given the observation $Y = y$.
The p-vector $b(\hat{\varrho}, \theta) := \mathbb{E}_\theta\big[\hat{\varrho}\big] - \varrho(\theta)$ is called bias of $\hat{\varrho}$ or of $\hat{\varrho}(Y)$, respectively.
The estimator $\hat{\varrho}(Y)$ is called unbiased, if $b(\hat{\varrho}, \theta) = 0$ for all $\theta \in \Theta$.

Lemma 1.11 (Bias-variance decomposition)

Under the assumptions of Definition 1.10, let $p = 1$ and let loss be the quadratic loss, i. e.,

$$loss(\theta, a) = (\varrho(\theta) - a)^2,\ a \in A \subseteq \mathbb{R}^1.$$

(a) The quadratic risk of an estimator $\hat{\varrho}(Y)$ with finite variance can be decomposed as follows:

$$\begin{aligned}\mathbb{E}_\theta\big[loss(\theta, \hat{\varrho})\big] &= \mathbb{E}_\theta^2[\hat{\varrho} - \varrho(\theta)] + Var_\theta\left(\hat{\varrho}\right)\\ &= b^2(\hat{\varrho}, \theta) + Var_\theta\left(\hat{\varrho}\right).\end{aligned}$$

(b) The quadratic risk of an unbiased, square-integrable, real-valued estimator equals its variance.

Proof: Part (b) follows immediately from Part (a). For proving (a), we calculate as follows.

$$\begin{aligned}\mathbb{E}_\theta\big[\text{loss}(\theta, \hat{\varrho})\big] &= \mathbb{E}_\theta\big[(\hat{\varrho} - \varrho(\theta))^2\big]\\ &= \mathbb{E}_\theta\big[(\hat{\varrho})^2 - 2\hat{\varrho}\varrho(\theta) + (\varrho(\theta))^2\big]\\ &= \mathbb{E}_\theta\big[(\hat{\varrho})^2\big] - 2\varrho(\theta)\mathbb{E}_\theta\big[\hat{\varrho}\big] + (\varrho(\theta))^2\\ &= \text{Var}_\theta\left(\hat{\varrho}\right) + \{\mathbb{E}_\theta^2[\hat{\varrho}] - 2\varrho(\theta)\mathbb{E}_\theta\big[\hat{\varrho}\big] + (\varrho(\theta))^2\}\\ &= \text{Var}_\theta\left(\hat{\varrho}\right) + \mathbb{E}_\theta^2[\hat{\varrho} - \varrho(\theta)],\ \text{since Var}_\theta\left(\hat{\varrho}\right) = \mathbb{E}_\theta\big[(\hat{\varrho})^2\big] - \mathbb{E}_\theta^2[\hat{\varrho}].\end{aligned}$$

■

Remark 1.12

As an exercise, one may generalize the statement of Lemma 1.11 to the case of $p > 1$.

Definition 1.13 (Wishful properties of estimators)

Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta}, \mathbb{R}, \mathcal{B}(\mathbb{R}), loss)$ denote an estimation problem, let $\varrho(\theta)$ denote the (derived) parameter of interest, and let $\hat{\varrho}$ be an estimation rule.

(a) The estimator $\hat{\varrho}(Y)$ is called unbiased, if $\mathbb{E}_\theta\left[\hat{\varrho}\right] = \varrho(\theta)$ for all $\theta \in \Theta$.

(b) An unbiased estimator $\hat{\varrho}^*(Y)$ is called efficient (or UMVU), if (for all $\theta \in \Theta$):

$$Var_\theta\left(\hat{\varrho}^*\right) = \inf_{\hat{\varrho}:\hat{\varrho}(Y)\ unbiased} Var_\theta\left(\hat{\varrho}\right).$$

(c) If $n \in \mathbb{N}$ is a sample size and $\mathcal{Y} \subseteq \mathbb{R}^n$, then $\hat{\varrho}(Y) = \hat{\varrho}_n(Y)$ is called consistent or strongly consistent, respectively, if $\hat{\varrho}(Y) \to \varrho(\theta)$ for $n \to \infty$ under θ in probability or almost surely, respectively. For this to make sense, all observables have to be defined on the same probability space; cf. Remark 1.3.

(d) The estimator $\hat{\varrho}(Y)$ is called asymptotically normally distributed (or asymptotically normal for short), if $0 < Var_\theta\left(\hat{\varrho}\right) < \infty$ and

$$\mathcal{L}\left(\frac{\hat{\varrho}(Y) - \mathbb{E}_\theta\left[\hat{\varrho}\right]}{\sqrt{Var_\theta\left(\hat{\varrho}\right)}}\right) \xrightarrow[n\to\infty]{w} \mathcal{N}(0,1) \text{ under } \theta.$$

Definition 1.14

A statistical model $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ is called dominated (by μ), if there exists a σ-finite measure μ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ such that for all $\theta \in \Theta$ the probability measure $\mathbb{P}_\theta$ is absolutely continuous with respect to μ (Notation: $\forall\theta \in \Theta : \mathbb{P}_\theta << \mu$). The μ-density

$$Z(y,\theta) := \frac{d\mathbb{P}_\theta}{d\mu}(y), \theta \in \Theta, y \in \mathcal{Y}$$

of $\mathbb{P}_\theta$ is called likelihood function (evaluated at the observation $Y = y$).
If the sample $Y = (Y_1, \ldots, Y_n)^\top$ consists of stochastically independent random variables $Y_1, \ldots, Y_n$, it holds that

$$Z(y,\theta) = \prod_{i=1}^n p^{(i)}(y_i,\theta),$$

where

$$\forall 1 \le i \le n : p^{(i)}(y_i,\theta) := \frac{dP_\theta^{(i)}}{d\mu_i}(y_i), \theta \in \Theta,$$

denotes the likelihood function of the i-th observational unit and $\mathbb{P}_\theta = \bigotimes_{i=1}^n P_\theta^{(i)}$ is a product measure, dominated by $\mu = \bigotimes_{i=1}^n \mu_i$. (In other words, μ_i denotes the dominating measure for a single observational unit.) Analogously, we let

$$\forall 1 \le i \le n : \ell^{(i)}(y_i,\theta) := \log p^{(i)}(y_i,\theta)$$

denote the log-Likelihood of the i-th observational unit and

$$L(y,\theta) = \log Z(y,\theta) = \sum_{i=1}^n \ell^{(i)}(y_i,\theta)$$

the log-likelihood of the entire sample, if that sample is again constituted of stochastically independent random variables $Y_1, \ldots, Y_n$, which will be the typical setup throughout the remainder.

Remark 1.15

(i) The family of all continuous distributions on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is (by definition) dominated by the n-dimensional Lebesgue measure λ^n. Every statistical model on a countable sample space $\mathcal{Y}$ is dominated by the counting measure. These are essentially the only two dominating measures which are relevant throughout the remainder.

(ii) The notation for (log-)likelihood functions of single observational units and of entire samples, respectively, is very diverse in the literature. The notation introduced in Definition 1.14 is similar to that in the textbook by Spokoiny and Dickhaus (2015).

Definition 1.16

Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ with $\Theta \subseteq \mathbb{R}^k$ be a dominated (by μ) statistical model with log-likelihood function $L(y,\theta)$ (of the entire sample).
If the function $\theta \mapsto L(y,\theta)$ is differentiable in θ_0 for almost all y, we call

$$y \mapsto \frac{\partial}{\partial\theta} L(y,\theta)|_{\theta=\theta_0} =: \dot{L}(\cdot,\theta_0) \quad \underline{\text{score function in } \theta_0},$$

where $\partial/(\partial\theta)$ denotes the gradient operator (vector of partial derivatives).
The $(k \times k)$-matrix

$$I(\theta_0) := \mathbb{E}_{\theta_0}\left[\dot{L}(\cdot,\theta_0)(\dot{L}(\cdot,\theta_0))^\top\right]$$

is called Fisher information in the point θ_0.

Example 1.17

Consider the Gaussian model $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (\mathcal{N}(\mu,\sigma^2))_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_{>0}})$ for a single, real-valued observational unit. The λ-density of $\mathcal{N}(\mu,\sigma^2)$ is given by

$$f_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-\mu)^2}{2\sigma^2}) = p(y,\theta);\ \theta = (\mu,\sigma^2)^\top.$$

We calculate the Fisher information in the point $\theta_0 = (\mu_0, \sigma_0^2)$ as follows:

$$\begin{aligned}
\ell(y,\theta) &= \ln(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{(y-\mu)^2}{2\sigma^2},\\
\frac{\partial\ell(\theta,y)}{\partial\mu} &= \frac{y-\mu}{\sigma^2},\\
\frac{\partial\ell(\theta,y)}{\partial\sigma^2} &= \frac{(y-\mu)^2-\sigma^2}{2\sigma^4} = \frac{(y-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\\
&\Rightarrow \quad \dot{\ell}(y,\theta_0)(\dot{\ell}(y,\theta_0))^\top = \begin{pmatrix} \frac{(y-\mu_0)^2}{\sigma_0^4} & \frac{(y-\mu_0)^3}{2\sigma_0^6} - \frac{(y-\mu)}{2\sigma_0^4} \\ \frac{(y-\mu_0)^3}{2\sigma_0^6} - \frac{(y-\mu_0)}{2\sigma_0^4} & \frac{[(y-\mu_0)^2-\sigma_0^2]^2}{4\sigma_0^8} \end{pmatrix}\\
&\Rightarrow \quad I(\theta_0) = \begin{pmatrix} \sigma_0^{-2} & 0 \\ 0 & \frac{1}{2\sigma_0^4} \end{pmatrix}.
\end{aligned}$$

Lemma 1.18

Let $Y_1, \ldots, Y_n$ denote observable, stochastically independent random variables, which induce statistical models (one per observational unit) with one and the same parameter space $\Theta \subseteq \mathbb{R}^k$ for all of these n (marginal) models. If, under these specifications, the (marginal) Fisher information I_j with respect to the dominating measure μ_j exists for all $1 \leq j \leq n$ on whole Θ, then there exists the joint Fisher information I induced by $Y = (Y_1, \ldots, Y_n)^\top$, and it holds for all $\theta \in \Theta$, that

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta).$$

Proof: The joint log-likelihood function is given by

$$L(y, \theta) = \sum_{j=1}^{n} \ell(y_j, \theta) \quad \text{with respect to } \bigotimes_{j=1}^{n} \mu_j.$$

By assumption, $L(y, \theta)$ is differentiable almost everywhere with score function

$$\dot{L}(y, \theta) = \sum_{j=1}^{n} \dot{\ell}(y_j, \theta).$$

It is an easy exercise to show that the score is centered, meaning that $\mathbb{E}_\theta\big[\dot{\ell}(Y_j, \theta)\big] = 0$ for all $1 \leq j \leq n$. Thus, we can calculate as follows:

$$\begin{aligned}
\mathbb{E}_\theta\big[\dot{L}(Y, \theta)(\dot{L}(Y, \theta))^\top\big] &= \mathbb{E}_\theta\left[\left(\sum_{j=1}^{n} \dot{\ell}(Y_j, \theta)\right)\left(\sum_{j=1}^{n} \dot{\ell}(Y_j, \theta)^\top\right)\right] \\
&= \sum_{k=1}^{n}\sum_{m=1}^{n} \mathbb{E}_\theta\big[\dot{\ell}(Y_k, \theta)(\dot{\ell}(Y_m, \theta))^\top\big] \\
&= \sum_{j=1}^{n} \mathbb{E}_\theta\big[\dot{\ell}(Y_j, \theta)(\dot{\ell}(Y_j, \theta))^\top\big],
\end{aligned}$$

which yields the assertion. ■

Theorem 1.19 (Cramér-Rao bound)

Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta \in \Theta})$ with $\Theta \subseteq \mathbb{R}^k$ for $k \in \mathbb{N}$ be a statistical model, let $\varrho : \Theta \to \mathbb{R}$ be differentiable in $\theta_0 \in \Theta \setminus \partial\Theta$ (i. e., in the interior of Θ), and let $\hat{\varrho}(Y)$ be an unbiased estimator for $\varrho(\theta)$. Assume that $\mathbb{P}_\theta << \mathbb{P}_{\theta_0}$ for all θ in a neighborhood of θ_0.

Furthermore, assume that the model is regular in the sense of Definition 2.7.1 in Spokoiny and Dickhaus (2015). Then it holds:

$$\mathbb{E}_{\theta_0}\big[(\hat{\varrho} - \varrho(\theta_0))^2\big] = \mathit{Var}_{\theta_0}(\hat{\varrho}) \geq \langle I(\theta_0)^{-1}\dot{\varrho}(\theta_0), \dot{\varrho}(\theta_0)\rangle_{\mathbb{R}^k}.$$

Proof: See Sections 2.7.1 and 2.7.2 in Witting (1985) as well as Section 2.7 in Spokoiny and Dickhaus (2015). ■

Example 1.20 (Gaussian shift model)
Let $Y = (Y_1, \ldots, Y_n)^\top$ be distributed as $\mathcal{N}(\mu, \sigma^2)^{\otimes n}$. In this, assume that $\mu \in \mathbb{R}$ is the parameter of interest, while $\sigma^2 > 0$ is a given, known constant.
Let $\hat{\mu}(Y) = \bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$. Then, $\hat{\mu}(Y)$ is an unbiased estimator of μ and it holds that $Var_\mu(\hat{\mu}) = \frac{\sigma^2}{n}$. According to Example 1.17 in connection with Lemma 1.18, it holds that $I(\mu) = \frac{n}{\sigma^2}$. Hence, $\hat{\mu}$ is Cramér-Rao efficient, because $\varrho = id$ here.

Remark 1.21
The Cramér-Rao bound is only sharp in exponential families, see Section 2.7.4 in Spokoiny and Dickhaus (2015).

1.3 Basics of test theory

In this section, we follow up on Example 1.6.(b) and study test problems, which are binary statistical decision problems: Given two disjoint, non-empty subsets $\mathcal{P}_0$ and $\mathcal{P}_1$ of $\mathcal{P} = (\mathbb{P}_\theta)_{\theta \in \Theta}$ with $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$, the goal is to make a decision (on the basis of data) whether $\mathbb{P}^Y$ belongs to $\mathcal{P}_0$ or to $\mathcal{P}_1$. If $\mathcal{P}$ is parametrized by θ in a one-to-one manner, one can equivalently formalize the decision task via θ and subsets Θ_0 and Θ_1 of Θ such that $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$.
Formal description of the test problem:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1 \quad \text{or}$$

$$H_0 : \mathbb{P}^Y \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : \mathbb{P}^Y \in \mathcal{P}_1.$$

We call H_0 and H_1 hypotheses. H_0 is called null hypothesis (or null for short), and H_1 is called alternative hypothesis (alternative for short). Often, one interprets H_0 and H_1 themselves directly as subsets of the parameter space, i. e., $H_0 \cup H_1 = \Theta$ and $H_0 \cap H_1 = \emptyset$. Now, a decision is sought between H_0 und H_1, on the basis of $y \in \mathcal{Y}$. The corresponding decision rule is called a statistical test.

Definition 1.22 (Statistical test)
A (non-randomized) statistical test is a measurable map

$$\phi : (\mathcal{Y}, \mathcal{B}(\mathcal{Y})) \to (\{0,1\}, 2^{\{0,1\}}).$$

Convention:

$$\phi(y) = 1 \iff \textit{Null hypothesis is rejected, decision in favor of } H_1,$$

$$\phi(y) = 0 \iff \textit{Null hypothesis is retained (not rejected).}$$

The set $\{y \in \mathcal{Y} : \phi(y) = 1\}$ is called rejection region (or: critical region) of ϕ, short notation: $\{\phi = 1\}$. The set $\{y \in \mathcal{Y} : \phi(y) = 0\}$ is called retention region of ϕ, short notation: $\{\phi = 0\} = \complement\{\phi = 1\}$.

Problem: Testing implies the possibility of making an error (wrong decision).

Error of the first kind (α-error, type I error): Decision in favor of H_1, although H_0 is true.
Error of the second kind (β-error, type II error): Retention of H_0, although H_1 is true.

In general, it is not possible to minimize the probabilities for both types of errors simultaneously. This is why typically the two types of error are treated asymmetrically as follows:

(i) Bounding the type I error probability by a given upper bound α (called significance level),

(ii) Subject to (i): Minimization of the type II error probability $\Rightarrow$ "optimal" level α test.

It follows that a statistically safeguarded decision (at level α) can only be taken in favor of H_1.

$\Rightarrow$ Mnemonic device: "The statement for which we want to find evidence has to be formulated as alternative H_1 !"

Notation 1.23

(i) The number $\beta_\phi(\theta) = \mathbb{E}_\theta[\phi] = \mathbb{P}_\theta(\phi(Y) = 1) = \int_{\mathcal{Y}} \phi d\mathbb{P}_\theta \in [0,1]$ denotes the rejection probability of a given test ϕ as a function of $\theta \in \Theta$. For $\theta \in \Theta_1$, $\beta_\phi(\theta)$ is called the power of ϕ under θ. For $\theta \in \Theta_0$, $\beta_\phi(\theta)$ is the type I error probability of ϕ under θ.

For fixed, given $\alpha \in (0,1)$, we call

(ii) a test ϕ with $\beta_\phi(\theta) \leq \alpha$ for all $\theta \in H_0$ a level α test,

(iii) a level α test ϕ unbiased, if $\beta_\phi(\theta) \geq \alpha$ for all $\theta \in H_1$.

(iv) the level α test ϕ_1 better than the level α test ϕ_2, if $\beta_{\phi_1}(\theta) \geq \beta_{\phi_2}(\theta)$ for all $\theta \in H_1$ and $\exists \theta^* \in H_1$ with $\beta_{\phi_1}(\theta^*) > \beta_{\phi_2}(\theta^*)$.

Throughout the remainder, we typically consider the class $\mathcal{M}$ of all level α tests, together with the risk function $R(\theta, \phi) = 1 - \beta_\phi(\theta)$ for $\theta \in \Theta_1$. Under these premisses, the test problem (regarded as a statistical decision problem) is already completely specified by the tuple $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta \in \Theta}, H_0)$.

Definition 1.24 (p-value)
Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model, and let ϕ be a test for the pair of hypotheses $\emptyset \neq H_0 \subset \Theta$ versus $H_1 = \Theta \setminus H_0$. Assume that ϕ is based on a test statistic $T : \mathcal{Y} \to \mathbb{R}$. Namely, assume that ϕ is characterized by rejection regions $\Gamma_\alpha \subset \mathbb{R}$ for each significance level $\alpha \in (0,1)$, such that $\phi(y) = 1 \iff T(y) \in \Gamma_\alpha$, for $y \in \mathcal{Y}$. Under these specifications, the p-value of a realization $y \in \mathcal{Y}$ with respect to ϕ is defined by

$$p_\phi(y) = \inf_{\{\alpha : T(y) \in \Gamma_\alpha\}} \mathbb{P}^*(T(Y) \in \Gamma_\alpha),$$

where the probability measure $\mathbb{P}^*$ is chosen such, that

$$\mathbb{P}^*\left(T(Y) \in \Gamma_\alpha\right) = \sup_{\theta \in H_0} \mathbb{P}_\theta(T(Y) \in \Gamma_\alpha)$$

holds true, if H_0 is a composite null hypothesis.

Remark 1.25

(i) If H_0 is one-elementary (i. e., "simple"), and $\mathbb{P}_{H_0} \equiv \mathbb{P}_{\theta_0}$ is a continuous probability measure, then it (typically) holds that

$$p_\phi(y) = \inf\{\alpha : T(y) \in \Gamma_\alpha\}.$$

(ii) p-values are often referred to as "observed significance levels".

(iii) Let Ω denote the domain of Y (corresponding to the target population; cf. Remark 1.3). The map $p_\phi(Y) : \Omega \to [0,1], \omega \mapsto p_\phi(Y(\omega))$, can be interpreted as a random variable. However, this map is typically denoted with a lower-case letter, to avoid confusion with (indexed) probability measures. Therefore, one has to deduce from the context whether $p_\phi \equiv p$ denotes a realized value (number) in $[0,1]$ or a random variable.

Definition 1.26

Under the assumptions of Definition 1.24, assume that $T(Y)$ is such, that the monotonicity condition

$$\forall \theta_0 \in H_0 : \forall \theta_1 \in H_1 : \forall c \in \mathbb{R} : \mathbb{P}_{\theta_0}(T(Y) > c) \leq \mathbb{P}_{\theta_1}(T(Y) > c) \tag{1.1}$$

is fulfilled. Then, we call ϕ a test of (generalized) Neyman-Pearson type, if there exists for each $\alpha \in (0,1)$ a constant c_α, such that

$$\phi(y) = \begin{cases} 1, & T(y) > c_\alpha, \\ 0, & T(y) \leq c_\alpha. \end{cases}$$

Remark 1.27

(a) The monotonicity condition (1.1) *has the interpretation that "the test statistic tends to larger values under the alternative" (as compared to the null).*

(b) The rejection regions pertaining to a test of Neyman-Pearson (N-P) type are given by $\Gamma_\alpha = (c_\alpha, \infty)$.

(c) In practice, the constants c_α are determined via the level condition $c_\alpha = \inf\{c \in \mathbb{R} : \mathbb{P}^*(T(Y) > c) \leq \alpha\}$, with $\mathbb{P}^*$ as in Definition 1.24 ("at the boundary of the null"). If H_0 is one-elementary and $\mathbb{P}_{H_0}$ is continuous, we have that $c_\alpha = F_T^{-1}(1-\alpha)$, where F_T denotes the cdf of $T(Y)$ under H_0.

(d) The fundamental lemma of test theory by Neyman und Pearson implies that uniformly (over all $\theta_1 \in H_1$) most powerful (non-randomized) level α tests for H_0 versus H_1 are necessarily of N-P type, if they exist.

Lemma 1.28

Let ϕ be a test of N-P type, and assume that $\mathbb{P}^*$ does not depend on α. Then, the p-value of a realization $y \in \mathcal{Y}$ with respect to ϕ can be computed as follows:

$$p_\phi(y) = \mathbb{P}^*(T(Y) \geq t^*) \text{ with } t^* := T(y).$$

Proof: The rejection regions $\Gamma_\alpha = (c_\alpha, \infty)$ are nested. Thus, $\inf\{\alpha : T(y) \in \Gamma_\alpha\}$ is attained in $[t^*, \infty)$. Due to the structure of this rejection region, it holds that $\mathbb{P}^*(T(Y) \in [t^*, \infty)) = \mathbb{P}^*(T(Y) \geq t^*)$. ■

Corollary 1.29

If H_0 is one-elementary, $\mathbb{P}_{H_0}$ is continuous, and ϕ is of N-P type, it holds for all $y \in \mathcal{Y}$ that $p_\phi(y) = 1 - F_T(t^*)$, with notation as in Remark 1.27 and Lemma 1.28.

Theorem 1.30 (Testing in terms of the p-value)

*Let $\alpha \in (0, 1)$ be a **fixed, pre-defined** significance level, and assume that $\mathbb{P}^*$ is continuous. Then, the duality*

$$\phi(y) = 1 \Longleftrightarrow p_\phi(y) < \alpha \tag{1.2}$$

holds true.

Proof: We only prove the assertion for tests of N-P type.

The function $t \mapsto \mathbb{P}^*(T(Y) > t)$ is monotonically decreasing in t. Furthermore, by construction of c_α (see Remark 1.27.c), we have that $\mathbb{P}^*(T(Y) > c_\alpha) \leq \alpha$ as well as $\mathbb{P}^*(T(Y) > c) > \alpha$ for all real constants $c < c_\alpha$. Thus, $p_\phi(y) < \alpha$ is equivalent to $t^* > c_\alpha$. The latter event characterizes the rejection of H_0, if ϕ is a test of N-P type. ■

Remark 1.31

(i) The advantage of p-values for testing is, that they can be computed without specifying a significance level. This is the reason why essentially all statistics software systems implement statistical hypothesis tests via the computation of p-values. However, this also implies the possibility of "cheating". Namely, if one violates the good statistical practice to specify all modalities of the experiment (including the specification of the significance level!) before collecting the data, one may be tempted to specify α a posteriori (after having carried out the experiement and having looked at the resulting p-value), in order to arrive at an a priori intended conclusion. This is why many statisticians refuse to formalize the test ϕ by means of (1.2).

(ii) The interpretation of the p-value is a subtle issue. The p-value essentially answers the question "How probable are the measured data, given that the null is true?". It does not answer the question "How probable is the validity of the null, given the measured data?", although the latter question may be more interesting for practitioners.

1.4 Confidence estimation and the correspondence theorem

There exist dualities between test problems (or tests) and confidence estimation problems (or confidence regions).

Definition 1.32
Assume that a statistical model $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathcal{P} = \{P_\theta : \theta \in \Theta\})$ *is given. Then, we call* $\mathcal{C} = (C(y) : y \in \mathcal{Y})$ *with* $C(y) \subseteq \Theta$ *for all* $y \in \mathcal{Y}$ *a family of confidence regions at confidence level* $1 - \alpha$ *for* $\theta \in \Theta$*, if* $\mathbb{P}_\theta(\{y : C(y) \ni \theta\}) \geq 1 - \alpha$ *holds true for all* $\theta \in \Theta$*.*

Theorem 1.33 (Correspondence theorem; see, e. g., Pages 162-163 of Lehmann and Romano (2005))

(a) Assume that, for each $\theta \in \Theta$*, a level* α *test* ϕ_θ *for the simple null hypothesis* $H_0 = \{\theta\}$ *is available. Define* $\phi = (\phi_\theta, \theta \in \Theta)$*. Then, the family* $\mathcal{C} \equiv \mathcal{C}(\phi)$*, defined by* $C(y) = \{\theta \in \Theta : \phi_\theta(y) = 0\}$*, constitutes a family of confidence regions at confidence level* $1 - \alpha$ *for* $\theta \in \Theta$*.*

(b) If $\mathcal{C}$ *constitutes a family of confidence regions at confidence level* $1 - \alpha$ *for* $\theta \in \Theta$*, we can define the family* $\phi = (\phi_\theta, \theta \in \Theta)$ *of point hypothesis tests via* $\phi_\theta(y) = 1 - \mathbf{1}_{C(y)}(\theta)$*. The so-defined multiple test* ϕ *is a multiple test at local level* α*, meaning that each* ϕ_θ *is a level* α *test for the simple null hypothesis* $H_0 = \{\theta\}$*.*

Proof:
Both in Part (a) and in Part (b), we have that $\forall \theta \in \Theta : \forall y \in \mathcal{Y} : \phi_\theta(y) = 0 \Longleftrightarrow \theta \in C(y)$. Thus, ϕ is a multiple test at local level α, if and only if

$$\begin{aligned} & \forall\, \theta \in \Theta : \quad \mathbb{P}_\theta(\{\phi_\theta = 0\}) \geq 1 - \alpha \\ \Leftrightarrow \quad & \forall\, \theta \in \Theta : \quad \mathbb{P}_\theta(\{y : C(y) \ni \theta\}) \geq 1 - \alpha \\ \Leftrightarrow \quad & \mathcal{C} \text{ constitutes a family of confidence regions at confidence level } 1 - \alpha. \end{aligned}$$

■

Remark 1.34

(a) Figure 1.1 illustrates the duality $\phi_\theta(y) = 0 \Leftrightarrow \theta \in C(y)$ for the special case that both $\mathcal{Y}$ and Θ are one-dimensional.

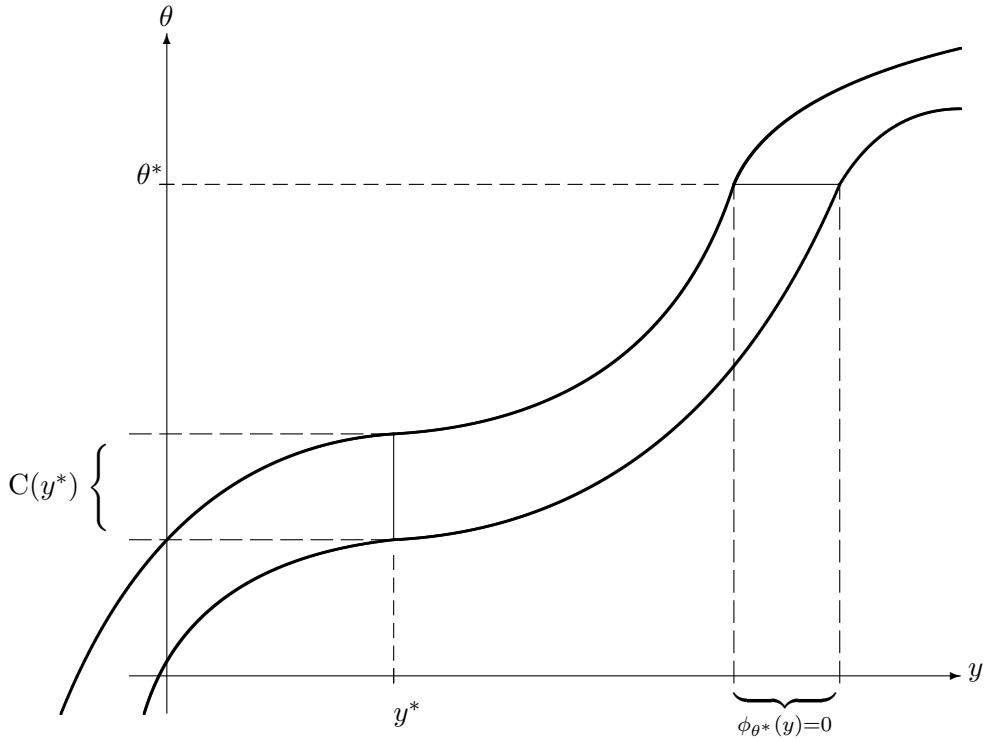


Figure 1.1: Illustration of the duality $\phi_\theta(y) = 0 \Leftrightarrow \theta \in C(y)$

(b) A single level α test ϕ for an arbitrary null hypothesis H can also be interpreted as a $(1-\alpha)$-confidence region. To this end, let

$$C(y) = \begin{cases} \Theta\,, & \text{if } \phi(y) = 0, \\ K = \Theta \setminus H\,, & \text{if } \phi(y) = 1. \end{cases}$$

One may compare this with the mnemonic device mentioned directly before Notation 1.23.

Conversely, a single confidence region $C(y)$ defines a level α test for a given null hypothesis $H \subset \Theta$. To this end, define $\phi(y) = \mathbf{1}_K(C(y))$, where $K = \Theta \setminus H$ and

$$\mathbf{1}_B(A) := \begin{cases} 1\,, & \text{if } A \subseteq B, \\ 0\,, & \text{otherwise,} \end{cases}$$

for arbitrary sets A and B.

Example 1.35

Assume that, under the Gaussian shift model $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\mu, \sigma^2)^{\otimes n})_{\mu \in \mathbb{R} = \Theta})$ mit known variance $\sigma^2 > 0$, we search for a subset of the real line with minimum length (i. e., Lebesgue

measure) amongst all subsets of $\mathbb{R}$ which cover the unknown parameter μ with a probability of at least $(1-\alpha)$ and only depend on the ($\mathbb{R}^n$-valued) data.
Solution: The statistic $\bar{Y}_n$ is sufficient for μ, meaning that it reflects all information that Y contains about μ. The distribution (under μ) of $\sqrt{n}(\bar{Y}_n-\mu)/\sigma$ is $\mathcal{N}(0,1)$. Hence, $\bar{Y}_n$ is under μ symmetrically distributed around the center μ, with exponentially declining distributional mass on both sides of that center. This implies that an optimal confidence region for μ needs to be of the form

$$C(y) = [\hat{\mu} - k(y), \hat{\mu} + k(y)]$$

with $\hat{\mu} \equiv \hat{\mu}(y) = \bar{y}_n$ and a suitable constant $k(y)$.
We compute $k(y)$ via the level condition as follows:

$$\begin{aligned}
& \mathbb{P}_\mu([\bar{Y}_n - k, \bar{Y}_n + k] \ni \mu) \stackrel{!}{=} 1-\alpha \\
\Leftrightarrow \quad & \mathbb{P}_\mu(\bar{Y}_n - k \le \mu \le \bar{Y}_n + k) = 1-\alpha \\
\Leftrightarrow \quad & \mathbb{P}_\mu(\sqrt{n}\frac{k}{\sigma} \ge \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \ge -\sqrt{n}\frac{k}{\sigma}) = 1-\alpha \\
\Leftrightarrow \quad & \mathbb{P}_\mu(-\sqrt{n}\frac{k}{\sigma} \le Z \le \sqrt{n}\frac{k}{\sigma}) = 1-\alpha, \text{ where } Z \sim \mathcal{N}(0,1) \\
\Leftrightarrow \quad & \Phi(\sqrt{n}\frac{k}{\sigma}) - \Phi(-\sqrt{n}\frac{k}{\sigma}) = 1-\alpha \\
\Leftrightarrow \quad & 2\Phi(\sqrt{n}\frac{k}{\sigma}) - 1 = 1-\alpha \\
\Leftrightarrow \quad & \Phi(\sqrt{n}\frac{k}{\sigma}) = 1 - \frac{\alpha}{2} \;\Leftrightarrow\; \sqrt{n}\frac{k}{\sigma} = z_{1-\alpha/2} \Leftrightarrow k = \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \\
\Rightarrow \quad & C(y) = \left[\bar{y}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \bar{y}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right],
\end{aligned}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ denotes the $1-\alpha/2$-quantile of $\mathcal{N}(0,1)$.

Remark 1.36

a) If $\sigma^2 > 0$ is unknown, we can apply the correspondence theorem (Theorem 1.33) to the t-Test (cf. Page 201 in Witting (1985), Test No. 2). The resulting $(1-\alpha)$-confidence interval for μ is given by

$$C(y) = \left[\bar{y}_n - \frac{\hat{\sigma}(y)}{\sqrt{n}} t_{n-1,1-\alpha/2},\ \bar{y}_n + \frac{\hat{\sigma}(y)}{\sqrt{n}} t_{n-1,1-\alpha/2}\right].$$

b) The calculations in Example 1.35 do not depend on the concrete structure of $\bar{Y}_n$, but only on the fact that $\sqrt{n}(\bar{Y}_n-\mu)/\sigma$ id (under μ) standard normally distributed. Hence, it can be carried out analogously for other models with normally distributed sufficient statistics.

c) For $\alpha = 5\%$, we get that $z_{1-\alpha/2} = \Phi^{-1}(0.975) \approx 1.96 \approx 2$. This is why the construction in Example 1.35 is often referred to as 2σ-rule.

In Definition 1.13, we have listed asymptotic normality as one of the wishful properties of a point estimator. A slight generalization of Example 1.35 reveals why this property is indeed wishful. Namely, as demonstrated in Theorem 1.37, asymptotic normality facilitates the construction of confidence regions.

Theorem 1.37
Let $(\mathcal{Y}^n, \mathcal{B}(\mathcal{Y})^{\otimes n}, (\mathbb{P}_\theta^{\otimes n})_{\theta\in\Theta\subseteq\mathbb{R}^k})$ denote a product model with suitable regularity properties, and assume that $\hat{\theta}_n(Y)$ is a $\sqrt{n}$-normal point estimator for $\theta \in \mathbb{R}^k$ in the sense that $\sqrt{n}(\hat{\theta}_n(Y)-\theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1}(\theta_0))$ under θ_0. Let $\varrho : \Theta \to \mathbb{R}$ be a continuously differentiable map with gradient $\dot{\varrho}(\theta) \neq 0$.
Then it holds:

$$\sqrt{n}\left\{\varrho(\hat{\theta}_n(Y)) - \varrho(\theta_0)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_{\theta_0}^2) \text{ under } \mathbb{P}_{\theta_0} \text{ with } \sigma_{\theta_0}^2 = \dot{\varrho}(\theta_0) I^{-1}(\theta_0) \dot{\varrho}(\theta_0)^\top.$$

If the Fisher information I is continuous in θ, a confidence interval for $\varrho(\theta_0)$ with asymptotic coverage probability $1-\alpha$ is given by

$$C(y) = \left[\varrho(\hat{\theta}_n(y)) \pm z_{1-\alpha/2}\, \hat{\sigma}_n/\sqrt{n}\right],$$

where

$$\hat{\sigma}_n^2 := \dot{\varrho}(\hat{\theta}_n(y)) I^{-1}(\hat{\theta}_n(y)) \left[\dot{\varrho}(\hat{\theta}_n(y))\right]^\top.$$

Proof: See Section 12.4.2 in Lehmann and Romano (2005). Notice that the gradient is written as a row vector here! ■

1.5 Inferential likelihood theory

Definition 1.38 (Maximum likelihood estimator (MLE))
Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a dominated statistical model with likelihood function $Z(y,\theta)$. Assume that the parameter space Θ is equipped with the σ-field $\mathcal{F}_\Theta$. A statistic $\hat{\theta}(Y)$ with $\hat{\theta} : (\mathcal{Y}, \mathcal{B}(\mathcal{Y})) \to (\Theta, \mathcal{F}_\Theta)$ is called maximum likelihood estimator (MLE) of θ, if

$$Z(y, \hat{\theta}(y)) = \sup_{t\in\Theta} Z(y,t)$$

holds true for all $\theta \in \Theta$ and for $\mathbb{P}_\theta$-almost all $y \in \mathcal{Y}$.

Remark 1.39

(a) Neither existence nor uniqueness of the MLE are guaranteed without further model assumptions.

(b) In the case of a re-parametrization (bijection) $\theta \mapsto \varrho(\theta)$, it holds that $\hat{\varrho}(Y) := \varrho(\hat{\theta}(Y))$ is the MLE for $\varrho(\theta)$, if the unique MLE $\hat{\theta}(Y)$ for θ exists.

Example 1.40

(a) Assume that $Y_1, \ldots, Y_n$ are i.i.d. with $Y_1 \sim$ Poisson(θ), and let $Y := (Y_1, \ldots, Y_n)^\top$ with values in $\mathbb{N}_0^n$. Assume that the value of the parameter $\theta > 0$ is unknown. We have that

$$Z(y, \theta) = \prod_{i=1}^{n} \exp(-\theta) \frac{\theta^{y_i}}{y_i!}$$

Noticing that $0! = 0$ by definition, we have that

$$\begin{aligned} L(y,\theta) &= \sum_{i=1}^{n} \{-\theta + y_i \ln(\theta) - \ln(y_i!)\} \\ &= -n\theta + \ln(\theta) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!) \\ \Rightarrow \frac{\partial}{\partial\theta} L(y,\theta) &= \dot{L}(y,\theta) = -n + \theta^{-1} \sum_{i=1}^{n} y_i \\ \Rightarrow \hat{\theta}(y) &= n^{-1} \sum_{i=1}^{n} y_i, \text{ since } \frac{\partial^2}{\partial\theta^2} L(y,\theta) < 0. \end{aligned}$$

Here, the MLE exists uniquely in Θ, if there exists an $i \in \{1, \ldots, n\}$ with $y_i > 0$.

(b) General regression model

Let $Y = (Y_1, \ldots, Y_n)^\top$ with values in $\mathbb{R}^n$. Assume for each $1 \leq i \leq n$ the representation $Y_i = g_\theta(x_i) + \varepsilon_i$, where $(x_i)_{1 \leq i \leq n}$ are deterministic, fixed "measurement locations", g_θ is a deterministic, real-valued function parametrized by $\theta \in \Theta \subseteq \mathbb{R}^k$ for $k \in \mathbb{N}$, and $(\varepsilon_i)_{1 \leq i \leq n}$ are random i.i.d. "measurement errors", fulfilling that $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$.

These model assumptions imply that $\forall 1 \leq i \leq n : Y_i \sim \mathcal{N}(g_\theta(x_i), \sigma^2)$ and $Y_i \perp\!\!\!\perp Y_j$ for all $1 \leq i \neq j \leq n$.

It is a worthwhile exercise to show that, under the above specifications,

$$\hat{\theta}(Y) = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - g_\theta(x_i))^2 \right\}.$$

Hence, under this model the MLE for θ coincides with the minimizer of the sum of squares of residuals, which is referred to as the least squares estimator for θ.

Theorem 1.41 (Asymptotics of the MLE in i.i.d. models)
Let $(\mathcal{Y}^n, \mathcal{B}(\mathcal{Y})^n, (P_\theta^{\otimes n})_{\theta\in\Theta})_{n\geq 1}$ with $\Theta \subseteq \mathbb{R}^k$ be a sequence of dominated (by $\mu^{\otimes n}$) product models with log-likelihood function $\ell(y_1, \theta) = \log(\frac{dP_\theta}{d\mu}(y_1))$ of a single observable, where we denote $\mathbb{P}_\theta = P_\theta^{\otimes n}$.
Let the following assumptions be fulfilled:

(a) Θ is compact and θ_0 is an interior point of Θ.

(b) $\forall \theta \neq \theta_0 : \mathbb{P}_\theta \neq \mathbb{P}_{\theta_0}$ (Identifiability)

(c) The function $\theta \mapsto \ell(y, \theta)$ is continuous on Θ and twice continuously differentiable in a neighborhood $\mathcal{U}$ of θ_0 for all $y \in \mathcal{Y}$.

(d) There exist functions $H_0, H_2 \in L_1(P_{\theta_0})$ and $H_1 \in L_2(P_{\theta_0})$ with

$$\sup_{\theta\in\Theta} |\ell(y, \theta)| \leq H_0(y) \text{ as well as } \sup_{\theta\in\mathcal{U}} \left| \frac{\partial^i}{\partial\theta^i} \ell(y, \theta) \right| \leq H_i(y),\ i = 1, 2,$$

for all $y \in \mathcal{Y}$.

(e) The Fisher information (pertaining to a single observable), given by

$$I(\theta_0) = \mathbb{E}_{\theta_0}\big[\dot{\ell}(\cdot, \theta_0)(\dot{\ell}(\cdot, \theta_0))^\top\big],$$

is positive definite.

Let $Y = (Y_1, \ldots, Y_n)^\top$ with values in $\mathcal{Y}^n$. Then, the MLE $\hat{\theta}_n(Y)$ is asymptotically normally distributed and asymptotically Cramér-Rao-efficient under $Y \sim \mathbb{P}_{\theta_0}$, meaning that

$$\sqrt{n}(\hat{\theta}_n(Y) - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)^{-1}) \text{ under } \theta_0 \text{ for } n \to \infty.$$

Proof: See Section 6.5 in Lehmann and Casella (1998). ■

Corollary 1.42
Under the assumptions of Theorem 1.41, $\hat{\theta}_n(Y)$ estimates θ consistently.

Remark 1.43
As we will see in later chapters, analogous results also exist for product models with unequal factors.

Definition 1.44 (Likelihood ratio test)
Let $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a dominated statistical model with likelihood function $Z(y, \theta)$. Assume a test problem, given by $H_0 = \Theta_0$ versus $H_1 = \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_i \neq \emptyset$, $i = 0, 1$. We call

$$\Lambda : \mathcal{Y} \to [1, \infty],\ \Lambda(y) := \frac{\sup_{\theta\in\Theta} Z(y, \theta)}{\sup_{\tilde{\theta}\in\Theta_0} Z(y, \tilde{\theta})}$$

the likelihood ratio statistic for testing H_0 versus H_1, and we call a test of the form

$$\phi(y) = \begin{cases} 1, & \text{if } \Lambda(y) > k, \\ 0, & \text{if } \Lambda(y) < k, \\ \gamma(y), & \text{if } \Lambda(y) = k, \end{cases}$$

a likelihood ratio (LR) test. In this, $k \geq 1$ is the critical value of the test and $\gamma(y) \in [0,1]$ is a randomization constant.

Remark 1.45

If $\hat{\theta}$ and $\hat{\theta}_0$, respectively, are maximum likelihood estimators for θ, where θ is allowed to vary in the full parameter space Θ and $\hat{\theta}_0$ is only allowed to vary in the restricted parameter space Θ_0, we have that

$$\Lambda(y) = \frac{Z(y, \hat{\theta}(y))}{Z(y, \hat{\theta}_0(y))}.$$

Theorem 1.46 (Asymptotics of LR tests in i.i.d. models)

Assume that the product model $(\mathcal{Y}^n, \mathcal{B}(\mathcal{Y})^n, (P_\theta^{\otimes n})_{\theta \in \Theta})$ satisfies that assumptions of Theorem 1.41 regarding the asymptotics of the MLE with log-likelihood function $\ell(\cdot, \theta)$. Assume further that the subset Θ_0 pertaining to the null hypothesis of interest lies in an r-dimensional subspace of $\Theta \subseteq \mathbb{R}^k$ with $0 \leq r < k$, where $r = 0$ refers to testing a point null hypothesis $\Theta_0 = \{\theta_0\}$. Then it holds that

$$2\log(\Lambda_n(Y)) = 2\left[\sup_{\theta \in \Theta} \sum_{i=1}^n \ell(Y_i, \theta) - \sup_{\tilde{\theta} \in \Theta_0} \sum_{i=1}^n \ell(Y_i, \tilde{\theta})\right] \xrightarrow{\mathcal{D}} \chi^2_{k-r}$$

whenever $Y = (Y_1, \ldots, Y_n)^\top$ is distributed according to $\mathbb{P}_{\theta_0} = P_{\theta_0}^{\otimes n}$ with $\theta_0 \in \Theta_0 \cap [\Theta \setminus \partial\Theta]$. In particular, the likelihood ratio test ϕ, given by

$$\phi(y) = \mathbf{1}_{\{\log(\Lambda_n(y)) > \chi^2_{(k-r);(1-\alpha)}/2\}}$$

with $\chi^2_{(k-r);(1-\alpha)}$ denoting the $(1-\alpha)$-quantile of χ^2_{k-r}, asymptotically has level $\alpha \in (0,1)$ on the set $\Theta_0 \cap [\Theta \setminus \partial\Theta]$.

Example 1.47 (Multinomial distributions)

Consider a sequence of n independent, homogeneous trials with k possible outcomes (each). Assume that the probability for outcome j with $1 \leq j \leq k-1$ in a single trial equals p_j, and define moreover $p_k := 1 - \sum_{j=1}^{k-1} p_j$.

Let N_j for $1 \leq j \leq k$ denote the random variable which counts the number of trials with outcome j. Then, the random vector $N = (N_1, \ldots, N_k)^\top$ is multinomially distributed with parameters n (total number of trials), k (total number of categories or classes, respectively), and $p = (p_1, \ldots, p_{k-1})^\top$ (vector of class probabilities). Treating n and k as fixed, given constants, we have that $\dim(\Theta) = k - 1$.

More precisely, we consider the parameter space

$$\Theta = \{(p_1, \ldots, p_{k-1})^\top \in [0,1]^{k-1} : \sum_{j=1}^{k-1} p_j \leq 1\}.$$

The likelihood statistic pertaining to this model is given by

$$Z(N,p) = \frac{n!}{\prod_{j=1}^k N_j!} \prod_{\ell=1}^k p_\ell^{N_\ell},$$

and the MLE for p is given by $\hat{p}_j = N_j/n$ for $1 \leq j \leq k-1$. It is canonical to define $\hat{p}_k = 1 - \sum_{j=1}^{k-1} \hat{p}_j$.

Now, assume that the value of p in unknown, and consider the point null hypothesis $\Theta_0 = \{\pi\}$ for a fixed, given vector $\pi \in \Theta$. This leads to the (log-)likelihood ratio statistics

$$\Lambda_n(N) = \frac{Z(N,\hat{p})}{Z(N,\pi)} \text{ and } \log(\Lambda_n(N)) = n \sum_{j=1}^k \hat{p}_j \log\left(\frac{\hat{p}_j}{\pi_j}\right),$$

where $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$. According to Theorem 1.46, it holds that $2\log(\Lambda_n(N)) \xrightarrow{\mathcal{D}} \chi^2_{k-1}$ as n tends to infinity.

In order to carry out the resulting asymptotic χ^2-test in practice, the following considerations can be helpful. Consider the function h, given by $h(x) = x\log(x/x_0)$ for a fixed, given real number $x_0 \in (0,1)$. Then, the Taylor expansion of h about x_0 is given by

$$h(x) = (x - x_0) + \frac{1}{2x_0}(x - x_0)^2 + o[(x - x_0)^2] \text{ for } x \to x_0.$$

Thus, for $\hat{p}$ "close to" π, we get that

$$2\log(\Lambda_n(N)) \approx Q_n \text{ with } Q_n = \sum_{j=1}^k \frac{(N_j - n\pi_j)^2}{n\pi_j}.$$

The statistic Q_n is called Pearson's chi-square statistic. More precisely, it holds that

$$2\log(\Lambda_n(N)) - Q_n \to 0 \text{ in probability}$$

under the null hypothesis, implying that Q_n is asymptotically χ^2_{k-1}-distributed under $p = \pi$. In practice, it is often more convenient to compute the value of Q_n than to compute the value of $2\log(\Lambda_n(N))$.

Remark 1.48

Asymptotic χ^2-tests can be generalized to models which result in a $(k \times \ell)$-contingency table as data material. Namely, if two categorical random variables X and Y are considered, where X can take exactly k different values and Y can take exactly ℓ different values, then the null hypothesis $X \perp\!\!\!\perp Y$ can be tested with an asymptotic χ^2-test, applied to the observed $(k \times \ell)$-contingency table. In this, the number of degrees of freedom is given by $(k-1)\cdot(\ell-1)$.

Chapter 2

Continuously distributed response variables

2.1 Multiple linear regression (ANCOVA)

Model 2.1 (Classical multiple linear regression, ANCOVA)
We consider the sample space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, and we model the observations $y_1, \dots, y_n$ as realizations of real-valued, stochastically independent random variables $Y_1, \dots, Y_n$ with the representation

$$\forall 1 \leq i \leq n: \quad Y_i = f(x_{i,1}, \dots, x_{i,k}) + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i. \tag{2.1}$$

The vector $\beta = (\beta_0, \beta_1, \dots, \beta_k)^\top \in \mathbb{R}^{k+1}$ is the parameter of interest. We let $p := k + 1$ denote the dimension of the corresponding parameter space. In the classical setup, we assume that $n > p$. We can write

$$\begin{aligned}
Y &:= (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n && \text{response vector,} \\
X &:= \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times p} && \text{design matrix,} \\
\varepsilon &:= (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n && \text{vector of error terms,} \\
\beta &\equiv (\beta_0, \beta_1, \dots, \beta_k)^\top \in \mathbb{R}^p && \text{parameter vector.}
\end{aligned}$$

With these notations, we can write (2.1) *in matrix form as follows:*

$$Y = X\beta + \varepsilon. \tag{2.2}$$

Furthermore, we make the following model assumptions.

(a) The design matrix has maximal rank, so that $X^\top X \in \mathbb{R}^{p\times p}$ is positively definite and hence invertible.

(b) The error terms are i.i.d., where $\mathbb{P}^{\varepsilon_1}$ is induced by the cdf F. We assume that $\mathbb{E}[\varepsilon_1] = 0$ and $0 < \sigma^2 := Var(\varepsilon_1) < \infty$, thus in particular homoscedasticity. The unknown cdf F is considered an (infinite-dimensional) nuisance parameter, meaning that it is not itself the target of statistical inference.

Optionally, we consider for likelihood-based inference occasionally a normal distribution assumption for the error terms:

(c) $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$.

Remark 2.2 (Modeling aspects)

(a) The statistical analysis under Model 2.1 is called analysis of covariance (ANCOVA).

(b) The assumptions regarding the design matrix X under Model 2.1 can be interpreted in two different ways: (i) the design matrix consists of deterministic "measurement locations", which have been pre-defined when planning (i. e., designing) the experiment, or (ii) the statistical analysis is carried out conditionally to the (realized values of the) design matrix. In the latter case, any potential randomness in the design points (entries of X) is not considered in the modeling approach. In contrast, models which incorporate a probabilistic model for the design matrix X are called regression models with random design or correlation models, respectively. Such correlation models require additional considerations. In particular, correlations between the stochastic covariates and the stochastic error terms have to be taken into account. Actually, even the parameter vector β is, under a correlation model, a feature of the joint distribution of covariates and response. In contrast, under Model 2.1, β is considered as a fixed, but unknown point in $\mathbb{R}^p$, and the stochasticity of the model is solely induced by the error terms. We will treat Cases (i) (deterministic design) and (ii) (conditioning on the design matrix) in complete analogy. In particular, we often notationally suppress the conditioning on X in Case (ii), to keep the notation simple.

(c) The assumption of uncorrelated error terms has to be validated after model fit by means of a so-called residual analysis, cf. Definition 2.4. If there is an apparent structure in the residuals (the values of the error terms estimated by the model), then this can be an indication that additional covariates should be included in the model.

(d) Categorical covariates should be encoded by using a set of so-called "dummy indicators", to avoid an implicit (and often inadequate) metrization of the discrete support of such covariates. More precisely, a categorical covariate with exactly ℓ possible values (represented by the integers $1, \ldots, \ell$, without loss of generality) is represented by a set of $(\ell - 1)$ dummy

indicators. In this, the j-th dummy indicator encodes the event that the corresponding covariate takes the value $j+1$, for $j = 1, \ldots, \ell - 1$. Hence, if all $(\ell - 1)$ dummy indicators are equal to zero, this has the interpretation that the corresponding categorical covariate takes the (reference) value 1.

(e) Interactions (interrelationships) amongst covariates are modeled with so-called interaction terms. These interaction terms constitute additional columns in the design matrix. In this, the interaction term for the interrelationship between covariates X_j and X_k is given by $x_j \cdot x_k$ (pointwise multiplication of the n-vectors of realized values of covariates j and k with respect to the n observational units).

Corollary 2.3

The assumptions of Model 2.1 imply that

$$\begin{aligned} \forall 1 \leq i \leq n: \quad \mathbb{E}_\beta[Y_i] &= \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k}, \\ \forall 1 \leq i \leq n: \quad \mathit{Var}(Y_i) &= \sigma^2, \\ \forall 1 \leq i \neq j \leq n: \quad \mathit{Cov}(Y_i, Y_j) &= \mathit{Cov}(\varepsilon_i, \varepsilon_j) = 0. \end{aligned}$$

If we assume in addition normally distributed error terms, we get that

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n).$$

Definition 2.4 (Residuals)

If an estimator $\hat{\beta}$ for the parameter vector β is available, we get the (plug-in) estimator $\widehat{\mathbb{E}[Y]} = X\hat{\beta}$ for the (conditional) expectation vector of Y. We define the components of $\widehat{\mathbb{E}[Y]}$ as $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_k x_{i,k}$ for $i = 1, \ldots, n$, and we define the residuals pertaining to $\hat{\beta}$ as the observed differences between the observed response values and their estimated (conditional) expected values. Thus, the n residuals are given by $\hat{\varepsilon}_i = y_i - \hat{y}_i$, $1 \leq i \leq n$.

Theorem 2.5

Under Model 2.1, the following assertions hold true.

(a) The least squares estimator (LSE) of the parameter vector β is given by

$$\hat{\beta} \equiv \hat{\beta}(Y) = (X^\top X)^{-1} X^\top Y,$$

implying the representation

$$\hat{\beta} - \beta = (X^\top X)^{-1} X^\top \varepsilon$$

for the (random) estimation error vector.

(b) Plugging the representation of $\hat{\beta}$ into the equation $\hat{Y} = X\hat{\beta}$, we furthermore get that

$$\hat{Y} = X(X^\top X)^{-1} X^\top Y =: HY$$

with the $(n \times n)$-matrix $H = X(X^\top X)^{-1}X^\top$. The matrix H is called prediction matrix or hat matrix. The matrix $X^+ = (X^\top X)^{-1}X^\top$ is called (Moore-Penrose) pseudo inverse of X.

(c) *If we assume normally distributed error terms, then the maximum likelihood estimator (MLE) of β coincides with the LSE for β.*

(d) *The moments of $\hat{\beta}$ up to the second order are given by*

$$\mathbb{E}_\beta\left[\hat{\beta}\right] = \beta \text{ and } Cov(\hat{\beta}) = \sigma^2(X^\top X)^{-1}.$$

Proof: The assertion (b) follows immediately, once (a) has been shown. We carry out the proof of assertion (a) geometrically, using methods of linear algebra. To this end, let $x_i' := (1, x_{i,1}, \ldots, x_{i,k})$ denote the i-th row of the design matrix. Notice that the least squares criterion

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^n (Y_i - x_i'\beta)^2\right\} = \arg\min_{\beta\in\mathbb{R}^p} ||Y - X\beta||_2^2$$

can equivalently be expressed by saying that $X\hat{\beta}$ is the L_2-projection of Y onto the vector space $\{z \in \mathbb{R}^n : z = X\gamma, \gamma \in \mathbb{R}^p\}$. This fact is illustrated by Figure 2.1.

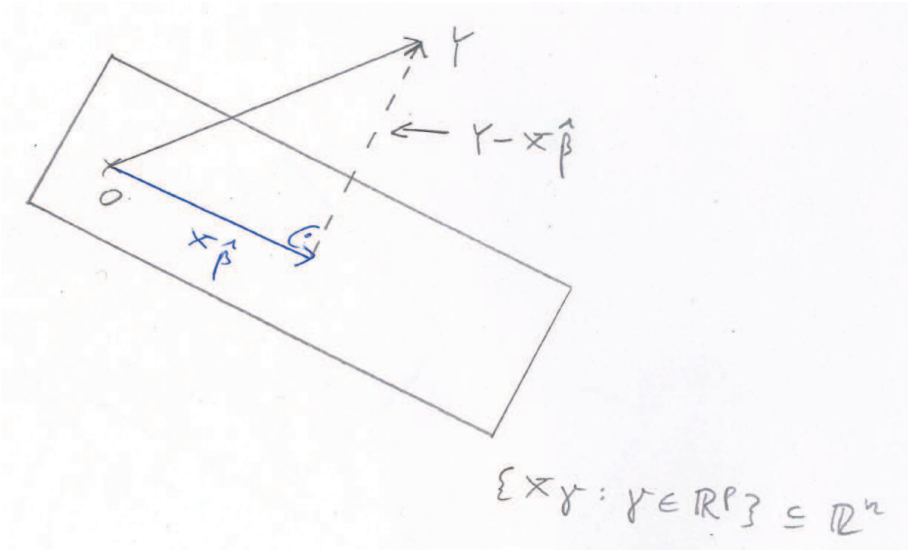


Figure 2.1: Orthogonal projection of Y into the space $\{X\gamma : \gamma \in \mathbb{R}^p\}$

Thus (see Figure 2.1), $\hat{\beta}$ can be characterized via

$$\begin{aligned}
&\forall\gamma \in \mathbb{R}^p: \quad \langle Y - X\hat{\beta}, X\gamma\rangle_{\mathbb{R}^n} = 0 \\
\Leftrightarrow\ &\forall\gamma \in \mathbb{R}^p: \quad Y^\top X\gamma = \hat{\beta}^\top X^\top X\gamma \qquad \text{(Bilinearity of } \langle\cdot,\cdot\rangle_{\mathbb{R}^n}\text{)} \\
\Leftrightarrow\ &\qquad\qquad Y^\top X = \hat{\beta}^\top X^\top X.
\end{aligned}$$

Multiplication of the latter equation from the right with $(X^\top X)^{-1}$ yields $Y^\top X(X^\top X)^{-1} = \hat{\beta}^\top$, hence $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ as desired, since $(X^\top X)^{-1}$ is a symmetric matrix.
Parts (c) and (d) are exercises. ■

Remark 2.6 (Geometric properties of $\hat{\beta}$)
Let $\hat{\beta} = X^+Y$ be as in Theorem 2.5. Then, the following assertions hold true.

(i) The estimated ((conditional) expected) values $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)^\top$ are orthogonal to the residuals $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^\top$, meaning that $\hat{y}^\top \hat{\varepsilon} = 0$.

(ii) The columns of X are orthogonal to the residuals, meaning that $X^\top \cdot \hat{\varepsilon} = 0$.

(iii) The residuals are on average equal to zero, meaning that $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ and $\bar{\hat{\varepsilon}} = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i = 0$, respectively.

(iv) The arithmetic mean of $\{\hat{y}_i : 1 \le i \le n\}$ equals the arithmetic mean of the response values $\{y_i : 1 \le i \le n\}$, meaning that $\bar{\hat{y}} = n^{-1}\sum_{i=1}^n \hat{y}_i = \bar{y} = n^{-1}\sum_{i=1}^n y_i$.

(v) The regression hyperplane contains the barycenter of the data, meaning that $\bar{y} = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j \bar{x}_j$, where $\bar{x}_j := n^{-1}\sum_{i=1}^n x_{i,j}$ for all $1 \le j \le k$.

Lemma 2.7 (Centering operator)
Let $\mathbf{1} := (1, \ldots, 1)^\top \in \mathbb{R}^n$ and $C := I_n - n^{-1}\mathbf{1} \cdot \mathbf{1}^\top \in \mathbb{R}^{n\times n}$. Then, the following assertions hold true.

(i) For any vector $a \in \mathbb{R}^n$, we have that

$$Ca = \begin{pmatrix} a_1 - \bar{a} \\ \vdots \\ a_n - \bar{a} \end{pmatrix}.$$

Thus, we call C centering operator.

(ii) C is symmetric and idempotent, i. e., $C^2 = C$.

(iii) For all $a \in \mathbb{R}^n$: $a^\top Ca = \sum_{i=1}^n (a_i - \bar{a})^2$.

Proof: Elementary linear algebra, left as an exercise. ■

Theorem and Definition 2.8
Let $\hat{\beta} = X^+Y$ be as in Theorem 2.5. Then, the following assertions hold true.

(a) Decomposition of spread:

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \\ \iff: SST &= SSR + SSE \\ \iff: s_y^2 &= s_{\hat{y}}^2 + s_{\hat{\varepsilon}}^2. \end{aligned}$$

(b) Due to Part (a),

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

holds true. Thus, the statistic

$$R^2 := \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

with values in $[0, 1]$ *quantifies the proportion of total variation in the response variables of the sample, which can be explained by the regression model.*
The value of R^2 *is called R-square value or coefficient of determination, respectively.*

(c) Letting $r_{a,b}$ *denote the empirical Pearson product-moment correlation coefficient of two data vectors* a *und* b*, it holds that* $R^2 = r_{y,\hat{y}}^2$.

Proof: For proving Part (a), we multiply the identity $y = \hat{y} + \hat{\varepsilon}$ from the left with the centering matrix C. This yields that $Cy = C\hat{y} + C\hat{\varepsilon}$. Since the residuals are already centered by construction (see Part (iii) of Remark 2.6), we have that $C\hat{\varepsilon} = \hat{\varepsilon}$, hence $Cy = C\hat{y} + \hat{\varepsilon}$ as well as $y^\top C = \hat{y}^\top C + \hat{\varepsilon}^\top$. We conclude that

$$\begin{aligned} y^\top CCy &= \left(\hat{y}^\top C + \hat{\varepsilon}^\top\right)(C\hat{y} + \hat{\varepsilon}) \\ &= \hat{y}^\top CC\hat{y} + \hat{y}^\top C\hat{\varepsilon} + \hat{\varepsilon}^\top C\hat{y} + \hat{\varepsilon}^\top\hat{\varepsilon} \\ &= \hat{y}^\top C\hat{y} + \hat{y}^\top\hat{\varepsilon} + \hat{\varepsilon}^\top\hat{y} + \hat{\varepsilon}^\top\hat{\varepsilon} \\ \iff \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \end{aligned}$$

due to Part (i) of Remark 2.6.

For proving Part (c), we make use of the definition of the empirical correlation coefficient and obtain that

$$\begin{aligned} r_{y,\hat{y}} &= \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}} \\ \Rightarrow r_{y,\hat{y}}^2 &= \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2}{SST \cdot SSR}. \end{aligned}$$

Comparing this with the definition of R^2, it remains to show that

$$\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2 = (SSR)^2 = \left[\sum_{i=1}^n (\hat{y}_i - \bar{y})^2\right]^2.$$

Thus, it suffices to show that

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

We multiply out the left-hand side and obtain that

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

Since the residuals are centered and orthogonal to the vector $\hat{y}$ (see Remark 2.6), the assertion follows. ■

Theorem 2.9 (Calculation rules for expectation vectors and covariance matrices)
Let Z_1 and Z_2 denote random vectors with values in $\mathbb{R}^d$, and let A and b, respectively, be a deterministic matrix and a deterministic vector, respectively, of appropriate dimensions. Denote $\mathbb{E}[Z_1] =: \mu$ and Cov$(Z_1) := \mathbb{E}[(Z_1 - \mu)(Z_1 - \mu)^\top] =: \Sigma$.
Then, the following assertions hold true.

(i) $\mathbb{E}[Z_1 + Z_2] = \mathbb{E}[Z_1] + \mathbb{E}[Z_2]$.

(ii) $\mathbb{E}[AZ_1 + b] = A\mu + b$.

(iii) Cov$(Z_1) = \mathbb{E}[Z_1 Z_1^\top] - \mu\mu^\top$.

(iv) Var$(b^\top Z_1) = b^\top \Sigma b = \sum_{i=1}^{d}\sum_{j=1}^{d} b_i b_j \sigma_{ij}$.

(v) Cov$(AZ_1 + b) = A\Sigma A^\top$.

(vi) $\mathbb{E}[Z_1^\top A Z_1] =$ tr$(A\Sigma) + \mu^\top A\mu$.

Proof: Satz B.1 in Fahrmeir et al. (2009). ■

Theorem 2.10 (Gauß-Markov)
Under Model 2.1, $\hat{\beta} = X^+ Y$ has minimum variance amongst all linear (in the data) and unbiased estimators of β. In particular,

$$\forall\, 0 \le j \le k : \text{Var}(\hat{\beta}_j) \le \text{Var}(\hat{\beta}_j^L) \tag{2.3}$$

holds true for every linear, unbiased estimator $\hat{\beta}^L \equiv \hat{\beta}^L(Y)$ of β.
Furthermore, we get for any linear combination $c^\top\beta$ with fixed $c \in \mathbb{R}^p$, that Var$(c^\top\hat{\beta}) \le$ Var$(c^\top\hat{\beta}^L)$, *where $\hat{\beta}^L$ is as in* (2.3).
Thus, $\hat{\beta}$ is BLUE (best linear unbiased estimator).

Proof: Every estimator $\hat{\beta}^L$ of $\beta \in \mathbb{R}^{p\times 1}$, which is linear in the data $Y \in \mathbb{R}^{n\times 1}$, can be written in the form $\hat{\beta}^L = AY$ with $A \in \mathbb{R}^{p\times n}$. Since $\mathbb{E}[\hat{\beta}^L] = \mathbb{E}[AY] = AX\beta$, unbiasedness implies that the condition

$$\begin{aligned} \forall\beta \in \mathbb{R}^p : AX\beta = \beta \quad &\iff \quad (AX - I_p)\beta = 0 \\ &\iff \quad AX = I_p \end{aligned}$$

needs to be fulfilled. This implies that $\text{rank}(A) = p$, and thus we can write A (w. l. o. g.) in the form $A = (X^\top X)^{-1}X^\top + B$, for some suitable matrix B. Plugging the latter representation of A into the identity $I_p = AX$ yields that $I_p = AX = (X^\top X)^{-1}X^\top X + BX = I_p + BX$, hence $BX = 0$. Thus,

$$\begin{aligned}
\text{Cov}(\hat{\beta}^L) &= \sigma^2 AA^\top \\
&= \sigma^2 \left[(X^\top X)^{-1}X^\top + B\right]\left[X(X^\top X)^{-1} + B^\top\right] \\
&= \sigma^2 \left[(X^\top X)^{-1}X^\top X(X^\top X)^{-1} + (X^\top X)^{-1}X^\top B^\top + BX(X^\top X)^{-1} + BB^\top\right] \\
&= \sigma^2 (X^\top X)^{-1} + \sigma^2 BB^\top \\
&= \text{Cov}(\hat{\beta}) + \sigma^2 BB^\top.
\end{aligned}$$

Since $BB^\top$ is non-negative definite, we have that $\text{Cov}(\hat{\beta}^L) - \text{Cov}(\hat{\beta}) = \sigma^2 BB^\top$ is non-negative definite, too. For a fixed vector $c \in \mathbb{R}^p$, noticing that

$$\begin{aligned}
\text{Var}(c^\top \hat{\beta}^L) &= c^\top \text{Cov}(\hat{\beta}^L) c \text{ as well as} \\
\text{Var}(c^\top \hat{\beta}) &= c^\top \text{Cov}(\hat{\beta}) c,
\end{aligned}$$

we conclude that $\text{Var}(c^\top \hat{\beta}^L) \geq \text{Var}(c^\top \hat{\beta})$. The inequality (2.3) follows by considering special vectors c with entries $c_i = \mathbf{1}_{\{i=j+1\}}$, $i = 1, \ldots, p$, for all $0 \leq j \leq k$. ■

We now turn to the task of constructing tests and confidence regions for the parameters $(\beta_j)_{0 \leq j \leq k}$, based on $\hat{\beta}$ and (characteristics of) its sampling distribution. Part (d) of Theorem 2.5 indicates that for this task an estimate of the (in general unknown) error variance σ^2 is required, because the covariance matrix of $\hat{\beta}$ depends on σ^2. Under the normal distribution assumption for the error terms considered in Part (c) of Model 2.1, the maximum likelihood approach can be employed for estimating σ^2.

Theorem 2.11 (Estimation of σ^2)
Consider Model 2.1.

(a) Under the normal distribution assumption for the error terms considered in Part (c) of Model 2.1, the MLE for σ^2 is given by $\widehat{\sigma^2}_{ML} = \hat{\varepsilon}^\top \hat{\varepsilon}/n$.

(b) In general, it holds that $\mathbb{E}[\hat{\varepsilon}^\top \hat{\varepsilon}] = (n-p)\sigma^2$, so that an unbiased (moment) estimator for σ^2 is given by $\widehat{\sigma^2} = \hat{\varepsilon}^\top \hat{\varepsilon}/(n-p)$, even without assuming normally distributed error terms.

Remark 2.12

(i) The unbiased estimator $\widehat{\sigma^2}$ is typically preferred in practice, but it has a larger variance than $\widehat{\sigma^2}_{ML}$.

(ii) Under the conditions of Part (a) of Theorem 2.11, the unbiased estimator $\widehat{\sigma^2}$ is the restricted MLE of σ^2. Namely, the restricted maximum likelihood (REML) approach maximizes the marginal likelihood function

$$Z(y, \sigma^2) = \int_{\mathbb{R}^p} Z(y, (\beta, \sigma^2)) d\beta,$$

which is obtaied by "integrating out" the parameter vector.

Proof: Let us prove Theorem 2.11. Part (a) is an exercise.
For proving Part (b), notice first that $\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (I_n - H)Y$. An exercise yields that the matrix $I_n - H$ is symmetric and idempotent with $\text{rank}(I_n - H) = n - p$. Thus, we get that $\mathbb{E}[\hat{\varepsilon}^\top \hat{\varepsilon}] = \mathbb{E}[Y^\top (I_n - H)Y]$. Calculation rule (vi) from Theorem 2.9 then yields that

$$\mathbb{E}[\hat{\varepsilon}^\top \hat{\varepsilon}] = \text{tr}((I_n - H)\sigma^2 I_n) + \beta^\top X^\top (I_n - H) X\beta.$$

Now, $\text{tr}((I_n - H)\sigma^2 I_n) = \sigma^2[\text{tr}(I_n) - \text{tr}(H)] = \sigma^2(n-p)$, because $\text{tr}(H) = \text{tr}(X(X^\top X)^{-1}X^\top) = \text{tr}(I_p) = p$. We conclude that

$$\begin{aligned}
\mathbb{E}[\hat{\varepsilon}^\top \hat{\varepsilon}] &= \sigma^2(n-p) + \beta^\top X^\top (I_n - X(X^\top X)^{-1}X^\top) X\beta \\
&= \sigma^2(n-p) + \beta^\top X^\top X\beta - \beta^\top X^\top X(X^\top X)^{-1}X^\top X\beta \\
&= \sigma^2(n-p) + \beta^\top X^\top X\beta - \beta^\top X^\top X\beta \\
&= \sigma^2(n-p),
\end{aligned}$$

as desired. ■

Corollary 2.13
Under Model 2.1, a (plug-in) estimator for Cov$(\hat{\beta})$, where $\hat{\beta} = X^+ Y$, is given by

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

Theorem 2.14 (Statistical properties of the residuals)
Under Model 2.1, consider the residuals $\hat{\varepsilon} \equiv \hat{\varepsilon}(Y) = Y - \hat{Y} = Y - HY = Y - X(X^\top X)^{-1}X^\top Y$ as random variables. Then, the following assertions hold true.

1) $\mathbb{E}[\hat{\varepsilon}] = 0$.

2) Cov$(\hat{\varepsilon}) = \sigma^2(I_n - H)$. In particular, the residuals exhibit heteroscedasticity, and they are non-trivially correlated.

3) The standardized residuals, given by

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \; 1 \leq i \leq n,$$

are homoscedastically distributed, if the assumptions of Model 2.1 are fulfilled.

Proof: The expectation vector of $\hat{\varepsilon}$ is calculated as $\mathbb{E}[\hat{\varepsilon}] = \mathbb{E}[Y] - X(X^\top X)^{-1}X^\top \mathbb{E}[Y] = X\beta - X(X^\top X)^{-1}X^\top X\beta = 0$.
For the covariance matrix of $\hat{\varepsilon}$, we get that $\mathrm{Cov}(\hat{\varepsilon}) = \mathrm{Cov}((I_n - H)Y) = (I_n - H)\sigma^2 I_n (I_n - H)^\top = \sigma^2(I_n - H)^\top = \sigma^2(I_n - H)$, since the matrix $(I_n - H)$ is symmetric and idempotent, cf. the proof of Part (b) of Theorem 2.11. ■

Theorem 2.15 (Multivariate central limit theorem)
Consider a sequence of ANCOVA models, indexed by the sample size n. In this, assume that the following two assumptions regarding the sequence $(X_n)_{n\geq p}$ of design matrices are fulfilled.

(i) $n^{-\frac{1}{2}} \max\limits_{1\leq i\leq n, 1\leq j\leq p} \left|x_{i,j}\right| \longrightarrow 0$ *for* $n \to \infty$.

(ii) $n^{-1}X_n^\top X_n \longrightarrow V$ *for a positive definite, symmetric matrix* $V \in \mathbb{R}^{p\times p}$.

Then, the following two assertions hold true.

(a) Let $a^\top = (a_1, \ldots, a_p)$ *denote an arbitrary, but fixed vector in* $\mathbb{R}^p$. *Letting* $\rho^2 = \sigma^2 a^\top V a$, *we have that*

$$\mathcal{L}\left(n^{-\frac{1}{2}} a^\top X_n^\top \varepsilon\right) \xrightarrow[n\to\infty]{w} \mathcal{N}(0, \rho^2).$$

(b) For $\hat{\beta}(n) = X_n^+ Y_n$, *we have that*

$$\mathcal{L}\left(\sqrt{n}\left\{\hat{\beta}(n) - \beta\right\}\right) \xrightarrow[n\to\infty]{w} \mathcal{N}_p\left(0, \sigma^2 V^{-1}\right).$$

Proof: Define $S_n := a^\top X_n^\top \varepsilon$. We notice, that

$$S_n = \sum_{j=1}^{p}\left(a_j \sum_{i=1}^{n} x_{i,j}\varepsilon_i\right) = \sum_{i=1}^{n}\varepsilon_i\left(\sum_{j=1}^{p} a_j x_{i,j}\right) =: \sum_{i=1}^{n} b_i\,\varepsilon_i$$

is a sum of stochastically independent, centered random variables. Furthermore, we get that

$$\begin{aligned}\mathrm{Var}\,(S_n) = \sigma^2 \sum_{i=1}^{n} b_i^2 &= \sigma^2 \sum_{i=1}^{n}\sum_{j,\ell=1}^{p} a_j a_\ell x_{i,j} x_{i,\ell} \\ &= \sigma^2 \sum_{j,\ell=1}^{p} a_j a_\ell (X_n^\top X_n)_{j,\ell} \\ &= \sigma^2 a^\top (X_n^\top X_n) a.\end{aligned}$$

If follows that $\mathrm{Var}\left(n^{-\frac{1}{2}} S_n\right) = n^{-1}\sigma^2 a^\top X_n^\top X_n a \longrightarrow \rho^2 = \sigma^2 a^\top V a$ for $n \to \infty$. Checking the Lindeberg condition by making use of Assumption (i) completes the proof of Part (a).
For proving Part (b), we recall the representation

$$\sqrt{n}\{\hat{\beta}(n) - \beta\} = \frac{1}{\sqrt{n}}(n^{-1}X_n^\top X_n)^{-1} X_n^\top \varepsilon,$$

which we have obtained in Part (a) of Theorem 2.5. By the Cramér-Wold device (see, e. g., Page 862 in Shorack and Wellner (1986)), it holds that

$$\mathcal{L}\left(\frac{1}{\sqrt{n}}X_n^\top \varepsilon\right) \xrightarrow[n\to\infty]{w} \mathcal{N}_p\left(0, \sigma^2 V\right).$$

Moreover, $(n^{-1}X_n^\top X_n)^{-1}$ converges to V^{-1} by Assumption (ii). Altogether, this yields that

$$\mathcal{L}\left(\frac{1}{\sqrt{n}}(n^{-1}X_n^\top X_n)^{-1}X_n^\top \varepsilon\right) \xrightarrow[n\to\infty]{w} \mathcal{N}_p\left(0, \sigma^2 V^{-1}\right),$$

as desired. ■

Theorem 2.16 (Distribution of quadratic forms in Gaussian variables)

1. *Let $X \sim \mathcal{N}_n(\mu, \Sigma)$, where Σ is symmetric and positive definite.*
   Then, $(X-\mu)^\top \Sigma^{-1}(X-\mu) \sim \chi^2_n$.

2. *Let $X \sim \mathcal{N}_n(0, I_n)$, let R be a symmetric, idempotent $(n \times n)$-matrix with rank$(R) = r$, and let B be a $(p \times n)$-matrix with $p \leq n$. Then, the following two assertions hold true.*

   (a) $X^\top R X \sim \chi^2_r$

   (b) From $BR = 0$ it follows that $X^\top R X$ is stochastically independent of BX.

3. *Let $X \sim \mathcal{N}_n(0, I_n)$, and let R and S be symmetric and idempotent $(n \times n)$-matrices with rank$(R) = r$, rank$(S) = s$, and $RS = 0$. Then, the following two assertions hold true.*

   (a) $X^\top R X$ and $X^\top S X$ are stochastically independent.

   (b) $\frac{s}{r}\frac{X^\top R X}{X^\top S X} \sim F_{r,s}$

Proof: We follow Satz B.6 in Fahrmeir et al. (2009).

Proof of 1. Denote by $\Sigma^{1/2}$ the symmetric and positive definite matrix with $\Sigma^{1/2} \cdot \Sigma^{1/2} = \Sigma$ and pertaining inverse matrix $\Sigma^{-1/2}$. Then, $Z := \Sigma^{-1/2}(X-\mu) \sim \mathcal{N}_n(0, I_n)$. The definition of the chi-square distribution yields that $Z^\top Z \sim \chi^2_n$ and hence the assertion.

Proof of 2. (a) Since R is idempotent and symmetric, there exists an orthonormal matrix P such that $R = P D_r P^\top$, where $D_r = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$; see Section 8.5.1.1 in Gentle (2017). Due to the orthonormality of P, the random vector $W := P^\top X$ has the same distribution as X, namely $W \sim \mathcal{N}_n(0, I_n)$. The assertion follows by noticing that

$$X^\top R X = X^\top R^2 X = (RX)^\top (RX) = (PD_rW)^\top (PD_rW) = W^\top D_r P^\top P D_r W$$

$$= W^\top D_r W = \sum_{i=1}^{r} W_i^2,$$

and by the definition of the chi-square distribution.
(b) Define $Z_1 := BX \sim \mathcal{N}_p(0, B^\top B)$ and $Z_2 := RX \sim \mathcal{N}_n(0, R)$. Since

$$\text{Cov}(Z_1, Z_2) = \text{Cov}(BX, RX) = B\,\text{Cov}(X)\,R^\top = BR = 0$$

and Z_1, Z_2 are jointly normal, we conclude that Z_1 and Z_2 are stochastically independent. This implies that $Z_1 = BX$ and $Z_2^\top \cdot Z_2 = X^\top RX$ are stochastically independent, too.

Proof of 3. (a) Define $Z_1 := SX \sim \mathcal{N}_n(0, S)$ and $Z_2 := RX \sim \mathcal{N}_n(0, R)$. It holds that

$$\text{Cov}(Z_1, Z_2) = S\,\text{Cov}(X)\,R = SR = S^\top R^\top = (RS)^\top = 0.$$

Due to joint normality of Z_1, Z_2, their uncorrelatedness implies their stochastic independence. In turn, the random variables $Z_1^\top Z_1$ and $Z_2^\top Z_2$ are stochastically independent, too. The assertion follows by the identities $X^\top SX = Z_1^\top Z_1$ and $X^\top RX = Z_2^\top Z_2$.
Part (b) is a consequence of Part (a) and of the assertion under Part 1., by means of the definition of Fisher's F-distribution.

■

Definition 2.17 (Linear hypotheses)
Let K be a deterministic $(r \times p)$-matrix with rank$(K) = r \leq p$. In the context of Model 2.1, we call such a K a contrast matrix, and we call a test problem of the form H_0: $K\beta = d$ versus H_1: $K\beta \neq d$ with a fixed, given vector $d \in \mathbb{R}^{r\times 1}$ a (two-sided) linear test problem. This means, that the linear hypothesis H_0 imposes $r \leq p$ linearly independent conditions / restrictions on the parameters of the ANCOVA model.

Example 2.18

(i) Test for a significant influence of one particular covariate on the (expected) response:

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

for a given $1 \leq j \leq k$.
$\Rightarrow$ *$K \in \mathbb{R}^{1\times p}$ with entries $K_i = \mathbf{1}_{\{i=j+1\}}$, and $d = 0 \in \mathbb{R}$.*

(ii) Testing the influence of a sub-vector $\beta^* = (\beta_1, \ldots, \beta_r)^\top$:

$$\Rightarrow K = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & & & \ddots & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \end{pmatrix} \in \mathbb{R}^{r\times p} \tag{2.4}$$

with entries $K_{i,\ell} = \mathbf{1}_{\{\ell=i+1\}}$, and $d = 0 \in \mathbb{R}^r$.

(iii) Test for equality of two regression coefficients:

$$H_0 : \beta_{j_1} - \beta_{j_2} = 0 \quad \textit{versus} \quad H_1 : \beta_{j_1} - \beta_{j_2} \neq 0, \quad \textit{with } 1 \leq j_1 \neq j_2 \leq k.$$

$\Rightarrow$ $K \in \mathbb{R}^{1\times p}$ *with entries* $K_i = \mathbf{1}_{\{i=j_1+1\}} - \mathbf{1}_{\{i=j_2+1\}}$, *i. e.,*
$K = (0, \ldots, 0, \underbrace{1}_{j_1+1\text{-}th}, 0, \ldots, 0, \underbrace{-1}_{j_2+1\text{-}th}, 0, \ldots, 0)$, *and* $d = 0 \in \mathbb{R}$.

Theorem 2.19

From now on, we refer to the test statistic $2\log\Lambda_n(Y)$ *of a likelihood ratio test as deviance.*

Under Model 2.1, the following assertions hold true.

(a) For computing the deviance for testing the linear hypothesis H_0*:* $K\beta = d$*, it is necessary to maximize the (log-)likelihood function of the model under the constraints encoded by* K *and* d*. This constrained maximization can be avoided by using instead of the deviance the Wald statistic*

$$W = (K\hat{\beta} - d)^\top (K\hat{V}K^\top)^{-1}(K\hat{\beta} - d), \tag{2.5}$$

where $\hat{\beta}$ *denotes the MLE for* β *in the full (unrestricted) model and* $\hat{V}$ *denotes the estimated covariance matrix of* $\hat{\beta}$*; cf. Corollary 2.13. The statistic* W *is asymptotically equivalent to the deviance* $2\log\Lambda_n(Y)$*. In particular,* $\mathcal{L}(W) \xrightarrow[H_0]{w} \chi^2_r$ *for* $n \to \infty$.

(b) If we assume normally distributed error terms (see the optional Assumption (c) of Model 2.1), then the deviance is an isotone transformation of $F = \frac{n-p}{r}\frac{\Delta SSE}{SSE}$*, where* $\Delta SSE = SSE_{H_0} - SSE$*. Furthermore, the statistic* F *is in this case exactly* F*-distributed under* H_0*, meaning that* $F \underset{H_0}{\sim} F_{r,n-p}$.

(c) Under the assumptions of Part (b), it holds that $W = rF$.

Proof: The asymptotic χ^2_r-distribution postulated in Part (a) under the null follows from the asymptotic normality of $\hat{\beta}$ established in Theorem 2.15. To understand this conclusion, we first explain it by means of the situation of Theorem 1.41. Namely, with the notation used there, we have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)^{-1}) \quad \text{under } \theta_0 \text{ for } n \to \infty,$$

under the stated regularity assumptions. Thus,

$$I(\theta_0)^{1/2} n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_r), \text{ with } r := \dim(\theta_0).$$

The Continuous Mapping Theorem then yields, that

$$(\hat{\theta}_n - \theta_0)^\top n I(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \chi^2_r.$$

If the Fisher information (pertaining to a single observational unit) $\theta \mapsto I(\theta)$ is continuous and hence $I(\hat{\theta}_n)$ is a consistent estimator of $I(\theta_0)$, Slutsky's lemma entails that

$$(\hat{\theta}_n - \theta_0)^\top n I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \chi^2_r, \tag{2.6}$$

where $nI(\theta_0)$ is the Fisher information of the product model.
Now, we refer to Remark 1.43 and Theorem 2.15, and connect (2.6) to our situation. We have to test $H_0 : K\beta - d = 0$, thus the parameter of interest is given by $\theta = K\beta - d$, $\theta_0 = 0$, and $\hat{\theta}_n = K\hat{\beta} - d$. Plugging these specifications into (2.6) and replacing $nI(\hat{\theta}_n)$ (inverse covariance matrix for large $n \to \infty$) by $(K\hat{V}K^\top)^{-1}$, we get that $\mathcal{L}(W) \xrightarrow{w} \chi^2_r$ for $n \to \infty$ under H_0.

For Part (b), we introduce in analogy to Remark 1.45 the abbreviations

$$\begin{aligned}
\hat{\beta}_{H_0} :&\quad \text{MLE in the reduced model (under the constraints encoded by } K \text{ and } d),\\
\widehat{\sigma^2}_{H_0} :&\quad \text{MLE of the error variance in the reduced model,}\\
\hat{Z}(y) :=&\quad Z(y, (\hat{\beta}, \widehat{\sigma^2}_{ML})),\\
\hat{Z}_{H_0}(y) :=&\quad Z(y, (\hat{\beta}_{H_0}, \widehat{\sigma^2}_{H_0})),
\end{aligned}$$

and compute the deviance as follows.

$$\begin{aligned}
2\log\Delta_n(y) &= 2\left[\ln(\hat{Z}(y)) - \ln(\hat{Z}_{H_0}(y))\right]\\
&= 2\left[-\frac{n}{2}\log(2\pi\widehat{\sigma^2}_{ML}) - \frac{SSE}{2\hat{\sigma}^2_{ML}} + \frac{n}{2}\log(2\pi\widehat{\sigma^2}_{H_0}) + \frac{SSE_{H_0}}{2\widehat{\sigma^2}_{H_0}}\right]\\
&= n\log\left(\frac{\widehat{\sigma^2}_{H_0}}{\widehat{\sigma^2}_{ML}}\right) = n\log\left(\frac{SSE_{H_0}}{SSE}\right) = n\log\left(\frac{\Delta SSE}{SSE} + 1\right).
\end{aligned}$$

For the verification of the $F_{r,n-p}$-distribution of $F = \frac{n-p}{r}\frac{\Delta SSE}{SSE}$ under H_0, we employ Part 3. of Theorem 2.16. To this end, it remains to show that

(i) $\Delta SSE/\sigma^2 \sim \chi^2_r$,

(ii) ΔSSE and SSE are stochastically independent.

(The argumentation is then completed by an exercise.) We show these remaining properties (i) and (ii) as well as the assertion of Part (c) in Corollary 2.21. ■

Theorem 2.20

Under the conditions of Parts (b) and (c) of Theorem 2.19, the following assertions hold true.

(i) The restricted MLE of θ is given by

$$\hat{\beta}_{H_0} = \hat{\beta} - (X^\top X)^{-1}K^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\hat{\beta} - d).$$

This estimator $\hat{\beta}_{H_0}$ fulfills the constraint $K\hat{\beta}_{H_0} = d$, because

$$K\hat{\beta}_{H_0} = K\hat{\beta} - K(X^\top X)^{-1}K^\top(K(X^\top X)^{-1}K^\top)^{-1} \cdot (K\hat{\beta} - d) = K\hat{\beta} - K\hat{\beta} + d = d.$$

Furthermore, $\hat{\beta}_{H_0} = \hat{\beta}$, if the unrestricted MLE $\hat{\beta}$ already fulfills the constraint.

(ii) Abbreviating $\Delta_{H_0} = (X^\top X)^{-1}K^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\hat{\beta} - d)$*, we get that* $SSE_{H_0} = \hat{\varepsilon}^\top\hat{\varepsilon} + \Delta_{H_0}^\top X^\top X\Delta_{H_0}$.

(iii) We obtain the representation $\Delta SSE = (K\hat{\beta} - d)^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\hat{\beta} - d)$*. Thus,* ΔSSE *is a quadratic form.*

Proof: Due to the assumption of normally distributed error terms (see Part (c) of Model 2.1), the restricted MLE $\hat{\beta}_{H_0}$ is also the restricted LSE. We may thus argue geometrically, as in the proof of Theorem 2.5.

Let $\gamma \in \mathbb{R}^p$ be any candidate vector fulfilling the constraint $K\gamma = d$. We first see that

$$\begin{aligned}||Y - X\gamma||_2^2 &= (Y - X\gamma)^\top(Y - X\gamma) = (Y - X\hat{\beta} + X(\hat{\beta} - \gamma))^\top(Y - X\hat{\beta} + X(\hat{\beta} - \gamma)) \\ &= ||Y - X\hat{\beta}||_2^2 + (\hat{\beta} - \gamma)^\top X^\top X(\hat{\beta} - \gamma), \text{ because}\end{aligned}$$

$(\hat{\beta}-\gamma)^\top X^\top(Y-X\hat{\beta}) = (Y-X\hat{\beta})^\top X(\hat{\beta}-\gamma) = (\hat{\beta}-\gamma)^\top(X^\top Y-X^\top X(X^\top X)^{-1}X^\top Y) = 0.$
Furthermore, it holds that

$$(\hat{\beta} - \gamma)^\top X^\top X(\hat{\beta} - \gamma) = ||X(\hat{\beta} - \hat{\beta}_{H_0})||_2^2 + ||X(\hat{\beta}_{H_0} - \gamma)||_2^2,$$

because we can calculate analogously

$$\begin{aligned}||X(\hat{\beta} - \hat{\beta}_{H_0})||_2^2 &= (X(\hat{\beta} - \hat{\beta}_{H_0}))^\top X(\hat{\beta} - \hat{\beta}_{H_0}) = (X(\hat{\beta} - \gamma + \gamma - \hat{\beta}_{H_0}))^\top X(\hat{\beta} - \gamma + \gamma - \hat{\beta}_{H_0}) \\ &= \left[(\hat{\beta} - \gamma)^\top + (\gamma - \hat{\beta}_{H_0})^\top\right] X^\top X \left[(\hat{\beta} - \gamma) + (\gamma - \hat{\beta}_{H_0})\right] \\ &= (\hat{\beta} - \gamma)^\top X^\top X(\hat{\beta} - \gamma) + 2(\hat{\beta} - \gamma)^\top X^\top X(\gamma - \hat{\beta}_{H_0}) \\ &\quad + (\gamma - \hat{\beta}_{H_0})^\top X^\top X(\gamma - \hat{\beta}_{H_0}),\end{aligned}$$

and with $||X(\hat{\beta}_{H_0} - \gamma)||_2^2 = (\hat{\beta}_{H_0} - \gamma)^\top X^\top X(\hat{\beta}_{H_0} - \gamma)$ it follows that

$$\begin{aligned}&||X(\hat{\beta} - \hat{\beta}_{H_0})||_2^2 + ||X(\hat{\beta}_{H_0} - \gamma)||_2^2 \\ &= (\hat{\beta} - \gamma)^\top X^\top X(\hat{\beta} - \gamma) + 2(\hat{\beta} - \gamma)^\top X^\top X(\gamma - \hat{\beta}_{H_0}) + 2(\gamma - \hat{\beta}_{H_0})^\top X^\top X(\gamma - \hat{\beta}_{H_0}) \\ &= (\hat{\beta} - \gamma)^\top X^\top X(\hat{\beta} - \gamma) + 2(\hat{\beta} - \hat{\beta}_{H_0})^\top X^\top X(\gamma - \hat{\beta}_{H_0}).\end{aligned}$$

But now, we notice that

$$\begin{aligned}2(\hat{\beta} - \hat{\beta}_{H_0})^\top X^\top X(\gamma - \hat{\beta}_{H_0}) &= 2\left[(X^\top X)^{-1}K^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\hat{\beta} - d)\right]^\top X^\top X(\gamma - \hat{\beta}_{H_0}) \\ &= 2(K\hat{\beta} - d)^\top(K(X^\top X)^{-1}K^\top)^{-1}K(X^\top X)^{-1}(X^\top X)(\gamma - \hat{\beta}_{H_0}) \\ &= 2(K\hat{\beta} - d)^\top(K(X^\top X)^{-1}K^\top)^{-1}K(\gamma - \hat{\beta}_{H_0}) = 0.\end{aligned}$$

Altogether, we get that for all $\gamma \in \mathbb{R}^p$ with $K\gamma = d$ the identity

$$||Y - X\gamma||_2^2 = ||Y - X\hat{\beta}||_2^2 + ||X(\hat{\beta} - \hat{\beta}_{H_0})||_2^2 + ||X(\hat{\beta}_{H_0} - \gamma)||_2^2.$$

holds true. Since the first to summands on the right-hand side of the latter identity are invariant in γ and X has maximum rank, we conclude that $\hat{\beta}_{H_0}$ uniquely minimizes the sum of squares of residuals under the linear constraint given by H_0. The calculation reveals moreover, that the decomposition formula

$$||Y - X\hat{\beta}_{H_0}||_2^2 = ||Y - X\hat{\beta}||_2^2 + ||X(\hat{\beta} - \hat{\beta}_{H_0})||_2^2,$$

holds, thus orthogonality of $(Y - X\hat{\beta})$ and $(X\hat{\beta} - X\hat{\beta}_{H_0})$ (Pythagoras!).
This relationship is illustrated in Figure 2.2, which is an extension of Figure 2.1 and can be found in a similar form on the cover of the book by Lehmann and Romano (2005).

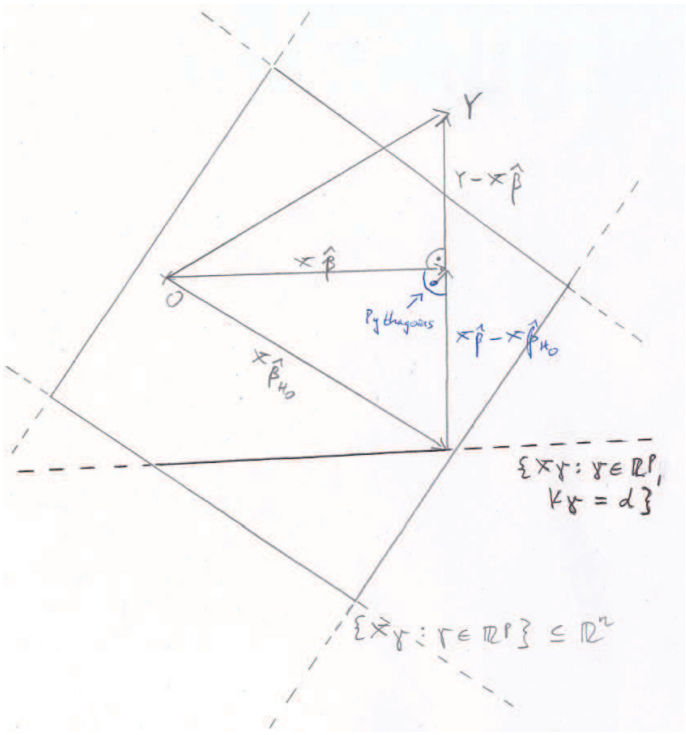


Figure 2.2: Geometric illustration of the restricted least squares estimator $\hat{\beta}_{H_0}$.

For proving Parts (ii) and (iii), we notice that $\hat{\beta}_{H_0} = \hat{\beta} - \Delta_{H_0}$ and compute

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

$$\hat{\varepsilon}_{H_0} = y - \hat{y}_{H_0} = y - \hat{y} + X\Delta_{H_0} = \hat{\varepsilon} + X\Delta_{H_0},$$

thus

$$\begin{aligned} SSE_{H_0} &= \hat{\varepsilon}_{H_0}^\top \hat{\varepsilon}_{H_0} = (\hat{\varepsilon} + X\Delta_{H_0})^\top (\hat{\varepsilon} + X\Delta_{H_0}) \\ &= \hat{\varepsilon}^\top \hat{\varepsilon} + \hat{\varepsilon}^\top X\Delta_{H_0} + \Delta_{H_0}^\top X^\top \hat{\varepsilon} + \Delta_{H_0}^\top X^\top X \Delta_{H_0} \\ &= \hat{\varepsilon}^\top \hat{\varepsilon} + \Delta_{H_0}^\top X^\top X \Delta_{H_0}, \end{aligned}$$

because $X^\top \hat{\varepsilon} = 0$ according to Part (ii) of Remark 2.6. Finally, this leads to

$$\begin{aligned} &\Delta SSE = \hat{\varepsilon}^\top \hat{\varepsilon} + \Delta_{H_0}^\top X^\top X \Delta_{H_0} - \hat{\varepsilon}^\top \hat{\varepsilon} = \Delta_{H_0}^\top X^\top X \Delta_{H_0} \\ &= \left[(X^\top X)^{-1} K^\top (K(X^\top X)^{-1} K^\top)^{-1} (K\hat{\beta} - d)\right]^\top X^\top X (X^\top X)^{-1} K^\top (K(X^\top X)^{-1} K^\top)^{-1} (K\hat{\beta} - d) \\ &= (K\hat{\beta} - d)^\top (K(X^\top X)^{-1} K^\top)^{-1} K (X^\top X)^{-1} K^\top (K(X^\top X)^{-1} K^\top)^{-1} (K\hat{\beta} - d) \\ &= (K\hat{\beta} - d)^\top (K(X^\top X)^{-1} K^\top)^{-1} (K\hat{\beta} - d), \end{aligned}$$

completing the argumentation. ■

Corollary 2.21

Under the conditions of Parts (b) and (c) of Theorem 2.19, the following assertions hold true under the linear hypothesis $H_0 : K\beta = d$.

(a) $\Delta SSE/\sigma^2 \sim \chi_r^2$

(b) $\Delta SSE \perp\!\!\!\perp SSE$. *Thus, by Part 3. of Theorem 2.16,* $F \underset{H_0}{\sim} F_{r,n-p}$.

(c) $W = rF = (n-p)\frac{\Delta SSE}{SSE}$

Proof: For proving Part (a), we use Part 1. of Theorem 2.16. To this end, we define $Z = K\hat{\beta}$. Under H_0, $\mathbb{E}[Z] = d$ and $\mathrm{Cov}(Z) = \sigma^2 K(X^\top X)^{-1} K^\top$. Since $\hat{\beta}$ is normally distributed, we even get that $Z \sim \mathcal{N}_r(d, \sigma^2 K(X^\top X)^{-1} K^\top)$, implying the assertion.

For proving Part (b), we notice that ΔSSE is a deterministic transformation of $\hat{\beta}$. Since $\hat{\beta}$ is stochastically independent of SSE, we conclude that ΔSSE is stochastically independent of SSE, too.

Finally, for proving Part(c), we calculate straightforwardly that

$$\begin{aligned} F &= \frac{n-p}{r}\frac{\Delta SSE}{SSE} = \frac{n-p}{r}\frac{(K\hat{\beta} - d)^\top (K(X^\top X)^{-1} K^\top)^{-1} (K\hat{\beta} - d)}{(n-p)\widehat{\sigma^2}} \\ &= \frac{(K\hat{\beta} - d)^\top (\widehat{\sigma^2} K(X^\top X)^{-1} K^\top)^{-1} (K\hat{\beta} - d)}{r} \\ &= \frac{(K\hat{\beta} - d)^\top (K\hat{V}K^\top)^{-1} (K\hat{\beta} - d)}{r} = \frac{W}{r}. \end{aligned}$$

■

Example 2.22 (Continuation of Example 2.18)

Let us calculate the concrete form of the F-statistic for three special test problems.

(i) *Test of (significant) influence of one particular covariate on the (expected) response:*
$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0, \quad 1 \leq j \leq k$ *fixed.*
According to Part (i) of Example 2.18, we have $K \in \mathbb{R}^{1\times p}$ *with entries* $K_i = \mathbf{1}_{\{i=j+1\}}$*, and* $d = 0$*. Plugging these specifications into the quantities considered in Corollary 2.21 yields that*

$$\Delta SSE = \frac{SSE}{n-p}\frac{(\hat{\beta}_j)^2}{\widehat{Var}(\hat{\beta}_j)}, \text{ thus}$$

$$F = (n-p)\frac{\Delta SSE}{SSE} = \frac{(\hat{\beta}_j)^2}{\widehat{Var}(\hat{\beta}_j)} \text{ as well as } F \underset{H_0}{\sim} F_{1,(n-p)}.$$

The resulting F-Test is equivalent to the two-sided t-test employing the test statistic

$$T = \frac{|\hat{\beta}_j|}{\widehat{SE}(\hat{\beta}_j)} \text{ with } \widehat{SE}(\hat{\beta}_j) := \sqrt{\widehat{Var}(\hat{\beta}_j)}.$$

(ii) *Test for the influence of a sub-vector* $\beta^* = (\beta_1, \ldots, \beta_r)^\top$*:*
Here, we employ $K \in \mathbb{R}^{r\times p}$ *with entries* $K_{i\ell} = \mathbf{1}_{\{\ell=i+1\}}$*, and* $d = 0 \in \mathbb{R}^r$*.*
This leads to

$$\Delta SSE = \frac{SSE\,(\hat{\beta}^*)^\top[\widehat{Cov}(\hat{\beta}^*)]^{-1}\hat{\beta}^*}{n-p} \text{ as well as } F = \frac{n-p}{r}\frac{\Delta SSE}{SSE} = \frac{(\hat{\beta}^*)^\top[\widehat{Cov}(\hat{\beta}^*)]^{-1}\hat{\beta}^*}{r}$$

with the corresponding null distribution $F \underset{H_0}{\sim} F_{r,(n-p)}$*.*

(iii) *Global test:*
Consider the test problem $H_0 : \beta_j = 0$ *for all* $1 \leq j \leq k$ *versus* $H_1 : \exists j \in \{1, \ldots, k\} : \beta_j \neq 0$*.*
In this case, $SSE_{H_0} = SST = \sum_{i=1}^n (Y_i - \overline{Y})^2$*, and by decomposition of spread*

$$\Delta SSE = SSE_{H_0} - SSE = SST - SSE = SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2.$$

$$\Rightarrow F = \frac{n-p}{k}\frac{\Delta SSE}{SSE} = \frac{n-p}{k}\frac{SSR}{SSE} = \frac{n-p}{k}\frac{R^2}{1-R^2} \text{ and } F \underset{H_0}{\sim} F_{k,(n-p)}.$$

Remark 2.23

F-Tests can equivalently be carried out as Hotelling's T^2*-tests. Namely, the following relationship holds true: If* $F \sim F_{r,s}$*, then* $\frac{r(s+r-1)}{s}F \sim T^2(r, s+r-1)$*.*
(Hotelling's T^2*-distribution, cf. Hotelling (1931))*

Corollary 2.24

By the correspondence theorem, Theorem 2.20, Corollary 2.21, and Example 2.22 entail the following confidence statements.

1) For fixed $1 \leq j \leq k$, a $(1-\alpha)$-confidence interval for β_j is given by

$$\left[\hat{\beta}_j - t_{n-p;1-\alpha/2} \cdot \widehat{SE}(\hat{\beta}_j) \ , \ \hat{\beta}_j + t_{n-p;1-\alpha/2} \cdot \widehat{SE}(\hat{\beta}_j)\right].$$

2) A confidence ellipsoid for a sub-vector $\beta^* = (\beta_1, \ldots, \beta_r)^\top$ at confidence level $(1-\alpha)$ is given by

$$\left\{\gamma \in \mathbb{R}^r : \quad (\hat{\beta}^* - \gamma)^\top \left[\widehat{Cov}(\hat{\beta}^*)\right]^{-1} (\hat{\beta}^* - \gamma) \leq r \cdot F_{r,n-p;1-\alpha}\right\}.$$

Further implications are:

3) Consider a future observation Y_0 with pertaining profile of covariates $\vec{X}_0 = \vec{x}_0$. Then, a $(1-\alpha)$-confidence interval for $\mu_0 := \mathbb{E}\big[Y_0 \mid \vec{X}_0 = \vec{x}_0\big]$ is given by

$$\left[\vec{x}_0\hat{\beta} - t_{n-p;1-\alpha/2} \cdot \hat{\sigma}\sqrt{\vec{x}_0(X^\top X)^{-1}\vec{x}_0^\top} \ , \ \vec{x}_0\hat{\beta} + t_{n-p;1-\alpha/2} \cdot \hat{\sigma}\sqrt{\vec{x}_0(X^\top X)^{-1}\vec{x}_0^\top}\right].$$

4) Under the circumstances of Part 3), a $(1-\alpha)$-prognosis interval for the response value y_0 itself is given by

$$\left[\vec{x}_0\hat{\beta} - t_{n-p;1-\alpha/2} \cdot \hat{\sigma}\sqrt{1 + \vec{x}_0(X^\top X)^{-1}\vec{x}_0^\top} \ , \ \vec{x}_0\hat{\beta} + t_{n-p;1-\alpha/2} \cdot \hat{\sigma}\sqrt{1 + \vec{x}_0(X^\top X)^{-1}\vec{x}_0^\top}\right].$$

Attention: We have defined profiles of covariates as row vectors!

Remark 2.25

The multivariate central limit theorem (Theorem 2.15) yields, that without the additional assumption in Part (c) of Model 2.1 we have the asymptotic / approximate distributional result that

$$\hat{\beta}(n) \underset{approx.}{\sim} \mathcal{N}_p(\beta, \widehat{\sigma^2}_n (X_n^\top X_n)^{-1}).$$

Hence, all results concerning tests and confidence regions, which we have derived under the working assumption of normally distributed error terms, stay at least approximately correct for large sample sizes (under suitable regularity assumptions) if non-Gaussian errors have to be considered. For small to moderate sample sizes, it is often advisable to approximate the null distribution of certain test statistics of interest by means of resampling methods, if the additional assumption in Part (c) of Model 2.1 is prone to be violated; cf., e. g., Chapter 6 of Dickhaus (2018), which also contains some considerations for models with random design.

2.2 Analysis of variance (ANOVA)

Model 2.26 (One-factorial design, ANOVA1)

We assume that we can observe response values $(y_{ij})_{\substack{1 \leq i \leq k, \\ 1 \leq j \leq n_i}}$, which we model as realizations of

jointly stochastically independent random variables $(Y_{ij})_{\substack{1\leq i\leq k,\\ 1\leq j\leq n_i}}$.
In this, we assume that $\forall 1 \leq i \leq k : \forall 1 \leq j \leq n_i : Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$.

We call the first dimension (corresponding to the index i*) the factor and the value of* $1 \leq i \leq k$ *the factor level. The integers* $(n_i)_{1\leq i\leq k}$ *denote the numbers of independent repetitions of the experiment per factor level. We can write the model in matrix form by considering the equivalent formulation*

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \forall 1 \leq i \leq k, \forall 1 \leq j \leq n_i,$$

with i.i.d. error terms $(\varepsilon_{ij})_{\substack{1\leq i\leq k,\\ 1\leq j\leq n_i}}$ *fulfilling that* $\varepsilon_{11} \sim \mathcal{N}(0, \sigma^2)$.
Hence, we have a regression model with exactly one categorical covariate, which is of the form

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix}.$$

Its matrix form is given by

$$Y = X\mu + \varepsilon,$$

where $\mu = (\mu_1, \dots, \mu_k)^\top$ *is the parameter vector and* $\varepsilon \sim \mathcal{N}_{n_\bullet}(0, \sigma^2 I_{n_\bullet})$ *mit* $n_\bullet := \sum_{i=1}^k n_i$ *is the stochastic component of the model. The design matrix of the model is given by*

$$X = (x_{ji})_{\substack{1\leq j\leq n_\bullet\\ 1\leq i\leq k}} \text{ with } x_{ji} = \mathbf{1}\{\text{observational unit } j \text{ belongs to factor level } i\}.$$

In the special case of $n_1 = \dots = n_k =: n$*, we call the model ANOVA1 model with balanced design.*

Notice that the model formulation in Model 2.26 has no intercept. It holds that $\operatorname{rank}(X) = k$, because the k columns of X are linearly independent.

The classical question of the ANOVA is: Do there exist (any) differences in the factor level-specific (theoretical) means μ_i, $1 \leq i \leq k$, or not? This has the interpretation of whether the factor has an influence on the (expected) response or not.
Mathematical formulation as a test problem:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k \text{ versus } H_1 : \exists 1 \leq i \neq \ell \leq k : \mu_i \neq \mu_\ell \tag{2.7}$$

Apart from this problem, other questions may be of interest, too, e. g.:
(MCA) Testing all pairwise mean differences
(MCB) Comparing the factor level-specific means with the (empirically) "best"

(referring to the largest empirical mean)
(MCC) Comparing the factor level-specific means with the mean of a pre-defined control group
These are classical multiple test problems, which are treated, for instance, in the textbook by Hochberg and Tamhane (1987).

Theorem 2.27 (Sum of squares decomposition)
Define the following quantities.

$$\forall 1 \leq i \leq k: \quad \overline{Y}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij} \qquad \text{(Empirical group means)}$$

$$\overline{Y}_{..} = n_\bullet^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} \qquad \text{(Grand mean)}$$

$$SSB = \sum_{i=1}^{k} n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2 \qquad \text{(Sum of squares between groups)}$$

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2 \qquad \text{(Sum of squares within groups)}$$

Then it holds: $SST = SSB + SSW$.

Proof: We calculate straightforwardly, that

$$\begin{aligned} SST &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{..})^2 \\ &= \sum_i \sum_j (Y_{ij} - \overline{Y}_{i.} + \overline{Y}_{i.} - \overline{Y}_{..})^2 \\ &= \sum_i \sum_j \left[(Y_{ij} - \overline{Y}_{i.})^2 + 2(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..}) + (\overline{Y}_{i.} - \overline{Y}_{..})^2 \right]. \end{aligned}$$

But now, we have that

$$\begin{aligned} \sum_i \sum_j (Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..}) &= \sum_i (\overline{Y}_{i.} - \overline{Y}_{..}) \sum_j (Y_{ij} - \overline{Y}_{i.}) \\ &= \sum_i (\overline{Y}_{i.} - \overline{Y}_{..})(n_i \overline{Y}_{i.} - n_i \overline{Y}_{i.}) = 0, \end{aligned}$$

which implies the assertion. ■

Obviously, evidence against the null hypothesis is gained if the spread between the groups is (much) bigger than the spread within the groups. This motivates the usage of the (scaled) ratio of SSB and SSW as test statistic for testing H_0 from (2.7).

Theorem 2.28
Utilizing the notation introduced in Theorem 2.27, the following assertions hold true.

(i) $SSW/\sigma^2 \sim \chi^2_{n_\bullet - k}$

(ii) Under H_0, $SSB/\sigma^2 \sim \chi^2_{k-1}$

(iii) SSW *is stochastically independent of* SSB.

(iv) Under H_0, *the statistic* $F = \frac{SSB/(k-1)}{SSW/(n_\bullet - k)}$ *possesses the* $F_{k-1,n_\bullet-k}$*-distribution.*

Hence, a level α *test for testing* (2.7) *is given by the following decision rule: Reject* H_0, *if the observed value of the* F*-statistic defined in Part (iv) exceeds* $F_{k-1,n_\bullet-k;1-\alpha}$ *(the* $(1-\alpha)$*-quantile of the null distribution).*

Proof: This is an exercise. ■

Definition 2.29 (Effect coding)
Under Model 2.26, define $\mu_0 := \mathbb{E}\left[\overline{Y}_{\cdot\cdot}\right] = n_\bullet^{-1} \sum_{i=1}^k \sum_{j=1}^{n_i} \mu_i = n_\bullet^{-1} \sum_{i=1}^k n_i \mu_i$ *and* $\alpha_i := \mu_i - \mu_0$ *for all* $1 \leq i \leq k$. *This introduces an "intercept"* μ_0 *into the ANOVA1 model. In terms of the quantities just defined, the model equations are given by*

$$\forall 1 \leq i \leq k: \quad \forall 1 \leq j \leq n_i: \quad Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij} \tag{2.8}$$

with homoscedastic, stochastically independent, centered, normally distributed error terms $(\varepsilon_{ij})_{\substack{1 \leq i \leq k, \\ 1 \leq j \leq n_i}}$.
In this, it is important to take account of the constraint that $\sum_{i=1}^k n_i \alpha_i = 0$, *in order to ensure maximum rank of the resulting design matrix (cf. the corresponding exercise). We call the representation in* (2.8) *the effect coding of the ANOVA1 model, and we call* α_i *the effect of factor level* i, *for* $1 \leq i \leq k$.
Convention:
To encode the constraint that $\sum_{i=1}^k n_i \alpha_i = 0$ *in the design matrix, we eliminate the superfluous parameter* α_k *by writing* $\alpha_k := -n_k^{-1} \sum_{i=1}^{k-1} n_i \alpha_i$.
With these specifications, the matrix form of the effect coding is given by

$$Y = X \, (\mu_0, \alpha_1, \ldots, \alpha_{k-1})^\top + \varepsilon$$

or

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{k-1,1} \\ \vdots \\ Y_{k-1,n_{k-1}} \\ Y_{k,1} \\ \vdots \\ Y_{k,n_k} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 \\ 1 & -\frac{n_1}{n_k} & -\frac{n_2}{n_k} & -\frac{n_\ell}{n_k} & \dots & -\frac{n_{k-1}}{n_k} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & -\frac{n_1}{n_k} & -\frac{n_2}{n_k} & -\frac{n_\ell}{n_k} & \dots & -\frac{n_{k-1}}{n_k} \end{bmatrix} \begin{pmatrix} \mu_0 \\ \alpha_1 \\ \vdots \\ \alpha_{k-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_{k,n_k} \end{pmatrix},$$

respectively, with the k parameters μ_0, $(\alpha_i)_{1\leq i\leq k-1}$.

Theorem 2.30

Under the specifications of Definition 2.29, the following assertions hold true.

(i) The least squares estimators (or, equivalently, the MLEs) for the unknown parameters are given by

$$\hat{\mu}_0 = \overline{Y}_{\cdot\cdot} \qquad \textit{and} \qquad \hat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot},\ 1 \leq i \leq k.$$

(ii) The F-statistic for testing the global hypothesis H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_{k-1} = 0$ coincides with the F-statistic given in Part (iv) of Theorem 2.28, i. e., $F = \frac{SSB/(k-1)}{SSW/(n_\bullet - k)}$.

Proof: For proving Part (i), we consider the (over-parametrized) design matrix

$$X_{(k+1)} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} \in \{0,1\}^{n_\bullet \times (k+1)} \text{ with } \operatorname{rank}(X_{(k+1)}) = k,$$

and we solve the system $X_{(k+1)}^\top X_{(k+1)}\ (\mu_0, \alpha_1, \ldots, \alpha_k)^\top = X_{(k+1)}^\top Y$ of normal equations under the linear restriction that $\sum_{i=1}^k n_i \alpha_i = 0$. We obtain that

$$X_{(k+1)}^\top X_{(k+1)} = \begin{pmatrix} n_\bullet & n_1 & n_2 & \ldots & n_{k-1} & n_k \\ n_1 & n_1 & 0 & \ldots & \ldots & 0 \\ n_2 & 0 & n_2 & 0 & \ldots & 0 \\ \vdots & 0 & \ldots & \ddots & \ldots & 0 \\ n_{k-1} & 0 & \ldots & 0 & n_{k-1} & 0 \\ n_k & 0 & \ldots & \ldots & 0 & n_k \end{pmatrix}.$$

This entails the normal equations

$$n_\bullet \mu_0 + \sum_{i=1}^k n_i \alpha_i = \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} \quad \text{and}$$

$$\forall 1 \leq i \leq k : n_i \mu_0 + n_i \alpha_i = \sum_{j=1}^{n_i} Y_{ij}.$$

Plugging the linear restriction into the first equation yields

$$n_\bullet \hat{\mu}_0 = \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} \Leftrightarrow \hat{\mu}_0 = \overline{Y}_{..},$$

thus

$$\forall 1 \leq i \leq k : \ n_i \left(\overline{Y}_{..} + \hat{\alpha}_i \right) = \sum_{j=1}^{n_i} Y_{ij} \Leftrightarrow \hat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{..},$$

as desired.

For proving Part (ii), notice that $SSE_{H_0} = SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{..})^2$ as well as

$$\begin{aligned} SSE &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_0 - \hat{\alpha}_i)^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} [Y_{ij} - \overline{Y}_{..} - (\overline{Y}_{i.} - \overline{Y}_{..})]^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2 = SSW \end{aligned}$$

hold true. By the sum of squares decomposition, this yields that $\Delta SSE = SSE_{H_0} - SSE = SST - SSW = SSB$, and Part (b) of Theorem 2.19 leads to $F = \frac{SSB/(k-1)}{SSW/(n_\bullet - k)}$, because here $r = k - 1$ (number of restrictions), $\dim(Y) = n_\bullet$ (total sample size), and the number of columns of the design matrix with full rank equals k. ■

Remark 2.31

The effect representation of the ANOVA1 model, given by

$$\forall 1 \leq i \leq k : \forall 1 \leq j \leq n_i : \quad Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij}, \tag{2.9}$$

does not fulfill the conditions of Model 2.1 in the first place, because the corresponding design matrix does not have full column rank. Only by taking into account the intrinsic constraint that $\sum_{i=1}^{k} n_i \alpha_i = 0$*, which results from the definition of* $(\alpha_i)_{1 \leq i \leq k}$*, leads to a proper regression model that fulfills the conditions of Model 2.1. Starting directly from the specification* (2.9)*, it is also possible to impose other constraints which guarantee full column rank of the design matrix. For instance, one may consider constraints of the type* $\sum_{i=1}^{k} c_i \alpha_i = c^\top \alpha = 0$*, where* $\sum_{i=1}^{k} c_i \neq 0$*. If one enforces orthogonality of the columns of the design matrix, the resulting representation is called the contrast coding. In summary, the following codings may be considered.*

1. *Dummy coding:*
   $x_{ji} = 1$*, if observational unit* j *belongs to factor level* i*, and* $x_{ji} = 0$ *otherwise, for* $1 \leq j \leq n_\bullet$ *and* $1 \leq i \leq k$*.*

2. *Effect coding:*
   $$\forall 1 \leq j \leq n_\bullet : x_{j1} = 1 \text{ and } x_{ji} = \begin{cases} 1, & \text{if factor level } i-1 \text{ applies}, \\ -\frac{n_{i-1}}{n_k}, & \text{if factor level } k \text{ applies}, \\ 0, & \text{otherwise,} \end{cases}$$
   for $2 \leq i \leq k$*. In this, we say that factor level* i *applies (in row* j*), if observational unit* j *belongs to factor level* i*.*

3. *Contrast coding:*
   Obtained by orthogonalization of the design matrix.

In all three cases, the resulting design matrix $X = (x_{ji})_{\substack{1 \leq j \leq n_\bullet, \\ 1 \leq i \leq k}}$ *is an* $(n_\bullet \times k)$*-matrix.*

We now turn to two-factorial designs.

Model 2.32 (Two-factorial design, ANOVA2)

We assume that we can observe response values $(y_{ijk})_{\substack{1 \leq i \leq I, \\ 1 \leq j \leq J, \\ 1 \leq k \leq n}}$*. In particular, we restrict our attention to balanced designs. We model these observations as realizations of* (Y_{ijk})*, and make the model assumption that*

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}\,, \quad 1 \leq i \leq I,\ 1 \leq j \leq J,\ 1 \leq k \leq n.$$

The total sample size is given by $n_{\bullet\bullet} := I \cdot J \cdot n$*. The first two dimensions are called first and second factor, respectively, with factor levels* $1 \leq i \leq I$ *for the first factor and* $1 \leq j \leq J$ *for the second factor. The third index* $1 \leq k \leq n$ *runs over the* n *repetitions per factor level combination. For the*

stochastic part of the model, we assume that all error terms (ε_{ijk}) *are i.i.d. with* $\varepsilon_{111} \sim \mathcal{N}(0, \sigma^2)$. *The ANOVA2 model is a multiple linear regression model with exactly two categorical covariates. The dimensionality of the pertaining parameter vectors is given by*

$$dim((\mu_{11}, \mu_{12}, \ldots, \mu_{1J}, \mu_{21}, \ldots, \mu_{I1}, \ldots, \mu_{IJ})^\top) = I \cdot J.$$

Example 2.33 (Schuchard-Ficher et al. (1980), Page 30)
Response: Sold quantity units of a certain margarine brand in different supermarkets
Factor 1: Price policy ("low price", "normal price", "high price")
Factor 2: Communication strategy ("direct mail", "press advertising")
Experimental plan: Six randomly chosen supermarkets, one supermarket per factor level combination, ten randomly chosen working days

This leads to $I = 3$ *(price policy strata),* $J = 2$ *(communication strategy strata), and* $n = 10$ *(repetitions per factor level combination).*

Definition and Lemma 2.34 (Effect coding of the balanced ANOVA2 model)
Under Model 2.32, we define

$$\mu_{i\bullet} := J^{-1} \sum_{j=1}^{J} \mu_{ij}\,, \quad 1 \le i \le I; \tag{2.10}$$

$$\mu_{\bullet j} := I^{-1} \sum_{i=1}^{I} \mu_{ij}\,, \quad 1 \le j \le J; \tag{2.11}$$

$$\mu_0 := (IJ)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}. \tag{2.12}$$

With these definitions, the following assertions hold true.

(i) We obtain the decomposition

$$\begin{aligned} \mu_{ij} &= \mu_0 + (\mu_{i\bullet} - \mu_0) + (\mu_{\bullet j} - \mu_0) + (\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_0) \\ &=: \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij}; \quad 1 \le i \le I,\ 1 \le j \le J \end{aligned}$$

as well as the effect representation

$$Y_{ijk} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}. \tag{2.13}$$

The parameter vector corresponding to the effect coding is given by

$$\tilde{\theta} := (\mu_0, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, (\alpha\beta)_{11}, \ldots, (\alpha\beta)_{IJ})^\top \in \mathbb{R}^{1+I+J+IJ}.$$

From (2.10) - (2.12) *it is apparent, that there exist* $(I + J + 1)$ *restrictions between the components of* $\tilde{\theta}$. *Thus, the over-parametrized* $[(I \cdot J \cdot n) \times (1 + I + J + IJ)]$*-design matrix*

pertaining to the effect coding has column rank $(I \cdot J)$, which is the same column rank as for the dummy coding. In particular, the design matrix pertaining to the effect coding does not have full column rank.

(ii) It holds that

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0.$$

The parameters $(\alpha_i)_{1\le i\le I}$ are called main effects of the first factor, the parameters $(\beta_j)_{1\le j\le J}$ are called main effects of the second factor, and the parameters $((\alpha\beta)_{ij})_{\substack{1\le i\le I,\\ 1\le j\le J}}$ are called interaction effects.

It does not (!!) necessarily hold that $(\alpha\beta)_{ij} = \alpha_i\beta_j$.

Proof: Part (i) is obvious. For proving Part (ii), we calculate straightforwardly, that

$$\begin{aligned}
\sum_{i=1}^{I} \alpha_i &= \sum_{i=1}^{I} (\mu_{i\bullet} - \mu_0) \\
&= \sum_{i=1}^{I} \left[J^{-1} \sum_{j=1}^{J} \mu_{ij} - (IJ)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij} \right] \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\mu_{ij}}{J} - I^{-1} \sum_{i=1}^{I} \left[\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\mu_{ij}}{J} \right] = 0,
\end{aligned}$$

because $\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}/J$ is a constant.

The other assertions of Part (ii) follow analogously. ■

Theorem 2.35

Under the conditions of Definition and Lemma 2.34, let

$$\theta := (\mu_0, \alpha_1, \ldots, \alpha_{I-1}, \beta_1, \ldots, \beta_{J-1}, (\alpha\beta)_{11}, \ldots, (\alpha\beta)_{(I-1)(J-1)})^\top \in \mathbb{R}^{IJ}.$$

Then, the following assertions hold true.

(a) The model equation of the ANOVA2 model with balanced design can equivalently be written as

$$Y = X\theta + \varepsilon \quad \textit{with}$$

$$Y = (Y_{111}, \ldots, Y_{IJn})^\top \in \mathbb{R}^{n_{\bullet\bullet}},$$

$$\varepsilon = (\varepsilon_{111}, \ldots, \varepsilon_{IJn})^\top \in \mathbb{R}^{n_{\bullet\bullet}},$$

and design matrix $X \in \mathbb{R}^{n_{\bullet\bullet} \times (I\cdot J)}$. Its entries $(X_{rs})_{\substack{1\le r\le n_{\bullet\bullet},\\ 1\le s\le I\cdot J}}$ are given by the following construction. We partition X into four sub-matrices $X_{\mu_0}, X_\alpha, X_\beta$, and $X_{(\alpha\beta)}$ with $n_{\bullet\bullet}$ rows each. Namely,

[1] X_{μ_0} *is a column vector, containing only ones (intercept column).*

[2] X_α *is an* $(n_{\bullet\bullet} \times (I-1))$*-matrix, whose columns correspond to the main effects of the first factor. In column* $1 \leq i \leq I-1$ *of* X_α*, it holds that*

$$X_\alpha^{(k,i)} = +1, \quad \textit{if observational unit } k \textit{ belongs to factor level } i,\ 1 \leq k \leq n_{\bullet\bullet},$$

$$X_\alpha^{(k,i)} = -1, \quad \textit{if observational unit } k \textit{ belongs to factor level } I,$$

$$X_\alpha^{(k,i)} = 0, \quad \textit{otherwise}.$$

[3] X_β *corresponds in analoguous manner to the main effects of the second factor.*

[4] $X_{(\alpha\beta)}$ *corresponds to the* $(I-1)\cdot(J-1)$ *interaction effects. For its entries, it holds (in self-explaining notation):* $X_{(\alpha\beta)}^{(k,ij)} = X_\alpha^{(k,i)} \cdot X_\beta^{(k,j)}, \quad 1 \leq i \leq I-1,$ $1 \leq j \leq J-1,\ 1 \leq k \leq n_{\bullet\bullet}$*. Hence,* X *has the following structure.*

$$X = \begin{array}{c} \\ 111 \\ \vdots \\ 11n \\ \vdots \\ 1J1 \\ \vdots \\ 1Jn \\ \vdots \\ IJ1 \\ \vdots \\ IJn \end{array} \begin{array}{c} \begin{array}{cccccccccc} \mu_0 & \alpha_1 & \dots & \alpha_{I-1} & \beta_1 & \dots & \beta_{J-1} & (\alpha\beta)_{11} & \dots & (\alpha\beta)_{(I-1),(J-1)} \end{array} \\ \begin{pmatrix} 1 & 1 & \dots & 0 & 1 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 0 & 1 & \dots & 0 & 1 & \dots & 0 \\ & \vdots & & & & & & & & \vdots \\ 1 & 1 & \dots & 0 & -1 & \dots & -1 & -1 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 0 & -1 & \dots & -1 & -1 & \dots & 0 \\ & \vdots & & & & & & & & \vdots \\ 1 & -1 & \dots & -1 & -1 & \dots & -1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & -1 & \dots & -1 & -1 & \dots & -1 & 1 & \dots & 1 \end{pmatrix} \\ \underbrace{}_{X_{\mu_0}} \ \underbrace{}_{X_\alpha} \ \underbrace{}_{X_\beta} \ \underbrace{}_{X_{(\alpha\beta)}} \end{array}$$

(b) The matrix X *has full column rank, and it is block-orthogonal in the sense that columns, which originate from different sub-matrices* $\{X_{\mu_0}, X_\alpha, X_\beta, X_{(\alpha\beta)}\}$*, are pairwise orthogonal to each other.*

Proof: Due to Part (ii) of Lemma 2.34, we have that $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$, $\beta_J = -\sum_{j=1}^{J-1} \beta_j$, and $\forall 1 \leq i \leq I-1 : \forall 1 \leq j \leq J-1 : \quad (\alpha\beta)_{Ij} = -\sum_{i=1}^{I-1}(\alpha\beta)_{ij}$, $(\alpha\beta)_{iJ} = -\sum_{j=1}^{J-1}(\alpha\beta)_{ij}$. Furthermore,

$$\begin{aligned} 0 &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\alpha\beta)_{ij} \\ &= \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}(\alpha\beta)_{ij} + \sum_{i=1}^{I}(\alpha\beta)_{iJ} + \sum_{j=1}^{J}(\alpha\beta)_{Ij} - (\alpha\beta)_{IJ} \\ &= \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}(\alpha\beta)_{ij} - (\alpha\beta)_{IJ} \Leftrightarrow (\alpha\beta)_{IJ} = \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}(\alpha\beta)_{ij}. \end{aligned}$$

With these identities, the representation claimed in Part (a) can be verified by means of a case distinction with respect to (i, j). Namely, one verifies that the model equations resulting from the representation claimed in Part (a) coincide with the model equations resulting from (2.13), by taking into consideration the aforementioned restrictions.

For proving Part (b), it suffices to show block-orthogonality of X. To this end, notice that due to balanced design the number of entries equal to "$+1$" equals the number of entries equal to "-1" in every column of X_α, X_β, and $X_{(\alpha\beta)}$, respectively. Thus, all these columns are orthogonal to the column vector X_{μ_0}. Moreover, we get that for any arbitrary combination (i, j) of main effects the inner column product

$$[X_\alpha^{(i)}]^\top X_\beta^{(j)} = \underbrace{n}_{\text{Stratum } (i,j)} - \underbrace{n}_{\text{Stratum } (i,J)} - \underbrace{n}_{\text{Stratum } (I,j)} + \underbrace{n}_{\text{Stratum } (I,J)} = 0.$$

The analogous consideration for a combination of an arbitrary main effect and an arbitrary interaction effect completes the argumentation. ■

Utilizing the block-orthogonality property of the design matrix X, which we have shown in Part (b) of Theorem 2.35, we can write the model equation of the ANOVA2 model with balanced design as follows:

$$Y = (X_{\mu_0} \quad X_\alpha \quad X_\beta \quad X_{(\alpha\beta)}) \begin{pmatrix} \mu_0 \\ \alpha \\ \beta \\ (\alpha\beta) \end{pmatrix} + \varepsilon,$$

with self-explaining vectors α, β, and $(\alpha\beta)$. This facilitates the estimation of the model parameters.

Theorem 2.36 (Parameter estimation for ANOVA2 under balanced design)
Under an ANOVA2 model with balanced design, the LSEs and MLEs, respectively, for the parameters μ_0, $\alpha = (\alpha_1, \ldots, \alpha_{I-1})^\top$, $\beta = (\beta_1, \ldots, \beta_{J-1})^\top$, and $(\alpha\beta) = ((\alpha\beta)_{ij})_{\substack{1 \leq i \leq I-1, \\ 1 \leq j \leq J-1}}$, are given

by

$$\hat{\mu}_0 = \overline{Y}_{\bullet\bullet\bullet} = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} Y_{ijk},$$

$$\hat{\alpha} = (X_\alpha^\top X_\alpha)^{-1} X_\alpha^\top Y = X_\alpha^+ Y,$$

$$\hat{\beta} = (X_\beta^\top X_\beta)^{-1} X_\beta^\top Y = X_\beta^+ Y,$$

$$\widehat{(\alpha\beta)} = (X_{(\alpha\beta)}^\top X_{(\alpha\beta)})^{-1} X_{(\alpha\beta)}^\top Y = X_{(\alpha\beta)}^+ Y.$$

Proof: Utilizing the partitioning $X = (X_{\mu_0} \quad X_\alpha \quad X_\beta \quad X_{(\alpha\beta)})$, the system

$$X^\top X(\mu_0, \alpha^\top, \beta^\top, (\alpha\beta)^\top)^\top = X^\top Y$$

of normal equations can be written as

$$\begin{pmatrix} X_{\mu_0}^\top X_{\mu_0} & X_{\mu_0}^\top X_\alpha & X_{\mu_0}^\top X_\beta & X_{\mu_0}^\top X_{(\alpha\beta)} \\ X_\alpha^\top X_{\mu_0} & X_\alpha^\top X_\alpha & X_\alpha^\top X_\beta & X_\alpha^\top X_{(\alpha\beta)} \\ X_\beta^\top X_{\mu_0} & X_\beta^\top X_\alpha & X_\beta^\top X_\beta & X_\beta^\top X_{(\alpha\beta)} \\ X_{(\alpha\beta)}^\top X_{\mu_0} & X_{(\alpha\beta)}^\top X_\alpha & X_{(\alpha\beta)}^\top X_\beta & X_{(\alpha\beta)}^\top X_{(\alpha\beta)} \end{pmatrix} \begin{pmatrix} \mu_0 \\ \alpha \\ \beta \\ (\alpha\beta) \end{pmatrix} = \begin{pmatrix} X_{\mu_0}^\top Y \\ X_\alpha^\top Y \\ X_\beta^\top Y \\ X_{(\alpha\beta)}^\top Y \end{pmatrix}.$$

Due to block-orthogonality of X, the system reduces to

$$X_{\mu_0}^\top X_{\mu_0} \mu_0 = X_{\mu_0}^\top Y,$$

$$X_\alpha^\top X_\alpha \alpha = X_\alpha^\top Y,$$

$$X_\beta^\top X_\beta \beta = X_\beta^\top Y,$$

$$X_{(\alpha\beta)}^\top X_{(\alpha\beta)} (\alpha\beta) = X_{(\alpha\beta)}^\top Y.$$

Hence, the four parameter estimation problems are de-coupled and the claimed representations of $\hat{\alpha}$, $\hat{\beta}$, and $\widehat{(\alpha\beta)}$ follow, because the involved sub-design matrices have full column rank each. Noticing that X_{μ_0} is a column vector with all its entries equal to one, we get for $\hat{\mu}_0$, that $n_{\bullet\bullet}\hat{\mu}_0 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} Y_{ijk}$, leading to $\hat{\mu}_0 = \overline{Y}_{\bullet\bullet\bullet}$, completing the proof. ■

Corollary and Definition 2.37

Under Model 2.32, let us define

(i) $\overline{Y}_{\bullet\bullet\bullet} = n_{\bullet\bullet}^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} Y_{ijk}$,

(ii) $\forall 1 \le i \le I: \quad \overline{Y}_{i\bullet\bullet} = (Jn)^{-1} \sum_{j=1}^{J} \sum_{k=1}^{n} Y_{ijk}$,

(iii) $\forall 1 \le j \le J: \quad \overline{Y}_{\bullet j\bullet} = (In)^{-1} \sum_{i=1}^{I} \sum_{k=1}^{n} Y_{ijk}$,

(iv) $\forall 1 \le i \le I: \forall 1 \le j \le J: \quad \overline{Y}_{ij\bullet} = n^{-1} \sum_{k=1}^{n} Y_{ijk}$.

Then it follows from Theorem 2.36, that LSEs and MLEs, respectively, for the parameters of the ANOVA2 in effect representation are given by

$$\begin{aligned}
\hat{\mu}_0 &= \overline{Y}_{\bullet\bullet\bullet}, \\
\forall 1 \leq i \leq I-1: \quad \hat{\alpha}_i &= \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}, \\
\forall 1 \leq j \leq J-1: \quad \hat{\beta}_j &= \overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}, \\
\forall 1 \leq i \leq I-1: \quad \forall 1 \leq j \leq J-1: \; \widehat{(\alpha\beta)}_{ij} &= \overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet};
\end{aligned}$$

cf. the definition in Part (i) of Definition and Lemma 2.34.

Finally, we consider the test theory under an ANOVA2 model with balanced design. In this, it is common practice to first test the interaction effects for statistical significance. The reason is, that the main effects are better to interpret if the interaction effects are removed from the model, and it is common practice to perform this removal if the null hypothesis of zero interaction effects is not rejected.

Theorem 2.38 (Significance testing for interaction effects)
Under Model 2.32, consider the linear hypothesis $H_{(\alpha\beta)}: (\alpha\beta)_{ij} = 0 \quad \forall 1 \leq i \leq I-1$, $\forall 1 \leq j \leq J-1$. *Then, the following assertions hold true.*

(a) The null hypothesis $H_{(\alpha\beta)}$ *can be encoded as* $K\theta = 0$, *with* θ *as in Theorem 2.35. The contrast matrix is given by*

$$K = \begin{pmatrix} \underbrace{\begin{matrix} 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{matrix}}_{I+J-1} & \underbrace{\begin{matrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & 0 & \dots & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 \end{matrix}}_{(I-1)(J-1)} \end{pmatrix}$$

with rank$(K) = (I-1)(J-1)$.

(b) For the residual sum of squares, we get that

$$\begin{aligned}
SSE_{H_{(\alpha\beta)}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2, \\
\Delta SSE_{H_{(\alpha\beta)}} &= SSE_{H_{(\alpha\beta)}} - SSE \\
&= n\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2
\end{aligned}$$

(c) For the test statistic, we get that

$$F_{(\alpha\beta)} = \frac{\Delta SSE_{H_{(\alpha\beta)}}/[(I-1)(J-1)]}{SSE/[IJ(n-1)]}$$
$$= \frac{n\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2/[(I-1)(J-1)]}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2/[IJ(n-1)]}$$

is distributed as $F_{(I-1)(J-1),IJ(n-1)}$ under $H_{(\alpha\beta)}$.

Proof: Part (a) is obvious.

The representation of $SSE_{H_{(\alpha\beta)}}$ claimed in Part (b) follows from Theorem 2.36 and Corollary 2.37, because $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\mu}_0$ are invariant with respect to the validity of $H_{(\alpha\beta)}$).

It remains to verify the representation of $\Delta SSE_{H_{(\alpha\beta)}}$. To this end, notice that Corollary 2.37 entails that

$$\begin{aligned} SSE &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \hat{\mu}_0 - \hat{\alpha}_i - \hat{\beta}_j - \widehat{(\alpha\beta)}_{ij})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}\{Y_{ijk} - \overline{Y}_{\bullet\bullet\bullet} - (\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}) - (\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}) \\ &\qquad -(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})\}^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2. \end{aligned}$$

Thus, SSE has $IJ(n-1)$ degrees of freedom. Utilizing the latter representation of SSE, we furthermore get that

$$\begin{aligned} SSE_{H_{(\alpha\beta)}} - SSE &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}\{(Y_{ijk} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2 - (Y_{ijk} - \overline{Y}_{ij\bullet})^2\} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2 \\ &\quad + 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})(\overline{Y}_{\bullet\bullet\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{ij\bullet}) \\ &= n\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2, \end{aligned}$$

as desired. For the second equality, we have used that $a^2 - b^2 = (a-b)^2 + 2b(a-b)$ for real numbers a and b.

Finally, Part (c) follows from the general test theory of multiple linear regression by setting $r \mathrel{\widehat{=}} (I-1)(J-1)$, $p \mathrel{\widehat{=}} IJ$, and $n_{\bullet\bullet} - p \mathrel{\widehat{=}} IJ(n-1)$. ■

Remark 2.39

In an analogous manner, one can show the following results for testing the main effects.

(a) For testing the linear hypothesis $H_\alpha : \alpha_i = 0$ *for all* $1 \leq i \leq I-1$,

$$SSE_{H_\alpha} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2 + Jn\sum_{i=1}^{I}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$$

and thus

$$F_\alpha = \frac{Jn\sum_{i=1}^{I}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2/(I-1)}{SSE/[IJ(n-1)]} \underset{H_\alpha}{\sim} F_{I-1,IJ(n-1).}$$

(b) For testing the linear hypothesis $H_\beta : \beta_j = 0$ *for all* $1 \leq j \leq J-1$,

$$SSE_{H_\beta} = SSE + In\sum_{j=1}^{J}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$$

and thus

$$F_\beta = \frac{In\sum_{j=1}^{J}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2/(J-1)}{SSE/[IJ(n-1)]} \underset{H_\beta}{\sim} F_{J-1,IJ(n-1).}$$

Altogether, we obtain the following tabular overview for Model 2.32:

Balanced ANOVA2	Numerator sum of squares	Degrees of freedom	F-statistic
Main effects of the first factor	$\Delta SSE_{H_\alpha} = Jn\sum_{i=1}^{I}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$	$I-1$	$\frac{\Delta SSE_{H_\alpha}/(I-1)}{SSE/[IJ(n-1)]}$
Main effects of the second factor	$\Delta SSE_{H_\beta} = In\sum_{j=1}^{J}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$	$J-1$	$\frac{\Delta SSE_{H_\beta}/(J-1)}{SSE/[IJ(n-1)]}$
Interaction effects	$\Delta SSE_{H_{(\alpha\beta)}} = n\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2$	$(I-1)(J-1)$	$\frac{\Delta SSE_{H_{(\alpha\beta)}}/[(I-1)(J-1)]}{SSE/[IJ(n-1)]}$
Full model	$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2$	$IJ(n-1)$	

Table 2.1: Table of the ANOVA2 with balanced design. The number of degrees of freedom of the null model (intercept only) equals $IJn-1$, which is the sum of the tabulated degrees of freedom.

Example 2.40

The quantities tabulated in Table 2.1 can be computed by hand for the margarine dataset from Schuchard-Ficher et al. (1980). This is a continuation of Example 2.33 (see presentation with R software).

Remark 2.41 (Designs with repeated measurements)

There exist other ANOVA models which do not assume i.i.d. error terms. For instance, one can consider cases where the response of every observational unit is measured repeatedly under different experimental conditions. In the case of exactly one factor, this leads to a model equation of the form

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad 1 \leq i \leq k,\ 1 \leq j \leq n$$

with balanced design, in which the error terms $(\varepsilon_{ij})_{\substack{1 \leq i \leq k, \\ 1 \leq j \leq n}}$ *are no longer stochastically independent, because the same* n *observational units are measured in every factor level.*

This requires a modification of the model for the (joint) distribution of the error terms. One reasonable distributional assumption is given by

$$\forall 1 \leq j \leq n: \quad \varepsilon_j := (\varepsilon_{1j}, \ldots, \varepsilon_{kj})^\top \sim \mathcal{N}_k(0, \Sigma),$$

$$\forall 1 \leq j_1 \neq j_2 \leq n: \quad \varepsilon_{j_1} \perp\!\!\!\perp \varepsilon_{j_2}.$$

Under these specifications, let us assume we want to test the global hypothesis ("no treatment effect"), given by

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \qquad \textit{or, equivalently,}$$

$$H_0 : \mu_i - \mu_{i+1} = 0 \quad \forall 1 \leq i \leq k-1.$$

To this end, denote by

$$\Delta_{ij} := Y_{ij} - Y_{i+1,j}, \quad 1 \leq i \leq k-1,\ 1 \leq j \leq n$$

the differences of consecutive measurement values of observational unit j*. Then it holds that*

$$\overline{\Delta} := (\overline{\Delta}_{1\bullet}, \ldots, \overline{\Delta}_{k-1,\bullet})^\top = (\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}, \ldots, \overline{Y}_{k-1,\bullet} - \overline{Y}_{k\bullet})^\top.$$

Letting S_Δ *denote the* $(k-1) \times (k-1)$*-sample covariance matrix of the differences* (Δ_{ij})*, we obtain that the test statistic*

$$T_\Delta^2 = n \overline{\Delta}^\top S_\Delta^{-1} \overline{\Delta}$$

is $T^2(k-1, n-1)$*-distributed under* H_0*.*

Other experimental designs can be treated similarly.

Chapter 3

Discretely distributed response variables

Definition 3.1 (Generalized linear model, GLM)
Let $(\mathcal{Y}^n, \mathcal{B}(\mathcal{Y})^{\otimes n}, \bigotimes_{i=1}^n P_{\theta_i})$ with $\theta_i \in \Theta \subseteq \mathbb{R}$ for all $1 \leq i \leq n$ denote a product model, induced by stochastically independent, observable response variables $Y_1, \ldots, Y_n$, each taking their values in $\mathcal{Y} \subseteq \mathbb{R}$. We assume that the value of θ_i is unknown for each $1 \leq i \leq n$.

Then, we call $(\mathcal{Y}^n, \mathcal{B}(\mathcal{Y})^{\otimes n}, \bigotimes_{i=1}^n P_{\theta_i})$ a generalized linear model (GLM), if the following conditions are satisfied.

1) *All Y_i's are defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$.*

2) *For every $1 \leq i \leq n$, P_{θ_i} is an element of the same natural exponential family, meaning that there exists a density (likelihood function) with respect to some dominating measure, and this likelihood function is of the form*

$$p(y_i, \theta_i) = a(\theta_i) b(y_i) \exp(y_i \cdot T(\theta_i)).$$

The term $T(\theta)$ is called natural parameter of the exponential family, $\theta \in \Theta$.

3) *For each observational unit $1 \leq i \leq n$, the values of p covariates (independent variables) are given. For the entire model, this entails a design matrix $X \in \mathbb{R}^{n \times p}$. In this, an intercept is encoded by a pseudo-covariate, which is constantly equal to one.*
The systematic component of the GLM is then given by a vector $\eta = (\eta_1, \ldots, \eta_n)^\top$ with

$$\forall 1 \leq i \leq n: \quad \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

for (unknown) regression coefficients $\beta_1, \ldots, \beta_p$. The term η_i is called linear predictor pertaining to observational unit $1 \leq i \leq n$. The matrix-vector representation of the systematic model component is given by $\eta = X\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^\top$.

4) Denote by $\mu_i := \mathbb{E}[Y_i|\vec{X}_i = \vec{x}_i]$ the (conditional) expected value of the i-th response variable. There exists a link function g, which links the systematic component and the stochastic component of the model, where the stochastic component is described by $\mu_1, \ldots, \mu_n$, which are characteristics of the (conditional) distributions of $Y_1, \ldots, Y_n$:

$$\forall 1 \leq i \leq n: \quad \eta_i = g(\mu_i) \Leftrightarrow g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}.$$

The canonical link function (canonical link, for short) transforms the (conditional) expected value of Y_i into the natural parameter. Hence, it satisfies the following relationship:

$$g(\mu_i) = T(\theta_i) \Leftrightarrow T(\theta_i) = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Example 3.2 (ANCOVA with normally distributed error terms)
Let $Y = (Y_1, \ldots, Y_n)^\top$ with $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, as under Model 2.1 with the additional (distributional) assumption made in Part (c) of Model 2.1. Then, by Corollary 2.3, it holds that $\mu_i = \sum_{j=1}^{p} \beta_j x_{ij}$ for all $1 \leq i \leq n$. Hence, this ANCOVA model constitutes a GLM with $g = id$ (identity link).

The formulation of GLMs allows for a unified treatment of product models, which are based on natural exponential families, by means of the inferential likelihood theory. In this, the targets of statistical inference for GLMs are the regression coefficientes $(\beta_j)_{1 \leq j \leq p}$, as in Chapter 2.

3.1 Poisson regression

In this section, we study the Poisson regression model for count data. We first embed this model into the scope of GLMs.

Lemma 3.3
The family of Poisson distributions with intensity parameter $\theta > 0$ constitutes a natural exponential family with natural parameter $T(\theta) = \log(\theta)$.

Proof: The family is dominated by the counting measure, with likelihood function given by

$$\begin{aligned} p(k, \theta) &= \frac{\theta^k}{k!} \exp(-\theta) \\ &= \exp(-\theta)\big(\frac{1}{k!}\big) \exp(k \cdot \log(\theta)) \\ &= a(\theta) \cdot b(k) \cdot \exp(k \cdot T(\theta)), \quad k \in \mathbb{N}_0. \end{aligned}$$

This matches the representation of the likelihood function of a natural exponential family, with $T = \log$. ■

The Poisson regression models stochastically independent observables, which have a count data structure. Assuming that the distribution of each of these observables can be described well by a Poisson distribution with covariate-dependent intensity parameter, Model 3.4 employs the general setup of GLMs for the modeling. According to Definition 3.1 and Lemma 3.3, the canonical link for the Poisson regression model is given by the natural logarithm.

Model 3.4 (Poisson regression with intercept)
Let $k = p - 1$ and denote by $\vec{X} = (X_1, \ldots, X_k)$ the vector of covariates of interest. The sample space for the response vector $Y = (Y_1, \ldots, Y_n)^\top$ is $(\mathbb{N}_0^n, 2^{\mathbb{N}_0^n})$. For each $1 \leq i \leq n$, we take a so-called size s_i as a foundation and make the model assumption, that Y_i possesses the (conditional) Poisson$(\lambda(\vec{x}_i) \cdot s_i)$-distribution, $1 \leq i \leq n$. In this, we make the structural assumption that

$$\log(\lambda(\vec{x}_i)) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \quad \textit{or, equivalently,}$$

$$\begin{aligned} \mathbb{E}[Y_i | \vec{X}_i = \vec{x}_i; s_i] &= \lambda(\vec{x}_i) \cdot s_i \\ &= \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln(s_i)), \quad 1 \leq i \leq n. \end{aligned}$$

Letting $\beta = (\beta_0, \ldots, \beta_k)^\top$, we obtain for the likelihood function of the entire sample, that

$$\begin{aligned} Z(y, \beta) &= \prod_{i=1}^{n} \frac{[\lambda(\vec{x}_i) s_i]^{y_i}}{y_i!} \exp(-\lambda(\vec{x}_i) s_i) \\ &= \prod_{i=1}^{n} \frac{[\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln(s_i))]^{y_i}}{y_i!} \cdot \exp(-\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln(s_i))) \end{aligned}$$

with pertaining (joint) log-likelihood function given by

$$\begin{aligned} L(y, \beta) &= \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln(s_i)) - \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln(s_i)) - \ln(y_i!) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \left(\eta_i + \ln(s_i)\right) - \exp\left(\eta_i + \ln(s_i)\right) - \ln(y_i!) \right\}, \ \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}. \end{aligned}$$

Remark 3.5

(i) The term $\ln(s_i)$ is referred to as offset. It leads to an individual intercept $\beta_0 + \ln(s_i)$ for each observational unit.

(ii) The canonical link $g = \log$ transforms the original parameter space $(0, \infty)$ of $\lambda(\vec{x}_i)$ to entire $\mathbb{R}$, which is the natural parameter space of the Poisson family.

(iii) The Poisson regression is a multiplicative model. Namely, it holds for the ratio of two incidence rates (referred to as relative risk, RR for short, in epidemiology) corresponding to two profiles $\vec{x}_A$ and $\vec{x}_B$ of covariates, which only differ in the value of one particular covariate j, that

$$RR = \frac{\lambda(\vec{x}_A)}{\lambda(\vec{x}_B)} = \exp(\log(\lambda(\vec{x}_A)) - \log(\lambda(\vec{x}_B)))$$

$$= \exp(\beta_j(x_{A,j} - x_{B,j})).$$

In particular, for a dichotomous covariate j: $RR = \exp(\beta_j)$

This technique of keeping all other covariates (except covariate j) constant is called "adjustment" or "to adjust".

(iv) The point estimation problem for the regression coefficients is solved via the maximum likelihood approach. There is no closed-form solution for the MLE, but the existence of a unique minimum of the negative (joint) log-likelihood function is guaranteed, because $-L(y, \beta)$ is a convex function of the parameter vector β. Thus, (numerical) convex optimization routines can be used to find the (numerical) value of the MLE or at least a precise approximation of this value.

(v) Nested models can be compared with a likelihood ratio test.

Example 3.6

Example 10.5 of Le (2003) reports a study, in which data of $n = 44$ physicians working for an emergency at a major hospital have been collected. The response variable of interest is in this example given by the number of complaints (per physician) received during the preceding year. The size per physician is given by the number of visits, and the four covariates of interest are given by revenue (in dollars per hour), the workload at the emergency service (in hours) as well as gender and residency training in emergency services (no/yes).
(For the analysis of this dataset, see the presentation with R and the handout.)

Theorem 3.7 (Multivariate central limit theorem for GLMs)
Let $\hat{\beta}(n)$ denote the MLE for the parameter vector $\beta = (\beta_1, \ldots, \beta_p)^\top$ of a GLM with canonical link at sample size n. If all p covariates have a compact support and the sequence $(X_n)_{n \geq p}$ of design matrices fulfills that $(X_n^\top X_n)^{-1} \to 0$ as n tends to infinity, then

$$\hat{\beta}(n) \underset{as.}{\sim} \mathcal{N}_p(\beta, F_n^{-1}(\beta)) \text{ with } F_n(\beta) = X_n^\top Cov_n(Y) X_n.$$

This asymptotic statement remains correct, if $F_n(\beta)$ is replaced by $F_n(\hat{\beta}(n))$.

Proof: Satz 2.2 in Kapitel 7 of Fahrmeir and Hamerle (1984); see also the corresponding exercise.

■

Remark 3.8

Regarding the quantities appearing in Theorem 3.7, the following representations hold true.

(a) $Cov_n(Y) = diag\left(\left[Var\left(Y_i|\vec{X}_i = \vec{x}_i\right)\right]_{1\leq i\leq n}\right)$. *In this,* $Var\left(Y_i|\vec{X}_i = \vec{x}_i\right)$ *depends on* β.

(b) In the special case of a Poisson regression, we have that

$$\begin{aligned}\forall 1 \leq i \leq n : Var\left(Y_i|\vec{X}_i = \vec{x}_i, s_i\right) &= \mu_i = \lambda(\vec{x}_i)s_i \\ &= \exp(\eta_i + \ln(s_i)) \\ &= \exp\left(\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij} + \ln(s_i)\right).\end{aligned}$$

Definition 3.9 (Saturated model)

For a given GLM based on response variables $Y = (Y_1, \ldots, Y_n)^\top$ *and their realizations* $y = (y_1, \ldots, y_n)^\top$, *we express the (joint) likelihood function in terms of* $\mu = (\mu_1, \ldots, \mu_n)^\top$, *and we indicate this by writing* $Z(y, \mu)$. *This representation is possible whenever the model can be parametrized by the expectation parameter.*

Hence, $\ln(Z(y, \hat{\mu}))$ *is the fitted value of the (joint) log-likelihood function of the given GLM.*

If we consider the latter value for all possible models in the model space, then there exists a maximum achievable value, which corresponds to optimal model fit. This maximum value is given by $\ln(Z(y, y))$. *The "model" pertaining to this maximum value assigns to every observational unit an own parameter (meaning that* $p = n$*), and it is called saturated model.*

Remark 3.10

(a) The saturated model is not actually a model (no abstraction, but merely a description of reality), because it merely encodes the data points contained in the sample in the language of a GLM. The saturated model is therefore not of interest in its own right, but merely a mathematical tool for the definition of a coefficient of determination for a GLM; see Definition 3.11.

(b) With the notation introduced in Definition 3.9, we get for the Poisson regression that

$$\begin{aligned}Z(y, \mu) &= \prod_{i=1}^{n}\frac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i) \text{ and} \\ \ln(Z(y, y)) &= \sum_{i:y_i>0}\{y_i\ln(y_i) - \ln(y_i!) - y_i\}.\end{aligned}$$

(c) It is possible to encode the saturated model in different ways (see the corresponding exercise for an example).

Definition 3.11 (Coefficient of determination of a GLM)
Utilizing the notation introduced in Definition 3.9, let $\ln(Z(y,y))$ *denote the log-likelihood value of saturated model,* $\ln(Z(y,\hat{\mu}))$ *the (fitted) log-likelihood value of a given model* $\mathcal{M}$ *with design matrix* $X \in \mathbb{R}^{n\times p}$ *for* $p < n$*, and* $\ln(Z(y,\hat{\beta}_0))$ *the log-likelihood value of the null model (intercept only). Define*

$$\begin{aligned} D(\hat{\mu}) &= 2\left[\ln(Z(y,y)) - \ln(Z(y,\hat{\mu}))\right], \\ D(\hat{\beta}_0) &= 2\left[\ln(Z(y,y)) - \ln(Z(y,\hat{\beta}_0))\right]. \end{aligned}$$

Then, the coefficient of determination of the model $\mathcal{M}$ *which corresponds to* $\hat{\mu}$ *is given by*

$$R^2 = 1 - \frac{D(\hat{\mu})}{D(\hat{\beta}_0)}.$$

If $\mathcal{M}$ *describes the data perfectly, then* $D(\hat{\mu}) = 0$*, thus* $R^2 = 1$*.*
If the goodness-of-fit of $\mathcal{M}$ *equals the goodness-of-fit of the null model, then* $D(\hat{\mu}) = D(\hat{\beta}_0)$*, thus* $R^2 = 0$*.*

Remark 3.12 (Overdispersion)
By construction, the Poisson distribution is one-parametric with

$$\mathbb{E}\left[\text{Poisson}\left(\lambda\right)\right] = \textit{Var}\left(\text{Poisson}\left(\lambda\right)\right) = \lambda.$$

In practice, this oftentimes poses a problem for the Poisson regression, namely, the problem of overdispersion. Overdispersion occurs, if $\textit{Var}\left(Y_i\right) > \lambda(\vec{x}_i)s_i$ *holds true.*
In particular, overdispersion is a problem for testing regression coefficients for statistical significance, because under-estimated standard deviations result in systematically too large test statistics (Z-scores).
One approach to address the problem of overdispersion is to assume that there exists an (additional) scale parameter ϕ*, such that* $\textit{Var}\left(Y_i|\vec{X}_i = \vec{x}_i\right) = \phi\lambda(\vec{x}_i)s_i$ *for all* $1 \leq i \leq n$*. Under this assumption, the following considerations can be made:*
We know from inferential likelihood theory, that (using the notations from Definition 3.11) $D(\hat{\mu}) \underset{as.}{\sim} \chi^2_{n-p}$ *holds true under a correctly specified model. By Pearson's argumentation (cf. Example 1.47), this entails that*

$$D(\hat{\mu}) \underset{as.}{\sim} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = \sum \frac{(O-E)^2}{E} =: Q(Y).$$

Assuming that $\hat{\mu}_i \approx \mu_i$ *holds true for all* $1 \leq i \leq n$*, this yields (under potential overdispersion with scale parameter* ϕ*)*

$$\mathbb{E}\left[\frac{Q(Y)}{\phi}\right] \approx n - p,$$

because $Q(Y)/\phi$ then behaves like a sum of squares of n standardized random variables, where p model parameters have been estimated, and $\mathbb{E}\left[\chi^2_\nu\right] = \nu$. Hence, we conclude that $\mathbb{E}\left[\frac{Q(Y)}{n-p}\right] \approx \phi$ may be a reasonable approximation.
These (heuristic) considerations result in two plausible estimation methods for the overdispersion parameter ϕ:

(a) Include ϕ as an additional parameter into the (joint) likelihood function, and optimize the (log-) likelihood function under the constraint that $D(\hat{\mu}) = n - p$ (quasi-likelihood method).

(b) Estimate ϕ by $\hat{\phi} = Q(y)/(n-p)$.

In both Cases (a) and (b), the point estimates of the regression coefficients remain unchanged, but their estimated standard deviations are multiplied by $\sqrt{\hat{\phi}}$.

3.2 Logistic regression

The second type of GLM that we study in this chapter is the logistic regression for binary responses. For the sake of embedding this into the GLM theory we first study properties of the family of Bernoulli distributions.

Lemma 3.13
The family of Bernoulli distributions with success parameter $p \in (0,1)$ constitutes a natural exponential family with natural parameter $T(p) = \log(\frac{p}{1-p}) =: \operatorname{logit}(p)$.

Proof: For each $p =: \theta \in (0,1)$ there exists a counting density (likelihood function) of Bernoulli(θ), which has the form

$$\begin{aligned} p(y,\theta) &= \theta^y (1-\theta)^{1-y} \\ &= (1-\theta)\left[\frac{\theta}{1-\theta}\right]^y \\ &= (1-\theta)\exp(y \operatorname{logit}(\theta)),\ y \in \{0,1\}. \end{aligned}$$

Using the notations of Definition 3.1, we thus can choose $a(\theta) = 1-\theta$, $b(y) \equiv 1$, and $T(\theta) = \log(\theta/(1-\theta)) = \operatorname{logit}(\theta)$, implying the assertion. ■

Model 3.14 (Logistic regression with intercept)
We consider the sample space $\left(\{0,1\}^n, 2^{\{0,1\}^n}\right)$ and model the (conditional) distributions of the stochastically independent response variables $Y = (Y_1, \ldots, Y_n)^\top$ as follows.

$$\forall 1 \le i \le n : Y_i | \vec{X}_i = \vec{x}_i \sim \text{Bernoulli}\left(p(\vec{x}_i)\right),$$

where we make the structural assumption

$$\operatorname{logit}(p(\vec{x}_i)) = \ln\left(\frac{p(\vec{x}_i)}{1-p(\vec{x}_i)}\right) = \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

for the covariate-dependent success probability $p(\vec{x}_i)$. In this, β_0 is the intercept and $\vec{x}_i = (x_{i1}, \ldots, x_{ik})$ is the profile of covariates of observational unit $1 \leq i \leq n$.

Remark 3.15

(a) By virtue of Lemma 3.13, the logistic regression model is a GLM with canonical link.

(b) Under Model 3.14, the first two (conditional) moments of Y_i are given by

$$\begin{aligned} \mathbb{E}_\beta\left[Y_i|\vec{X}_i = \vec{x}_i\right] &= p(\vec{x}_i) = \mathbb{P}_\beta\left(Y_i = 1|\vec{X}_i = \vec{x}_i\right), \\ \mathit{Var}_\beta\left(Y_i|\vec{X}_i = \vec{x}_i\right) &= p(\vec{x}_i)\left[1 - p(\vec{x}_i)\right]. \end{aligned}$$

(c) For $p \in (0,1)$, it holds that

$$g(p) = \operatorname{logit}(p) = z \in \mathbb{R} \iff p = g^{-1}(z) = \frac{1}{1+\exp(-z)}.$$

Exploiting this relationship, we can equivalently write the structural assumption made in Model 3.14 as follows.

$$\begin{aligned} \forall 1 \leq i \leq n:\ p(\vec{x}_i) &= \mathbb{E}_\beta\left[Y_i|\vec{X}_i = \vec{x}_i\right] \\ &= \frac{1}{1+\exp(-\eta_i)} \\ &= \frac{1}{1+\exp(-\beta_0 - \sum_{j=1}^{k}\beta_j x_{ij})}. \end{aligned}$$

The (inverse) function $g^{-1}: \mathbb{R} \to (0,1),\ z \mapsto [1+\exp(-z)]^{-1}$ is called the (standard) logistic function.

(d) The logistic regression model is a multiplicative model. To see this, consider two profiles $\vec{x}_A$ and $\vec{x}_B$ of covariates, which only differ in the value of one particular covariate $1 \leq j \leq k$. Then,

$$\frac{p(\vec{x}_A)}{1-p(\vec{x}_A)} = \exp\left(\operatorname{logit}(p(\vec{x}_A))\right)$$

denotes the (conditional) odds for the occurrence of the target event under profile of covariates $\vec{x}_A$. Consequently, we obtain for the logarithmic odds ratio (OR), that

$$\begin{aligned} \log(\mathrm{OR}) &= \operatorname{logit}\left(p(\vec{x}_A)\right) - \operatorname{logit}\left(p(\vec{x}_B)\right) \\ &= \beta_j(x_{A,j} - x_{B,j}), \end{aligned}$$

thus

$$\begin{aligned} \mathrm{OR} &= \exp(\beta_j(x_{A,j} - x_{B,j})) \\ &= \frac{\exp(\beta_j x_{A,j})}{\exp(\beta_j x_{B,j})} = [\exp(\beta_j)]^{x_{A,j} - x_{B,j}} . \end{aligned}$$

If covariate j is dichotomous, $\mathrm{OR} = \exp(\beta_j)$.

(e) Under Model 3.14, the (joint) likelihood function of the entire sample is given by

$$\begin{aligned} Z(y, \beta) &= \prod_{i=1}^{n} p(\vec{x}_i)^{y_i} (1 - p(\vec{x}_i))^{1-y_i} = \prod_{i=1}^{n} \frac{[\exp(\eta_i)]^{y_i}}{1 + \exp(\eta_i)} \\ &= \prod_{i=1}^{n} \frac{\left[\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)\right]^{y_i}}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}, \end{aligned}$$

because for every observational unit $1 \leq i \leq n$ it holds that

$$p(y_i, \beta) = \begin{cases} p(\vec{x}_i) = [1 + \exp(-\eta_i)]^{-1}, & \text{if } y_i = 1, \\ 1 - p(\vec{x}_i) = [1 + \exp(+\eta_i)]^{-1}, & \text{if } y_i = 0. \end{cases}$$

The (joint) log-likelihood function of the entire sample is given by

$$L(y, \beta) = \sum_{i=1}^{n} \{y_i \log(p(\vec{x}_i)) + (1 - y_i) \log(1 - p(\vec{x}_i))\},$$

where

$$p(\vec{x}_i) = \left[1 + \exp\left(-\beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)\right]^{-1},$$

which elucidates the dependence of $L(y, \beta)$ on $\beta = (\beta_0, \ldots, \beta_k)^\top$.

Application 3.16 (Case-control studies)

In the first place, the logistic regression is a statistical technique to analyze prospectively the (conditional) probabilities of the occurrence of a given target event for the observational units under consideration, given their profiles of covariates. Suitable study designs for this analysis are (prospective) cohort studies or cross-sectional studies. However, especially in epidemiology, also (retrospective) case-control studies are conducted. In a case-control study, the binary (disease) status $Y_i = y_i$ of the response of every observational unit $1 \leq i \leq n$ is already known at the begin of the study, and the sampling refers retrospectively to the corresponding profile $\vec{X}_i = \vec{x}_i$ of covariates. Hence, under such a retrospective study design we can only analyze the probabilities $\mathbb{P}_\beta \left[\vec{X}_i = \vec{x}_i | Y_i = y_i\right]$, $1 \leq i \leq n$.

By applying Bayes' theorem, it is possible to analyze case-control studies with logistic regression, too. The only drawback is that the interpretability of the intercept is lost when doing so.

To explain this, consider the following notations.

Z_i : *Indicator for "Inclusion into the case-control study" (no / yes),* $1 \leq i \leq n$.

$\pi_1 := \mathbb{P}(Z_i = 1|Y_i = 1)$ *Inclusion probability for cases, not depending on* $1 \leq i \leq n$.

$\pi_0 := \mathbb{P}(Z_i = 1|Y_i = 0)$ *analogous inclusion probability for controls*

Furthermore, consider the following assumptions.

(1) The sampling probabilities only depend on the value of the response, and not on the profile of covariates, meaning that $\mathbb{P}(Z_i = 1|Y_i = \ell, \vec{X}_i = \vec{x}_i) = \pi_\ell$, $\ell \in \{0, 1\}$.

(2) Logistic model for the (conditional) distribution of each response variable Y_i*:*

$$\mathbb{P}_\beta(Y_i = 1|\vec{X}_i = \vec{x}_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_{ij}\right)}$$

Under these specifications, Bayes' theorem entails that

$$\begin{aligned}\mathbb{P}_\beta(Y_i = 1|Z_i = 1, \vec{X}_i = \vec{x}_i) &= \frac{\pi_1 \mathbb{P}_\beta(Y_i = 1|\vec{X}_i = \vec{x}_i)}{\sum_{\ell=0}^1 \pi_\ell \mathbb{P}_\beta(Y_i = \ell|\vec{X}_i = \vec{x}_i)} \\ &= \frac{\exp\left(\beta_0^* + \sum_{j=1}^k \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0^* + \sum_{j=1}^k \beta_j x_{ij}\right)}, 1 \leq i \leq n,\end{aligned}$$

where $\beta_0^* := \beta_0 + \log(\pi_1/\pi_0)$. *Hence, we can take into account the response status-specific inclusion probabilities by a simple re-definition of the model's intercept.*
For analyzing $p(\vec{x}_i)$ *we can thus fit, under a case-control study design, a logistic regression model in complete analogy to the situation under a prospective study design, in which only the intercept loses its (original) interpretation. If no reliable information at all about* π_0 *and* π_1 *is available, the interpretability of the intercept is lost completely. However, inference about* $p(\vec{x}_i)$ *can nevertheless be made as in the case of a prospective study design.*

Application 3.17 (Receiver Operating Characteristic (ROC) curve)
Once a logistic regression model has been fitted, it is near at hand to utilize the estimated regression coefficients and the resulting estimated (conditional) success probabilities for the classification of new observational units, for which only their profiles of covariates is known. In medicine, this task is called diagnosis. This boils down to defining a threshold p^* *for* $\hat{p}(\vec{x}_{new})$*, so that we diagnose* $\hat{y}_{new} = 1$ *whenever* $\hat{p}(\vec{x}_{new}) \geq p^*$ *holds true. Since the estimation of the regression coefficients had to cope with stochastic fluctuations in the "training sample" which cannot (and should not) fully be captured by the model, one can not expect a perfect diagnostic quality of the fitted model.*

For the determination of p^* it is common practice to perform a so-called ROC analysis. For this, we order the estimated failure probabilities derived from the training sample $((y_1, \vec{x}_1), \ldots, (y_n, \vec{x}_n))$. We denote these estimated failure probabilities by $(\hat{q}(\vec{x}_i))_{1 \leq i \leq n}$, with

$$\forall 1 \leq i \leq n : \hat{q}(\vec{x}_i) := 1 - \hat{p}(\vec{x}_i) = \mathbb{P}_{\hat{\beta}}(Y_i = 0 | \vec{X}_i = \vec{x}_i).$$

Their ordered values are given by $\hat{q}_{1:n} \leq \hat{q}_{2:n} \leq \ldots \leq \hat{q}_{n:n}$. Furthermore, we define $n_0 := |\{1 \leq i \leq n : y_i = 0\}|$ as well as $n_1 := n - n_0 = |\{1 \leq i \leq n : y_i = 1\}|$.

The so-called ROC curve is the graph of a random walk with n steps in the unit square, starting in $(0, 0)$ and ending in $(1, 1)$. In every step $1 \leq \ell \leq n$, the random walk makes a jump of width $1/n_0$ to the right, if $\hat{q}_{\ell:n}$ corresponds to $y_{\ell:n} = 0$, and it makes an upward jump of height $1/n_1$, if $\hat{q}_{\ell:n}$ corresponds to $y_{\ell:n} = 1$. In this, the response values are permuted according to the ordering of the estimated failure probabilities $(\hat{q}(\vec{x}_i))_{1 \leq i \leq n}$. It is easy to see that with this jumping rule the random walk always arrives in the point $(1, 1)$ (the upper right corner of the unit square) after the n-th step. Based on the aforementioned construction, we choose the threshold $p^* \in \{1 - \hat{q}_{1:n}, \ldots, 1 - \hat{q}_{n:n}\}$, which corresponds to the step ℓ^* of the random walk, in which the random walk has been closest to the point $(0, 1)$ (the upper left corner in the unit square). This rule minimizes the estimated (or: empirical) weighted misclassification probability.

Example:

Assume that exactly $n_0 = 2$ failures ($y_i = 0$) and exactly $n_1 = 3$ successes ($y_i = 1$) have been observed in the training sample. The total size of the training sample is thus given by $n = n_0 + n_1 = 5$. Assume further, that $y_{1:5} = y_{2:5} = y_{4:5} = 1$ as well as $y_{3:5} = y_{5:5} = 0$.

As illustrated in Figure 3.1, we would choose $p^* = 1 - \hat{q}_{2:5}$ in this example (the exact values of the five estimated success probabilities are omitted here, because they are not needed to draw the graph in Figure 3.1). Applying this threshold to the training sample, we would correctly classify two third of the cases (corresponding to $y_i = 1$) and all the controls (corresponding to $y_i = 0$), because

$$\begin{aligned}
\hat{p}_{1:5} = 1 - \hat{q}_{1:5} > p^* &\quad \Rightarrow \quad \hat{y}_{1:5} = 1 = y_{1:5}, \\
\hat{p}_{2:5} = 1 - \hat{q}_{2:5} = p^* &\quad \Rightarrow \quad \hat{y}_{2:5} = 1 = y_{2:5}, \\
\hat{p}_{3:5} = 1 - \hat{q}_{3:5} < p^* &\quad \Rightarrow \quad \hat{y}_{3:5} = 0 = y_{3:5}, \\
\hat{p}_{4:5} = 1 - \hat{q}_{4:5} < p^* &\quad \Rightarrow \quad \hat{y}_{4:5} = 0 \neq y_{4:5} = 1, \\
\hat{p}_{5:5} = 1 - \hat{q}_{5:5} < p^* &\quad \Rightarrow \quad \hat{y}_{5:5} = 0 = y_{5:5}.
\end{aligned}$$

In biometrics and epidemiology (theory of diagnostic tests), the following quantities are conside-

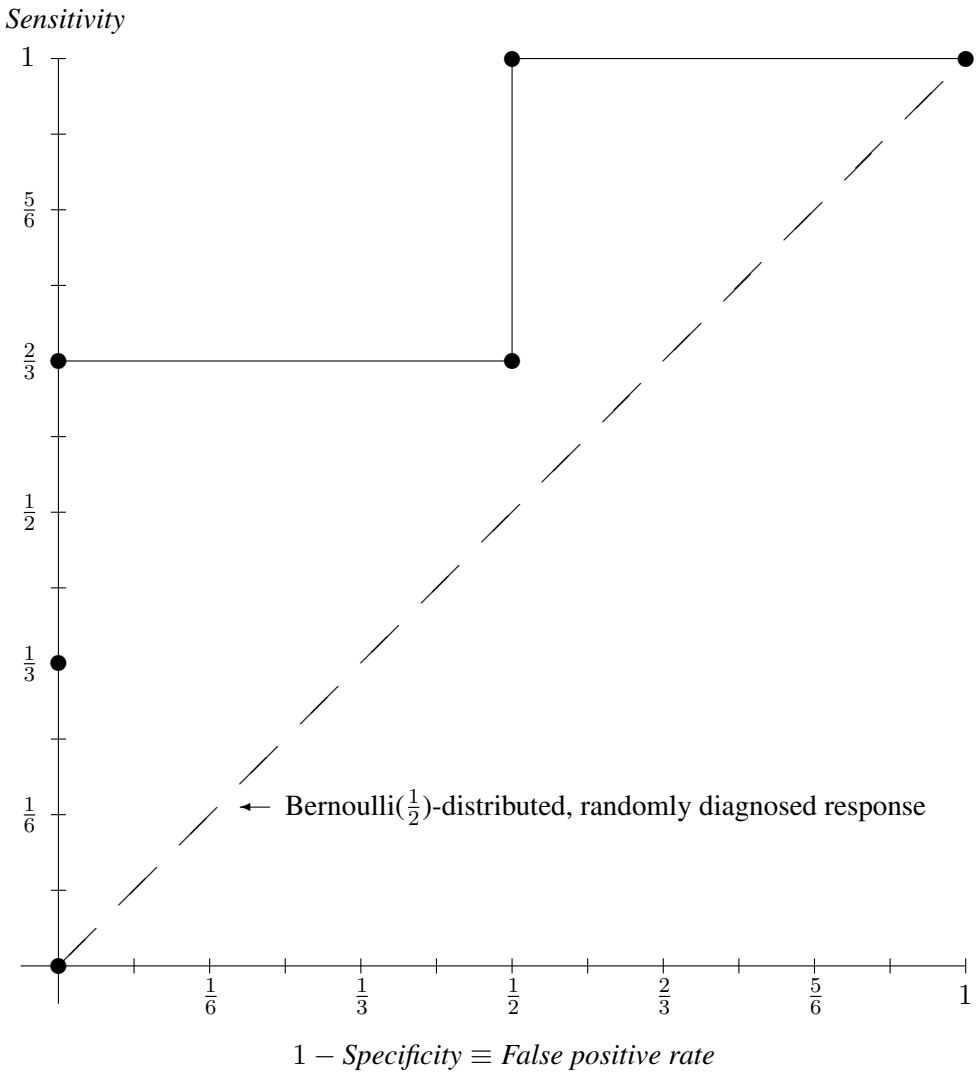


Figure 3.1: Example of an ROC curve

red for a binary response (disease status) Y:

$$\mathbb{P}_{\textit{test procedure}}(\hat{Y} = 1|Y = 1) \quad \textit{is called sensitivity},$$

$$\mathbb{P}_{\textit{test procedure}}(\hat{Y} = 0|Y = 0) \quad \textit{is called specificity.}$$

Furthermore,

$$\mathbb{P}_{\textit{test procedure}}(\hat{Y} = 1|Y = 0) = 1 - \textit{specificity} \qquad \textit{is called false positive rate},$$

$$\mathbb{P}_{\textit{test procedure}}(\hat{Y} = 0|Y = 1) = 1 - \textit{sensitivity} \qquad \textit{is called false negative rate.}$$

Hence, as indicated in Figure 3.1, in an ROC analysis the empirical ("in-sample") false positive rate is plotted against the empirical ("in-sample") sensitivity (also referred to as the true positive rate). In our toy example, we obtain an empirical false positive rate of zero (all controls in the training sample are correctly classified when applying the learned diagnosis rule) and an empirical sensitivity of $2/3$ (two third of all cases in the training sample are correctly classified when applying the learned diagnosis rule).
The area under the ROC curve (ROC-AUC) is a summarizing measure for the diagnostic quality of the classification procedure. As a baseline for comparison, one can take $\mathrm{AUC}_{\mathrm{Guessing}} = 1/2$, which corresponds to a completely random (uniformly on $\{0, 1\}$ distributed) assignment of the diagnosed response, without taking into account the given profile of covariates (main diagonal in the unit square, cf. Figure 3.1).

Remark 3.18 (Probit model)
Although $g = \mathrm{logit}$ is the canonical link for the logistic regression, there exist alternative proposals for the choice of the link function. To motivate these alternative proposals, notice that the function $G_{\mu,\tau} : \mathbb{R} \to (0, 1)$, given by

$$G_{\mu,\tau}(x) = \frac{\exp\left((x - \mu)/\tau\right)}{1 + \exp\left((x - \mu)/\tau\right)} \text{ for } \mu \in \mathbb{R} \text{ and } \tau > 0,$$

is the cdf of the so-called logistic distribution on $\mathbb{R}$ with location parameter μ and dispersion parameter $\tau > 0$.
This means, that the logistic function

$$x \mapsto \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{1 + \exp(x)}$$

is the cdf of the standardized logistic distribution with $\mu = 0$ and $\tau = 1$. Consequently, the structural assumption of the logistic regression is given by

$$p(\vec{x}_i) = \mathbb{E}\left[Y_i|\vec{X}_i = \vec{x}_i\right] = \frac{1}{1 + \exp(-\eta_i)} = G_{0,1}(\eta_i),\ 1 \leq i \leq n.$$

Many other (families of) probability distributions have similar properties and their cdfs can be used as inverse link function in analogous manner. If Φ (the cdf of the standard normal distribution) is chosen, the resulting model is called probit model.

Remark 3.19

The classical logistic regression model is a single-layer feed forward neural network with logistic activation function; cf. Section 3.1.3 in Bishop (1995).

Chapter 4

Survival analysis, Cox regression

In this chapter, we consider the last type of response data listed in Table 0.1, namely, survival (time-to-event) data. As a preparation, we first introduce some basic notions from the field of survival analysis.

Definition 4.1 (Basic notions of survival analysis)
Let T denote a non-negative, real-valued random variable, which is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $F : [0, \infty) \to [0, 1]$ denote the cdf of T, so that $F(t) = \mathbb{P}(T \leq t)$, $t \in [0, \infty)$. We think of T as a (random) time span until the occurrence of a given target event. Then we call

a) the function $S : [0, \infty) \to [0, 1]$, given by $S(t) = \mathbb{P}(T > t) = 1 - F(t)$, the survival function of T.

b) the function $\Lambda : [0, \infty) \to [0, \infty]$, given by

$$\Lambda(t) = \int_0^t \frac{F(ds)}{S(s-)},$$

the cumulative hazard function of T.

As argued by Gill and Johansen (1990), there exists a sequence $(t_i^{(n)})_{1 \leq i \leq k(n)}$ of partitions of the interval $(0, t]$, such that

$$\begin{aligned} \Lambda(t) &= \lim_{n \to \infty} \sum_{i=1}^{k(n)} \left[1 - \frac{S(t_i^{(n)})}{S(t_{i-1}^{(n)})} \right] \\ &= \lim_{n \to \infty} \sum_{i=1}^{k(n)} \mathbb{P}\big(T \leq t_i^{(n)} | T > t_{i-1}^{(n)}\big), \end{aligned}$$

where $t_0^{(n)} \equiv 0$ for all $n \in \mathbb{N}$.

c) If the distribution of T is continuous with Lebesgue density f, then we call

$$\lambda(t) := \frac{d\Lambda(t)}{dt} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

the hazard function or the incidence density of T, respectively. In the latter case, it holds for all $t > 0$, that

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \,|T > t)}{\Delta t} \,.$$

For this reason, $\lambda(t)$ is also referred to as the instantaneous failure rate or the instantaneous risk, respectively.

d) For two (sub-) populations A and B with corresponding incidence densities λ_A and λ_B, we call the function RR, given by

$$\mathrm{RR}(t) = \frac{\lambda_A(t)}{\lambda_B(t)},$$

relative risk.

Remark 4.2

Under the conditions of Part c) of Definition 4.1, the following statements hold true.

(i) Regarding the relationship between hazard function and survival function, it holds that

$$\begin{aligned} \Lambda(t) \quad &= \quad \int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1-F(s)}ds \\ &\underset{u:=F(s)}{=} \int_0^{F(t)} \frac{du}{1-u} = -\ln(1-u)\Big|_0^{F(t)} = -\ln(S(t)) \\ \iff S(t) &= \exp(-\Lambda(t)) \iff F(t) = 1 - \exp(-\Lambda(t)). \end{aligned}$$

(ii) For "small" $\Delta t > 0$, the quantity

$$\begin{aligned} \mathrm{R}(t, \Delta t\,) &:= \mathbb{P}(t < T \leq t + \Delta t \,|T > t) \\ &= \frac{F(t + \Delta t\,) - F(t)}{1 - F(t)} = \frac{S(t) - S(t + \Delta t\,)}{S(t)} \\ &\approx \Lambda(t + \Delta t\,) - \Lambda(t) \qquad (*) \end{aligned}$$

denotes the (conditional) risk for the occurrence of the target event within the time interval $(t, t + \Delta t\,]$, given that the time point t has been target event-free. In this, the approximation (*) follows from the identity

$$\frac{S(t) - S(t + \Delta t\,)}{S(t)} = 1 - \exp(\Lambda(t) - \Lambda(t + \Delta t\,))$$

and from the first-order Taylor approximation $1 - \exp(-x) \approx x$ for "small" $x > 0$.

(iii) If T is exponentially distributed with intensity parameter $\theta > 0$, then

$$\lambda(t) = \frac{\theta \exp(-\theta t)}{\exp(-\theta t)} \equiv \theta$$

is the constant incidence rate of T.

A first main task of survival analysis consists in the estimation of the survival function (for a homogeneous target population). In this, the problem of censored data is often encountered in practice. Censoring occurs for a given observational unit, if it is not possible to decide until the end of the study period whether the target event has occurred for that observational unit or not. Censoring renders the naive estimator for the survival function, which is based on the empirical cdf of the sample, biased.

Possible reasons for censoring are:

- Loss to follow-up (patient moves away, etc.)
- Dropout (e. g., unexpected side effects occur during a therapy)
- End of the study (e. g., funding runs out)
- Patient dies by a cause that is unrelated to the target event of interest.

A refined estimator for the survival function, which takes into account potential censoring, is the Kaplan-Meier estimator, which is also referred to as product-limit estimator; see Kaplan and Meier (1958).

Definition 4.3 (Kaplan-Meier estimator)
Let a random sample consisting of n observational units from a (homogeneous) target population with (unknown) survival function S (with respect to a given target event) be given.
Denote by $t_1 < t_2 < \ldots < t_k$ with $k \leq n$ the ordered, distinct observation time points in the sample. In this, the term "observation time point" refers to any time point at which a target event or a censoring occurs. Denote by d_i, $1 \leq i \leq k$, the number of observed target events at observation time point t_i. In this, t_i is expressed with reference to the inclusion time points of the observational units in the sample. Moreover, denote by n_i, $1 \leq i \leq k$, the number of observational units in the sample which have still been under risk immediately before time point t_i.

With these specifications, the Kaplan-Meier estimator for S based on the sample of size n is given by

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), t \geq 0.$$

Remark 4.4

(i) If no censoring at all occurs in the sample, then $\hat{S}(t) = 1 - \hat{F}_n(t) = 1 - n^{-1} \sum_{i=1}^{n} \mathbf{1}_{[0,t]}(t_i)$ for all $t \geq 0$, where $\hat{F}_n$ denotes the empirical cdf of the n event time points.

(ii) A heuristic justification for $\hat{S}$ is given by the chain rule for events (or: chain factorization of probabilities).

Example 4.5

Table 4.1 contains the survival times (in weeks) of ten patients with superficial bladder cancer after the (respective) start of a chemotherapy. In this, censored observations are marked with a cross ("+"):

$$63, 59, 57+, 37, 33, 21+, 11, 57, 44+, 37.$$

t_i	d_i	n_i	Factor $= \frac{n_i - d_i}{n_i}$	$\hat{S}(t_i)$
0	-	10	—	1.000
11	1	10	0.900	0.900
21	0	9	1.000	0.900
33	1	8	0.875	0.788
37	2	7	0.714	0.563
44	0	5	1.000	0.563
57	1	4	0.750	0.422
59	1	2	0.500	0.211
63	1	1	0.000	0.000

Table 4.1: Table for the computation of the Kaplan-Meier estimator

The resulting Kaplan-Meier curve (graph of the estimated survival function) is displayed in Figure 4.1.

Definition 4.6 (Nonparametric likelihood function)

Let $X_1, \ldots, X_n$ be real-valued i.i.d. random variables, let $x = (x_1, \ldots, x_n)^\top$ stand for a realization of $(X_1, \ldots, X_n)^\top$, and denote by $\mathcal{M}$ the set of all cdfs on $\mathbb{R}$.
Then, the function $Z : \mathcal{M} \times \mathbb{R}^n \to [0, 1]$, given by

$$\begin{aligned}(F, x) \mapsto Z(F, x) &:= \prod_{i=1}^{n} [F(x_i) - F(x_i-)] \\ &= \prod_{i=1}^{n} \mathbb{P}_F(\{x_i\}),\end{aligned}$$

is called the nonparametric likelihood function at x. In this, $\mathbb{P}_F$ denotes the probability measure induced by F.

Remark 4.7

Obviously, the only candidates for maximizing the nonparametric likelihood function for given data are discrete probability distributions which distribute their entire mass amongst the observed data points $(x_i)_{1 \le i \le n}$.

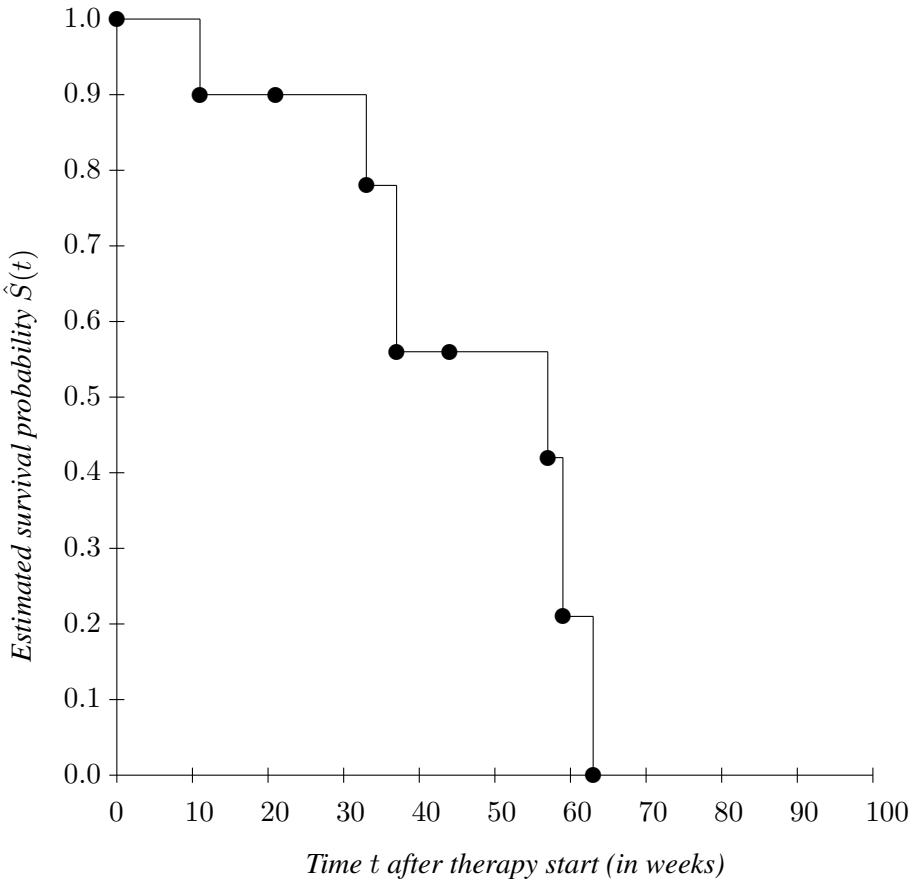


Figure 4.1: Exemplary Kaplan-Meier curve

Theorem 4.8 (Statistical properties of $\hat{S}$)

Let $T_1, \ldots, T_n$ be non-negative, real-valued random variables with $T_i \sim T$ for all $1 \leq i \leq n$, where T denotes a random variable with unknown cdf F and pertaining survival function $S = 1 - F$.

Furthermore, let $C = (C_1, \ldots, C_n)^\top$ be a vector of non-negative, real-valued random variables with (joint) distribution $\mathcal{L}_C$, which does not depend on F.

We assume that the $(T_i)_{1 \leq i \leq n}$ are conditionally stochastically independent given C. Moreover, we assume that we can observe $Y = (Y_1, \ldots, Y_n)^\top$, where $Y_i = \min(T_i, C_i)$ for all $1 \leq i \leq n$. Finally, we assume that censoring information is available via indicators $\delta_i := \mathbf{1}_{\{T_i \leq C_i\}}$ for $1 \leq i \leq n$.

Under these assumptions, the following assertions hold true.

(a) Utilizing the notations from Definition 4.3, $\hat{S}$ is the nonparametric maximum likelihood estimator (NPMLE) of S.

(b) The estimator $\hat{S}$ is uniformly consistent on intervals $[0, t]$ with $S(t) > 0$.

For all time points $t \geq 0$ with $S(t) > 0$, we have:

(c) If the $(C_i)_{1\leq i\leq n}$ are i.i.d. with cdf G of C_1, then

$$0 \leq \mathbb{E}\left[\hat{S}(t)\right] - S(t) \leq (1-S(t))\left\{1 - S(t)(1-G(t))\right\}^n.$$

In particular, the bias of $\hat{S}(t)$ tends to zero for increasing sample size n. There exist other bounds for the bias and even exact formulas (see Stute (1994)).

(d) A central limit theorem holds true for $\hat{S}(t) \equiv \hat{S}(n,t)$, and the variance of $\hat{S}(t)$ can be approximated by the Greenwood formula:

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \sum_{i:t_i\leq t} \frac{d_i}{n_i(n_i-d_i)}.$$

Proof: Let $\vec{t} := (t_1,\ldots,t_n)^\top$ and $\vec{c} := (c_1,\ldots,c_n)^\top$. For proving Part (a), we notice that the representation

$$Z((S,\mathcal{L}_C),(\vec{t},\vec{c})) = Z((S,\mathcal{L}_C),\vec{c}) \cdot Z(S,\vec{t}|C=\vec{c})$$

is valid. In this, $Z((S,\mathcal{L}_C),\vec{c}) = \mathcal{L}_C(\{\vec{c}\})$ and hence does not contain any information about the survival function S of interest. Thus, it suffices to maximize the conditional nonparametric likelihood function $Z(S,\vec{t}|C=\vec{c})$. To this end, we calculate that

$$\begin{aligned} Z(S,\vec{t}|C=\vec{c}) &= \prod_{i:\delta_i=1} \mathbb{P}_S(\{t_i\}) \prod_{i:\delta_i=0} S(c_i) \\ &= \prod_{i=1}^n \left[\mathbb{P}_S(\{y_i\})\right]^{\delta_i} \left[S(y_i)\right]^{1-\delta_i} \\ &= \prod_{j=1}^k \lambda_j^{d_j}(1-\lambda_j)^{n_j-d_j} \end{aligned} \tag{4.1}$$

as detailed in the corresponding exercise. In this, $k \leq n$ is as in Definition 4.3, $\mathbb{P}_S$ is the probability measure induced by S, and

$$\lambda_j = \mathbb{P}_S\left(T = t_{j:k} | T \geq t_{j:k}\right),\ 1 \leq j \leq k.$$

Maximizing (4.1) with respect to $\lambda_1,\ldots,\lambda_k$, we get that $\hat{\lambda}_j = d_j/n_j$ is the estimated value of λ_j, for all $1 \leq j \leq k$. Altogether, this entails that

$$\hat{S}_{\mathrm{NPMLE}}(t) = \prod_{j:t_j\leq t} \frac{n_j - d_j}{n_j}$$

as desired, because (see the corresponding exercise)

$$\forall 0 \leq t \leq t_{k:k} : S(t) = \prod_{j:t_{j:k}\leq t} (1-\lambda_j).$$

The assertion of Part (b) is Theorem IV.3.1 in the book by Andersen et al. (1993). The asymptotic normality of $\hat{S}(t)$ claimed in Part (d) is a consequence of Theorem IV.3.2 in Andersen et al. (1993), and the Greenwood formula follows by virtue of the Delta method (see exercise). ■

While the aforementioned considerations referred to homogeneous target populations, Sir David Cox (1972) developed a (regression-type) model enabling to perform survival analysis for heterogeneous target populations (taking into account covariates), too.

Model 4.9 (Cox proportional hazards model, Cox (1972))
We consider observable, stochastically independent response variables $Y_1, \ldots, Y_n$ *with* $Y_i = \min(T_i, C_i)$ *for all* $1 \leq i \leq n$*, as in Theorem 4.8.*
We model the distribution of the times-to-event $(T_i)_{1 \leq i \leq n}$ *of interest as dependent on covariates, and in this we make the structural assumption that*

$$\forall 1 \leq i \leq n : \lambda(t|\vec{X}_i = \vec{x}_i) = \lambda_0(t) \exp(\eta_i) \iff \log\left(\lambda(t|\vec{X}_i = \vec{x}_i)\right) = \log\left(\lambda_0(t)\right) + \eta_i. \quad (4.2)$$

In (4.2)*,* $\eta_i = \sum_{j=1}^{k} \beta_j x_{ij}$ *with* $\vec{x}_i = (x_{i1}, \ldots, x_{ik})$ *denotes the linear predictor of observational unit* $1 \leq i \leq n$*, as in previous chapters. In particular, it holds for the comparison with the baseline profile of covariates (*$\vec{x}_{Baseline} \equiv \vec{0}$*), that* $\mathrm{RR}(t|\vec{X}_i = \vec{x}_i) \equiv \mathrm{RR}(\vec{x}_i) = \exp(\eta_i)$*. The function* λ_0 *remains unspecified and is called baseline hazard.*
The (main) targets of statistical inference are the regression coefficients $\beta = (\beta_1, \ldots, \beta_k)^\top$*.*

Remark 4.10

(a) Model 4.9 is a semiparametric, multiplicative model.

(b) The map $t \mapsto \log(\lambda_0(t))$ *can be interpreted as an "intercept function". For a meaningful interpretation of* $\vec{x}_{Baseline} = \vec{0}$*, all covariates should be centered at their empirical means ("standardization"), before they are included in the model.*

(c) If a plausible assumption about λ_0 *can be made (e. g., Weibull or Gompertz), it is possible to modify Model 4.9 to become a fully parametric model, such that the (parametric) inferential likelihood theory can be applied.*

(d) The proportionality assumption made in (4.2) *can and should be checked by investigating covariate-specific Kaplan-Meier curves (see Application 4.11 below). It is remarkable that Model 4.9 is often a suitable model in practice.*

(e) In the case of an unspecified baseline hazard (semiparametric situation), the estimation of the regression coefficients is performed by maximizing the partial likelihood function of the sample, which results from a conditioning on the observed event data points.

To explain this further, assume for simplicity that exactly m *target events at distinct time points* $t_{1:m} < t_{2:m} < \ldots < t_{m:m}$ *have been observed in the sample, for some* $m \leq n$*. In the sequel, we denote the linear predictor corresponding to the observational unit with target event occurrence at time point* $t_{i:m}$ *with* $\eta_{i:m}$*,* $1 \leq i \leq m$*.*

According to Part (ii) of Remark 4.2, it then holds that

$$\begin{aligned} \mathrm{R}(t_{i:m}, \Delta t \,|\vec{X}_i = \vec{x}_i) &= \Delta t \, \lambda(t_{i:m}|\vec{X}_i = \vec{x}_i) \\ &= \Delta t \, \lambda_0(t_{i:m}) \exp(\eta_{i:m}) \end{aligned}$$

for infinitesimally small Δt, for all $1 \leq i \leq m$. Denoting by R_i the set of all observational units under risk immediately before $t_{i:m}$,

$$\sum_{\ell \in R_i} \Delta t \, \lambda_0(t_{i:m}) \exp(\eta_\ell) = \Delta t \, \lambda_0(t_{i:m}) \sum_{\ell \in R_i} \exp(\eta_\ell)$$

denotes for $1 \leq i \leq m$ the conditional probability for the occurrence of the target event within the interval $(t_{i:m}, t_{i:m} + \Delta t\,]$ for <u>any</u> observational unit in R_i, where conditioning refers to survival until immediately before $t_{i:m}$. Based on these considerations, the partial likelihood function is defined as the product of the covariate-specific individual observation probabilities of all m target event, conditionally to the observed event time points. Thus,

$$Z_{partial}(\beta, y) = \prod_{i=1}^{m} \frac{\exp(\eta_{i:m})}{\sum_{\ell \in R_i} \exp(\eta_\ell)}$$

with pertaining partial Log-likelihood function

$$\ln(Z_{partial}(\beta, y)) = \sum_{i=1}^{m} \eta_{i:m} - \ln\left(\sum_{\ell \in R_i} \exp(\eta_\ell) \right).$$

The obvious advantage of using the partial likelihood function for inference about β is that the baseline hazard drops out of $Z_{partial}$ completely. The case of ties amongst $(t_{i:m})_{1 \leq i \leq m}$ can be treated by (combinatorial) modifications of the partial likelihood function.

Application 4.11 (Tests for proportional hazards)
Let $\vec{x}_A$ be a given profile of covariates, and assume that the conditions of Part (c) of Definition 4.1 are fulfilled. Then, the model equation (4.2)*, together with the relationship $S(t) = \exp(-\Lambda(t))$ from Part (i) of Remark 4.2, yields that*

$$\begin{aligned} S_A(t) &= \exp(-\Lambda_A(t)) = \exp\left(-\int_0^t \lambda_A(s) ds \right) \\ &= \exp\left(-\int_0^t \lambda_0(s) ds \, \exp(\eta_A) \right) = [S_0(t)]^{\exp(\eta_A)}, \end{aligned}$$

where S_A is the survival function under profile of covariates $\vec{x}_A$, and S_0 is the baseline survival function. Hence,

$$\log(-\log(S_A(t))) = \eta_A + \log(-\log(S_0(t))).$$

For two different profiles $\vec{x}_A$ and $\vec{x}_B$ of covariates, this entails the relationship

$$\log(-\log(S_A(t))) - \log(-\log(S_B(t))) = \eta_A - \eta_B.$$

For two different strata of profiles of covariates, we can thus employ a visual inspection of the log-minus log-transformed estimated survival functions (i. e., Kaplan-Meier curves) for the purpose of model diagnosis. A formal significance test for proportional hazards can be constructed based on the scaled Schoenfeld residuals (see, e. g., Grambsch and Therneau (1994)).

Example 4.12
Prisoner data from Rossi et al. (1980), see presentation with R software.

Definition 4.13 (Pseudo-coefficient of determination for partial likelihood)
Since the partial (and not the full) likelihood function is used for model fit under the circumstances of Part (e) of Remark 4.10, it is in this case more complicated than under an ANCOVA or a GLM to evaluate the amount of spread in the response which can be explained by the model. One approximation of this amount is provided by the formula of Maddala (Maddala, 1983, Page 39). To this end, let

$$D(\hat{\beta}) = 2\left[\log(Z_{partial}(\hat{\beta}, y)) - \log(Z_{partial}(\hat{\beta}_0, y))\right],$$

where $\hat{\beta}_0$ corresponds to the null model (intercept only), in analogy to Definition 3.11. Then, a pseudo-coefficient of determination is given by

$$\tilde{R}^2_{Maddala} := 1 - \exp\left(-\frac{D(\hat{\beta})}{n}\right).$$

There exist many other proposals for pseudo-coefficients of determination for partial likelihood in the literature. Some of them are based on a comparison with the saturated model, as in the case of a GLM (see Definition 3.11).

Application 4.14 (Time-dependent covariates)
Apart from static covariates (like, for instance, place of birth or body height for adults), oftentimes also dynamic covariates have to be considered in a survival-analytic context. The values of such dynamic covariates (can) change over time. Examples are cumulative pollutant exposure or the therapy indicator in a crossover study. For the statistical analysis of the influence (on the response) of such time-dependent covariates in the (partial) likelihood approach, it is required to have access to the current values of the covariates at every event time point $t_{i:m}$ for all observational units under risk (corresponding to $\ell \in R_i$).
Under the assumptions of Part (e) of Remark 4.10, a modified version of the partial likelihood function is then given by

$$Z_{partial}(\beta, y) = \prod_{i=1}^{m} \frac{\exp(\eta_{i:m,i})}{\sum_{\ell \in R_i} \exp(\eta_{\ell,i})},$$

where $\eta_{\ell,i} = \sum_{j=1}^{k} \beta_j x_{\ell j,i}$ denotes the (time-dependent) value of the linear predictor of observational unit ℓ at time point $t_{i:m}$, and $\eta_{i:m,i}$ corresponds to that observational unit, for which the

target even occurs at time point $t_{i:m}$ (with generalizations with respect to the handling of ties which are analogous to the considerations in Part (e) of Remark 4.10).
Since this modification can imply a substantial increase in computational complexity, it is advisable to carry out a pre-test for time dependency for covariates which are prone to be dynamic. Let X_1 denote such a covariate. The original proportional hazards model (for simplicity only containing this single covariate) makes the assumption that

$$\lambda(t|X_{i1} = x_{i1}) = \lambda_0(t) \exp(\beta_1 x_{i1}),\ 1 \leq i \leq n.$$

Now, in order to check for time dependency of the influence of X_1 on the response, we additionally include a derived dynamic covariate X_2 into the model, for instance $X_2 = X_1 t$ or $X_2 = X_1 \log(t)$, and consider the extended model, which makes the assumption that

$$\lambda(t|X_{i1} = x_{i1}) = \lambda_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2}),\ 1 \leq i \leq n.$$

By testing the hypothesis $H_0 : \beta_2 = 0$, which can be done with standard methods (Wald test, t-test), it is possible to judge whether the influence of the covariate X_1 of interest exhibits time dependency.

Example 4.15 (Leukemia data, Example 11.6 in Le (2003))
Example 11.6 in Le (2003) reports a controlled efficacy study for a certain drug against leukemia. In this study, $n = 42$ patients were randomly assigned to either the therapy arm ($x_{i1} = 1$) or the placebo arm ($x_{i1} = 0$).
The target event (response) in this study was the remission time (time with reduced symptoms or free of symptoms) until relapse. Censoring occurred, if no relapse was observed (for a particular observational unit) within the study time.
The following two questions were of interest:

(1) Is the drug efficacious in prolonging the remission time as compared to placebo (yes / no)?

(2) Does the drug have an accumulated effect (over time), meaning that the treatment duration plays an important role?

$\rightsquigarrow$ see presentation with R software and handouts.

Chapter 5

Bayesian analysis of linear models

In this chapter, we take a brief look into Bayesian methods to analyze (multiple) linear regression models and GLMs.

We start with the classical ANCOVA model introduced in Model 2.1. Namely, we study the model given by $Y = X\beta + \varepsilon$ with all assumptions made in Model 2.1, including the assumption of normally distributed error terms. In Bayesian notation, we can write

$$Y|\tilde{\beta} = \beta, \tilde{\sigma}^2 = \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n),$$

where $\tilde{\beta}$ and $\tilde{\sigma}$ are treated as random objects.

Bayesian inference is facilitated in the presence of conjugate distributional classes. To this end, we first study some distribution theory around the (multivariate) normal distribution.

Definition 5.1 (Inverse gamma distribution)
Letting Z_1 denote a non-negative, real-valued random variable with $Z_1 \sim \mathrm{Gamma}(\alpha, r)$, we call $Z_2 := 1/Z_1$ inverse gamma-distributed, and we write $Z_2 \sim \mathrm{IG}(\alpha, r)$. The Lebesgue density of Z_2 is given by

$$f_{Z_2}(z) = \frac{\alpha^r}{\Gamma(r)} z^{-(r+1)} \exp(-\alpha/z) \mathbf{1}_{(0,\infty)}(z), \text{ and it holds that}$$

$$\mathbb{E}[Z_2] = \frac{\alpha}{r-1} \text{ as well as } Var(Z_2) = \frac{\alpha^2}{(r-1)^2(r-2)}, \text{ provided that } r > 2.$$

Definition 5.2 (Normal-inverse gamma distribution)
Let $\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2 \sim \mathcal{N}_p\left(m, \sigma^2 M\right)$ for hyperparameters m und M. Furthermore, let $\tilde{\sigma}^2 \sim \mathrm{IG}(\alpha, r)$ with hyperparameters α und r. Then, the joint distribution of $\tilde{\beta}$ and $\tilde{\sigma}^2$ is called normal-inverse gamma distribution with parameters m, M, α, and r. The (joint) density of this distribution is

given by

$$\begin{aligned} f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta,\sigma^2) &= f_{\tilde{\beta}|\tilde{\sigma}^2=\sigma^2}(\beta)\, f_{\tilde{\sigma}^2}(\sigma^2) \\ &= \left[(2\pi)^{\frac{p}{2}}\sigma^p|M|^{\frac{1}{2}}\right]^{-1}\exp\left(-\frac{1}{2\sigma^2}(\beta-m)^\top M^{-1}(\beta-m)\right)\times \\ &\qquad \frac{\alpha^r}{\Gamma(r)}(\sigma^2)^{-(r+1)}\exp\left(-\frac{\alpha}{\sigma^2}\right),\ \beta\in\mathbb{R}^p,\sigma^2>0. \end{aligned}$$

We write that $(\tilde{\beta},\tilde{\sigma}^2)\sim \mathrm{NIG}(m,M,\alpha,r)$.
Dropping factors in the representation of the joint density that neither depend on β *nor on* σ^2*, we get that*

$$\begin{aligned} f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta,\sigma^2) &\propto \sigma^{-p}\exp\left(-\frac{1}{2\sigma^2}(\beta-m)^\top M^{-1}(\beta-m)\right)(\sigma^2)^{-(r+1)}\exp\left(-\frac{\alpha}{\sigma^2}\right) \\ &= (\sigma^2)^{-\frac{p}{2}-r-1}\exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m)-\alpha\right]\right). \end{aligned} \tag{5.1}$$

Corollary 5.3
If we assume that $(\tilde{\beta},\tilde{\sigma}^2)\sim \mathrm{NIG}(m,M,\alpha,r)$*, then it holds that*

$$\mathbb{E}\big[\tilde{\beta}|\tilde{\sigma}^2=\sigma^2\big]=m,\ \mathit{Cov}\big(\tilde{\beta}|\tilde{\sigma}^2=\sigma^2\big)=\sigma^2 M,$$

$$\mathbb{E}\left[\tilde{\sigma}^2\right]=\frac{\alpha}{r-1}\quad \textit{if } r>1,\ \textit{and}$$

$$\mathit{Var}\left(\tilde{\sigma}^2\right)=\frac{\alpha^2}{(r-1)^2(r-2)}\quad \textit{if } r>2.$$

Moreover, unconditional moments of $\tilde{\beta}$ *for* $r>1$ *are given by*

$$\begin{aligned} \mathbb{E}\big[\tilde{\beta}\big] &= \mathbb{E}\left[\mathbb{E}\big[\tilde{\beta}|\tilde{\sigma}^2\big]\right]=m \text{ and} \\ \mathit{Cov}\big(\tilde{\beta}\big) &= \mathbb{E}\big[\mathit{Cov}\big(\tilde{\beta}|\tilde{\sigma}^2\big)\big]+\mathit{Cov}\big(\mathbb{E}\big[\tilde{\beta}|\tilde{\sigma}^2\big]\big) \\ &= \mathbb{E}\big[\tilde{\sigma}^2\big]M=\frac{\alpha}{r-1}M \end{aligned}$$

according to the covariance decomposition formula; see, e. g., (Ross, 2002, Page 392).
Exploiting furthermore that $f_{\tilde{\sigma}^2|\tilde{\beta}=\beta}(\sigma^2)\propto f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta,\sigma^2)$ *we see that*

$$\tilde{\sigma}^2|\tilde{\beta}=\beta\sim \mathrm{IG}\left(\alpha+\frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m),r+\frac{p}{2}\right).$$

Finally, we compute the unconditional (marginal) distribution of $\tilde{\beta}$*. To this end, we notice that, by normalization of the* $\mathrm{IG}\left(\alpha+\frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m),r+\frac{p}{2}\right)$*-distribution,*

$$\begin{aligned} &\int_0^\infty (\sigma^2)^{-\frac{p}{2}-r-1}\exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m)-\alpha\right]\right)d\sigma^2 \\ &=\Gamma(r+\frac{p}{2})\left\{\alpha+\frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m)\right\}^{-r-\frac{p}{2}}. \end{aligned}$$

Hence, it follows from (5.1) *that*

$$\begin{aligned} f_{\tilde{\beta}}(\beta) &\propto \Gamma(r+\frac{p}{2})\left\{\alpha+\frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m)\right\}^{-r-\frac{p}{2}} \\ &\propto \left\{1+\frac{1}{2r}(\beta-m)^\top\left[\frac{\alpha}{r}M\right]^{-1}(\beta-m)\right\}^{-r-\frac{p}{2}}, \end{aligned} \tag{5.2}$$

by integrating out the joint density of $(\tilde{\beta},\tilde{\sigma}^2)$ *with respect to* σ^2.
The representation in (5.2) *corresponds to the Lebesgue density of a multivariate t-distribution with* $2r$ *degrees of freedom, location parameter* m*, and dispersion parameter* $\alpha M/r$.

Thus, $\tilde{\beta}\sim t(2r,m,\alpha M/r)$.

Remark 5.4

The covariance matrix of a multivariate t-distribution with dispersion parameter Σ *and* ν *degrees of freedom is given by* $\frac{\nu}{\nu-2}\Sigma$*, provided that* $\nu>2$.

Theorem 5.5

The family of normal-inverse gamma distributions as priors for the parameters $(\tilde{\beta},\tilde{\sigma}^2)$ *is conjugate to the family of normal distributions for the likelihood of the response* Y *in the classical ANCOVA given by Model 2.1.*

More precisely, the following statement holds true: If $Y|\tilde{\beta}=\beta,\tilde{\sigma}^2=\sigma^2\sim\mathcal{N}_n\left(X\beta,\sigma^2 I_n\right)$ *and* $(\tilde{\beta},\tilde{\sigma}^2)\sim\mathrm{NIG}(m,M,\alpha,r)$*, then*

$$(\tilde{\beta},\tilde{\sigma}^2)|Y=y\sim\mathrm{NIG}(m^*,M^*,\alpha^*,r^*) \text{ with}$$

$$M^*=(X^\top X+M^{-1})^{-1},\qquad m^*=M^*(M^{-1}m+X^\top y),$$
$$r^*=r+\frac{n}{2},\qquad \alpha^*=\alpha+\frac{1}{2}\left(y^\top y+m^\top M^{-1}m-(m^*)^\top(M^*)^{-1}m^*\right).$$

Proof: By straightforward calculation,

$$\begin{aligned} f_{(\tilde{\beta},\tilde{\sigma}^2)|Y=y}(\beta,\sigma^2) &\propto f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta,\sigma^2)\cdot Z\left((\beta,\sigma^2),y\right) \\ &\propto (\sigma^2)^{-\frac{p}{2}-r-1}\exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m)-\alpha\right]\right)\times \\ &\qquad \sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta)\right). \end{aligned}$$

The latter expression can be brought into the desired form by algebraic manipulations, which are detailed in an exercise. ■

Corollary 5.6

Under the assumptions of Theorem 5.5, the following assertions hold true.

(i) $\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2, Y = y \sim \mathcal{N}_p(\mu_\beta, \Sigma_\beta)$ *with*

$$\begin{aligned} \Sigma_\beta &= \left(\frac{1}{\sigma^2}X^\top X + \frac{1}{\sigma^2}M^{-1}\right)^{-1} \text{ and} \\ \mu_\beta &= \Sigma_\beta\left(\frac{1}{\sigma^2}X^\top y + \frac{1}{\sigma^2}M^{-1}m\right). \end{aligned}$$

(ii) $\tilde{\sigma}^2|\tilde{\beta} = \beta, Y = y \sim \mathrm{IG}(\alpha', r')$ *with*

$$\begin{aligned} \alpha' &= \alpha + \frac{1}{2}(y - X\beta)^\top(y - X\beta) + \frac{1}{2}(\beta - m)^\top M^{-1}(\beta - m) \text{ and} \\ r' &= r + \frac{n+p}{2}. \end{aligned}$$

Proof: It is known from Theorem 5.5, that $(\tilde{\beta}, \tilde{\sigma}^2)|Y = y \sim \mathrm{NIG}(m^*, M^*, \alpha^*, r^*)$. Furthermore, Definition 5.2 and Corollary 5.3 have provided the following characterizations of $\mathrm{NIG}(m^*, M^*, \alpha^*, r^*)$:

(a) $\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2, Y = y \sim \mathcal{N}_p\left(m^*, \sigma^2 M^*\right)$

(b) $\tilde{\sigma}^2|\tilde{\beta} = \beta, Y = y \sim \mathrm{IG}\left(\alpha^* + \frac{1}{2}(\beta - m^*)^\top (M^*)^{-1}(\beta - m^*), r^* + p/2\right)$

It remains to identify the parameters:

(i)

$$\begin{aligned} \sigma^2 M^* = \sigma^2\left(X^\top X + M^{-1}\right)^{-1} &= \left(\sigma^{-2}X^\top X + \sigma^{-2}M^{-1}\right)^{-1} = \Sigma_\beta. \\ m^* = M^*\left(M^{-1}m + X^\top y\right) &= \sigma^{-2}\Sigma_\beta\left(M^{-1}m + X^\top y\right) \\ &= \Sigma_\beta\left(\sigma^{-2}M^{-1}m + \sigma^{-2}X^\top y\right) = \mu_\beta. \end{aligned}$$

(ii) It is easy to verify that $r^* + p/2 = r + n/2 + p/2 = r + \frac{n+p}{2} = r'$. Furthermore,

$$\begin{aligned} \alpha^* + (\beta - m^*)^\top (M^*)^{-1}(\beta - m^*)/2 &= \alpha + \left(y^\top y + m^\top M^{-1}m - (m^*)^\top (M^*)^{-1}m^*\right)/2 \\ &\quad + \frac{1}{2}(\beta - m^*)^\top (M^*)^{-1}(\beta - m^*) =: \alpha''. \end{aligned}$$

We compare α'' with α':
$2(\alpha'' - \alpha') = y^\top y + m^\top M^{-1}m - (m^*)^\top (M^*)^{-1}m^* + (\beta - m^*)^\top (M^*)^{-1}(\beta - m^*)$
$-(y - X\beta)^\top (y - X\beta) - (\beta - m)^\top M^{-1}(\beta - m) = 0.$

Notice for the last equality:

$$\begin{aligned} (M^*)^{-1}m^* &= M^{-1}m + X^\top y, \\ (m^*)^\top (M^*)^{-1} &= \left(M^{-1}m + X^\top y\right)^\top, \end{aligned}$$

because $(M^*)^{-1}$ is symmetric. The equality follows by straightforward verification. ■

Remark 5.7

We know from Corollary 1.9, that the conditional expectation $\mathbb{E}\big[\tilde{\beta}|Y\big]$ is the Bayes-optimal point estimator for β under quadratic loss. This entails that

$$\hat{\beta}_{Bayes} = \mathbb{E}\big[\tilde{\beta}|Y\big] = (X^\top X + M^{-1})^{-1}(M^{-1}m + X^\top Y).$$

Defining the matrix $A := \big(X^\top X + M^{-1}\big)^{-1} X^\top X$, it thus holds that $\hat{\beta}_{Bayes} = (I_p - A)m + A\hat{\beta}$, where $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ denotes the LSE for β, which is at the same time the MLE for β under the assumption of normally distributed error terms.

Conceptually, it is possible to apply the presented concepts regarding Bayesian inference from the classical ANCOVA model to generalized linear models, too.

Definition 5.8 (Bayesian GLMs)

Let $Z(y,\beta) = \prod_{i=1}^n p(y_i,\beta)$ denote the (joint) likelihood function of a GLM, where we suppress notationally the dependence on the covariates, which enter the model via the regression coefficients. Moreover, let $f_{\tilde{\beta}}$ be a prior density with respect to the Lebesgue measure λ^p on $\mathbb{R}^p$.
Then we call

(i) $f_{\tilde{\beta}|Y=y}$*, given by*

$$f_{\tilde{\beta}|Y=y}(\beta) = \frac{f_{\tilde{\beta}}(\beta)Z(y,\beta)}{\int_{\mathbb{R}^p} f_{\tilde{\beta}}(\beta)Z(y,\beta)d\beta} \propto f_{\tilde{\beta}}(\beta)Z(y,\beta)$$

posterior density of $\tilde{\beta}$ (wrt. λ^p), given the observed data $Y = y$.

(ii) $\mathbb{E}\big[\tilde{\beta}|Y=y\big] = \int_{\mathbb{R}^p} \beta f_{\tilde{\beta}|Y=y}(\beta)d\beta$ *posterior mean (expected value) of $\tilde{\beta}$ (given $Y = y$).*

(iii) $Cov\big(\tilde{\beta}|Y=y\big) = \int_{\mathbb{R}^p} \big(\beta - \mathbb{E}\big[\tilde{\beta}|Y=y\big]\big)\big(\beta - \mathbb{E}\big[\tilde{\beta}|Y=y\big]\big)^\top f_{\tilde{\beta}|Y=y}(\beta)d\beta$
posterior covariance matrix of $\tilde{\beta}$ (given $Y = y$).

(iv) $\hat{\beta}_{post.} := \underset{\beta\in\mathbb{R}^p}{\operatorname{argmax}}\, f_{\tilde{\beta}|Y=y}(\beta) = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmax}} \Big\{\ln(f_{\tilde{\beta}}(\beta)) + \ln(Z(y,\beta))\Big\}$
maximum a posteriori (MAP) estimator for β.

Remark 5.9

Analytic calculation of the posterior mean $\mathbb{E}\big[\tilde{\beta}|Y=y\big]$ and the posterior covariance matrix $Cov\big(\tilde{\beta}|Y=y\big)$, respectively, is only possible in a very few special cases. Furthermore, numerical integration in $\mathbb{R}^p$ is only stable / reliable for small to moderate dimensions p. This is why one often resorts to MAP estimators instead of posterior mean estimators.

Theorem and Definition 5.10 (Ridge estimator)

Under the circumstances of Definition 5.8, assume that we choose a Gaussian prior, meaning that $\tilde{\beta} \sim \mathcal{N}_p(m, M)$. Then, the following assertions hold true.

(a)

$$\begin{aligned}\hat{\beta}_{post.} &= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmax}} \left\{ \ln(Z(y,\beta)) - \frac{1}{2}(\beta - m)^\top M^{-1}(\beta - m) \right\} \\ &=: \underset{\beta\in\mathbb{R}^p}{\operatorname{argmax}} \ln(Z_{post.}(y,\beta)).\end{aligned}$$

This allows for interpreting the logarithmic posterior density $\ln(Z_{post.}(y,\cdot))$ as a penalized log-likelihood function. In this, the penalty term $(\beta-m)^\top M^{-1}(\beta-m)$ penalizes deviations from the prior mean (expected value) m.

(b) For the special choice of $m = 0$ and $M = \tau^2 I_p$, the MAP estimator $\hat{\beta}_{post.}$ equals the so-called ridge estimator with shrinkage parameter $\lambda := [2\tau^2]^{-1}$, and the penalty term simplifies to

$$\lambda \cdot (\beta^\top \beta) = \lambda \|\beta\|_2^2.$$

(c) Define the penalized Fisher information matrix $F_{post.}(\beta)$ by

$$(F_{post.}(\beta))_{i,j} = -\mathbb{E}\left[\frac{\partial^2 \ln(Z_{post.}(y,\beta))}{\partial\beta_i \partial\beta_j} \right],\ 1 \le i,j \le p.$$

Then it holds for $n \to \infty$, that

$$\hat{\beta}_{post.}(n) \underset{as.}{\sim} \mathcal{N}_p\left(\beta, F_{post.}^{-1}(\hat{\beta}_{post.})\right),$$

where $F_{post.}$ depends on the sample size n via $Z(y,\beta)$.

Proof: See Section 4.6 in Fahrmeir et al. (2009). ■

Another popular shrinkage method is given by the LASSO (least absolute shrinkage and selection operator). This method can be embedded into the Bayesian theory by considering double-exponential priors for the vector of regression coefficients.

Definition 5.11 (Double-exponential distribution)
The double-exponential distribution (sometimes also referred to as Laplace distribution) with scale parameter $\lambda > 0$ is a continuous probability distribution on $\mathbb{R}$ with Lebesgue density f_λ, given by

$$f_\lambda(t) = \frac{\lambda}{2}\exp(-\lambda|t|), t \in \mathbb{R}.$$

Theorem 5.12
Under the circumstances of Definition 5.8, assume that we choose the prior density as $f_{\tilde{\beta}}(\beta) = \prod_{j=1}^p f_\lambda(\beta_j)$, meaning that we assume a priori stochastically independent, identically Laplace(λ)-distributed regression coefficients.

Then, the resulting MAP estimator for β is an L_1-norm penalized MLE.

Proof: For the posterior density of $\tilde{\beta}$, we get that

$$f_{\tilde{\beta}|Y=y}(\beta) \propto Z(y,\beta) \cdot \prod_{j=1}^{p} \exp(-\lambda|\beta_j|).$$

Thus,

$$\begin{aligned}
\hat{\beta}_{post.} &= \arg\max_{\beta\in\mathbb{R}^p} f_{\tilde{\beta}|Y=y}(\beta) \\
&= \arg\max_{\beta\in\mathbb{R}^p} \{\ln\left(Z(y,\beta)\right) - \lambda \sum_{j=1}^{p} |\beta_j|\} \\
&= \arg\max_{\beta\in\mathbb{R}^p} \{\ln\left(Z(y,\beta)\right) - \lambda \cdot \|\beta\|_1\},
\end{aligned}$$

and the assertion follows. ■

Remark 5.13

(a) Equivalently,

$$\hat{\beta}_{post.} = \arg\min_{\beta\in\mathbb{R}^p} \{-\ln\left(Z(y,\beta)\right) + \lambda \|\beta\|_1\}.$$

In the case of a Gaussian likelihood, $-\ln\left(Z(y,\beta)\right)$ is an isotone transformation of the residual sum of squares.

(b) In the classical ANCOVA model with unknown error variance σ^2, it is convenient to choose the conditional (to σ^2) prior distribution $\mathcal{L}(\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2)$ as $[Laplace(\lambda/\sigma)]^{\otimes p}$. This guarantees that the posterior density is unimodal for every prior inverse gamma distribution of $\tilde{\sigma}^2$ (see Park and Casella (2008)).

(c) The penalization or regularization parameter λ can either be chosen explicitly (e. g., by cross validation or by maximizing the marginal likelihood), or one can introduce an additional level of hierarchy into the Bayesian modelling. This additional level of hierarchy consists in choosing a hyper-prior for the random variable $\tilde{\lambda}$. Park and Casella (2008) recommend a prior gamma distribution for $\tilde{\lambda}^2$.

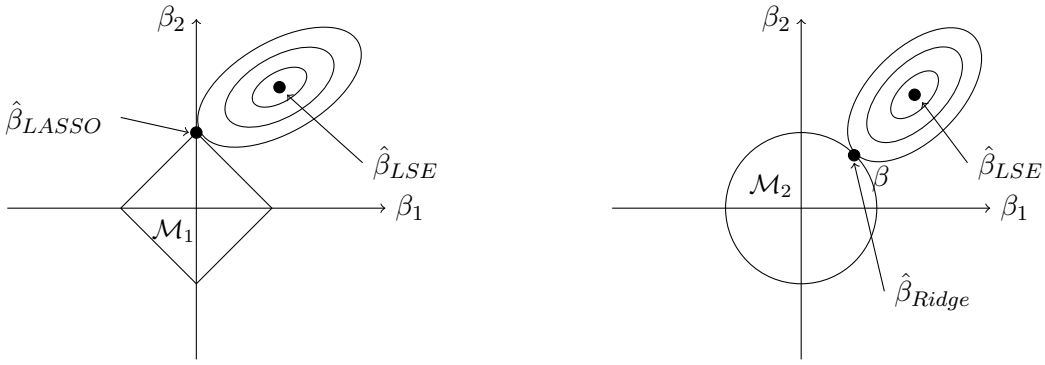
(d) Under the circumstances of Part (b), meaning a classical ANCOVA, the MAP estimator $\hat{\beta}_{post.}$ from Theorem 5.12 equals the LASSO estimator from Tibshirani (1996).

(e) In contrast to L_2-regularization, the L_1-regularization often implicitly performs a variable selection. To see this, let us re-write the target criteria of the two regularization approaches as follows:

$$\begin{aligned}
\hat{\beta}_{LASSO} &= \arg\min_{\mathcal{M}_1}\{-\ln\left(Z(y,\beta)\right)\}, \quad \mathcal{M}_1 := \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq C_1\}, \\
\hat{\beta}_{Ridge} &= \arg\min_{\mathcal{M}_2}\{-\ln\left(Z(y,\beta)\right)\}, \quad \mathcal{M}_2 := \{\beta \in \mathbb{R}^p : \|\beta\|_2^2 \leq C_2\},
\end{aligned}$$

where $C_1 \equiv C_1(\lambda)$ and $C_2 \equiv C_2(\lambda)$ are transformations of the respective regularization parameter.

In the case of an ANCOVA, the term $-\ln(Z(y,\beta))$ is equivalent to the residual sum of squares $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2$, whose contour lines (wrt. β) are ellipsoids. For $p = 2$, we get the following two graphs (adapted from Tibshirani (1996)).



(f) There exist a plethora of other (more general) regularization or penalization techniques for likelihood-based inference in (generalized) linear models, for instance bridge regression (L_q-norm regularization with general $q \geq 0$) or the "elastic net" (penalty term is a weighted average of L_1- and L_2-norm regularization terms). The derivation of (asymptotic) distribution theory for penalized MLEs or LSEs, respectively, is a topic of modern research in mathematical statistics.

Application 5.14 (Markov chains and Markov Chain Monte Carlo methods)

A general approach towards Bayesian inference in complicated (analytically not tractable) models is given by so-called Markov Chain Monte Carlo (MCMC) methods; see, e. g., Liang et al. (2010). Such methods deliver algorithms, which allow one to generate pseudo-samples from the corresponding posterior distributions on the computer.

See the respective presentation (with examples in R).

Notes regarding the presentation "Markov chains on finite state spaces"

Let $(\Omega, 2^\Omega)$ be a state space with $|\Omega| = m$. The transition kernel $\mathcal{K}(x,\cdot)$ is a probability measure on $\mathcal{F} = 2^\Omega$, i. e., $\mathcal{K}(x,\cdot) : \mathcal{F} \to [0,1]$. If (X_n) is time-homogeneous, then $\mathcal{K}(x,\cdot)$ is already completely specified by specifying all point masses $\mathcal{K}(x,y)$ for $x, y \in \Omega$. Hence, $\mathcal{K}$ can be characterized by an $(m \times m)$-matrix P with $P(x,y) = \mathbb{P}\big(X_1 = y | X_0 = x\big)$.

Example: $m = 4$ *and* $P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}$

a) Is P recurrent? Is P irreducible?

$(1,1) : (1) - (2) - (1)$ $\surd$

$(1,2) : (1) - (2)$ $\surd$

$(1,3) : (1) - (2) - (3)$ $\surd$

$(1,4) : (1) - (2) - (3) - (4)$ $\surd$

$(2,1) : \surd$

$(2,2) : (2) - (3) - (2)$ $\surd$

$(2,3) : \surd$

$(2,4) : (2) - (3) - (4)$ $\surd$

$(3,1) : (3) - (2) - (1)$ $\surd$

$(3,2) :$ $\surd$

$(3,3) : (3) - (4) - (3)$ $\surd$

$(3,4) : \surd$

$(4,1) : (4) - (3) - (2) - (1)$ $\surd$

$(4,2) : (4) - (3) - (2)$ $\surd$

$(4,3) : \surd$

$(4,4) : (4) - (3) - (4)$ $\surd$

b) Invariant measure (initial distribution)

$$\mu \overset{!}{=} \mu P \Longleftrightarrow (\mu_1, \mu_2, \mu_3, \mu_4) = (\mu_1, \mu_2, \mu_3, \mu_4) \cdot \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

This is equivalent to the following system of linear equations.

$$\mu_1 = \frac{\mu_2}{3} \qquad [1]$$

$$\mu_2 = \mu_1 + \frac{2\mu_3}{3} \qquad [2]$$

$$\mu_3 = \frac{2\mu_2}{3} + \mu_4 \qquad [3]$$

$$\mu_4 = \frac{\mu_3}{3}. \qquad [4]$$

$$[1] \textit{ in } [2]\textit{:}\ \mu_2 = \frac{\mu_2}{3} + \frac{2\mu_3}{3} \iff \mu_2 = \mu_3. \qquad [5]$$

Equations [5], [4], *and* [1] *imply that* $\mu_1 = \mu_4 = \frac{\mu_2}{3}$.

$$\begin{aligned}\sum_{i=1}^{4} \mu_i \equiv 1 \quad &\Longrightarrow \quad 2\mu_2 + \frac{2}{3}\mu_2 = 1 \\ &\Longleftrightarrow \quad \frac{8}{3}\mu_2 = 1 \iff \mu_2 = \frac{3}{8} \\ &\Longrightarrow \quad (\mu_1, \mu_2, \mu_3, \mu_4) = \left(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\right).\end{aligned}$$

Cross check:

$$\left(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\right) \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} = \left(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\right) \quad \surd$$

But: This chain is periodic (with period 2*). Therefore, convergence to the invariant measure does not happen for every* $\mu^{(0)}$*! This example is an Ehrenfest chain with* $N = 3$*. For general* N*, the corresponding transition matrix is given by*

$$\begin{aligned} P(x,y) &= \frac{x-1}{N}, y = x-1, \\ P(x,y) &= \frac{N-x+1}{N}, y = x+1, \\ P(x,y) &= 0, \textit{ otherwise.} \end{aligned}$$

New chain with $m = 3$*:*

$$P = \begin{pmatrix} \frac{4}{10} & \frac{4}{10} & \frac{2}{10} \\ \frac{3}{10} & \frac{4}{10} & \frac{3}{10} \\ \frac{2}{10} & \frac{4}{10} & \frac{4}{10} \end{pmatrix}$$

This P *is recurrent and irreducible (*$\surd$*), and* P *is aperiodic (*$\surd$*).*

$$\mu^* = \left(\frac{3}{10}, \frac{4}{10}, \frac{3}{10}\right), \textit{ because}$$

$$\left(\frac{3}{10}, \frac{4}{10}, \frac{3}{10}\right) \begin{pmatrix} \frac{4}{10} & \frac{4}{10} & \frac{2}{10} \\ \frac{3}{10} & \frac{4}{10} & \frac{3}{10} \\ \frac{2}{10} & \frac{4}{10} & \frac{4}{10} \end{pmatrix} = \left(\frac{3}{10}, \frac{4}{10}, \frac{3}{10}\right) \cdot \surd$$

List of Tables

List of Figures

Literature

Andersen, K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical models based on counting processes*. Springer series in statistics. Springer.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol. 57*(1), 289–300.

Bickel, P. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics 9*, 1196–1217.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. The Clarendon Press, Oxford University Press, New York. With a foreword by Geoffrey Hinton.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), pp. 187–220.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability.* Springer Texts in Statistics. New York, NY: Springer.

Dickhaus, T. (2018). *Theory of nonparametric tests*. Cham: Springer.

Dudoit, S. and M. J. van der Laan (2008). *Multiple testing procedures with applications to genomics.* Springer Series in Statistics. Springer, New York.

Efron, B. (1977, July). Bootstrap methods: Another look at the jackknife. Technical Report 37, Department of Statistics, Stanford University.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*, 1–26.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap.* Monographs on Statistics and Applied Probability. 57. New York, NY: Chapman & Hall.

Fahrmeir, L. and A. Hamerle (1984). *Multivariate statistische Verfahren. Unter Mitarbeit von Walter Häußler, Heinz Kaufmann, Peter Kemény, Christian Kredler, Friedemann Ost, Heinz Pape, Gerhard Tutz.* Berlin-New York: Walter de Gruyter.

Fahrmeir, L., T. Kneib, and S. Lang (2009). *Regression. Models, methods and applications. (Regression. Modelle, Methoden und Anwendungen.) 2nd ed.* Statistik und ihre Anwendungen. Berlin: Springer.

Finner, H. (1994). *Testing Multiple Hypotheses: General Theory, Specific Problems, and Relationships to Other Multiple Decision Procedures*. Habilitationsschrift. Fachbereich IV, Universität Trier.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh and London.

Freedman, D. A. (1981). Bootstrapping Regression Models. *Annals of Statistics 9*, 1218–1228.

Gaenssler, P. and W. Stute (1977). *Wahrscheinlichkeitstheorie.* Hochschultext. Berlin-Heidelberg-New York: Springer-Verlag.

Gentle, J. E. (2017). *Matrix algebra. Theory, computations and applications in statistics. Second Edition.* Springer Texts Stat. Cham: Springer.

Georgii, H.-O. (2007). *Stochastics. Introduction to probability theory and statistics. (Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik.) 3rd ed.* de Gruyter Lehrbuch. Berlin: de Gruyter.

Gill, R. D. and S. Johansen (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics 18*(4), pp. 1501–1555.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*(3), pp. 515–526.

Hall, P. (1988). Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics 16*(3), 927–953.

Hall, P. (1992). *The bootstrap and Edgeworth expansion.* Springer Series in Statistics, New York.

Hall, P. and S. R. Wilson (1991). Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics 47*(2), 757–762.

Hewitt, E. and K. Stromberg (1975). *Real and abstract analysis. A modern treatment of the theory of functions of a real variable. 3rd printing.* Graduate Texts in Mathematics. 25. New York - Heidelberg - Berlin: Springer-Verlag.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple comparison procedures.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. New York etc.: John Wiley & Sons, Inc.

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Stat. 2*, 360–378.

Janssen, A. (1998). *Zur Asymptotik nichtparametrischer Tests, Lecture Notes. Skripten zur Stochastik Nr. 29.* Gesellschaft zur Förderung der Mathematischen Statistik, Münster.

Janssen, A. (2005). Resampling Student's t-type statistics. *Ann. Inst. Stat. Math. 57*(3), 507–529.

Janssen, A. and T. Pauls (2003). How do bootstrap and permutation tests work? *Ann. Stat. 31*(3), 768–806.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Le, C. T. (2003). *Introductory biostatistics.* Hoboken, NJ: Wiley.

Lehmann, E. and G. Casella (1998). *Theory of point estimation. 2nd ed.* Springer Texts in Statistics. New York, NY: Springer.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses. 3rd ed.* Springer Texts in Statistics. New York, NY: Springer.

Liang, F., C. Liu, and R. J. Carroll (2010). *Advanced Markov chain Monte Carlo methods. Learning from past samples*. Wiley Series in Computational Statistics. John Wiley & Sons, Ltd., Chichester.

Loève, M. (1977). *Probability theory I. 4th ed.* Graduate Texts in Mathematics. 45. New York - Heidelberg - Berlin: Springer-Verlag. XVII, 425 p. DM 45.00; $ 19.80 .

Maddala, G. (1983). *Limited-dependent and qualitative variables in econometrics*. Econometric Society monographs. Cambridge University Press.

Park, T. and G. Casella (2008). The Bayesian lasso. *J. Am. Stat. Assoc. 103*(482), 681–686.

Pauls, T. (2003). *Resampling-Verfahren und ihre Anwendungen in der nichtparametrischen Testtheorie.* Books on Demand GmbH, Norderstedt.

Pauly, M. (2009). *Eine Analyse bedingter Tests mit bedingten Zentralen Grenzwertsätzen für Resampling-Statistiken.* Ph. D. thesis, Heinrich Heine Universität Düsseldorf.

Pitman, E. (1937). Significance Tests Which May be Applied to Samples From any Populations. *Journal of the Royal Statistical Society 4*(1), 119–130.

Ross, S. M. (2002). *A first course in probability. Sixth edition*. Prentice-Hall, Inc.

Rossi, P., R. Berk, and K. Lenihan (1980). *Money, work, and crime: experimental evidence*. Quantitative studies in social relations. Academic Press.

Schuchard-Ficher, C., K. Backhaus, U. Humme, W. Lohrberg, W. Plinke, and W. Schreiner (1980). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.* Berlin Heidelberg New York: Springer-Verlag. VII, 346 S. 63 Abb., 146 Tab. DM 36.00; $ 21.30 .

Shorack, G. R. and J. A. Wellner (1986). *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics. New York, NY: Wiley.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics 9*(6), 1187–1195.

Spokoiny, V. and T. Dickhaus (2015). *Basics of modern mathematical statistics.* Berlin: Springer.

Student (1908). The probable error of a mean. *Biometrika 6*, 1–25.

Stute, W. (1990). Bootstrap of the linear correlation model. *Statistics 21*(3), 433–436.

Stute, W. (1994). The bias of Kaplan-Meier integrals. *Scandinavian Journal of Statistics 21*(4), pp. 475–484.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B 58*(1), 267–288.

Westfall, P. H. and S. Young (1992). *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, New York.

Witting, H. (1985). *Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang.* Stuttgart: B. G. Teubner.

Witting, H. and U. Müller-Funk (1995). *Mathematische Statistik II. Asymptotische Statistik: Parametrische Modelle und nichtparametrische Funktionale.* Stuttgart: B. G. Teubner.

Witting, H. and G. Nölle (1970). *Angewandte Mathematische Statistik. Optimale finite und asymptotische Verfahren.* Leitfäden der angewandten Mathematik und Mechanik. Bd. 14. Stuttgart: B.G. Teubner.