Methoden der Statistik

Vorlesungsskript

Thorsten Dickhaus
Humboldt-Universität zu Berlin
Wintersemester 2012 / 2013
Version: 8. Februar 2013

Vorbemerkungen

Das Material zu diesem Skript habe ich zum Teil im Rahmen meiner Vertretungsprofessur an der Technischen Universität Clausthal im Sommersemester 2011 zusammengestellt. Weitere wichtige Quellen waren das Skript über inferentielle Likelihoodtheorie von Prof. Guido Giani (Deutsches Diabetes-Zentrum Düsseldorf) und das Skript über Wahrscheinlichkeitsrechnung und Statistik von Dr. Wolfgang Meyer, Forschungszentrum Jülich, sowie die Arbeiten im GALA-Projekt, die auch Niederschlag in meiner Diplomarbeit an der Fachhochschule Aachen, Abteilung Jülich gefunden haben. Allen Lehrenden, die mich in Jülich und Düsseldorf begleitet haben, möchte ich herzlich danken.

Für die Manuskripterstellung danke ich Mareile Große Ruse und Konstantin Schildknecht.

Übungsaufgaben und R-Programme zu diesem Kurs stelle ich auf Anfrage gerne zur Verfügung. Einige Referenzen dazu finden sich im Text an den zugehörigen Stellen.

Verzeichnis der Abkürzungen und Symbole

$B(p,q)$	Betafunktion, $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$
$\mathcal{B}(\Omega)$	Borelmengen von Ω
$\lceil x \rceil$	Kleinste ganze Zahl größer oder gleich x
χ^2_ν	Chi-Quadrat Verteilung mit ν Freiheitsgraden
$\complement M$	Komplement der Menge M
δ_a	Dirac-Maß im Punkte a
$\stackrel{\mathcal{D}}{=}$	Gleichheit in Verteilung
F_X	Verteilungsfunktion einer reellwertigen Zufallsvariable X
$\lfloor x \rfloor$	Größte ganze Zahl kleiner oder gleich x
$\Gamma(\cdot)$	Gammafunktion, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt,\ x > 0$
$\mathrm{im}(X)$	Bildbereich einer Zufallsgröße X
iid.	independent and identically distributed
$\mathbf{1}_M$	Indikatorfunktion einer Menge M
$\inf M$	Infimum der Menge M
$\mathcal{L}(X)$	Verteilungsgesetz einer Zufallsvariable X
LFC	Least Favorable Configuration
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Parametern μ und σ^2
Φ	Verteilungsfunktion der $\mathcal{N}(0,1)$-Verteilung

$\varphi(\cdot)$	Verteilungsdichte der $\mathcal{N}(0,1)$-Verteilung
$\xrightarrow{w}$	schwache Konvergenz
sp(A)	Spur der Matrix A
$\sup M$	Supremum der Menge M
supp(F)	Träger der Verteilungsfunktion F
$A^\top$	Transponierte der Matrix A (analog für Vektoren)
UNI$[a,b]$	Gleichverteilung auf dem Intervall $[a,b]$

Inhaltsverzeichnis

Kapitel 1

Grundlagen

1.1 Entscheiden unter Unsicherheit, statistische Modelle

Beim Übergang von der *Wahrscheinlichkeitstheorie* zur *mathematischen Statistik* sind zwei wichtige Änderungen zu „verkraften“:

(1) Die Modellbildung erfolgt typischerweise auf dem „Ausgaberaum“ (Wertebereich) von Zufallsgrößen, nicht auf deren Definitionsbereich („Grundraum“).

(2) Statt eine einzige „richtige“ Wahrscheinlichkeitsverteilung für die Zufallsgröße X aus dem Grundraum $(\Omega, \mathcal{F}, \mathbb{P})$ herzuleiten, wird eine Familie von indizierten Wahrscheinlichkeitsmaßen $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ betrachtet und es wird zu ermitteln versucht, für welches ϑ das Maß $\mathbb{P}_\vartheta$ die (unbekannte oder nur teilweise bekannte) Verteilung von X gemäß gewisser Kriterien am besten / hinreichend gut beschreibt oder für welche ϑ die Verteilung $\mathbb{P}_\vartheta$ „kompatibel“ mit Realisierungen x von X (Beobachtungen, Stichproben) ist.

Wir werden etwas konkreter: In der *Wahrscheinlichkeitstheorie* ist das grundlegende Objekt der Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$. Zufallsvariablen sind messbare Abbildungen $X : \Omega \to \Omega'$. Typischerweise berechnet man $\mathcal{L}(X) \equiv \mathbb{P}^X = \mathbb{P} \circ X^{-1}$, ein Wahrscheinlichkeitsmaß auf Ω', genannt die „Verteilung von X“.
Veranschaulichen wir uns dies durch ein elementares Beispiel, das des doppelten Würfelwurfs. Hier ist $\Omega = \{1, \ldots, 6\}^2$, $\mathcal{F} = 2^\Omega$ und $\mathbb{P} = (\text{UNI}\{1, \ldots, 6\})^2$. Sei $X : \Omega \to \{2, \ldots, 12\} = \Omega'$ die Augensumme. Dann ist für $j \in \Omega'$

$$\begin{aligned} \mathbb{P}^X(\{j\}) &= \mathbb{P}(X = j) \\ &= \mathbb{P}(\{\omega \in \Omega : X(\omega) = j\}), \end{aligned}$$

z. B. $\mathbb{P}^X(\{7\}) = \mathbb{P}(X = 7) = \mathbb{P}(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}) = 6/36 = 1/6$.
In der *Statistik* lautet die Aufgabe nun indes, Rückschlüsse (Inferenz) auf $\mathbb{P}$ bzw. $\mathbb{P}^X$ nur aufgrund von Beobachtungen $X = x$ zu machen. Zum Beispiel könnte man sich die Frage stellen, ob die

beiden Würfel tatsächlich „fair“ sind und dazu das obige Experiment oft wiederholen und die Ausgänge in einer Strichliste festhalten.
Bezeichne daher formal X eine Zufallsgröße, die den möglichen Ausgang eines Experimentes beschreibt.[1] Da man die statistischen Schlüsse über ϑ nur vermittels der Stichprobe $X = x$ zieht, liegt es nahe, den Bildraum von X nunmehr zum grundlegenden Objekt zu machen. Sei also von nun an Ω der zu X gehörige Stichprobenraum, d. h., die Menge aller möglichen Realisierungen von X und $\mathcal{F} \subseteq 2^\Omega$ eine σ-Algebra über Ω. Die Elemente von $\mathcal{F}$ heißen messbare Teilmengen von Ω oder Ereignisse.
Bezeichne $\mathbb{P}^X$ die Verteilung von X. Es gelte $\mathbb{P}^X \in \mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$. Der Wert ϑ kann als der unbekannte und unbeobachtbare Zustand der Natur interpretiert werden.

Definition 1.1 (Statistisches Experiment / Modell)
Ein Tripel $(\Omega, \mathcal{F}, \mathcal{P})$ mit $\Omega \neq \emptyset$ eine nichtleere Menge, $\mathcal{F} \subseteq 2^\Omega$ eine σ-Algebra über Ω und $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $\mathcal{F}$ heißt statistisches Experiment bzw. statistisches Modell.
Falls $\Theta \subseteq \mathbb{R}^k, k \in \mathbb{N}$, so heißt $(\Omega, \mathcal{F}, \mathcal{P})$ parametrisches statistisches Modell, $\vartheta \in \Theta$ Parameter und Θ Parameterraum.

Appell: Obschon der eigentliche „Grundraum“ (der Definitionsbereich von X, die „Zielpopulation“) in der zentralen Definition 1.1 nicht mehr explizit auftaucht und auch nur an einigen wenigen Stellen im Skript für mathematische Zwecke gebraucht (und dann mit Ω^{-1} bezeichnet) wird, so sollte man sich insbesondere in der Praxis doch stets und ständig auch über Ω^{-1} im Klaren sein („Repräsentativität“) !

Beispiel 1.2

a) In einem großen industriellen Produktionsprozess interessiert der Ausschussanteil, d.h., der Anteil fehlerhafter Produktionstücke. Es wird zu diesem Zweck eine Stichprobe vom Umfang n zufällig aus den gefertigen Produktionsstücken entnommen. Die Zahl $n \in \mathbb{N}$ ist von der Geschäftsführung vorgegeben worden. Ihr wird nach Beendigung dieser Qualitätsprüfung mitgeteilt, wie viele der n geprüften Teile sich als Ausschuss erwiesen haben.

$$\Omega, \{0, \ldots, n\}, \mathcal{F} = 2^\Omega \textit{ (Potenzmenge)}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta} = (\textit{Bin}(n, p))_{0 \leq p \leq 1}, \Theta = [0, 1] \ni p = \vartheta.$$

b) Man nehme an, das Merkmal “Intelligenzquotient” sei in einer Zielpopulation (z.B der Bevölkerung Frankreichs) normalverteilt. Man ist aus demoskopischen Gründen an Erwartungswert und Varianz dieser Normalverteilung interessiert. Dazu führen n zufällig ausgewählte EinwohnerInnen Frankreichs einen Intelligenztest unabhängig voneinander unter standardisierten, kontrollierten Bedingungen durch. Für jede(n) TeilnehmerIn ergibt sich daraus ein Wert

[1] Witting (1985): „Wir denken uns das gesamte Datenmaterial zu einer „Beobachtung“ x zusammengefasst.“

ihres/seines Intelligenzquotienten.

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

Kritikpunkte: Der IQ kann weder negativ noch unendlich groß werden, noch kann jeder Wert aus einem Intervall angenommen werden, da die Berechnungsformel nur auf rationalen Zahlen basiert.

Hier ist das statistische Modell also nur eine näherungsweise Beschreibung des tatsächlichen Vorgangs in der Natur! Allgemein ist jedes Modell (nur) eine Abstraktion der Wirklichkeit.

c) *In einem landwirtschaftlichen Forschungsinstitut werden k unterschiedliche Weizensorten auf jeweils n Feldstücken angebaut. Man ist an Unterschieden im mittleren Ertrag der Sorten interessiert. Dazu nimmt man an, alle (k mal n) Ertragsmessungen seien stochastisch unabhängig und jeweils normalverteilt mit einem Sorten-spezifischen Mittelwert $\mu_i, 1 \leq i \leq k$. Die Variabilität der Messungen sei rein technisch bedingt und daher bei allen (k mal n) Messungen identisch sowie bekannt. Ein etwaiger "Feldeffekt" auf den Ertrag existiere nicht bzw. sein von vernachlässigbarer Größenordnung.*

$$\begin{aligned} \Omega = \mathbb{R}^{n\cdot k}, \quad \mathcal{F} = \mathcal{B}(\mathbb{R}^{n\cdot k}), \quad \Theta = \mathbb{R}^k, \vartheta = (\mu_1, \ldots, \mu_k)^T =: \vec{\mu} \\ (\mathbb{P}_\vartheta)_{\vartheta\in\Theta} \qquad = \qquad \bigotimes_{i=1}^{n} \mathcal{N}_k(\vec{\mu}, \sigma^2 \cdot I_k), \sigma^2 > 0 \text{ bekannt} \\ \widehat{=} \qquad \mathcal{N}_{n\cdot k}\left[\begin{pmatrix} \vec{\mu} \\ \vdots \\ \vec{\mu} \end{pmatrix}, \sigma^2 I_{n\cdot k}\right]. \end{aligned}$$

Die Messwerte werden hier typischerweise in Matrixform vorliegen.

Statistische Inferenz beschäftigt sich damit, Aussagen über die wahre Verteilung $\mathbb{P}^X$ bzw. den wahren Parameter ϑ zu gewinnen. Speziell formalisieren wir dies durch Entscheidungsprobleme.

Definition 1.3

Es sei $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Eine Entscheidungsregel ist eine messbare Abbildung $\delta : \Omega \to (A, \mathcal{A})$. Der Messraum $(A, \mathcal{A})$ heißt Aktionsraum. Jede Funktion $L : \Theta \times A \to \mathbb{R}_{\geq 0}$, die messbar im zweiten Argument ist, heißt eine Verlustfunktion. Das Tupel $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, A, \mathcal{A}, L)$ heißt ein statistisches Entscheidungsproblem.

Das Risiko einer Entscheidungsregel δ bei Vorliegen des Parameters ϑ ist der (unter ϑ) erwartete Verlust von δ, also

$$R(\vartheta, \delta) := \mathbb{E}_\vartheta\big[L(\vartheta, \delta)\big] = \int_\Omega L(\vartheta, \delta(x))\mathbb{P}_\vartheta(dx).$$

Beispiel 1.4

(a) *Punktschätzung:*

Sei $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), ((\mathcal{N}(\vartheta, 1))^n)_{\vartheta\in\Theta=\mathbb{R}})$.

Unsere Aufgabe sei, einen rellen Wert $\hat{\vartheta} = \hat{\vartheta}(x)$ *anzugeben, der den unbekannten Parameter* ϑ *aus der Realisierung* $x = (x_1, \ldots, x_n)$ *"möglichst präzise schätzt."*

Wir formalisieren dies als statistisches Entscheidungsproblem, indem wir zu $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ *den Aktionsraum* $(A, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *sowie den quadratischen Verlust* $L(\vartheta, a) = (\vartheta - a)^2, a \in A = \mathbb{R}$*, hinzufügen. Betrachten wir nun speziell* $\hat{\vartheta}(x) = \bar{x}_n = n^{-1}\sum_{j=1}^n x_j$*, so errechnen wir*

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

da $\mathbb{E}_\vartheta\big[\bar{X}_n^2\big] = (\mathbb{E}_\vartheta\big[\bar{X}_n\big])^2 + Var_\vartheta\left(\bar{X}_n\right)$ *ist und* $Var_\vartheta\left(\bar{X}_n\right) = n^{-2}\sum_{j=1}^n Var_\vartheta\left(X_j\right) = 1/n$ *gilt.*

(b) *Hypothesentest:*

Unter dem Modell aus (a) *möchten wir entscheiden, ob* ϑ *in einem vorgebenen Teilbereich* $\Theta_0 \subset \mathbb{R}$ *liegt oder in* $\Theta_1 := \mathbb{R} \setminus \Theta_0$ *(sowohl* Θ_0 *als auch* Θ_1 *seien nicht-leer).*
Der Aktionsraum besteht hier nur aus zwei Elementen, $A = \{a_0, a_1\}$*. O.B.d.A. kann also* $(A, \mathcal{A}) = (\{0, 1\}, 2^{\{0,1\}})$ *gewählt werden. Eine sinnvolle Verlustfunktion ist gegeben durch:*

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

für nicht-negative reelle Konstanten ℓ_1 *und* ℓ_2*.*

$$\Rightarrow R(\vartheta, \delta) = \begin{cases} \ell_1\mathbb{P}_\vartheta(\delta(X) = 1), & \text{falls } \vartheta \in \Theta_0, \\ \ell_2\mathbb{P}_\vartheta(\delta(X) = 0), & \text{falls } \vartheta \in \Theta_1. \end{cases}$$

Die sogenannte "TypI-Fehlerwahrscheinlichkeit" wird also mit ℓ_1 *und die sogenannte "TypII-Fehlerwahrscheinlichkeit" mit* ℓ_2 *gewichtet. Es ist auch möglich,* $\ell_1 = \ell_1(\vartheta)$ *und* $\ell_2 = \ell_2(\vartheta)$ *vom Wert des Parameters abhängig zu machen, um "schwere" Fehlentscheidungen stärker zu "bestrafen".*
Um eine Entscheidungsregel auszuwählen bedarf es nun Vergleichskriterien zwischen konkurrierenden Entscheidungsregeln. Da das Risiko vom unbekannten Parameter abhängt, kann eine lokal (auf $\Theta^* \subset \Theta$*) "gute" Entscheidungsregel in Bereichen außerhalb von* Θ^* *durchaus sehr schlechte Eigenschaften haben.*

Definition 1.5

Es sei $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, A, \mathcal{A}, L)$ ein statistisches Entscheidungsproblem. Ferner sei $\mathcal{M}$ eine Menge (konkurrierender) Entscheidungsregeln, also eine Menge von Abbildungen von Ω nach $(A, \mathcal{A})$.

a) Die Entscheidungsregel δ_1 heißt besser als die Entscheidungsregel δ_2, falls $\forall\vartheta \in \Theta$: $R(\vartheta, \delta_1) \leq R(\vartheta, \delta_2)$ gilt und falls ein $\vartheta_0 \in \Theta$ existiert mit $R(\vartheta_0, \delta_1) < R(\vartheta_0, \delta_2)$. Eine Entscheidungsregel $\delta^* \in \mathcal{M}$ heißt zulässig in $\mathcal{M}$, wenn es in $\mathcal{M}$ keine bessere Entscheidungsregel gibt.

b) $\delta^* \in \mathcal{M}$ heißt gleichmäßig beste Entscheidungsregel in $\mathcal{M}$, falls

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

c) Eine Entscheidungsregel δ^* heißt minimax in $\mathcal{M}$, falls

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

d) Der Parameterraum Θ trage die σ-Algebra $\mathcal{F}_\Theta$, die Verlustfunktion L sei produktmessbar und $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathcal{F}$.

Sei π ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathcal{F}_\Theta)$, dass die Unsicherheit über den Parameter vor Experimentbeginn ausdrückt (a priori-Verteilung von ϑ). Das mit π assoziierte Bayesrisiko von $\delta \in \mathcal{M}$ ist gegeben durch

$$\begin{aligned} R_\pi(\delta) &:= \mathbb{E}_\pi\big[R(\theta, \delta)\big] \\ &:= \int_\Theta R(\vartheta, \delta)\pi(d\vartheta) \\ &= \int_\Theta \int_\Omega L(\vartheta, \delta(x))\mathbb{P}_\vartheta(dx)\pi(d\vartheta) \end{aligned}$$

$\delta^* \in \mathcal{M}$ heißt Bayesregel oder Bayes-optimal in $\mathcal{M}$ (bezüglich π), falls

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

Bemerkung 1.6

(1) Das Bayesrisiko kann auch als insgesamt zu erwartender Verlust interpretiert werden. Betrachte dazu den Messraum $(\Omega \times \Theta, \mathcal{F} \otimes \mathcal{F}_\Theta)$ und das Wahrscheinlichkeitsmaß $\tilde{\mathbb{P}}$ auf $(\Omega \times \Theta, \mathcal{F} \otimes \mathcal{F}_\Theta)$, definiert durch $\tilde{\mathbb{P}}(dx, d\vartheta) = \mathbb{P}_\vartheta(dx)\pi(d\vartheta)$ (die gemeinsame Verteilung von Beobachtung und Parameter).

Bezeichnen wir mit X und θ die Koordinatenprojektionen von $\Omega \times \Theta$ auf Ω bzw. Θ, so gilt damit

$$R_\pi(\delta) = \mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta, \delta(X))\big].$$

(2) Ist $\forall \vartheta \in \Theta$ das Maß $\mathbb{P}_\vartheta$ absolutstetig bezüglich μ und π absolutstetig bezüglich ν mit Dichten $f_{X|\theta=\vartheta}$ bzw. f_θ und ist ferner $f_{X|\theta} : \Omega \times \Theta \to \mathbb{R}_{\geq 0}$ $(\mathcal{F} \otimes \mathcal{F}_\Theta)$-messbar, so definieren wir die a posteriori-Verteilung des Parameters (in Zeichen: $\mathbb{P}^{\theta|X=x}$) vermittels der folgenden ν-Dichte:

$$f_{\theta|X=x}(\vartheta) = \frac{f_\theta(\vartheta) \cdot f_{X|\theta=\vartheta}(x)}{\int_\Theta f_{X|\theta=\tilde{\vartheta}}(x) f_\theta(\tilde{\vartheta}) \nu(d\tilde{\vartheta})}$$

(Bayesformel für Dichten).

(3) Erhalten wir bei Wahl einer parametrischen Klasse von a priori-Verteilungen für ein statistisches Modell dieselbe Klasse (nur mit "upgedateten" Parametern) als a posteriori-Verteilungen zurück, so nennt man die entsprechenden Verteilungsklassen konjugiert.

Für komplexere Modelle ohne konjugierte Verteilungsklassen ist die Berechnung von a posteriori-Verteilungen in der Regel nur numerisch möglich; es kommen dabei sogenannte Markov Chain Monte Carlo (MCMC)-Algorithmen zum Einsatz. In der Praxis sind Bayesianische Methoden sehr beliebt.

Beispiel 1.7

(a) Unter dem statistischen Modell aus Beispiel 1.4(a) (Normalverteilungen mit unbekanntem Erwartungswert ϑ und bekannter Varianz $\sigma^2 = 1$, n-faches Produktexperiment) greifen wir das statistische Entscheidungsproblem $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), ((\mathcal{N}(\vartheta, 1))^n)_{\vartheta \in \mathbb{R}}, \mathbb{R}, \mathcal{B}(\mathbb{R}), L)$ der Punktschätzung mit $L(\vartheta, a) = (\vartheta - a)^2$ wieder auf und betrachten die drei Entscheidungsregeln

$$\begin{aligned} \hat{\vartheta}_1(x) &= n^{-1} \sum_{i=1}^n x_i =: \bar{x}_n, \\ \hat{\vartheta}_2(x) &= \bar{x}_n + 1/2 \text{ und} \\ \hat{\vartheta}_3(x) &\equiv 17. \end{aligned}$$

Wegen $R(\vartheta, \hat{\vartheta}_1) = 1/n < 1/n + 1/4 = R(\vartheta, \hat{\vartheta}_2)$ ist $\hat{\vartheta}_1$ besser als $\hat{\vartheta}_2$ und $\hat{\vartheta}_2$ damit unzulässig. Allerdings ist weder $\hat{\vartheta}_1$ besser als $\hat{\vartheta}_3$ noch umgekehrt. $\hat{\vartheta}_3$ ist zulässig, da $R(\vartheta, \hat{\vartheta}_3) = 0$ für $\vartheta = 17$ und L nicht-negativ ist.

(b) Unter den generellen Gegebenheiten von Beispiel 1.4(b) (Hypothesentest) seien sowohl Θ_0 als auch Θ_1 jeweils einelementig ("einfach"), also $\Theta = \{\vartheta_0, \vartheta_1\}$. Damit ist eine jede a priori-Verteilung π durch die Angabe von $\pi_0 := \pi(\{\vartheta_0\})$ und $\pi_1 := \pi(\{\vartheta_1\})$ festgelegt. Die Wahrscheinlichkeitsmaße $\mathbb{P}_{\vartheta_0}$ und $\mathbb{P}_{\vartheta_1}$ mögen Dichten $f_{X|\theta=\vartheta_0} =: p_0$ und $f_{X|\theta=\vartheta_1} =: p_1$ bezüglich eines Maßes μ (z.B. $\mu = \mathbb{P}_0 + \mathbb{P}_1$) besitzen. π besitzt offensichtlich eine Zähldichte.

Nach der Bayesformel ist die a posteriori-Verteilung festgelegt durch

$$\tilde{\mathbb{P}}(\theta = \vartheta_j | X = x) = \frac{\pi_j p_j(x)}{\sum_{\ell=0}^1 \pi_\ell p_\ell(x)}, j = 0, 1 \quad (\tilde{\mathbb{P}}^X - \text{fast überall}).$$

Erinnerung: Absolutstetigkeit

$(\Omega, \mathcal{F})$ ein Messraum, $\mathbb{P}_\vartheta$ und μ zwei Maße auf $(\Omega, \mathcal{F})$.

$$\mathbb{P}_\vartheta \text{ ist absolutstetig bezüglich } \mu :\Leftrightarrow \mu(B) = 0 \Rightarrow \mathbb{P}_\vartheta(B) = 0.$$

Also:

$$\mathbb{P}_\vartheta \text{ absolutstetig bezüglich } \mu \;\Leftrightarrow\; \{N : N \text{ Nullmenge bzgl. } \mathbb{P}_\vartheta\} \supseteq \{\tilde{N} : \tilde{N} \text{ Nullmenge bzgl. } \mu\}.$$

Satz von Radon-Nikodym:

$\mathbb{P}_\vartheta$ absolutstetig bezüglich $\mu \Leftrightarrow \mathbb{P}_\vartheta$ besitzt eine μ-Dichte.

Beweis von "$\Leftarrow$" durch Widerspruch:

Falls $\mathbb{P}_\vartheta$ nicht absolutstetig bezüglich μ ist, so $\exists \tilde{N} \in \mathcal{F} : \tilde{N}$ Nullmenge von μ, aber nicht Nullmenge von $\mathbb{P}_\vartheta \Rightarrow$

$$\int_{\tilde{N}} f d\mu = 0 \neq \mathbb{P}_\vartheta(\tilde{N})$$

für alle als Dichte in Frage kommenden Funktionen $f \Rightarrow \mathbb{P}_\vartheta$ besitzt keine μ-Dichte.

Satz 1.8 (Kriterium für Bayes-Optimalität)

Eine Regel δ^* ist Bayes-optimal, falls $\delta^*(X) = \underset{a \in A}{\operatorname{argmin}} \mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta, a)|X\big]$ $\tilde{\mathbb{P}} - f.s.$, d.h.

$$\mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta, \delta^*(x))|X = x\big] \leq \mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta, a)|X = x\big] \forall a \in A \text{ und für } \tilde{\mathbb{P}}^X\text{-fast alle } x \in \Omega.$$

Beweis: Sei δ eine beliebige Entscheidungsregel. Dann ist

$$R_\pi(\delta) = \mathbb{E}_{\tilde{\mathbb{P}}}\big[\mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta, \delta(X))|X\big]\big] \geq \mathbb{E}_{\tilde{\mathbb{P}}}\big[\mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta, \delta^*(X))|X\big]\big] = R_\pi(\delta^*).$$

■

Korollar 1.9

Sei das statistische Entscheidungsproblem (Schätzproblem) $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta \subseteq \mathbb{R}}, \mathbb{R}, \mathcal{B}(\mathbb{R}), L)$ gegeben.

(a) Für $L(\vartheta, a) = (\vartheta - a)^2$ ist die bedingte Erwartung $\mathbb{E}_{\tilde{\mathbb{P}}}\big[\theta|X\big]$ (also der a posteriori-Mittelwert) Bayes- optimaler Schätzer von ϑ bezüglich der a priori-Verteilung π.

(b) Für $L(\vartheta, a) = |\vartheta - a|$ ist jeder a posteriori-Median, d.h. jedes $\hat{\vartheta}_\pi$ mit $\tilde{\mathbb{P}}(\theta \leq \hat{\vartheta}_\pi|X) \geq \frac{1}{2}$ und $\tilde{\mathbb{P}}(\theta \geq \hat{\vartheta}_\pi|X) \geq \frac{1}{2}$ Bayes-optimaler Schätzer (falls die a posteriori-Verteilung existiert).

Beweis: L_2-Projektionseigenschaft der bedingten Erwartung, L_1-Minimierungseigenschaft des (eines) Medians. ■

Beispiel 1.10 (Fortsetzung von 1.7(b))
Nach Satz 1.8 muss die Minimalstelle von $\mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta,a)|X=x\big]$ bestimmt werden, um die optimale Entscheidungsregel zu finden. Der Parameterraum $\Theta=\{\vartheta_0,\vartheta_1\}$ ist diskret, also ist

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbb{P}}}\big[L(\theta,a)|X=x\big] &= \sum_{j=0}^{1} L(\vartheta_j,a)\tilde{\mathbb{P}}(\theta=\vartheta_j|X=x)\\ &= L(\vartheta_0,a)\cdot\tilde{\mathbb{P}}(\theta=\vartheta_0|X=x)+L(\vartheta_1,a)\cdot\tilde{\mathbb{P}}(\theta=\vartheta_1|X=x)\\ &= \frac{\ell_1\cdot a\cdot\pi_0p_0(x)+\ell_2(1-a)\pi_1p_1(x)}{\pi_0p_0(x)+\pi_1p_1(x)}\end{aligned}$$

Der Nenner ist offenbar unabhängig von a. Die Minimierung des Zählers bezüglich $a\in\{0,1\}$ erfolgt durch $a=0$, falls $\ell_1\pi_0p_0(x)>\ell_2\pi_1p_1(x)$ ist und durch $a=1$, falls $\ell_2\pi_1p_1(x)>\ell_1\pi_0p_0(x)$ ist. Also folgt:

$$\delta^*(x)=\begin{cases}0, & \text{falls } \ell_1\pi_0p_0(x)>\ell_2\pi_1p_1(x)\\ 1, & \text{falls } \ell_2\pi_1p_1(x)>\ell_1\pi_0p_0(x)\\ \text{beliebig}, & \text{falls } \ell_2\pi_1p_1(x)=\ell_1\pi_0p_0(x)\end{cases}$$

ist Bayes-Klassifikator (Bayestest) für das Problem 1.7(b).
Ist speziell $\ell_1=\ell_2$ gewählt, so heißt das ϑ_j, für das wir uns entscheiden, „maximum a posteriori (MAP)“-Schätzer.

1.2 Grundlagen der Schätztheorie

Definition 1.11
Es sei $(\Omega,\mathcal{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $p\in\mathbb{N}$, $\varrho(\vartheta)$ mit $\varrho:\Theta\to\mathbb{R}^p$ ein (abgeleiteter) Parameter und L eine Verlustfunktion.
Das statistische Entscheidungsproblem $(\Omega,\mathcal{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta},\mathbb{R}^p,\mathcal{B}(\mathbb{R}^p),L)$ heißt Schätzproblem für $\varrho(\vartheta)$.
Eine Entscheidungsregel $\hat{\varrho}:\Omega\to\mathbb{R}^p$ heißt Schätzvorschrift, die Zufallsgröße $\hat{\varrho}(X)$ heißt Schätzer für $\varrho(\vartheta)$ und der Wert $\hat{\varrho}(x)\in\mathbb{R}^p$ heißt Schätzwert für $\varrho(\vartheta)$ gegeben die Beobachtung $X=x$.
$b(\hat{\varrho},\vartheta):=\mathbb{E}_\vartheta\big[\hat{\varrho}\big]-\varrho(\vartheta)$ heißt Verzerrung (englisch: bias) von $\hat{\varrho}$ bzw. $\hat{\varrho}(X)$.
Der Schätzer $\hat{\varrho}(X)$ heißt erwartungstreu bzw. unverzerrt, falls $\forall\vartheta\in\Theta: b(\hat{\varrho},\vartheta)=0$.

Lemma 1.12 (Bias-Varianz-Zerlegung)
Unter den Gegebenheiten von Definition 1.11 sei $p=1$ und L der quadratische Verlust, d.h.

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

(a) Das quadratische Risiko eines Schätzers $\hat{\varrho}(X)$ mit endlicher Varianz lässt sich zerlegen in

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

(b) Das quadratische Risiko eines erwartungstreuen, quadratintegrierbaren, reellwertigen Schätzers ist seine Varianz.

Beweis: Teil (b) ist eine unmittelbare Konsequenz aus Teil (a). Zum Beweis von (a) rechnen wir

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

■

Definition 1.13 (Wünschenswerte Eigenschaften von Schätzern)
Sei $(\Omega,\mathcal{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta},\mathbb{R},\mathcal{B}(\mathbb{R}),L)$ ein Schätzproblem, $\varrho(\vartheta)$ der interessierende (abgeleitete) Parameter und $\hat{\varrho}$ eine Schätzvorschrift.

(a) Der Schätzer $\hat{\varrho}(X)$ heißt erwartungstreu, falls $\mathbb{E}_\vartheta\big[\hat{\varrho}\big] = \varrho(\vartheta)\ \forall\vartheta\in\Theta$ gilt.

(b) Falls $\hat{\varrho}^*(X)$ erwartungstreu ist, so heißt $\hat{\varrho}^*(X)$ effizient (bzw. UMVU), falls ($\forall\vartheta\in\Theta$):

$$\mathit{Var}_\vartheta\left(\hat{\varrho}^*\right) = \inf_{\hat{\varrho}:\hat{\varrho}(X)\ \text{erwartungstreu}} \mathit{Var}_\vartheta\left(\hat{\varrho}\right).$$

(c) Ist $n\in\mathbb{N}$ ein Stichprobenumfang und $\Omega\subseteq\mathbb{R}^n$, so heißt $\hat{\varrho}(X)=\hat{\varrho}_n(X)$ konsistent bzw. stark konsistent, falls $\hat{\varrho}(X)\to\varrho(\vartheta)$ für $n\to\infty$ $\mathbb{P}_\vartheta$-stochastisch bzw. $\mathbb{P}_\vartheta$-fast sicher.

(d) Der Schätzer $\hat{\varrho}(X)$ heißt asymptotisch normalverteilt, falls $0<\mathbb{E}_\vartheta\big[(\hat{\varrho})^2\big]<\infty$ und

$$\mathcal{L}\left(\frac{\hat{\varrho}(X)-\mathbb{E}_\vartheta\big[\hat{\varrho}\big]}{\sqrt{\mathit{Var}_\vartheta\left(\hat{\varrho}\right)}}\right) \xrightarrow[n\to\infty]{w} \mathcal{N}(0,1)\ \text{unter}\ \mathbb{P}_\vartheta.$$

Definition 1.14
Ein statistisches Modell $(\Omega,\mathcal{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt dominiert (vom Maß μ), falls es ein σ-endliches Maß μ auf $\mathcal{F}$ gibt, so dass für alle $\vartheta\in\Theta$ das Wahrscheinlichkeitsmaß $\mathbb{P}_\vartheta$ absolutstetig bezüglich μ ist (in Zeichen: $\forall\vartheta\in\Theta:\mathbb{P}_\vartheta << \mu$). Die durch ϑ parametrisierte Radon-Nikodym-Dichte

$$l(\vartheta,x) := \frac{d\mathbb{P}_\vartheta}{d\mu}(x), \vartheta\in\Theta, x\in\Omega$$

heißt Likelihoodfunktion, wobei sie meistens für festgehaltenes (beobachtetes) $x \in \Omega$ *als Funktion von* $\vartheta \in \Theta$ *aufgefasst wird.*

Anmerkung: Die Familie aller stetigen Verteilungen auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ ist dominiert von λ^n. Jedes statistische Modell auf einem abzählbaren Stichprobenraum Ω ist dominiert vom Zählmaß.

Definition 1.15

Es sei $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ *mit* $\Theta \subseteq \mathbb{R}^k$ *ein von* μ *dominiertes statistisches Modell mit Likelihoodfunktion* $l(\vartheta, x)$.
Falls $\vartheta \mapsto \ln(l(\vartheta, x))$ *für* μ*-fast alle* x *in* ϑ_0 *differenzierbar ist, nennen wir*

$$x \mapsto \frac{\partial}{\partial\vartheta} \ln(l(\vartheta, x))|_{\vartheta=\vartheta_0} =: \dot{l}(\cdot, \vartheta_0) \quad \text{Score-Funktion},$$

wobei $\partial/(\partial\vartheta)$ *den Gradient-Operator (Vektor der partiellen Ableitungen) bezeichnet.*
Die $(k \times k)$*-Matrix*

$$I(\vartheta_0) := \mathbb{E}_{\vartheta_0}\big[\dot{l}(\cdot, \vartheta_0)(\dot{l}(\cdot, \vartheta_0))^\top\big]$$

heißt Fisher-Information im Punkte ϑ_0.

Beispiel 1.16

Wir betrachten das Normalverteilungsmodell $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (\mathcal{N}(\mu, \sigma^2))_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_{>0}})$. *Die* λ*-Dichte von* $\mathcal{N}(\mu, \sigma^2)$ *ist gegeben durch*

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) = l(\vartheta, x);\ \vartheta = (\mu, \sigma^2)^t.$$

Wir berechnen die Fisher-Information im Punkte $(\mu_0, \sigma_0^2) =: \vartheta_0$ *und erhalten*

$$\begin{aligned}
\ln(l(\vartheta, x)) &= \ln(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{(x-\mu)^2}{2\sigma^2}, \\
\frac{\partial \ln(l(\vartheta, x))}{\partial\mu} &= \frac{x-\mu}{\sigma^2}, \\
\frac{\partial \ln(l(\vartheta, x))}{\partial\sigma^2} &= \frac{(x-\mu)^2 - \sigma^2}{2\sigma^4} = \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \\
&\Rightarrow \dot{l}(x, \vartheta_0)(\dot{l}(x, \vartheta_0))^t = \begin{pmatrix} \frac{(x-\mu_0)^2}{\sigma_0^4} & \frac{(x-\mu_0)^3}{2\sigma_0^6} - \frac{(x-\mu)}{2\sigma_0^4} \\ \frac{(x-\mu_0)^3}{2\sigma_0^6} - \frac{(x-\mu_0)}{2\sigma_0^4} & \frac{[(x-\mu_0)^2-\sigma_0^2]^2}{4\sigma_0^8} \end{pmatrix} \\
&\Rightarrow I(\vartheta_0) = \begin{pmatrix} \sigma_0^{-2} & 0 \\ 0 & \frac{1}{2\sigma_0^4} \end{pmatrix}.
\end{aligned}$$

Lemma 1.17

Es seien $X_1, \ldots, X_n$ *Zufallsgrößen, die stochastisch unabhängige Experimente mit ein und derselben Parametermenge* $\Theta \subseteq \mathbb{R}^k$ *induzieren. Existiert für alle* $1 \le j \le n$ *die jeweilige Fisher-*

Information I_j auf ganz Θ, so existiert die gemeinsame, von $X = (X_1, \ldots, X_n)$ erzeugte Fisher-Information I und es gilt für alle $\vartheta \in \Theta$:

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

Beweis: Die gemeinsame Log-Likelihoodfunktion ist gegeben durch

$$\ln(l(\vartheta, x)) = \sum_{j=1}^{n} \ln(l_j(\vartheta, x_j)) \quad \text{bezüglich } \bigotimes_{j=1}^{n} \mu_j.$$

Nach Voraussetzung ist $\ln(l(\vartheta, x))$ zudem fast überall differenzierbar mit Score-Funktion

$$\dot{l}(x, \vartheta) = \sum_{j=1}^{n} \dot{l}_j(x_j, \vartheta).$$

Nach Übungsaufgabe gilt zudem $\mathbb{E}_\vartheta\big[\dot{l}_j(X_j, \vartheta)\big] = 0 \quad \forall 1 \leq j \leq n$. Damit errechnen wir:

$$\begin{aligned}
\mathbb{E}_\vartheta\big[\dot{l}(X, \vartheta)(\dot{l}(X, \vartheta))^\top\big] &= \mathbb{E}_\vartheta\left[\left(\sum_{j=1}^{n} \dot{l}_j(X_j, \vartheta)\right)\left(\sum_{j=1}^{n} \dot{l}_j(X_j, \vartheta)^\top\right)\right] \\
&= \sum_{k=1}^{n}\sum_{m=1}^{n} \mathbb{E}_\vartheta\big[\dot{l}_k(X_k, \vartheta)(\dot{l}_m(X_m, \vartheta))^\top\big] \\
&= \sum_{j=1}^{n} \mathbb{E}_\vartheta\big[\dot{l}_j(X_j, \vartheta)(\dot{l}_j(X_j, \vartheta))^\top\big].
\end{aligned}$$

■

Satz 1.18 (Cramér-Rao-Schranke)
Seien $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}^k, k \in \mathbb{N}$ ein statistisches Modell, $\varrho : \Theta \to \mathbb{R}$ differenzierbar in $\vartheta_0 \in \Theta \setminus \partial\Theta$ und $\hat{\varrho}(X)$ ein erwartungstreuer Schätzer für $\varrho(\vartheta)$. Für alle ϑ in einer Umgebung von ϑ_0 gelte $\mathbb{P}_\vartheta << \mathbb{P}_{\vartheta_0}$.
Ferner sei die Likelihoodfunktion $l(\vartheta, x)$ $L_2(\mathbb{P}_{\vartheta_0})$-differenzierbar in ϑ_0, d.h.

$$\exists g : \Theta \times \Omega \to \mathbb{R}^k \text{ mit } \lim_{\vartheta\to\vartheta_0} \frac{\mathbb{E}_{\vartheta_0}\big[|l(\vartheta, \cdot) - l(\vartheta_0, \cdot) - < g(\vartheta_0, \cdot), \vartheta - \vartheta_0 > |^2\big]}{|\vartheta - \vartheta_0|^2} = 0.$$

Falls die Fisher-Information $I(\vartheta_0)$ im Punkte ϑ_0 endlich und strikt positiv definit ist, so gilt:

$$\mathbb{E}_{\vartheta_0}\big[(\hat{\varrho} - \varrho(\vartheta_0))^2\big] = \mathit{Var}_{\vartheta_0}(\hat{\varrho}) \geq < I(\vartheta_0)^{-1}\dot{\varrho}(\vartheta_0), \dot{\varrho}(\vartheta_0) > .$$

Beweis: Satz 2.124 in Witting (1985). ■

Beispiel 1.19
Sei $X = (X_1, \ldots, X_n)$ nach $(\mathcal{N}(\mu, \sigma^2))^n$ verteilt. Dabei sei $\mu \in \mathbb{R}$ der Parameter von Interesse und $\sigma^2 > 0$ bekannt.
Sei $\hat{\mu}(X) = \bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Dann ist $\hat{\mu}(X)$ erwartungstreu und es gilt $\mathit{Var}_\mu(\hat{\mu}) = \frac{\sigma^2}{n}$ und $I(\mu) = \frac{n}{\sigma^2}$ nach Beispiel 1.16 mit Lemma 1.17. Also ist $\hat{\mu}$ Cramér-Rao effizient, denn $\varrho = id$.

1.3 Grundlagen der Testtheorie

1.3.1 Allgemeine Testtheorie

Wir greifen Beispiel 1.4.(b) noch einmal auf und studieren Testprobleme als binäre statistische Entscheidungsprobleme: Gegeben zwei disjunkte, nicht-leere Teilmengen $\mathcal{P}_0, \mathcal{P}_1$ von $\mathcal{P} = (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ mit $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$ ist eine Entscheidung darüber gesucht, ob $\mathbb{P}^X$ zu $\mathcal{P}_0$ oder $\mathcal{P}_1$ gehört. Falls $\mathcal{P}$ durch ϑ eineindeutig identifiziert ist, kann die Entscheidungsfindung auch vermittels ϑ und Teilmengen Θ_0 und Θ_1 von Θ mit $\Theta_0 \cap \Theta_1 = \emptyset$ und $\Theta_0 \cup \Theta_1 = \Theta$ formalisiert werden.
Formale Beschreibung des Testproblems:

$$H_0 : \vartheta \in \Theta_0 \quad \text{versus} \quad H_1 : \vartheta \in \Theta_1 \quad \text{oder}$$

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

Die $H_i, i = 1, 2$ nennt man Hypothesen. H_0 heißt Nullhypothese, H_1 Alternativhypothese / Alternative. Oft interpretiert man H_0 und H_1 auch direkt selbst als Teilmengen des Parameterraums, d. h., $H_0 \cup H_1 = \Theta$ und $H_0 \cap H_1 = \emptyset$. Zwischen H_0 und H_1 ist nun aufgrund von $x \in \Omega$ eine Entscheidung zu treffen. Die dazu benötigte Entscheidungsregel nennt man einen statistischen Test.

Definition 1.20 (Statistischer Test)
Ein (nicht-randomisierter) statistischer Test ist eine messbare Abbildung

$$\varphi : (\Omega, \mathcal{F}) \to (\{0, 1\}, 2^{\{0,1\}}).$$

Konvention:

$$\varphi(x) = 1 \iff \textit{Nullhypothese wird verworfen, Entscheidung für } H_1,$$

$$\varphi(x) = 0 \iff \textit{Nullhypothese wird nicht verworfen.}$$

$\{x \in \Omega : \varphi(x) = 1\}$ *heißt Ablehnbereich (oder auch kritischer Bereich) von* φ*, kurz:* $\{\varphi = 1\}$.
$\{x \in \Omega : \varphi(x) = 0\}$ *heißt Annahmebereich von* φ*, kurz:* $\{\varphi = 0\} = \complement\{\varphi = 1\}$.

Problem: Testen beinhaltet mögliche Fehlentscheidungen.
Fehler 1. Art (α-Fehler, type I error): Entscheidung für H_1, obwohl H_0 wahr ist.
Fehler 2. Art (β-Fehler, type II error): Nicht-Verwerfung von H_0, obwohl H_1 wahr ist.
In der Regel ist es nicht möglich, die Wahrscheinlichkeiten für die Fehler 1. und 2. Art gleichzeitig zu minimieren. Daher findet in der frequentistischen Statistik eine asymmetrische Betrachtungsweise von Testproblemen statt.

(i) Begrenzung der Fehlerwahrscheinlichkeit 1. Art durch eine vorgegebene obere Schranke α (Signifikanzniveau, englisch: level),

(ii) Unter der Maßgabe (i) Minimierung der Wahrscheinlichkeit für Fehler 2. Art $\Rightarrow$ „optimaler“ Test.

Eine (zum Niveau α) statistisch abgesicherte Entscheidung kann also immer nur zu Gunsten von H_1 getroffen werden $\Rightarrow$ Merkregel: „Was nachzuweisen ist stets als Alternative H_1 formulieren!“.

Bezeichnungen 1.21

(i) $\beta_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi] = \mathbb{P}_\vartheta(\varphi(X) = 1) = \int_\Omega \varphi d\mathbb{P}_\vartheta$ *bezeichnet die Ablehnwahrscheinlichkeit eines vorgegebenen Tests* φ *in Abhängigkeit von* $\vartheta \in \Theta$. *Für* $\vartheta \in \Theta_1$ *heißt* $\beta_\varphi(\vartheta)$ *Gütefunktion von* φ *an der Stelle* ϑ. *Für* $\vartheta \in \Theta_0$ *ergibt* $\beta_\varphi(\vartheta)$ *die Typ I-Fehlerwahrscheinlichkeit von* φ *unter* $\vartheta \in \Theta_0$.

Für $\alpha \in (0, 1)$ *vorgegeben heißt*

(ii) ein Test φ *mit* $\beta_\varphi(\vartheta) \leq \alpha$ *für alle* $\vartheta \in H_0$ *Test zum Niveau* α,

(iii) ein Test φ *zum Niveau* α *unverfälscht, falls* $\beta_\varphi(\vartheta) \geq \alpha$ *für alle* $\vartheta \in H_1$.

(iv) ein Test φ_1 *zum Niveau* α *besser als ein zweiter Niveau-*α *Test* φ_2, *falls* $\beta_{\varphi_1}(\vartheta) \geq \beta_{\varphi_2}(\vartheta)$ *für alle* $\vartheta \in H_1$ *und* $\exists \vartheta^* \in H_1$ *mit* $\beta_{\varphi_1}(\vartheta^*) > \beta_{\varphi_2}(\vartheta^*)$.

Wir betrachten in der Folge in aller Regel die Menge $\mathcal{M}$ *der Niveau* α*-Tests mit der Risikofunktion* $R(\vartheta, \varphi) = 1 - \beta_\varphi(\vartheta),\ \vartheta \in \Theta_1$. *Unter diesen Prämissen ist das Testproblem dann bereits vollständig spezifiziert durch* $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, H_0)$.

Beispiel 1.22 (Einseitiger Binomialtest)
Von den 13 *Todesfällen unter* 55*- bis* 64*-jährigen Arbeitern eines Kernkraftwerkes in Jahre 1995 waren* 5 *auf einen Tumor zurückzuführen.*
Die Todesursachenstatistik 1995 weist aus, dass Tumore bei etwa $1/5$ *aller Todesfälle die Todesursache in der betreffenden Altersklasse (in der Gesamtbevölkerung) darstellen. Ist die beobachtete Häufung von tumorbedingten Todesfällen unter den Arbeitern im Kernkraftwerk signifikant auffällig zum Niveau* $\alpha = 5\%$ *oder noch “kompatibel” mit den Gegebenheiten in der Gesamtpopulation?*
Bezeichne dazu die Zufallsvariable X *die Anzahl der Tumortoten unter* $n = 13$ *Todesfällen von AKW-MitarbeiterInnen. Wir modellieren* $\Omega = \{0, \ldots, n = 13\},\ \mathcal{F} = 2^\Omega, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta} = (Bin(13, p))_{p\in[0,1]}$ *und haben* $H_0 = \{p \leq 1/5\}$ *zu testen.*
Betrachten wir speziell nicht-randomisierte Tests φ *der Form* $\varphi(x) = 1 \Leftrightarrow x > c_\alpha$ *mit kritischen Bereichen* $\Gamma_\alpha = (c_\alpha, \infty)$. *Um die Einhaltung des Signifikanzniveaus* $\alpha = 5\%$ *sicherzustellen, muss* $\sup_{0\leq p\leq 1/5} \mathbb{P}_p(X > c_\alpha) \leq \alpha$ *bzw. äquivalent dazu* $\inf_{0\leq p\leq 1/5} \mathbb{P}_p(X \leq c_\alpha) \geq 1 - \alpha$ *gelten.*

Für festes $k \in \Omega$ ist

$$\mathbb{P}_p(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = l(p,k) \text{ und } \mathbb{P}_p(X \leq k) = \sum_{l=0}^{k} \binom{n}{l} p^l (1-p)^{n-l} =: F(p,k).$$

Eine einfache Kurvendiskussion zeigt, dass $\forall k \in \Omega : F(p,k)$ fallend auf $\Theta_0 = [0, 1/5]$ ist.
Damit ist für alle $k \in \Omega$ $\inf\limits_{0 \leq p \leq 1/5} \mathbb{P}_p(X \leq k) = \mathbb{P}_{1/5}(X \leq k)$ und c_α wird so bestimmt, dass

$$c_\alpha = \min\{k \in \Omega : \sum_{\ell=0}^{k} \binom{n}{\ell} (\frac{1}{5})^\ell (\frac{4}{5})^{n-\ell} \geq 1 - \alpha\},$$

damit die Typ II-Fehlerwahrscheinlichkeit möglichst klein wird.
Wir erhalten:

$$\sum_{\ell=0}^{4} \binom{13}{\ell} (\frac{1}{5})^\ell (\frac{4}{5})^{13-\ell} \approx 0{,}901 \text{ und } \sum_{\ell=0}^{5} \binom{13}{\ell} (\frac{1}{5})^\ell (\frac{4}{5})^{13-\ell} \approx 0{,}9700.$$

Damit wird $c_\alpha = 5$ gewählt und H_0 kann bei der tatsächlich beobachteten Datenlage $x = 5$ nicht verworfen werden.

Definition 1.23 (p-Wert)
Sei $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell und sei φ ein Test für das Hypothesenpaar $\emptyset \neq H_0 \subset \Theta$ versus $H_1 = \Theta \setminus H_0$, der auf einer Prüfgröße $T : \Omega \to \mathbb{R}$ basiert. φ sei charakterisiert durch die Angabe von Ablehnbereichen $\Gamma_\alpha \subset \mathbb{R}$ für jedes Signifikanzniveau $\alpha \in (0,1)$, so dass $\varphi(x) = 1 \iff T(x) \in \Gamma_\alpha$ für $x \in \Omega$ gilt. Dann ist der p-Wert einer Realisierung $x \in \Omega$ bezüglich φ definiert als

$$p_\varphi(x) = \inf_{\{\alpha : T(x) \in \Gamma_\alpha\}} \mathbb{P}^*(T(X) \in \Gamma_\alpha),$$

wobei das Wahrscheinlichkeitsmaß $\mathbb{P}^*$ so gewählt ist, dass

$$\mathbb{P}^*(T(X) \in \Gamma_\alpha) = \sup_{\vartheta \in H_0} \mathbb{P}_\vartheta(T(X) \in \Gamma_\alpha)$$

gilt, falls H_0 eine zusammengesetzte Nullhypothese ist.

Bemerkung 1.24

(i) Falls H_0 einelementig („einfach“) und $\mathbb{P}_{H_0} \equiv \mathbb{P}_{\vartheta_0}$ ein stetiges Wahrscheinlichkeitsmaß ist, so gilt (in aller Regel)

$$p_\varphi(x) = \inf\{\alpha : T(x) \in \Gamma_\alpha\}.$$

(ii) p-Werte werden häufig auch als „beobachtete Signifikanzniveaus“ bezeichnet.

(iii) Sei Ω^{-1} der Urbildraum von X. Die Abbildung $p_\varphi(X) : \Omega^{-1} \to [0,1], \omega \mapsto p_\varphi(X(\omega))$, lässt sich als Zufallsvariable auffassen. Leider wird sie dennoch üblicherweise mit Kleinbuchstabe bezeichnet, um Verwechslungen mit (indizierten) Wahrscheinlichkeitsmaßen vorzubeugen. Es muss also häufig aus dem Kontext heraus interpretiert werden, ob $p_\varphi \equiv p$ einen realisierten Wert aus $[0,1]$ oder eine Zufallsvariable meint.

Definition 1.25

Unter den Voraussetzungen von Definition 1.23 sei die Teststatistik $T(X)$ derart, dass die Monotoniebedingung

$$\forall \vartheta_0 \in H_0 : \forall \vartheta_1 \in H_1 : \forall c \in \mathbb{R} : \mathbb{P}_{\vartheta_0}(T(X) > c) \leq \mathbb{P}_{\vartheta_1}(T(X) > c) \tag{1.1}$$

gilt. Dann heißt φ ein Test vom (verallgemeinerten) Neyman-Pearson Typ, falls für alle $\alpha \in (0,1)$ eine Konstante c_α existiert, so dass

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

Bemerkung 1.26

(a) Die Monotoniebedingung (1.1) wird häufig so umschrieben, dass „die Teststatistik unter Alternativen zu größeren Werten neigt“.

(b) Die zu einem Test vom Neyman-Pearson (N-P) Typ gehörigen Ablehnbereiche sind gegeben als $\Gamma_\alpha = (c_\alpha, \infty)$.

(c) Die Konstanten c_α werden in der Praxis bestimmt über $c_\alpha = \inf\{c \in \mathbb{R} : \mathbb{P}^*(T(X) > c) \leq \alpha\}$ mit $\mathbb{P}^*$ wie in Definition 1.23 („am Rande der Nullhypothese“). Ist H_0 einelementig und $\mathbb{P}_{H_0}$ stetig, so gilt $c_\alpha = F_T^{-1}(1-\alpha)$, wobei F_T die Verteilungsfunktion von $T(X)$ unter H_0 bezeichnet.

(d) Fundamentallemma der Testtheorie von Neyman und Pearson: Unter (leicht verschärftem) (1.1) ist ein Test vom N-P Typ gleichmäßig (über alle $\vartheta_1 \in H_1$) bester Test für H_0 versus H_1.

Lemma 1.27

Sei φ ein Test vom N-P Typ und $\mathbb{P}^*$ unabhängig von α. Dann gilt für die Berechnung des p-Wertes einer Realisierung $x \in \Omega$ bezüglich φ, dass

$$p_\varphi(x) = \mathbb{P}^*(T(X) \geq t^*) \text{ mit } t^* := T(x).$$

Beweis: Die Ablehnbereiche $\Gamma_\alpha = (c_\alpha, \infty)$ sind geschachtelt. Demnach wird $\inf\{\alpha : T(x) \in \Gamma_\alpha\}$ offensichtlich in $[t^*, \infty)$ angenommen. Aufgrund der Struktur dieses Ablehnbereiches gilt ferner $\mathbb{P}^*(T(X) \in [t^*, \infty)) = \mathbb{P}^*(T(X) \geq t^*)$. ■

Anmerkung: Ist H_0 einelementig, $\mathbb{P}_{H_0}$ stetig und φ vom N-P Typ, so gilt mit den Bezeichnungen aus Bemerkung 1.26 und Lemma 1.27 für alle $x \in \Omega$, dass $p_\varphi(x) = 1 - F_T(t^*)$.

Satz 1.28 (Testen mit dem p-Wert)
*Sei $\alpha \in (0,1)$ ein **fest vorgegebenes** Signifikanzniveau und $\mathbb{P}^*$ stetig.Dann gilt die Dualität*

$$\varphi(x) = 1 \Longleftrightarrow p_\varphi(x) < \alpha.$$

Beweis: Wir beweisen das Resultat hier nur für Tests vom N-P Typ. Da die Funktion $t \mapsto \mathbb{P}^*(T(X) > t)$ monoton fallend in t ist und aufgrund der Konstruktion von c_α (siehe 1.26.c) $\mathbb{P}^*(T(X) > c_\alpha) \leq \alpha$ sowie für alle $\mathbb{R} \ni c < c_\alpha : \mathbb{P}^*(T(X) > c) > \alpha$ gelten muss, ist $p_\varphi(x) < \alpha$ gleichbedeutend mit $t^* > c_\alpha$. Das führt bei einem Test vom N-P Typ aber gerade zur Ablehnung von H_0. ■

Bemerkung 1.29

(i) Der Vorteil von p-Werten für das Testen ist, dass sie unabhängig von einem a priori festgesetzten Signifikanzniveau α ausgerechnet werden können. Dies ist der Grund, warum alle gängigen Statistik-Softwaresysteme statistische Hypothesentests über die Berechnung von p-Werten implementieren. Aus puristischer Sicht birgt das jedoch Probleme, da man mit dieser Art des Testens tricksen kann. Hält man aich nämlich nicht an die gute statistische Praxis, alle Rahmenbedingungen des Experimentes (einschließlich des Signifikanzniveaus!) vor Erhebung der Daten festzulegen, so kann man der Versuchung erliegen, α erst a posteriori (nach Durchführung des Experimentes und Anschauen des resultierenden p-Wertes) zu setzen, um damit zu einer intendierten Schlussfolgerung zu kommen. Deswegen lehnen viele Statistiker die in satz 1.28 gezeigte Art des Testens strikt ab.

(ii) Die Interpretation des p-Wertes ist zu bedenken. Der p-Wert gibt eine Antwort auf die Frage: „Wie wahrscheinlich sind die gemessenen Daten, gegeben dass die Nullhypothese stimmt?“ und nicht auf die Frage „Wie wahrscheinlich ist es, dass die Nullhypothese wahr ist, gegeben die gemessenen Daten?“, obschon letztere Frage manchmal interessanter erscheinen mag und Praktiker ab und an dazu tendieren, den p-Wert dahingehend umzudeuten.

1.3.2 Tests für Parameter der Normalverteilung

Satz 1.30 (Multivariate Normalverteilung)
Seien $X_1, \ldots, X_d$ iid. standardnormalverteilte Zufallsvariablen. Dann heißt $X = (X_1, \ldots, X_d)^t$ standardnormalverteilt im $\mathbb{R}^d$.
Ist ferner $\Sigma = QQ^t \in \mathbb{R}^{m \times m}$ mit $Q \in \mathbb{R}^{m \times d}$ eine positiv definite, symmetrische Matrix und $Y = QX + \mu,\ \mu \in \mathbb{R}^m$, so heißt $Y = (Y_1, \ldots, Y_m)^t$ allgemein normalverteilt im $\mathbb{R}^m$, in Zeichen: $Y \sim \mathcal{N}_m(\mu, \Sigma)$. Es gilt:

a) Y hat die λ^m-Dichte

$$\varphi_{\mu,\Sigma}(y) = (2\pi)^{-m/2} |\det \Sigma|^{-1/2} \exp(-\frac{1}{2}(y-\mu)^t \Sigma^{-1}(y-\mu)).$$

b)

$$\forall 1 \leq j \leq m: \ \mathbb{E}[Y_j] = \mu_j, \quad \forall 1 \leq i,j \leq m: \mathit{Cov}\,(Y_i, Y_j) = \Sigma_{i,j}.$$

Beweis: Siehe Kapitel 3.1 in Fahrmeir and Hamerle (1984). ■

Satz 1.31 (Affine Transformationen)
Sei $Y \sim \mathcal{N}_m(\mu, \Sigma),\ k \leq m,\ A \in \mathbb{R}^{k \times m}$ eine Matrix mit maximalem Rang und $b \in \mathbb{R}^k$. Dann hat der Zufallsvektor $Z = AY + b$ die k-dimensionale Normalverteilung $\mathcal{N}_k(A\mu + b, A\Sigma A^t)$.

Beweis: Satz 9.5 in Georgii (2007). ■

Lemma 1.32
Ist X standardnormalverteilt auf $\mathbb{R}^1$, so hat X^2 die Gamma-Verteilung $\Gamma_{\frac{1}{2},\frac{1}{2}}$.

Beweis: Übung ■

Korollar 1.33
Seien $X_1, \ldots, X_n$ iid. auf $\mathbb{R}^1$ mit $\mathcal{L}(X_1) = \mathcal{N}(0,1)$. Dann ist

$$\sum_{i=1}^{n} X_i^2 \sim \Gamma_{\frac{1}{2},\frac{n}{2}} = \chi_n^2.$$

Beweis: Nach Lemma 1.32 ist $X_1 \sim \Gamma_{\frac{1}{2},\frac{1}{2}}$. Faltungsstabilität der Familie der Gammaverteilungen bezüglich des zweiten Parameters (siehe Aufgabe 4.6) liefert die Aussage. ■

Anmerkung: Die Verteilung von $\sum_{i=1}^n X_i^2$ wurde erstmals 1863 in der Dissertation von Ernst Abbe (später Carl Zeiss Jena) hergeleitet.

Lemma 1.34
Seien $\alpha, r, s > 0$ und X, Y stochastisch unabhängige Zufallsvariablen mit $X \sim \Gamma_{\alpha,r}$ und $Y \sim \Gamma_{\alpha,s}$. Dann sind $S = X + Y$ und $R = \frac{X}{X+Y}$ stochastisch unabhängig mit $S \sim \Gamma_{\alpha,r+s}$ und $R \sim$ Beta(r,s).

Beweis: Übung ■

Satz und Definition 1.35
Seien $X_1, \ldots, X_m,\ Y_1, \ldots, Y_n$ iid. standardnormalverteilt auf $\mathbb{R}^1$. Dann hat der Quotient

$$F_{m,n} := m^{-1} \sum_{i=1}^{m} X_i^2 \ / \ (n^{-1} \sum_{j=1}^{n} Y_j^2)$$

die folgende Verteilungsdichte bezüglich λ:

$$f_{m,n}(x) = \frac{m^{m/2}n^{n/2}}{B(m/2,n/2)} \frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}} \mathbf{1}_{(0,\infty)}(x).$$

Beweis: Übung ■

Die Verteilung von $F_{m,n}$ heißt Fisher'sche F-Verteilung mit m und n Freiheitsgraden (nach Sir R. A. Fisher, 1890-1962).

Korollar und Definition 1.36
Seien $X, Y_1, \ldots, Y_n$ iid. auf $\mathbb{R}$ mit $X \sim \mathcal{N}(0,1)$. Dann hat

$$T = \frac{X}{\sqrt{n^{-1}\sum_{j=1}^n Y_j^2}} \text{ die } \lambda\text{-Dichte } t \mapsto \tau_n(t) = (1+\frac{t^2}{n})^{-\frac{n+1}{2}} \{B(1/2,n/2)\sqrt{n}\}^{-1}.$$

Die Verteilung von T heißt Studentische t-Verteilung mit n Freiheitsgraden.

Beweis: Nach Satz 1.35 ist $T^2 \sim F_{1,n}$. Nach Transformationssatz hat daher $|T| = \sqrt{T^2}$ die Dichtefunktion $t \mapsto f_{1,n}(t^2) \cdot 2t,\ t > 0$. Wegen der Symmetrie von $\mathcal{N}(0,1)$ ist aber auch T symmetrisch um 0 verteilt, d.h., T und $-T$ haben die gleiche Verteilung. Also hat T die Verteilungsdichte $t \mapsto f_{1,n}(t^2) \cdot |t| = \tau_n(t)$. ■

Satz 1.37 (Student (1908))
Im Gaußmodell $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}_{\mu,\sigma^2})^n)_{\vartheta=(\mu,\sigma^2)\in\Theta:=\mathbb{R}\times(0,\infty)}$ gilt für alle $\vartheta \in \Theta$:

(a)

$$\bar{X}_n = n^{-1}\sum_{j=1}^n X_j \text{ und } S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

sind stochastisch unabhängig.

(b) $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ *und* $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

(c)

$$T_n := \frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \sim t_{n-1}.$$

Anmerkung: W. S. Gosset publizierte 1908 unter dem Pseudonym "Student", da sein Arbeitgeber, die Guinness-Brauerei, ihren Mitarbeitern die Veröffentlichung wissenschaftlicher Arbeiten verbot.
Damit folgen die auf dem Handout (Seiten 200-204 aus Witting (1985)) wiedergegebenen Standardtests für die Parameter der Normalverteilung allesamt aus der allgemeinen Testtheorie.

1.3.3 Bereichsschätzungen und der Korrespondenzsatz

Es gibt Dualitäten zwischen Testproblemen / Tests und (Bereichs-)Schätzproblemen / Konfidenzintervallen.

Definition 1.38
Gegeben sei ein statistisches Modell $(\Omega, \mathcal{F}, \mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\})$. *Dann heißt* $\mathcal{C} = (C(x) : x \in \Omega)$ *mit* $C(x) \subseteq \Theta \, \forall \, x \in \Omega$ *eine Familie von Konfidenzbereichen zum Konfidenzniveau* $1 - \alpha$ *für* $\vartheta \in \Theta :\Longleftrightarrow \forall \, \vartheta \in \Theta : \mathbb{P}_\vartheta \left(\{x : C(x) \ni \vartheta\} \right) \geq 1 - \alpha$.

Satz 1.39 (Korrespondenzsatz, siehe z.B. Lehmann and Romano (2005) oder Witting, 1985)

(a) Liegt für jedes $\vartheta \in \Theta$ *ein Test* φ_ϑ *zum Niveau* α *vor und wird* $\varphi = (\varphi_\vartheta, \, \vartheta \in \Theta)$ *gesetzt, so ist* $\mathcal{C}(\varphi)$, *definiert über* $C(x) = \{\vartheta \in \Theta : \varphi_\vartheta(x) = 0\}$, *eine Familie von Konfidenzbereichen zum Konfidenzniveau* $1 - \alpha$.

(b) Ist $\mathcal{C}$ *eine Familie von Konfidenzbereichen zum Konfidenzniveau* $1 - \alpha$ *und definiert man* $\varphi = (\varphi_\vartheta, \, \vartheta \in \Theta)$ *über* $\varphi_\vartheta(x) = 1 - \mathbf{1}_{C(x)}(\vartheta)$, *so ist* φ *ein Test zum allgemeinen lokalen Niveau* α, *d. h., zum Niveau* α *für jedes* $\vartheta \in \Theta$.

Beweis:
Sowohl in (a) als auch in (b) erhält man $\forall \, \vartheta \in \Theta : \forall \, x \in \Omega : \varphi_\vartheta(x) = 0 \Longleftrightarrow \vartheta \in C(x)$. Also ist φ ein Test zum allgemeinen lokalen Niveau α genau dann, wenn

$$
\begin{aligned}
& \forall \, \vartheta \in \Theta : \quad \mathbb{P}_\vartheta \left(\{\varphi_\vartheta = 0\} \right) \geq 1 - \alpha \\
\Leftrightarrow \quad & \forall \, \vartheta \in \Theta : \quad \mathbb{P}_\vartheta \left(\{x : C(x) \ni \vartheta\} \right) \geq 1 - \alpha \\
\Leftrightarrow \quad & \mathcal{C} \text{ ist Familie von Konfidenzbereichen zum Konfidenzniveau } 1 - \alpha.
\end{aligned}
$$

■

Bemerkung 1.40

(a) Die Dualität $\varphi_\vartheta(x) = 0 \Leftrightarrow \vartheta \in C(x)$ *lässt sich schön grafisch veranschaulichen, falls* Ω *und* Θ *eindimensional sind.*

(b) Ein einzelner Test φ *zum Niveau* α *für eine Hypothese* H *kann interpretiert werden als* $(1 - \alpha)$*-Konfidenzbereich. Setze dazu*

$$
C(x) = \begin{cases} \Theta \, , & \text{falls } \varphi(x) = 0, \\ K = \Theta \backslash H \, , & \text{falls } \varphi(x) = 1. \end{cases}
$$

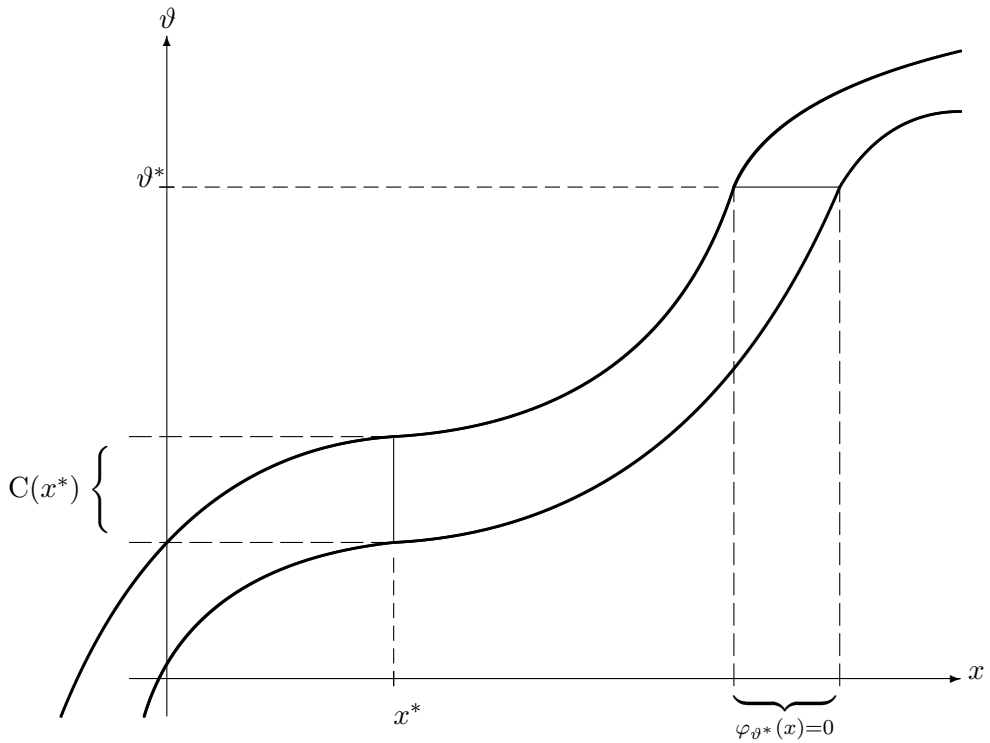


Abbildung 1.1: Dualität $\varphi_\vartheta(x) = 0 \Leftrightarrow \vartheta \in C(x)$

Umgekehrt liefert jeder Konfidenzbereich $C(x)$ einen Test zum Niveau α für eine Hypothese $H \subset \Theta$. Setze hierzu $\varphi(x) = \mathbf{1}_K(C(x))$, wobei

$$\mathbf{1}_B(A) := \begin{cases} 1\,, & \textit{falls } A \subseteq B, \\ 0\,, & \textit{sonst.} \end{cases}$$

für beliebige Mengen A und B.

Beispiel 1.41

Im Gaußmodell $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), ((\mathcal{N}(\mu, \sigma^2))^n)_{\mu \in \mathbb{R}=\Theta})$ mit bekannter Varianz $\sigma^2 > 0$ sei ein möglichst kleiner (bezüglich des Lebesguemaßes) Teilbereich der reellen Achse gesucht, der den unbekannten Erwartungswert μ mit einer Wahrscheinlichkeit von $(1-\alpha)$ überdeckt und der nur von $x \in \mathbb{R}^n$ abhängen darf.

Lösung: Die Statistik $\bar{X}_n$ ist suffizient für μ, beinhaltet also sämtliche Information, die X über μ liefert. Die Verteilung von $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ ist $\mathcal{N}(0,1)$. Damit ist $\bar{X}_n$ unter μ symmetrisch um μ verteilt mit exponentiell abfallender Verteilungsmasse zu beiden Seiten. Also ist ein optimaler Konfidenzbereich von der Form

$$C(x) = [\hat{\mu} - k(x), \hat{\mu} + k(x)] \textit{ mit } \hat{\mu} \equiv \hat{\mu}(x) = \bar{x}_n.$$

Wir müssen zur Berechnung von $k(x)$ das Konfidenzniveau $(1-\alpha)$ garantieren:

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

Bemerkung 1.42

a) Ist $\sigma^2 > 0$ <u>unbekannt</u>, so liefert der Korrespondenzsatz 1.39, angewendet auf den t-Test (Seite 152 b), Test Nr. 2) in Verbindung mit dem Kommentar über zweiseitige Tests auf Seite 152 e), dass ein optimaler $(1-\alpha)$-Konfidenzbereich für μ gegeben ist durch

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

b) Die Rechnung unter Beispiel 1.41 hängt nicht von der konkreten Bauart von $\bar{X}_n$, sondern lediglich von der Tatsache ab, dass $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ standardnormalverteilt ist. Sie kann also analog für andere Modelle mit normalverteilten suffizienten Statistiken durchgeführt werden.

In Definition 1.13 haben wir asymptotische Normalität als eine wünschenswerte Eigenschaft von Punktschätzern kennengelernt. In Anknüpfung an Beispiel 1.41 in Verbindung mit Bemerkung 1.42b) erhalten wir das folgende Resultat.

Satz 1.43

Sei $(\Omega^n, \mathcal{F}^n, (\mathbb{P}_\vartheta^n)_{\vartheta \in \Theta \subseteq \mathbb{R}^k})$ ein Produktmodell mit Regularitätseigenschaften und $\hat{\vartheta}_n(X)$ ein asymptotisch normalverteilter Punktschätzer für $\vartheta \in \mathbb{R}^k$ in dem Sinne, dass $\sqrt{n}(\hat{\vartheta}_n(X) - \vartheta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1}(\vartheta_0))$ unter ϑ_0. Sei $\varrho : \Theta \to \mathbb{R}$ eine stetig differenzierbare Abbildung mit Gradient $\dot{\varrho}(\vartheta) \neq 0$.

<u>Dann gilt:</u>

$$\sqrt{n}\left\{\varrho(\hat{\vartheta}_n(X)) - \varrho(\vartheta_0)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\vartheta_0}) \text{ unter } \mathbb{P}_{\vartheta_0} \text{ mit } \sigma^2_{\vartheta_0} = \dot{\varrho}(\vartheta_0) I^{-1}(\vartheta_0) \dot{\varrho}(\vartheta_0)^t.$$

Ist die Fisher-Information stetig, so ist ein Konfidenzintervall für $\varrho(\vartheta_0)$ mit asymptotischer Überdeckungswahrscheinlichkeit $(1-\alpha)$ gegeben durch

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

mit

$$\hat{\sigma}_n^2 := \dot{\varrho}(\hat{\vartheta}_n(x)) I^{-1}(\hat{\vartheta}_n(x)) \left[\dot{\varrho}(\hat{\vartheta}_n(x))\right]^t.$$

Beweis: Abschnitt 12.4.2 in Lehmann and Romano (2005). ■

Kapitel 2

Deskriptive Statistik

2.1 Univariate Merkmale

Siehe entsprechende Beamer-Präsentation.

2.2 Multivariate Merkmale

Siehe entsprechende Beamer-Präsentation.

Kapitel 3

Lineare Modelle und inferentielle Likelihoodtheorie

3.1 Einführung und Beispiele

Die Regressionsrechnung beschäftigt sich mit der Analyse von (systematischen) Zusammenhängen einer (univariaten) *Zielgröße (Response-Variable)* Y und einer Menge von k *erklärenden Variablen (Kovariablen, Regressoren)* $X_1, \ldots, X_k$. Anders als z. B. in der Physik, die deterministische Gesetzmäßigkeiten der Form $y = f(x_1, \ldots, x_k)$ mit $x = (x_1, \ldots, x_k)$ als „Eingabe“ und y als „Ausgabe“ zum Gegenstand hat, legt die Statistik zufällige „Störungen“ zu Grunde (Messfehler / -ungenauigkeiten etc.). Damit ist die Ausgabe / Response Y also eine Zufallsvariable, deren Verteilung von den Kovariablen abhängt.
Ziel der Regressionsanalyse ist die Untersuchung des Einflusses der erklärenden Variablen auf den Mittelwert der Zielgröße. Wir modellieren also

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

Die Variablen $Y_1, \ldots, Y_n$, die eine Stichprobe der Zielgröße beschreiben, lassen sich dann stets in eine systematische und eine stochastische Komponente zerlegen:

$$Y_i = \mathbb{E}[Y_i|X_{i,1} = x_{i,1}, \ldots, X_{i,k} = x_{i,k}] + \varepsilon_i = f(x_{i,1}, \ldots, x_{i,k}) + \varepsilon_i \text{ mit } \mathbb{E}[\varepsilon_i] = 0,\ 1 \leq i \leq n.$$

Der Vektor $\vec{x}_i = (x_{i,1}, \ldots, x_{i,k})$ heißt „Kovariablenprofil“ bei der i-ten Messung und ε_i heißt Fehlerterm bei der i-ten Messung. Die (verallgemeinerten) *linearen* Regressionsmodelle wählen speziell f als eine lineare Funktion in den Werten (Realisierungen) der Kovariablen.

Definition 3.1 ((Verallgemeinertes) Lineares Modell)
Seien $\vec{X} := (X_1, \ldots, X_k)$, $\vec{x} := (x_1, \ldots, x_k)$ *und* $\eta := g(\mathbb{E}[Y|\vec{X} = \vec{x}])$.

Modellannahme: $\eta = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$. *Dabei heißt* g *die Link-Funktion,* β_0 *der Intercept,* $(\beta_1, \ldots, \beta_k)^t$ *der Vektor der Regressionskoeffizienten (die Parameter des Modells!) und* $X_1, \ldots, X_k$ *werden auch als unabhängige Variablen und* Y *als die abhängige Variable bezeichnet.*

Schema 3.2 (Übersicht über GLMs)
GLM steht für „generalized linear model" bzw. verallgemeinertes lineares Modell.

Modell	Skalenniveau von Y	Link-Funktion g
ANCOVA (MLR)	stetig	id. ($\vec{X}$ hat stetige Komponenten)
ANOVA	stetig	id. ($\vec{X}$ rein kategoriell)
log-linear	$Y \in (0, \infty)$	$\eta := \mathbb{E}[\log(Y) \vert \vec{X} = \vec{x}]$
Poisson	$Y \in \mathbb{N}$	$\log$
logistisch	dichotom	$\text{logit}(x) = \ln(x/(1-x))$
Cox (proportional hazards)	Zeitspanne	$\eta = \ln(h(t))$, $h(t)$ Hazardfunktion

Tabelle 3.1: Übersicht über verallgemeinerte lineare Regressionsmodelle

Beispiel 3.3 (Realdaten)
Mietspiegeldaten, siehe R*-Skript.*

3.2 Inferentielle Likelihoodtheorie

Definition 3.4 (Maximum Likelihood-Schätzer)
Es sei $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ *ein von* μ *dominiertes Modell mit Likelihoodfunktion* $l(\vartheta, x)$. *Der Parameterraum* Θ *trage die* σ*-Algebra* $\mathcal{F}_\Theta$. *Eine Statistik* $\hat{\vartheta}(X)$ *mit* $\hat{\vartheta} : (\Omega, \mathcal{F}) \to (\Theta, \mathcal{F}_\Theta)$ *heißt Maximum-Likelihood-Schätzer (MLE) von* ϑ*, falls*

$$l(\hat{\vartheta}(x), x) = \sup_{\tilde{\vartheta} \in \Theta} l(\tilde{\vartheta}, x)$$

für $\mathbb{P}_\vartheta$*-fast alle* $x \in \Omega$ *und alle* $\vartheta \in \Theta$ *gilt.*

Bemerkung 3.5

(a) Weder Existenz noch Eindeutigkeit eines MLE sind ohne weitere Modellannahmen sichergestellt.

(b) Bei einer Re-Parametrisierung $\vartheta \mapsto \varrho(\vartheta)$ *ist natürlich* $\hat{\varrho}(X) := \varrho(\hat{\vartheta}(X))$ *der MLE für* $\varrho(\vartheta)$*, falls der MLE* $\hat{\vartheta}(X)$ *existiert.*

Beispiel 3.6

(a) $X_1, \ldots, X_n$ *iid. mit* $X_1 \sim$ *Poisson*(λ), $X := (X_1, \ldots, X_n)$ *mit Werten in* $\mathbb{N}_0^n$. *Der Parameter* $\lambda > 0$ *sei unbekannt.*

$$l(\lambda, x) = \prod_{i=1}^{n} \exp(-\lambda) \frac{\lambda^{x_i}}{x_i!}$$

$$\Rightarrow \quad \ln(l(\lambda, x)) = \sum_{i=1}^{n} -\lambda + x_i \ln(\lambda) - \ln(x_i!) = -n\lambda + \ln(\lambda) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!)$$

$$\Rightarrow \quad \frac{\partial}{\partial \lambda} \ln(l(\lambda, x)) = \dot{l}(x, \lambda) = -n + \lambda^{-1} \sum_{i=1}^{n} x_i$$

$$\Rightarrow \quad \hat{\lambda}(x) = n^{-1} \sum_{i=1}^{n} x_i, \text{ da } \frac{\partial^2}{\partial \lambda^2} \ln(l(\lambda, x)) < 0.$$

(b) *Allgemeines Regressionsmodell*

Sei $Y = (Y_1, \ldots, Y_n)$. *Für jedes* $1 \leq i \leq n$ *gelte* $Y_i = g_\vartheta(x_i) + \varepsilon_i$. *Dabei sind die* $(x_j)_{1 \leq j \leq n}$ *deterministische, fest vorgegeben "Messstellen",* g_ϑ *eine deterministische, vom interessierenden Parameter* $\vartheta \in \Theta \subseteq \mathbb{R}^k,\ k \in \mathbb{N}$, *parametrisierte Funktion und die* $(\varepsilon_j)_{1 \leq j \leq n}$ *zufällige iid. "Messfehler", für die* $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$ *mit* $\sigma^2 > 0$ *gelte.*

Damit gilt $\forall 1 \leq i \leq n : Y_i \sim \mathcal{N}(g_\vartheta(x_i), \sigma^2)$ *und* $Y_i \perp Y_j\ \forall 1 \leq i \neq j \leq n$.

Übungsaufgabe $\Rightarrow\ \hat{\vartheta}(Y) = \underset{\vartheta \in \Theta}{\operatorname{argmin}} \{\sum_{i=1}^{n} (Y_i - g_\vartheta(x_i))^2\}$, *also gleich dem Parameter, der die Fehlerquadratsumme minimiert.*

Satz 3.7 (Asymptotik des MLE)
Es sei $(\Omega^n, \mathcal{F}^n, (\mathbb{P}_\vartheta^n)_{\vartheta \in \Theta})_{n \geq 1}$ *mit* $\Theta \subseteq \mathbb{R}^k$ *eine Folge dominierter (von* μ^n*) Produktexperimente mit eindimensionaler Loglikelihoodfunktion* $\ln(l(\vartheta, x)) = \ln(\frac{d\mathbb{P}_\vartheta}{d\mu}(x))$.
Es gelte:

(a) Θ *ist kompakt und* ϑ_0 *liegt im Inneren von* Θ.

(b) $\forall \vartheta \neq \vartheta_0 :\ \mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta_0}$ *(Identifizierbarkeit)*

(c) $\vartheta \mapsto \ln(l(\vartheta, x))$ *ist stetig auf* Θ *und zweimal stetig differenzierbar in einer Umgebung* $\mathcal{U}$ *von* ϑ_0 *für alle* $x \in \Omega$.

(d) Es gibt $H_0, H_2 \in L_1(\mathbb{P}_{\vartheta_0})$ *und* $H_1 \in L_2(\mathbb{P}_{\vartheta_0})$ *mit*

$$\sup_{\vartheta \in \Theta} |\ln(l(\vartheta, x))| \leq H_0(x) \text{ sowie } \sup_{\vartheta \in \mathcal{U}} \left| \frac{\partial^i}{\partial \vartheta^i} \ln(l(\vartheta, x)) \right| \leq H_i(x),\ i = 1, 2, \forall x \in \Omega.$$

(e) Die Fisher-Information zu einer Beobachtung, also

$$I(\vartheta_0) = \mathbb{E}_{\vartheta_0}\left[\dot{l}(\cdot,\vartheta_0)(\dot{l}(\cdot,\vartheta_0))^t\right]$$

ist positiv definit.

Dann ist der MLE $\hat{\vartheta}_n(X)$, wobei $X = (X_1,\ldots,X_n)$ mit Werten in Ω^n, unter $\mathbb{P}^n_{\vartheta_0}$ asymptotisch normalverteilt:

$$\sqrt{n}(\hat{\vartheta}_n(X) - \vartheta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\vartheta_0)^{-1}) \text{ unter } \mathbb{P}^n_{\vartheta_0} \text{ für } n \to \infty.$$

Beweis: Siehe Abschnitt 6.5 in Lehmann and Casella (1998). ■

Korollar 3.8

Unter den Voraussetzungen von Satz 3.7 ist $\hat{\vartheta}_n(X)$ konsistent und asymptotisch Cramér-Rao-effizient.

Definition 3.9 (Likelihood-Quotienten-Test)

Es sei $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein dominiertes statistisches Modell mit Likelihoodfunktion $l(\vartheta, x)$. Das interessierende Testproblem sei gegeben durch $H_0 = \Theta_0$ gegen $H_1 = \Theta_1$, $\Theta_0 \neq \Theta_1 \neq \emptyset$, $\Theta_0 + \Theta_1 = \Theta$. Wir bezeichnen

$$\Lambda : \Omega \to [1,\infty], \Lambda(\cdot) := \frac{\sup_{\vartheta\in\Theta} l(\vartheta,\cdot)}{\sup_{\tilde{\vartheta}\in\Theta_0} l(\tilde{\vartheta},\cdot)}$$

als Likelihood-Ratio-Statistik und jeden Test der Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } \Lambda(x) > k \\ 0, & \text{falls } \Lambda(x) < k \\ \gamma(x), & \text{falls } \Lambda(x) = k \end{cases}$$

für $k \geq 1$ und $\gamma(x) \in [0,1]$ als einen Likelihood-Quotienten Test.

Bemerkung 3.10

Sind $\hat{\vartheta}$ bzw. $\hat{\vartheta}_0$ Maximum-Likelihood-Schätzer für ϑ, wobei $\vartheta \in \Theta$ bzw. Θ_0 variieren darf, so ist

$$\Lambda(x) = \frac{l(\hat{\vartheta}(x), x)}{l(\hat{\vartheta}_0(x), x)}.$$

Satz 3.11

Das Produktmodell $(\Omega^n, \mathcal{F}^n, (\mathbb{P}^n_\vartheta)_{\vartheta\in\Theta})$ erfülle die Voraussetzungen von Satz 3.7 über die Asymptotik von Maximum-Likelihood-Schätzern mit eindimensionaler Likelihoodfunktion $l(\vartheta, x_1)$. Die Hypothesenmenge Θ_0 liege in einem r-dimensionalen Unterraum von $\Theta \subseteq \mathbb{R}^k$ mit $0 \leq r < k$, wobei $r = 0$ dem Testen von Punkthypothesen $\Theta_0 = \{\vartheta_0\}$ entspricht. Dann gilt

$$2\log(\Lambda_n(X)) = 2\left[\sup_{\vartheta\in\Theta}\sum_{i=1}^n \ln(l(\vartheta, X_i)) - \sup_{\tilde{\vartheta}\in\Theta_0}\sum_{i=1}^n \ln(l(\tilde{\vartheta}, X_i))\right] \xrightarrow{\mathcal{D}} \chi^2_{k-r}$$

unter jedem $\mathbb{P}_{\vartheta_0}$ mit $\vartheta_0 \in \Theta_0 \cap [\Theta \setminus \partial\Theta]$.
Insbesondere besitzt der Likelihood-Quotienten-Test

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

mit $\chi^2_{(k-r);(1-\alpha)}$ dem $(1-\alpha)$-Quantil von χ^2_{k-r} damit auf der Menge $\Theta_0 \cap [\Theta \setminus \partial\Theta]$ asymptotisch das Niveau $\alpha \in (0,1)$.

Beispiel 3.12 (Multinomialverteilung)
Wir betrachten eine Folge von n stochastisch unabhängigen, gleichartigen Versuchen mit (jeweils) k möglichen Ausgängen.
Dabei trete der Ausgang i für $1 \leq i \leq k-1$ bei einem einzelnen Versuch mit Wahrscheinlichkeit p_i auf und wir definieren ferner $p_k := 1 - \sum_{i=1}^{k-1} p_i$.
Sei N_j, $1 \leq j \leq k$, die Zufallsvariable, die die Anzahl an Versuchen beschreibt, deren Ausgang gleich j ist. Dann heißt $N = (N_1, \ldots, N_k)$ multinomialverteilt mit Parametern $n, k-1$ und $p = (p_1, \ldots, p_{k-1})$ wobei wir n und k als fest vorgegeben betrachten wollen, so dass $\dim(\Theta) = k-1$ gilt.
Genauer gilt

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

Als Likelihoodstatistik erhalten wir

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

und als MLE ergibt sich analog zum Binomialmodell $\hat{p}_j = N_j/n,\ 1 \leq j \leq k-1$.
Betrachten wir nun die Punkthypothese $\Theta_0 = \{\pi\}$ für einen fest vorgegebenen Vektor $\pi \in \Theta$, so ergibt sich als Likelihood-Ratio-Statistik

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

wobei wir wieder $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$ setzen, und es gilt $2\log(\Lambda_n(N)) \xrightarrow{\mathcal{D}} \chi^2_{k-1}$ nach Satz 3.11.
Zur Durchführung des resultierenden asymptotisch χ^2-Tests kann die folgende Überlegung hilfreich sein. Betrachten wir die Funktion h, gegeben durch $h(x) = x\log(x/x_0)$ für eine fest vorgegebene reelle Zahl $x_0 \in (0,1)$. Dann ist die Taylor-Entwicklung von $h(x)$ um x_0 gegeben durch

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

und damit ist für $\hat{p}$ „nahe bei“ π

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

Die Statistik Q_n heißt Pearson'sche Chi-Quadrat Statistik. Es gilt präziser

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

unter der Nullhypothese, so das auch Q_n asymptotisch eine χ^2_{k-1}-Verteilung unter $p = \pi$ besitzt.

Bemerkung 3.13
Asymptotische χ^2-Tests lassen sich auf die Situation von $(k \times \ell)$-Feldertafeln verallgemeinern. Sind zwei kategorielle Zufallsvariablen X und Y gegeben, wobei X genau k Werte und Y genau ℓ Werte annehmen kann, so kann die Hypothese $X \perp Y$ mit einem asymptotischen χ^2-Test in der beobachteten $(k \times \ell)$-Kontingenztafel getestet werden. Die Anzahl der Freiheitsgrade berechnet sich dabei zu $(k-1) \cdot (\ell - 1)$.

3.3 Multiple lineare Regression (ANCOVA)

Modell 3.14 (Klassische multiple lineare Regression, ANCOVA)
Wir betrachten den Stichprobenraum $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ und modellieren die Beobachtungen $(y_1, \ldots, y_n)$ als Realisierungen von reellwertigen stochastisch unabhängigen Zufallsvariablen $(Y_1, \ldots, Y_n)$ mit

$$\forall 1 \le i \le n: \quad Y_i = f(x_{i,1}, \ldots, x_{i,k}) + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i. \tag{3.1}$$

Der Vektor $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^t$ ist der Parameter von Interesse. Wir setzen $p := k + 1$, kürzen ab:

$$Y := (Y_1, \ldots, Y_n)^t \in \mathbb{R}^n: \qquad \textit{Response-Vektor}$$

$$X := \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times p}: \qquad \textit{Design-Matrix}$$

$$\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)^t \in \mathbb{R}^n: \qquad \textit{Vektor der Fehlerterme}$$

$$\beta \equiv (\beta_0, \beta_1, \ldots, \beta_k)^t \in \mathbb{R}^p: \qquad \textit{Parametervektor}$$

und erhalten als Matrixschreibweise von (3.1)

$$Y = X\beta + \varepsilon. \tag{3.2}$$

Ferner machen wir die folgenden Modellannahmen:

(a) Die Designmatrix habe maximalen Rang, so dass $X^t X \in \mathbb{R}^{p \times p}$ positiv definit und invertierbar ist.

(b) Die Fehlerterme seien iid., wobei $\mathbb{P}^{\varepsilon_1}$ induziert sei durch F. Es gelte $\mathbb{E}[\varepsilon_1] = 0$ und $0 < \sigma^2 := Var(\varepsilon_1) < \infty$, also insbesondere Homoskedastizität. Die unbekannte Verteilungsfunktion F sei ein Störparameter, also nicht selbst Ziel der statistischen Inferenz.

Optional machen wir an einigen Stellen zusätzlich eine Normalverteilungsannahme an die Fehlerterme:

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

Bemerkung 3.15

(a) Die gemachten Modellannahmen unter 3.14 implizieren, dass entweder (i) die Designmatrix aus deterministischen, von der Experimentplanung her fest vorgegebenen „Messstellen" besteht oder (ii) die gesamte Analyse bedingt auf die (realisierten Werte der) Designmatrix durchgeführt wird, der eventuell vorhandene Zufall in den Designpunkten also im Modell nicht berücksichtigt wird. Möchte man stattdessen ein wahrscheinlichkeitstheoretisches Modell für die Designmatrix X mitmodellieren, so spricht man von einem Regressionsmodell mit zufälligem Design bzw. von einem Korrelationsmodell. Für die Bearbeitung von Korrelationsmodellen bedarf es gesonderter Auswertungstechniken. Insbesondere müssen dann Korrelationen der stochastischen Kovariablen mit den Fehlervariablen in die Auswertungsmethodik eingerechnet werden. Wir behandeln hier die Fälle (i) und (ii) völlig analog, unterdrücken also insbesondere in der Notation häufig das Bedingen auf die Designmatrix, um die Notation übersichtlich zu halten.

(b) Die Annahme der Unkorreliertheit der Fehlerterme ist durch eine sogenannte „Residualanalyse" nachtäglich zu validieren, vgl. Definition 3.17. Ist noch Struktur in den Residuen (den durch das Modell geschätzten Fehlertermen) zu erkennen, so kann das ein Hinweis darauf sein, dass weitere Kovariablen zusätzlich in das Modell aufzunehmen sind.

(c) Kategorielle Kovariablen sollten durch eine Menge von sogenannten „Dummy-Indikatoren" kodiert werden, um nicht implizit eine (inadäquate) Metrisierung auf dem diskreten Wertebereich solcher Kovariablen zu induzieren. Eine kategorielle Kovariable mit ℓ möglichen Ausprägungen wird dabei durch $(\ell - 1)$ Indikatoren repräsentiert. Der j-te Dummy-Indikator kodiert dabei das Ereignis, dass die Kategorie $(j + 1)$ bei der zugehörigen Kovariablen vorliegt, $j = 1, \ldots, (\ell - 1)$. Sind also alle $(\ell - 1)$ Indikatoren gleich Null, so entspricht dies der (Referenz-)Kategorie 1 der zugehörigen kategoriellen Kovariable.

(d) Wechselwirkungen zwischen Kovariablen werden durch Interaktionsterme modelliert. Der Interaktionsterm für die Wechselwirkung zwischen Kovariablen X_i und X_j wird dabei gebildet als $x_i \cdot x_j$ (Multiplikation der jeweiligen Ausprägungen).

Korollar 3.16

Aus den Modellvoraussetzungen unter 3.14 folgt bereits

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

Nehmen wir zusätzlich normalverteilte Fehlerterme an, so ist

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

Definition 3.17 (Residuen)

Ist ein Schätzer $\hat{\beta}$ des Parametervektors verfügbar, so erhalten wir einen (naheliegenden plug-in) Schätzer für den (bedingten) Erwartungswertvektor von Y durch $\widehat{\mathbb{E}[Y]} = X\hat{\beta}$. Wir definieren die Komponenten von $\widehat{\mathbb{E}[Y]}$ als $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_k x_{i,k}$, $i = 1, \ldots, n$ und die sogenannten Residuen als beobachtete Abweichungen der tatsächlich beobachteten Werte der Responsevariablen von den Schätzwerten ihrer (bedingten) Erwartungswerte, also $\hat{\varepsilon}_i = y_i - \hat{y}_i, 1 \leq i \leq n$.

Satz 3.18

Unter Modell 3.14 gilt:

(a) Der Kleinste Quadrate (KQ)-Schätzer des Parametervektors β ist gegeben durch

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

und damit folgt außerdem die Darstellung

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

(b) Durch Einsetzen von $\hat{\beta}$ in $\hat{Y} = X\hat{\beta}$ ergibt sich ferner

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

mit der $(n \times n)$-Matrix $H = X(X^t X)^{-1} X^t$. Die Matrix H wird als Prädiktionsmatrix bzw. hat matrix bezeichnet und $X^+ = (X^t X)^{-1} X^t$ heißt auch (Moore-Penrose) Pseudoinverse von X.

(c) Nehmen wir speziell normalverteilte Fehlerterme an, so stimmt der Maximum-Likelihood-Schätzer von β mit dem angegebenen KQ-Schätzer überein.

(d) Für den Schätzer $\hat{\beta}$, gegeben durch $\hat{\beta} = (X^t X)^{-1} X^t Y$ gilt

$$\mathbb{E}[\hat{\beta}] = \beta \;\; \textit{und} \;\; \mathit{Cov}(\hat{\beta}) = \sigma^2 (X^t X)^{-1}.$$

Beweis: Die Aussage unter (b) folgt sofort, sobald (a) gezeigt ist. Wir führen den Beweis zu (a) geometrisch bzw. mit Methoden der linearen Algebra. Dazu kürzen wir ab $x_i' := (1, x_{i,1}, \ldots, x_{i,k})$ (x_i' bezeichnet also die i-te Zeile der Designmatrix) und beachten, dass das kleinste Quadrate-Kriterium

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

äquivalent dadurch beschrieben werden kann, dass $X\hat{\beta}$ die L_2-Projektion von Y auf den Vektorraum $\{z \in \mathbb{R}^n : z = X\gamma, \gamma \in \mathbb{R}^p\}$ darstellt. Somit kann $\hat{\beta}$ charakterisiert werden durch

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

Multiplikation von rechts mit $(X^tX)^{-1}$ liefert $Y^tX(X^tX)^{-1} = \hat{\beta}^t$ und folglich wie gewünscht $\hat{\beta} = (X^tX)^{-1}X^tY$, da $(X^tX)^{-1}$ symmetrisch ist. Teile (c) und (d) sind Übungsaufgaben. ■

Bemerkung 3.19 (Geometrische Eigenschaften von $\hat{\beta}$)
Sei $\hat{\beta} = X^+Y$ wie unter Satz 3.18. Dann gilt

(i) Die geschätzten ((bedingten) Erwartungs-) Werte $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)^t$ sind orthogonal zu den Residuen $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^t$, d. h. $\hat{y}^t\,\hat{\varepsilon} = 0$.

(ii) Die Spalten von X sind orthogonal zu den Residuen, d. h. $X^t\,\hat{\varepsilon} = 0$.

(iii) Die Residuen sind im Mittel gleich Null, d. h. $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ bzw. $\bar{\hat{\varepsilon}} = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i = 0$.

(iv) Der arithmetische Mittelwert der $\hat{y}_i$ ist gleich dem Mittelwert der beobachteten Response-Werte y_i selbst, d. h. $\bar{\hat{y}} = n^{-1}\sum_{i=1}^n \hat{y}_i = \bar{y} = n^{-1}\sum_{i=1}^n y_i$.

(v) Die Regressionshyperebene geht durch den Schwerpunkt der Daten, d. h. $\bar{y} = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j \bar{x}_j$, wobei $\forall 1 \le j \le k : \bar{x}_j := n^{-1}\sum_{i=1}^n x_{i,j}$.

Lemma 3.20 (Zentrierungsoperator)
Sei $\mathbf{1} := (1, \ldots, 1)^t \in \mathbb{R}^n$ und $C := I_n - n^{-1}\mathbf{1} \cdot \mathbf{1}^t \in \mathbb{R}^{n \times n}$. Dann gilt

(i) Ist $a \in \mathbb{R}^n$ ein beliebiger Vektor, so ist

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

C ist also der Zentrierungsoperator.

(ii) C *ist symmetrisch und idempotent, d. h.* $C^2 = C$.

(iii) $a^t C a = \sum_{i=1}^n (a_i - \bar{a})^2, a \in \mathbb{R}^n$.

Beweis: Elementare Lineare Algebra, zur Übung. ■

Satz und Definition 3.21

Für $\hat{\beta} = X^+ Y$ *wie unter Satz 3.18 gilt*

(a) <u>*Streuungszerlegung:*</u>

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

(b) Da nach (a)

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

gilt, gibt

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

mit Werten in $[0, 1]$ *den Anteil der Gesamtvariation der Responsevariablen in der Stichprobe an, der durch das Regressionsmodell erklärt werden kann.*
Wir nennen R^2 <u>*Bestimmtheitsmaß*</u> *(englisch: R-square value / coefficient of determination).*

(c) Bezeichnet $r_{a,b}$ *den empirischen Pearson'schen Produktmomentkorrelationskoeffizienten zweier Datenvektoren* a *und* b*, so gilt* $R^2 = r_{y,\hat{y}}^2$.

Beweis: Zum Beweis von Teil (a) multiplizieren wir die Identität $y = \hat{y} + \hat{\varepsilon}$ von links mit der Zentrierungsmatrix C und erhalten $Cy = C\hat{y} + C\hat{\varepsilon}$. Da die Residuen bereits zentriert sind (vgl. Bemerkung 3.19.(iii)) gilt $C\hat{\varepsilon} = \hat{\varepsilon}$ und folglich $Cy = C\hat{y} + \hat{\varepsilon}$ bzw. $y^t C = \hat{y}^t C + \hat{\varepsilon}^t$. Damit folgt

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

wegen Bemerkung 3.19.(i).

Zum Nachweis von (c) beachten wir die Definition des Korrelationskoeffizienten und erhalten

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

Nach Definition von R^2 bleibt zu zeigen, dass $\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2 = (SSR)^2 = \left[\sum_{i=1}^n (\hat{y}_i - \bar{y})^2\right]^2$. Demnach genügt es zu zeigen, dass

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

Wir multiplizieren die linke Seite aus und erhalten

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

Da die Residuen zentriert und zu dem Vektor $\hat{y}$ orthogonal sind, folgt die Behauptung. ■

Satz 3.22 (Rechenregeln für Erwartungswertvektoren und Kovarianzmatrizen)
Seien Z_1 und Z_2 Zufallsvektoren mit Werten im $\mathbb{R}^d$ und A bzw. b geeignet dimensionierte, deterministische Matrix bzw. Vektor sowie $\mathbb{E}[Z_1] =: \mu$ und Cov$(Z_1) := \mathbb{E}[(Z_1 - \mu)(Z_1 - \mu)^t] =: \Sigma$. Dann gilt

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

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

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

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

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

Satz 3.23 (Satz von Gauß-Markov)
Es werde Modell 3.14 zu Grunde gelegt. Unter allen linearen und erwartungstreuen Schätzern für β besitzt $\hat{\beta} = X^+ Y$ minimale Varianz, d. h.,

$$\forall\, 0 \leq j \leq k : \text{Var}(\hat{\beta}_j) \leq \text{Var}(\hat{\beta}_j^L) \tag{3.3}$$

für jeden linearen, erwartungstreuen Schätzer $\hat{\beta}^L \equiv \hat{\beta}^L(Y)$. Damit ist $\hat{\beta}$ BLUE (best linear unbiased estimator).
Ferner gilt auch für jede Linearkombination $c^t\beta$, dass Var$(c^t\hat{\beta}) \leq$ Var$(c^t\hat{\beta}^L)$ mit $\hat{\beta}^L$ wie oben.

Beweis: Jeder in den Daten $Y \in \mathbb{R}^{n\times 1}$ lineare Schätzer $\hat{\beta}^L$ für $\beta \in \mathbb{R}^{p\times 1}$ ist von der Form $\hat{\beta}^L = AY$ mit $A \in \mathbb{R}^{p\times n}$. Es ist $\mathbb{E}[\hat{\beta}^L] = \mathbb{E}[AY] = AX\beta$, also muss für Erwartungstreue die Bedingung

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

erfüllt sein. Damit ist insbesondere rang$(A) = p$ und wir können A o. B. d. A. in die Form $A = (X^tX)^{-1}X^t + B$ bringen. Setzen wir diese Form in $I_p = AX$ ein, so folgt $I_p = AX = (X^tX)^{-1}X^tX + BX = I_p + BX \Rightarrow BX = 0$. Damit ist

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

Da BB^t nicht-negativ definit ist, ergibt sich also, dass auch $\text{Cov}(\hat{\beta}^L) - \text{Cov}(\hat{\beta}) = \sigma^2 BB^t \geq 0$. Damit folgt insbesondere für $c \in \mathbb{R}^p$ unter Beachtung von

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

dass $\text{Var}(c^t\hat{\beta}^L) \geq \text{Var}(c^t\hat{\beta})$. Ungleichung (3.3) folgt schließlich durch Betrachten spezieller Vektoren c mit Einträgen $c_i = \mathbf{1}_{\{i=j+1\}}$, $i = 1, \ldots, p$, für alle $0 \leq j \leq k$. ■

Um Tests und Konfidenzbereiche für die unbekannten Parameter $(\beta_j)_{0\leq j\leq k}$ konstruieren zu können, bedarf es nach Satz 3.18.(d) noch einer Schätzung der (in aller Regel unbekannten) Fehlervarianz σ^2. Zum Beispiel unter Normalverteilungsannahme 3.14.(c) an die Fehlerterme kann dazu die Maximum-Likelihood-Methode verwendet werden.

Satz 3.24 (Schätzung von σ^2)

(a) Unter Modell 3.14 mit Zusatzannahme 3.14.(c) gilt $\widehat{\sigma^2}_{ML} = \hat{\varepsilon}^t\hat{\varepsilon}/n$.

(b) Auch ohne Normalverteilungsannahme gilt unter Modell 3.14 $\mathbb{E}[\hat{\varepsilon}^t\hat{\varepsilon}] = (n-p)\sigma^2$*, so dass ein erwartungstreuer (Momenten-) Schätzer für* σ^2 *gegeben ist durch* $\widehat{\sigma^2} = \hat{\varepsilon}^t\hat{\varepsilon}/(n-p)$.

Bemerkung 3.25

(i) Der erwartungstreue Schätzer $\widehat{\sigma^2}$ *wird in der Praxis bevorzugt, hat jedoch größere Varianz als* $\widehat{\sigma^2}_{ML}$.

(ii) Der Schätzer $\widehat{\sigma^2}$ *ist restringierter MLE unter den Gegebenheiten von Satz 3.24.(a). Bei der REML-Schätzmethode wird die marginale Likelihoodfunktion*

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

die nach „Ausintegrieren" des Parametervektors zu Stande kommt, maximiert.

Beweis: Wir beweisen Satz 3.24. Teil (a) ist zur Übung. Zum Beweis von Teil (b) bemerken wir zunächst, dass $\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (I_n - H)Y$ gilt. Nach Übungsaufgabe ist die Matrix $I_n - H$ symmetrisch und idempotent mit rang$(I_n - H) = n - p$. Wir erhalten also $\mathbb{E}[\hat{\varepsilon}^t\hat{\varepsilon}] = \mathbb{E}[Y^t(I_n - H)Y]$. Unter Ausnutzung von Rechenregel 3.22.(vi) ergibt sich damit

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

Nun ist sp$((I_n - H)\sigma^2 I_n) = \sigma^2[\text{sp}(I_n) - \text{sp}(H)] = \sigma^2(n-p)$, da sp$(H) = \text{sp}(X(X^tX)^{-1}X^t) =$ sp$(I_p) = p$ ist. Also folgt

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

wie gewünscht. ■

Korollar 3.26

Ein (plug-in) Schätzer für Cov$(\hat{\beta})$ *mit* $\hat{\beta} = X^+Y$ *ist gegeben durch*

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

Satz 3.27 (Statistische Eigenschaften der Residuen)

Wir fassen unter Modell 3.14 die Residuen $\hat{\varepsilon} \equiv \hat{\varepsilon}(Y) = Y - \hat{Y} = Y - HY = Y - X(X^tX)^{-1}X^tY$ *als Zufallsgrößen auf. Dann gilt*

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

2) *Cov*$(\hat{\varepsilon}) = \sigma^2(I_n - H)$*, also insbesondere Heteroskedastizität, und es herrschen nicht-triviale Korrelationen.*

3) *Die standardisierten Residuen*

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

sind homoskedastisch verteilt, falls die Modellannahmen unter Modell 3.14 erfüllt sind.

Beweis: $\mathbb{E}[\hat{\varepsilon}] = \mathbb{E}[Y] - X(X^tX)^{-1}X^t\mathbb{E}[Y] = X\beta - X(X^tX)^{-1}X^tX\beta = 0.$
Cov$(\hat{\varepsilon})$ = Cov$((I_n - H)Y) = (I_n - H)\sigma^2 I_n (I_n - H)^t = \sigma^2(I_n - H)^t$, da die Matrix $(I_n - H)$ symmetrisch und idempotent ist, vgl. Beweis von Satz 3.24.(b). ■

Satz 3.28 (Multivariater zentraler Grenzwertsatz)
Wir betrachten Folgen von ANCOVA-Modellen, indiziert mit dem Stichprobenumfang n. Dabei seien die folgenden beiden Voraussetzungen an die Folge von Designmatrizen $(X_n)_{n\ge p}$ erfüllt.

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

(ii) $n^{-1}X_n^tX_n \longrightarrow V$ *für eine positiv-definite, symmetrische Matrix* $V \in \mathbb{R}^{p\times p}$.

Dann sind die folgenden beiden Aussagen richtig.

(a) Es sei $a^t = (a_1, \ldots, a_p)$ *ein beliebig ausgewählter, aber fest vorgegebener Vektor im* $\mathbb{R}^p$. *Dann gilt mit* $\rho^2 = \sigma^2 a^t V a$, *dass*

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

(b) Für $\hat{\beta}(n) = X_n^+ Y_n$ *gilt, dass*

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

Beweis: Sei $S_n := a^tX_n^t\varepsilon$. Wir beachten, dass

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

eine Summe stochastisch unabhängiger, zentrierter Zufallsvariablen ist. Ferner gilt

$$\begin{aligned}\text{Var}\,(S_n) = \sigma^2\sum_{i=1}^{n}b_i^2 &= \sigma^2\sum_{i=1}^{n}\sum_{j,\ell=1}^{p}a_ja_\ell x_{i,j}x_{i,\ell}\\ &= \sigma^2\sum_{j,\ell=1}^{p}a_ja_\ell(X_n^tX_n)_{j,\ell}\\ &= \sigma^2a^t(X_n^tX_n)a.\end{aligned}$$

Damit folgt Var $\left(n^{-\frac{1}{2}} S_n\right) = n^{-1}\sigma^2 a^t X_n^t X_n a \longrightarrow \rho^2 = \sigma^2 a^t V a$ für $n \to \infty$.
Überprüfung der Lindeberg-Bedingung unter Verwendung von Annahme (i) komplettiert den Beweis von Aussage (a).
Zum Beweis von Aussage (b) beachten wir die unter Satz 3.18.(a) berechnete Darstellung

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

Nach Cramér-Wold device (siehe z.B. Shorack and Wellner (1986), Seite 862) gilt

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

Da nach Annahme (ii) ferner $(n^{-1}X_n^t X_n)^{-1}$ gegen V^{-1} konvergent ist, gilt insgesamt

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

■

Satz 3.29 (Verteilung quadratischer Formen)

1. *Sei $X \sim \mathcal{N}_n(\mu, \Sigma)$ mit Σ symmetrisch und positiv definit.*
   Dann gilt $(X-\mu)^t \Sigma^{-1}(X-\mu) \sim \chi_n^2$.

2. *Sei $X \sim \mathcal{N}_n(0, I_n)$, R eine symmetrische, idempotente $(n \times n)$-Matrix mit rang$(R) = r$ und B eine $(p \times n)$-Matrix mit $p \leq n$. Dann gilt*

   (a) *$X^t R X \sim \chi_r^2$*

   (b) *Aus $BR = 0$ folgt: $X^t R X$ ist stochastisch unabhängig von BX.*

3. *Seien $X \sim \mathcal{N}_n(0, I_n)$ und R sowie S symmetrische und idempotente $(n \times n)$-Matrizen mit rang$(R) = r$, rang$(S) = s$ und $RS = 0$. Dann gilt*

   (a) *$X^t R X$ und $X^t S X$ sind stochastisch unabhängig.*

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

Beweis: Satz B.6 in Fahrmeir et al. (2009).

zu 1. Sei $\Sigma^{1/2}$ die symmetrische und positiv definite Matrix mit $\Sigma^{1/2} \cdot \Sigma^{1/2} = \Sigma$ und inverser Matrix $\Sigma^{-1/2}$. Dann ist $Z := \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}_n(0, I_n)$. Aus der Definition der Chi-Quadrat-Verteilung folgt $Z^t Z \sim \chi_n^2$ und damit die Behauptung.

zu 2. (a) Da R idempotent und symmetrisch ist, existiert eine orthogonale Matrix P mit $R = PD_rP^t$, wobei $D_r = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$. Weil P orthogonal ist, ist mit X auch auch $W := P^tX$ gemäß $\mathcal{N}_n(0, I_n)$ verteilt. Die Aussage ergibt sich nun unter Verwendung von

$$X^tRX = X^tR^2X = (RX)^t(RX) = (PD_rW)^t(PD_rW) = W^tD_rP^tPD_rW$$

$$= W^tD_rW = \sum_{i=1}^{r} W_i^2$$

und der Definition der Chi-Quadrat-Verteilung.

(b) Es ist $Z_1 := BX \sim \mathcal{N}_n(0, B^tB)$ und $Z_2 := RX \sim \mathcal{N}_n(0, R)$. Aus

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

und der Normalverteilungseigenschaft folgt die stochastische Unabhängigkeit von Z_1 und Z_2. Damit sind aber auch $Z_1 = BX$ und $Z_2^t \cdot Z_2 = X^tRX$ stochastisch unabhängig.

zu 3. (a) Hier setzen wir $Z_1 := SX \sim \mathcal{N}_n(0, S)$ und $Z_2 := RX \sim \mathcal{N}_n(0, R)$. Wir berechnen wieder

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

Erneut folgt aufgrund der Normalverteilungseigenschaft aus der Unkorreliertheit die stochastische Unabhängigkeit von Z_1 und Z_2 und damit die stochastische Unabhängigkeit von $Z_1^tZ_1$ und $Z_2^tZ_2$. Die Behauptung ergibt sich nun aus den Identitäten $X^tSX = Z_1^tZ_1$ und $X^tRX = Z_2^tZ_2$. Teil (b) ist eine einfache Folgerung aus 3.(a) und 1.

■

Definition 3.30 (Lineare Hypothesen)
Unter Modell 3.14 sei K eine deterministische $(r \times p)$-Matrix mit rang$(K) = r \leq p$. Wir nennen K Kontrastmatrix und jedes Testproblem der Form H_0: $K\beta = d$ versus H_1: $K\beta \neq d$ mit $d \in \mathbb{R}^{r\times 1}$ fest vorgegeben ein (zweiseitiges) lineares Testproblem. Das bedeutet, dass unter der linearen Hypothese H_0 insgesamt $r \leq p$ linear unabhängige Bedingungen an die Parameter des ANCOVA-Modells gestellt sind.

Beispiel 3.31

(i) Test auf signifikanten Zusammenhang einer bestimmten Kovariable mit der Response:

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

für ein vorgegebenes $1 \leq j \leq k$.
$\Rightarrow$ $K \in \mathbb{R}^{1\times p}$ *mit Einträgen* $K_i = \mathbf{1}_{\{i=j+1\}}$, *und* $d = 0$.

(ii) Test eines Subvektors $\beta^* = (\beta_1, \ldots, \beta_r)^t$:

$$\Rightarrow K = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & & & \ddots & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \end{pmatrix} \in \mathbb{R}^{r \times p} \tag{3.4}$$

mit Einträgen $K_{i\ell} = \mathbf{1}_{\{\ell = i+1\}}$ und $d = 0$.

(iii) Test auf Gleichheit zweier Regressionskoeffizienten:

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

$\Rightarrow \quad K \in \mathbb{R}^{1 \times p}$ *mit Einträgen* $K_i = \mathbf{1}_{\{i = j_1+1\}} - \mathbf{1}_{\{i = j_2+1\}}$,
also $K = (0, \ldots, 0, \underbrace{1}_{j_1+1\text{-te}}, 0, \ldots, 0, \underbrace{-1}_{j_2+1\text{-te}}, 0, \ldots, 0)$ *und* $d = 0$.

Satz 3.32
Unter Modell 3.14 gilt:

(a) Zur Berechnung der Teststatistik eines Likelihood-Quotienten-Tests (also der Devianz) zum Prüfen der lineare Hypothese H_0: $K\beta = d$ muss eine Maximierung der (Log-) Likelihoodfunktion unter den mittels K und d kodierten linearen Nebenbedingungen durchgeführt werden. Dieser rechenintensive Schritt kann vermieden werden durch Verwendung der Wald-Statistik

$$W = (K\hat{\beta} - d)^t (K\hat{V}K^t)^{-1} (K\hat{\beta} - d), \tag{3.5}$$

wobei $\hat{\beta}$ den MLE im vollen Modell und $\hat{V}$ die geschätzte Kovarianzmatrix von $\hat{\beta}$ bezeichnen. Die Statistik W ist asymptotisch äquivalent zur Devianz $2 \log \Lambda_n(Y)$ und es gilt insbesondere $\mathcal{L}(W) \xrightarrow[H_0]{w} \chi^2_r$ für $n \to \infty$.

(b) Treffen wir die Zusatzannahme 3.14.(c), so ist die Devianz eine isotone Transformation von $F = \frac{n-p}{r} \frac{\Delta SSE}{SSE}$ mit $\Delta SSE = SSE_{H_0} - SSE$. Ferner ist F unter H_0 dann exakt F-verteilt: $F \underset{H_0}{\sim} F_{r,n-p}$.

(c) Unter den Gegebenheiten von Teil (b) gilt $W = rF$.

Beweis: Zum Beweis der asymptotischen χ^2_r-Verteilung in Teil (a) kehren wir zurück zur Asymptotik des MLE, vgl. Satz 3.7. Es gilt

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\vartheta_0)^{-1}) \quad \text{unter } \mathbb{P}^n_{\vartheta_0} \text{ für } n \to \infty$$

unter den genannten Regularitätsannahmen.

$$\Rightarrow \quad I(\vartheta_0)^{1/2} n^{1/2} (\hat{\vartheta}_n - \vartheta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_r), \text{ wobei } r := \dim(\vartheta_0).$$

Das Continuous Mapping Theorem liefert damit, dass

$$(\hat{\vartheta}_n - \vartheta_0)^t n I(\vartheta_0)(\hat{\vartheta}_n - \vartheta_0) \xrightarrow{\mathcal{D}} \chi^2_r.$$

Ist die Fisher-Information stetig und $\hat{I}(\hat{\vartheta}_n)$ ein konsistenter Schätzer für $I(\vartheta_0)$, so gilt auch

$$(\hat{\vartheta}_n - \vartheta_0)^t n \hat{I}(\hat{\vartheta}_n)(\hat{\vartheta}_n - \vartheta_0) \xrightarrow{\mathcal{D}} \chi^2_r. \qquad \text{[A]}$$

Beachte: $nI(\vartheta_0)$ ist Fisher-Information des Produktmodells!
In unserem Fall ist $H_0 : K\beta - d = 0$ zu prüfen, also ist der Parameter von Interesse $\vartheta = K\beta - d$, $\vartheta_0 = 0$ und $\hat{\vartheta}_n = K\hat{\beta} - d$. Einsetzen dieser Terme in [A] liefert $\mathcal{L}(W) \xrightarrow[H_0]{w} \chi^2_r$ für $n \to \infty$. Weitere Details siehe Abschnitt 12.4.2 in Lehmann and Romano (2005).
Für Teil (b) kürzen wir in Anlehnung an Bemerkung 3.10 ab

$\hat{\beta}_{H_0}$: MLE im reduzierten Modell (unter den durch K und d kodierten Nebenbedingungen),

$\widehat{\sigma^2}_{H_0}$: MLE der Fehlervarianz im reduzierten Modell,

$\hat{l}(y) :=$ $l((\hat{\beta}, \widehat{\sigma^2}_{ML}), y)$,

$\hat{l}_{H_0}(y) :=$ $l((\hat{\beta}_{H_0}, \widehat{\sigma^2}_{H_0}), y)$

und berechnen

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

Zur Herleitung der $F_{r,n-p}$-Verteilung von $F = \frac{n-p}{r}\frac{\Delta SSE}{SSE}$ verwenden wir Satz 3.29.3. Dazu müssen wir noch zeigen:

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

(ii) ΔSSE und SSE sind stochastisch unabhängig.

(Übungsaufgabe 22.(b) komplettiert die Beweisführung.) Wir zeigen die Eigenschaften (i) und (ii) sowie die Aussage von Teil (c) als Korollar 3.34. ■

Satz 3.33

Unter den Gegebenheiten von Satz 3.32.(b) und (c) gilt:

(i) $\hat{\beta}_{H_0} = \hat{\beta} - (X^tX)^{-1}K^t(K(X^tX)^{-1}K^t)^{-1}(K\hat{\beta} - d)$.
Dieses $\hat{\beta}_{H_0}$ *erfüllt* $K\hat{\beta}_{H_0} = d$, *da* $K\hat{\beta}_{H_0} = K\hat{\beta} - K(X^tX)^{-1}K^t(K(X^tX)^{-1}K^t)^{-1}$. $(K\hat{\beta} - d) = K\hat{\beta} - K\hat{\beta} + d = d$. *Ferner ist* $\hat{\beta}_{H_0} = \hat{\beta}$, *falls* $\hat{\beta}$ *bereits die Nebenbedingungen erfüllt.*

(ii) Mit der Abkürzung $\Delta_{H_0} = (X^tX)^{-1}K^t(K(X^tX)^{-1}K^t)^{-1}(K\hat{\beta} - d)$ *gilt* $SSE_{H_0} = \hat{\varepsilon}^t\hat{\varepsilon} + \Delta_{H_0}^t X^t X \Delta_{H_0}$.

(iii) $\Delta SSE = (K\hat{\beta} - d)^t(K(X^tX)^{-1}K^t)^{-1}(K\hat{\beta} - d)$*, also eine quadratische Form.*

Beweis: Wegen Zusatzannahme 3.14.(c) ist $\hat{\beta}_{H_0}$ hier sowohl KQ-Schätzer als auch MLE. Für jeden Kandidaten-Vektor $\gamma \in \mathbb{R}^p$, der die Nebenbedingung $K\gamma = d$ erfüllt, errechnen wir zunächst

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

$(\hat{\beta} - \gamma)^t X^t(Y - X\hat{\beta}) = (Y - X\hat{\beta})^t X(\hat{\beta} - \gamma) = (\hat{\beta} - \gamma)^t(X^tY - X^tX(X^tX)^{-1}X^tY) = 0.$
Ferner gilt:

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

denn wir errechnen in analoger Weise

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

und mit $||X(\hat{\beta}_{H_0} - \gamma)||_2^2 = (\hat{\beta}_{H_0} - \gamma)^t X^t X(\hat{\beta}_{H_0} - \gamma)$ ist folglich insgesamt

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

Nun ist aber

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

und zusammenfassend ergibt sich also $\forall\gamma \in \mathbb{R}^p$ mit $K\gamma = d$:

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

Da die ersten beiden Summanden auf der rechten Seite invariant in γ sind und X vollen Rang hat, erhalten wir, dass $\hat{\beta}_{H_0}$ die Fehlerquadratsumme unter der linearen Nebenbedingung eindeutig minimiert. Darüber hinaus zeigt die Rechnung nämlich, dass die Zerlegungsformel

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

also die Orthogonalität von $(Y - X\hat{\beta})$ und $(X\hat{\beta} - X\hat{\beta}_{H_0})$ gilt (Pythagoras!).
Wir setzen nun $\hat{\beta}_{H_0} = \hat{\beta} - \Delta_{H_0}$ ein und berechnen

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

und damit

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

da $X^t\hat{\varepsilon} = 0$ nach Bemerkung 3.19.(ii). Damit ist

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

■

Korollar 3.34
Unter den Gegebenheiten von Satz 3.32.(b) und (c) gilt unter der linearen Hypothese $H_0 : K\beta = d$*:*

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

(b) $\Delta SSE \perp SSE$

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

Beweis: Für Teil (a) benutzen wir Satz 3.29.1. Wir definieren $Z = K\hat{\beta}$. Unter H_0 gilt dann $\mathbb{E}[Z] = d$ und $\mathrm{Cov}(Z) = \sigma^2 K(X^tX)^{-1}K^t$ und, da $\hat{\beta}$ normalverteilt ist, gilt sogar $Z \sim \mathcal{N}_r(d, \sigma^2 K(X^tX)^{-1}K^t)$.

Zum Nachweis von Teil (b) beachten wir, dass ΔSSE eine Funktion (alleine) von $\hat{\beta}$ ist. Da $\hat{\beta} \perp SSE$ ist, ist somit auch $\Delta SSE \perp SSE$.
Schließlich rechnen wir für den Nachweis von Teil (c), dass

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

■

Beispiel 3.35 (Fortführung von Beispiel 3.31)
Für drei spezielle Testprobleme berechnen wir die konkrete Gestalt der F-Statistik.

(i) *Test auf (signifikanten) Einfluss einer bestimmten Kovariable auf die Response:*
$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0, \quad 1 \leq j \leq k$ *fest vorgegeben.*
Wir haben $K \in \mathbb{R}^{1\times p}$ *mit Einträgen* $K_i = \mathbf{1}_{\{i=j+1\}}$ *nach Beispiel 3.31.(i) und* $d = 0$.
Einsetzen liefert

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

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

Dieser F-Test ist äquivalent zum zweiseitigen t-Test mit der Teststatistik

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

(ii) *Test eines Subvektors* $\beta^* = (\beta_1, \ldots, \beta_r)$*:*
Hier ist $K \in \mathbb{R}^{r\times p}$ *mit Einträgen* $K_{i\ell} = \mathbf{1}_{\{\ell=i+1\}}, d = 0$.
Damit ist

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

mit der Verteilungseigenschaft $F \underset{H_0}{\sim} F_{r,(n-p)}$.

(iii) *Globaltest:*
$H_0 : \beta_j = 0 \ \forall 1 \leq j \leq k$ *versus* $H_1 : \exists j \in \{1, \ldots, k\} : \beta_j \neq 0$.
Hier ist $SSE_{H_0} = SST = \sum_{i=1}^n (Y_i - \overline{Y})^2$ *und damit nach Streuungszerlegung*

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

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

Anmerkung: F-Tests können auch als Hotelling's T^2-Tests ausgeführt werden. Es gilt nämlich: Ist $F \sim F_{r,s}$, so ist $\frac{r(s+r-1)}{s}F \sim T^2(r, s+r-1)$ (Hotelling's T^2-Verteilung, Hotelling (1931)).

Korollar 3.36

Nach dem Korrespondenzsatz erhalten wir als Folgerung von Satz 3.33, Korollar 3.34 und unter Verwendung von Beispiel 3.35:

1) *Für ein fest vorgegebenes $1 \leq j \leq k$ ist ein $(1-\alpha)$-Konfidenzintervall für β_j gegeben durch*

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

2) *Ein Konfidenzellipsoid für einen Subvektor $\beta^* = (\beta_1, \ldots, \beta_r)^t$ zum Konfidenzniveau $(1-\alpha)$ ist gegeben durch*

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

Weitere einfache Folgerungen sind:

3) *Für eine zukünftige Beobachtung Y_0 mit zugehörigem Kovariablenprofil $\vec{X}_0 = \vec{x}_0$ ist ein $(1-\alpha)$-Konfidenzintervall für $\mu_0 := \mathbb{E}\left[Y_0 \mid \vec{X}_0 = \vec{x}_0\right]$ gegeben durch*

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

4) *Unter den Gegebenheiten von Teil 3) ist ein $(1-\alpha)$-Prognoseintervall für den wohl zu beobachtenden Responsewert y_0 selbst gegeben durch*

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

Beachte: Wir haben Kovariablenprofile als Zeilenvektoren definiert!

Bemerkung 3.37

Der multivariate zentrale Grenzwertsatz 3.28 besagt, dass auch ohne die Zusatzannahme 3.14.(c) zumindest asymptotisch/approximativ gilt

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

Damit bleiben alle Resultate zu Test- und Bereichsschätzproblemen auch ohne die Annahme normalverteilter Fehlerterme für große Stichprobenumfänge zumindest approximativ gültig. Sind die Stichprobenumfänge indes nur moderat, bieten sich stattdessen Resamplingverfahren an, wenn 3.14.(c) nicht angenommen werden kann.

3.4 Varianzanalyse (ANOVA)

Modell 3.38 (Einfaktorieller Versuchsplan, ANOVA1)

Wir beobachten Response-Werte $(y_{ij})_{\substack{1\le i\le k,\\ 1\le j\le n_i}}$ *, die wir als Realisierungen gemeinsam stochastisch unabhängiger Zufallsvariablen* $(Y_{ij})_{\substack{1\le i\le k,\\ 1\le j\le n_i}}$ *modellieren.*

Dabei gelte $\forall 1 \le i \le k: \quad \forall 1 \le j \le n_i: \quad Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$.

Wir bezeichnen die erste Dimension als den Faktor und den Wert von $1 \le i \le k$ *als die Faktorstufe. Die Zahlen* $(n_i)_{1\le i\le k}$ *geben die Anzahl der unabhängigen Versuchswiederholungen pro Faktorstufe an. In Matrixform lässt sich dieses Modell notieren, wenn wir äquivalent schreiben*

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

wobei die Fehlerterme $(\varepsilon_{ij})_{\substack{1\le i\le k,\\ 1\le j\le n_i}}$ *iid. sind mit* $\varepsilon_{11} \sim \mathcal{N}(0, \sigma^2)$.

Damit ist also ein Regressionsmodell mit einer kategoriellen Kovariable der Form

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

gegeben, in Matrixform

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

mit $\mu = (\mu_1, \dots, \mu_k)^t$ *als unbekanntem Parametervektor und* $\varepsilon \sim \mathcal{N}_{n_\bullet}(0, \sigma^2 I_{n_\bullet})$ *mit* $n_\bullet := \sum_{i=1}^n n_i$.

Gilt speziell $n_1 = \dots = n_k =: n$*, so nennen wir das Modell ANOVA1-Modell mit balanciertem Design.*

Beachte: Das Modell hat keinen Intercept! Es gilt $\operatorname{rang}(X) = k$, da die k Spalten von X linear unabhängig sind.

Die klassische Fragestellung der Varianzanalyse lautet: Existieren (irgendwelche) Unterschiede in den Faktorstufen-spezifischen Mittelwerten $\mu_i, 1 \le i \le k$, oder nicht?

Interpretation: Hat der Faktor einen Einfluss auf die Response oder nicht?

Mathematische Formulierung als Testproblem:

$H_0: \mu_1 = \mu_2 = \dots = \mu_k \quad$ versus $\quad H_1: \exists 1 \le i \ne \ell \le k: \ \mu_i \ne \mu_\ell$.

Darüber hinaus lassen sich viele andere Fragestellungen formulieren, z.B.:

(MCA) Prüfung aller paarweisen Mittelwertdifferenzen

(MCB) Vergleich der Faktorstufen-spezifischen Mittelwerte mit dem (empirisch) „besten“ (größten)

(MCC) Vergleich der Faktorstufen-spezifischen Mittelwerte mit dem einer ausgezeichneten Kontrollgruppe (control)
Dies sind klassische multiple Testprobleme, die in einer eigenen Spezialvorlesung behandelt werden.

Satz 3.39 (Quadratsummenzerlegung)
Wir setzen

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

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

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

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

Dann gilt: *SST = SSB + SSW.*

Beweis:

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

Nun ist aber

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

■

Offenbar spricht es gegen die Nullhypothese, wenn die Streuung zwischen den Gruppen größer ist als die Streuung innerhalb der Gruppen. Dies motiviert, das (skalierte) Verhältnis von SSB und SSW als Teststatistik zum Prüfen von H_0 zu verwenden.

Satz 3.40
Mit den Bezeichnungen von Satz 3.39 gilt:

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

(ii) Unter H_0 *ist* $SSB/\sigma^2 \sim \chi^2_{k-1}$

(iii) SSW *ist stochastisch unabhängig von* SSB*.*

(iv) Für $F = \frac{SSB/(k-1)}{SSW/(n_\bullet - k)}$ *gilt:* $F \underset{H_0}{\sim} F_{k-1,n_\bullet - k}$*.*

Insgesamt kann H_0 *also verworfen werden, falls der Wert der beobachteten F-Statistik den Wert* $F_{k-1,n_\bullet - k;1-\alpha}$ *übersteigt.*

Beweis: Zur Übung. ■

Definition 3.41 (Effektdarstellung)
Definiert man $\mu_0 := \mathbb{E}\left[\overline{Y}_{\cdot\cdot}\right] = n_\bullet^{-1} \sum_{i=1}^k \sum_{j=1}^{n_i} \mu_i = n_\bullet^{-1} \sum_{i=1}^k n_i \mu_i$ *und* $\alpha_i := \mu_i - \mu_0$*,* $1 \leq i \leq k$*, so hat man einen „Intercept“* μ_0 *in das ANOVA1-Modell eingeführt und die entstehende Darstellung lautet nun*

$$\forall 1 \leq i \leq k: \quad \forall 1 \leq j \leq n_i: \quad Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij} \tag{3.41.1}$$

mit homoskedastischen, stochastisch unabhängigen, zentriert normalverteilten Fehlertermen $(\varepsilon_{ij})_{\substack{1 \leq i \leq k, \\ 1 \leq j \leq n_i}}$.
Wichtig ist dabei, dass die Nebenbedingung $\sum_{i=1}^k n_i \alpha_i = 0$ *berücksichtigt wird, damit die resultierende Designmatrix vollen Rang besitzt (vgl. Aufgabe 23!). Wir nennen* (3.41.1) *die Effektdarstellung des ANOVA1-Modells und* α_i *den Effekt der Faktorstufe* i*,* $1 \leq i \leq k$*.*
Konvention: Um die Nebenbedingung $\sum_{i=1}^k n_i \alpha_i = 0$ *in der Designmatrix zu kodieren, einigen wir uns darauf,* $\alpha_k := -n_k^{-1} \sum_{i=1}^{k-1} n_i \alpha_i$ *vorzugeben.*

Damit lautet die Effektdarstellung in Matrixform $Y = X\ (\mu_0, \alpha_1, \ldots, \alpha_{k-1})^t + \varepsilon$ *bzw.*

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

mit den k *unbekannten Parametern* $\mu_0, (\alpha_i)_{1 \le i \le k-1}$.

Satz 3.42

(i) Unter den Gegebenheiten von Definition 3.41 sind die KQ-Schätzer (bzw. MLEs) für die unbekannten Parameter gegeben durch

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

(ii) Die F-Statistik zum Prüfen der Globalhypothese H_0*:* $\alpha_1 = \alpha_2 = \ldots = \alpha_{k-1} = 0$ *ist identisch zu der in Satz 3.40 (iv) angegebenen, also* $F = \frac{SSB/(k-1)}{SSW/(n_\bullet - k)}$.

Beweis: Zum Beweis von (i) betrachten wir die (überparametrisierte) Designmatrix

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

und lösen die Normalengleichungen $X_{(k+1)}^t X_{(k+1)}\ (\mu_0, \alpha_1, \ldots, \alpha_k)^t = X_{(k+1)}^t Y$ unter der linearen Restriktion $\sum_{i=1}^k n_i \alpha_i = 0$. Wir erhalten

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

und folglich für die Normalengleichungen

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

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

Einsetzen der linearen Restriktion ergibt

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

und damit

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

wie gewünscht.
Zum Nachweis von (ii) beachten wir, dass $SSE_{H_0} = SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{..})^2$ und

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

Nach Quadratsummenzerlegung ist damit $\Delta SSE = SSE_{H_0} - SSE = SST - SSW = SSB$ und Satz 3.32 (b) aus der allgemeinen ANCOVA-Theorie ergibt $F = \frac{SSB/(k-1)}{SSW/(n_\bullet - k)}$, denn hier ist $r = k - 1$ (Anzahl Restriktionen), der Gesamtstichprobenumfang $\dim(Y) = n_\bullet$ und die Anzahl der Spalten der Designmatrix mit vollem Rang ist gleich k. ■

Bemerkung 3.43

Das varianzanalytische Modell in Effektdarstellung, also

$$\forall 1 \leq i \leq k : \forall 1 \leq j \leq n_i : \quad Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij} \tag{3.43.1}$$

ist als Regressionsmodell zunächst einmal ungeeignet, da die Designmatrix nicht vollen Rang besitzt. Erst durch Berücksichtigung der intrinsischen Nebenbedingung $\sum_{i=1}^{k} n_i \alpha_i = 0$*, die aus der Definition der* $(\alpha_i)_{1 \leq i \leq k}$ *resultiert, haben wir ein Modell erhalten, das den Bedingungen von 3.14 genügt.*

Geht man indes direkt von (3.43.1) *aus, so lassen sich auch noch andere Nebenbedingungen finden, die vollen Spaltenrang der Design-Matrix garantieren, etwa* $\sum_{i=1}^{k} c_i \alpha_i = c^t \alpha = 0$ *mit* $\sum_{i=1}^{k} c_i \neq 0$*. Erzwingt man Orthogonalität der Spalten der Design-Matrix, so nennt man die entstehende Darstellung Kontrastkodierung. Zusammenfassend haben wir also*

1. *Dummy-Kodierung:*
   $x_{ji} = 1$*, falls Faktorstufe* i *vorliegt und 0 sonst,* $1 \leq j \leq n_\bullet, 1 \leq i \leq k$*.*

2. *Effekt-Kodierung:*

$$\forall 1 \leq j \leq n_\bullet : x_{j1} = 1 \text{ sowie } x_{ji} = \begin{cases} 1, & \text{falls Faktorstufe } i \text{ vorliegt}, \\ -1, & \text{falls Faktorstufe } k \text{ vorliegt}, \\ 0, & \text{sonst}, \end{cases} \quad 2 \leq i \leq k.$$

3. *Kontrast-Kodierung:*
   Orthogonalisierung der Design-Matrix.

In jedem der drei Fälle ist die entstehende Designmatrix $X = (x_{ji})_{\substack{1 \leq j \leq n_\bullet, \\ 1 \leq i \leq k}}$ *eine* $(n_\bullet \times k)$*-Matrix.*

Wenden wir uns nun zweifaktoriellen Versuchsplänen zu.

Modell 3.44 (Zweifaktorieller Versuchsplan, ANOVA2)

Wir beobachten Response-Werte $(y_{ijk})_{\substack{1 \leq i \leq I, \\ 1 \leq j \leq J, \\ 1 \leq k \leq n}}$ *(also hier nur balancierte Designs) und modellieren sie als Realisierungen von* (Y_{ijk}) *mit Modellannahme*

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

Der Gesamtstichprobenumfang ist $n_{\bullet\bullet} := I \cdot J \cdot n$*. Die ersten beiden Dimensionen heißen erster bzw. zweiter Faktor mit Faktorstufen* $1 \leq i \leq I$ *bzw.* $1 \leq j \leq J$ *und* $1 \leq k \leq n$ *durchläuft die Wiederholungen pro Faktorstufenkombination. Ferner nehmen wir an, dass alle* ε_{ijk} *iid. sind mit* $\varepsilon_{111} \sim \mathcal{N}(0, \sigma^2)$*. Das ANOVA2-Modell entspricht einem multiplen linearen Regressionsmodell mit zwei kategoriellen Kovariablen. Die Dimensionalität des Parametervektors ist*

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

Beispiel 3.45 (Schuchard-Ficher et al. (1980), Seite 30)
Response: Abgesetzte Mengeneinheit einer Margarinen-Marke in verschiedenen Supermärkten
Faktor 1: Preispolitik („Niedrigpreispolitik“, „Normalpreispolitik“, „Hochpreispolitik“)
Faktor 2: Kommunikationsstrategie („Postwurfsendungen“, „Anzeigenwerbung“)
Versuchsplan: Sechs zufällig ausgewählte Supermärkte, ein Supermarkt pro Faktorstufenkombination, zehn zufällig ausgewählte Werktage

Damit ergibt sich $I = 3$ *(Preispolitik-Strata),* $J = 2$ *(Kommunikationsstrategie-Strata) und* $n = 10$ *(jeweilige Wiederholungen).*

Definition und Lemma 3.46 (Effektdarstellung)
Wir definieren unter Modell 3.44

$$\mu_{i\bullet} := J^{-1} \sum_{j=1}^{J} \mu_{ij}\,, \quad 1 \le i \le I; \tag{3.46.1}$$

$$\mu_{\bullet j} := I^{-1} \sum_{i=1}^{I} \mu_{ij}\,, \quad 1 \le j \le J; \tag{3.46.2}$$

$$\mu_0 := (IJ)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}. \tag{3.46.3}$$

(i) Damit erhalten wir die Zerlegung

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

sowie die Effektdarstellung

$$Y_{ijk} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}. \tag{3.46.4}$$

Der neue Parametervektor ist

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

Aus (3.46.1) *bis* (3.46.3) *ist klar, dass* $(I + J + 1)$ *Restriktionen zwischen den Komponenten von* $\tilde{\vartheta}$ *herrschen. Damit hat die zur Effektdarstellung gehörige, überparametrisierte* $[(I \cdot J \cdot n) \times (1 + I + J + IJ)]$*-Designmatrix wie die der Dummy-Darstellung den Rang* $(I \cdot J)$*, also nicht vollen Spaltenrang.*

(ii) Es gilt

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

Die $(\alpha_i)_{1\le i\le I}$ heißen Haupteffekte des ersten Faktors, die $(\beta_j)_{1\le j\le J}$ heißen Haupteffekte des zweiten Faktors und die $((\alpha\beta)_{ij})_{\substack{1\le i\le I,\\ 1\le j\le J}}$ Interaktions- bzw. Wechselwirkungseffekte.
Es gilt nicht (!!) zwingend $(\alpha\beta)_{ij} = \alpha_i\beta_j$.

Beweis: zu (ii):

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

da $\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}/J$ konstant ist. Die anderen Aussagen folgen in analoger Weise. ■

Satz 3.47

Sei unter den Gegebenheiten von 3.46

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

(a) Dann lässt sich die Modellgleichung des ANOVA2-Modells mit balanciertem Design äquivalent schreiben als

$$Y = X\vartheta + \varepsilon \quad \textit{mit}$$

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

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

und der Designmatrix $X \in \mathbb{R}^{n_{\bullet\bullet}\times(I\cdot J)}$ mit Einträgen $(X_{rs})_{\substack{1\le r\le n_{\bullet\bullet},\\ 1\le s\le I\cdot J}}$ gegeben durch die folgende Konstruktion. Wir teilen X in vier Teilmatrizen $X_{\mu_0}, X_\alpha, X_\beta$ und $X_{(\alpha\beta)}$ mit jeweils $n_{\bullet\bullet}$ Zeilen auf:

[1] X_{μ_0} ist ein Spaltenvektor, bestehend aus lauter Einsen (Intercept-Spalte).

[2] X_α ist eine $(n_{\bullet\bullet}\times(I-1))$-Matrix, deren Spalten zu den Haupteffekten des ersten Faktors korrespondieren. In Spalte $1\le i\le I-1$ von X_α gilt:

$X_\alpha^{(k,i)} = +1$, *falls Faktorstufe i bei Beobachtungseinheit k vorliegt,* $1\le k\le n_{\bullet\bullet}$,

$X_\alpha^{(k,i)} = -1$, *falls Faktorstufe I vorliegt*,

$X_\alpha^{(k,i)} = 0$, *sonst*.

[3] X_β *korrespondiert in analoger Weise zu den Haupteffekten des zweiten Faktors.*

[4] $X_{(\alpha\beta)}$ *korrespondiert zu den* $(I-1)\cdot(J-1)$ *Interaktionseffekten. Für ihre Einträge gilt in selbsterklärender Notation:* $X_{(\alpha\beta)}^{(k,ij)} = X_\alpha^{(k,i)} \cdot X_\beta^{(k,j)}, \quad 1 \le i \le I-1,$ $1 \le j \le J-1,\ 1 \le k \le n_{\bullet\bullet}$. *Demnach hat* X *folgende Struktur:*

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

(b) X *hat vollen Rang und ist blockorthogonal in dem Sinne, dass Spalten, die aus verschiedenen Teilmatrizen* $\{X_{\mu_0}, X_\alpha, X_\beta, X_{(\alpha\beta)}\}$ *stammen, zueinander paarweise orthogonal sind.*

Beweis: Nach Lemma 3.46.(ii) ist $\alpha_I = -\sum_{i=1}^{I-1}\alpha_i,\ \beta_J = -\sum_{j=1}^{J-1}\beta_j$ und $\forall 1 \le i \le I-1:$ $\forall 1 \le j \le J-1: \quad (\alpha\beta)_{Ij} = -\sum_{i=1}^{I-1}(\alpha\beta)_{ij},\ (\alpha\beta)_{iJ} = -\sum_{j=1}^{J-1}(\alpha\beta)_{ij}.$
Ferner gilt:

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

Damit lässt sich die unter (a) behauptete Darstellung sofort vermittels Fallunterscheidung bezüglich (i,j) und Verifikation der Übereinstimmung mit (3.46.4) unter Berücksichtigung der obigen Restriktionen beweisen.
Zum Nachweis von (b) genügt es, Blockorthogonalität zu zeigen.
Wegen balanciertem Design ist in jeder Spalte von X_α bzw. X_β bzw. $X_{(\alpha\beta)}$ die Anzahl der Einträge „$+1$“ gleich der Anzahl der Einträge „-1“ und damit sind alle diese Spalten orthogonal zum Spaltenvektor X_{μ_0}. Für eine beliebige Kombination (i,j) von Haupteffekten ist das innere Spaltenprodukt

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

Analog rechnet man für die Kombination eines Haupteffektes mit einem Interaktionseffekt. ■

Unter Ausnutzung der unter Satz 3.47.(b) gezeigten Blockorthogonalitätseigenschaft der ANOVA2-Designmatrix im Falle balancierter Designs lässt sich die Modellgleichung in diesen Fällen wie folgt notieren:

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

mit selbsterklärenden Vektoren α, β und $(\alpha\beta)$. Dies hat günstige Auswirkungen auf die Parameterschätzungen:

Satz 3.48 (Parameterschätzungen für ANOVA2 unter balanciertem Design)
In ANOVA2-Modellen mit balanciertem Design sind KQ-Schätzer bzw. MLEs für die Parameter $\mu_0, \alpha = (\alpha_1, \ldots, \alpha_{I-1})^t, \beta = (\beta_1, \ldots, \beta_{J-1})^t$ *und* $(\alpha\beta) = (\alpha\beta_{ij})_{\substack{1 \le i \le I-1, \\ 1 \le j \le J-1}}$ *gegeben durch*

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

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

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

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

Beweis: Wir haben $X = (X_{\mu_0} \quad X_\alpha \quad X_\beta \quad X_{(\alpha\beta)})$ und damit lassen sich die Normalengleichungen $X^t X(\mu_0, \alpha^t, \beta^t, (\alpha\beta)^t)^t = X^t Y$ darstellen als

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

Wegen Blockorthogonalität reduziert sich dieses System zu

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

$$X_{\alpha}^t X_{\alpha}\alpha = X_{\alpha}^t Y,$$

$$X_{\beta}^t X_{\beta}\beta = X_{\beta}^t Y,$$

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

Damit entkoppeln sich die Parameterschätzprobleme und die Gestalt von $\hat{\alpha}, \hat{\beta}$ und $\widehat{(\alpha\beta)}$ folgen sofort, da die entsprechenden Teil-Designmatrizen jeweils vollen Spaltenrang haben.
Da X_{μ_0} ein Spaltenvektor aus lauter Einsen ist, gilt für $\hat{\mu}_0$, dass

$$n_{\bullet\bullet}\hat{\mu}_0 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n} Y_{ijk} \Leftrightarrow \hat{\mu}_0 = \overline{Y}_{\bullet\bullet\bullet} \text{ gelten muss.}$$

■

Korollar und Definition 3.49
Unter Modell 3.44 definieren wir

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

Dann folgt aus Satz 3.48, dass KQ-Schätzer bzw. MLEs für die ANOVA2-Parameter in Effektdarstellung gegeben sind durch

$$\hat{\mu}_0 = \overline{Y}_{\bullet\bullet\bullet}$$

$$\forall 1 \le i \le I-1: \quad \hat{\alpha}_i = \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet},$$

$$\forall 1 \le j \le J-1: \quad \hat{\beta}_j = \overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet},$$

$$\forall 1 \le i \le I-1: \quad \forall 1 \le j \le J-1: \ \widehat{(\alpha\beta)}_{ij} = \overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}$$

(vgl. auch die Definition unter 3.46.(i)).

Abschließend behandeln wir noch die Testtheorie im ANOVA2-Modell mit balanciertem Design. Dabei ist es üblich, zunächst die Interaktionseffekte auf statistische Signifikanz zu prüfen, da sich die Haupteffekte besser interpretieren lassen, wenn das Modell von den Wechselwirkungstermen befreit wird.

Satz 3.50

Unter Modell 3.44 sei die lineare Hypothese $H_{(\alpha\beta)} : (\alpha\beta)_{ij} = 0 \quad \forall 1 \leq i \leq I-1$, $\forall 1 \leq j \leq J-1$ *zu testen. Dann gilt*

(a) $H_{(\alpha\beta)}$ *ist darstellbar als* $K\vartheta = 0$ *(mit* ϑ *wie in Satz 3.47) mit Kontrastmatrix*

$$K = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & 0 & \dots & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

$$\underbrace{\qquad\qquad}_{I+J-1} \qquad \underbrace{\qquad\qquad\qquad\qquad}_{(I-1)(J-1)}$$

mit rang$(K) = (I-1)(J-1)$.

(b)

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

(c)

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

ist unter $H_{(\alpha\beta)}$ $\quad F_{(I-1)(J-1),IJ(n-1)}$*-verteilt.*

Beweis: Teil (a) ist offensichtlich.

Die Gestalt von $SSE_{H_{(\alpha\beta)}}$ unter Teil (b) folgt aus Satz 3.48 und Korollar 3.49 ($\hat{\alpha}, \hat{\beta}$ und $\hat{\mu}_0$ sind invariant gegenüber der Gültigkeit von $H_{(\alpha\beta)}$).

Zum Nachweis der Gestalt von $\Delta SSE_{H_{(\alpha\beta)}}$ beachten wir, dass nach Korollar 3.49 gilt

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

Offenbar hat SSE also $IJ(n-1)$ Freiheitsgrade. Damit errechnen wir weiter

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

wie gewünscht.

Damit folgt Teil (c) aus der allgemeinen Theorie der multiplen linearen Regression ($r \mathrel{\widehat{=}} (I-1)(J-1),\ p \mathrel{\widehat{=}} IJ,\ n_{\bullet\bullet} - p \mathrel{\widehat{=}} IJ(n-1)$). ■

Bemerkung 3.51

In analoger Weise erhält man die folgenden Resultate zum Testen der Haupteffekte.

(a) Für die lineare Hypothese $H_\alpha : \alpha_i = 0\ \forall 1 \le i \le I-1$ ist

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

und damit

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

(b) Für $H_\beta : \beta_j = 0 \forall 1 \le j \le J-1$ ist

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

und damit

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

Insgesamt erhalten wir folgende tabellarische Situation für Modell 3.44:

Balanciertes ANOVA2	Fehlerquadratsummenterm	Freiheitsgrade	F-Statistik
Haupteffekte des 1. Faktors	$\Delta SSE_{H_\alpha} = Jn\sum_{i=1}^{I}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$	$I-1$	$\frac{\Delta SSE_{H_\alpha}/(I-1)}{SSE/[IJ(n-1)]}$
Haupteffekte des 2. Faktors	$\Delta SSE_{H_\beta} = In\sum_{j=1}^{J}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$	$J-1$	$\frac{\Delta SSE_{H_\beta}/(J-1)}{SSE/[IJ(n-1)]}$
Interaktions-effekte	$\Delta SSE_{H_{(\alpha\beta)}} = n\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2$	$(I-1)(J-1)$	$\frac{\Delta SSE_{H_{(\alpha\beta)}}/[(I-1)(J-1)]}{SSE/[IJ(n-1)]}$
Residual-streuung	$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2$	$IJ(n-1)$	

Tabelle 3.2: Tabelle der ANOVA2 mit balanciertem Design

Beispiel 3.52

Margarine-Datensatz von Schuchard-Ficher et al. (1980), Fortführung von Beispiel 3.45 (siehe Präsentation mit R).

Bemerkung 3.53 (Versuchspläne mit Messwiederholungen)

Es gibt andere varianzanalytische Versuchspläne als das Design mit iid. Fehlertermen.

Beispielsweise könnte die Response jeder Beobachtungseinheit unter verschiedenen experimentellen Bedingungen (wiederholt) gemessen werden.

Im einfaktoriellen Fall führt dies auf eine Modellgleichung der Form

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

mit balanciertem Design, wobei die Fehlerterme $(\varepsilon_{ij})_{\substack{1\leq i\leq k,\\ 1\leq j\leq n}}$ nun indes nicht mehr stochastisch unabhängig sind, da in jeder Faktorstufe die selben n Versuchsobjekte bemessen werden.
Wir müssen die Annahme an die Verteilung der Fehlerterme also modifizieren:

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

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

Wir testen die Globalhypothese („kein Treatment-Effekt“)

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \qquad \text{bzw. äquivalenterweise}$$

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

Seien dazu

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

die Differenzen aufeinanderfolgender Messwerte des Versuchsobjekts j. Dann gilt offenbar

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

Mit der $(k-1) \times (k-1)$-Stichproben-Kovarianzmatrix S_Δ der Differenzwerte Δ_{ij} folgt dann, dass

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

unter H_0 $T^2(k-1, n-1)$-verteilt ist.

In ähnlicher Weise können andere Versuchspläne bearbeitet werden.

3.5 Poisson-Regression

Als erstes verallgemeinertes lineares Regressionsmodell betrachten wir das Modell der Poisson-Regression für Zähldaten. Vorbereitend dazu studieren wir noch allgemeine Charakteristika von verallgemeinerten linearen Modellen (englisch: generalized linear models, GLMs).

Definition 3.54 (Verallgemeinertes lineares Modell, GLM)
Sei $(\Omega^n, \mathcal{F}^n, \bigotimes_{i=1}^n \mathbb{P}_{\vartheta_i})$ mit unbekanntem $\vartheta_i \in \Theta \subseteq \mathbb{R}\, \forall 1 \leq i \leq n$ ein Produktexperiment , induziert von stochastisch unabhängigen, beobachtbaren Responsevariablen $Y_1, \ldots, Y_n$ mit Werten (jeweils) in $\Omega \subseteq \mathbb{R}$. Dann heißt $(\Omega^n, \mathcal{F}^n, \bigotimes_{i=1}^n \mathbb{P}_{\vartheta_i})$ verallgemeinertes lineares Modell (GLM), falls gilt:

1) *Für jedes $1 \leq i \leq n$ ist $\mathbb{P}_{\vartheta_i}$ ein Element einer natürlichen Exponentialfamilie, d.h. es existiert eine Dichte (Likelihoodfunktion) bezüglich eines dominierenden Maßes und sie hat die Form*

$$l(\vartheta_i, y_i) = a(\vartheta_i) b(y_i) \exp(y_i \cdot T(\vartheta_i)).$$

Der Term $T(\vartheta)$ heißt natürlicher Parameter der Exponentialfamilie, $\vartheta \in \Theta$.

2) *Für jede Beobachtungseinheit $1 \leq i \leq n$ sind die Werte von p Kovariablen (unabhängigen Variablen) gegeben, insgesamt also eine Designmatrix $X \in \mathbb{R}^{n \times p}$. Ein Intercept wird dabei durch eine Quasi-Kovariable, die konstant den Wert 1 annimmt, kodiert.*
Die systematische Komponente des GLM ist dann gegeben durch einen Vektor $\eta = (\eta_1, \ldots, \eta_n)^t$ mit

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

für (unbekannte) Regressionskoeffizienten $\beta_1, \ldots, \beta_p$. Der Term η_i heißt linearer Prädiktor von Beobachtungseinheit $1 \leq i \leq n$. Insgesamt ist also $\eta = X\beta,\ \beta = (\beta_1, \ldots, \beta_p)^t$.

3) *Sei $\mu_i := \mathbb{E}[Y_i | \vec{X}_i = \vec{x}_i]$ der (bedingte) Erwartungswert der i-ten Responsevariable. Es existiert eine Link-Funktion g, die die systematische Komponente und die durch μ_i beschriebene stochastische Komponente der Verteilung von Y_i koppelt:*

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

Die kanonische Link-Funktion (der kanonische Link) transformiert den (bedingten) Erwartungswert von Y_i auf den natürlichen Parameter, erfüllt also die Beziehung

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

Beispiel 3.55 (ANCOVA mit normalverteilten Fehlertermen)
Sei $Y = (Y_1, \ldots, Y_n)$ mit $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$ wie unter Modell 3.14 mit Zusatzanname 3.14 (c). Dann gilt nach Korollar 3.16, dass $\mu_i = \sum_{j=1}^{p} \beta_j x_{ij}$ $\forall 1 \leq i \leq n$ gilt, d.h. $g = id$. (identity link).

Die Formulierung von GLMs erlaubt eine einheitliche Behandlung von Produktmodellen, denen natürliche Exponentialfamilien zu Grunde liegen, mit Hilfe der inferentiellen Likelihoodtheorie. Ziel der statistischen Inferenz für GLMs sind wieder die Regressionskoeffizienten $(\beta_j)_{1 \leq j \leq p}$.

Lemma 3.56
Die Familie der Poissonverteilungen mit Intensitätsparameter $\vartheta > 0$ bildet eine natürliche Exponentialfamilie im natürlichen Parameter $T(\vartheta) = \log(\vartheta)$.

Beweis: Die Familie ist dominiert vom Zählmaß mit

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

■

Die Poisson-Regression modelliert stochastisch unabhängige Beobachtungseinheiten, die Zähldaten-Struktur haben und als jeweils Poisson-verteilt mit Kovariablen-abhängigen Intensitätsparametern angenommen werden, als GLM. Nach Definition 3.54 und Lemma 3.56 ist der kanonische Link dabei durch den natürlichen Logarithmus gegeben.

Modell 3.57 (Poisson-Regression mit Intercept)
Sei $k = p - 1$ und $\vec{X} = (X_1, \ldots, X_k)$ der Vektor der interessierenden Kovariablen. Der Stichprobenraum für die Response $Y = (Y_1, \ldots, Y_n)^t$ ist $(\mathbb{N}_0^n, 2^{\mathbb{N}_0^n})$. Für jedes $1 \leq i \leq n$ legen wir einen Zählrahmen (englisch: size) s_i zugrunde und machen die Modellannahme, dass Y_i Poisson($\lambda(\vec{x}_i) \cdot s_i$) verteilt ist, $1 \leq i \leq n$. Dabei gelte die Strukturannahme

$$\log(\lambda(\vec{x}_i)) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \quad \text{bzw.}$$

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

Für das Gesamtexperiment ergibt sich als (Log-)Likelihoodfunktion mit $\beta = (\beta_1, \ldots, \beta_k)^t$

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

sowie

$$\ln(l(\beta, y)) = \sum_{i=1}^{n} \{ y_i (\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln(s_i)) - \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \ln(s_i)) - \ln(y_i!) \}.$$

Bemerkung 3.58

(i) Der Term $\ln(s_i)$ wird als Offset bezeichnet, da er jeder Beobachtungseinheit einen individuellen Intercept zuweist.

(ii) Der kanonische Link $g = \log$ transformiert den ursprünglichen Wertebereich $(0, \infty)$ von $\lambda(\vec{x}_i)$ auf ganz $\mathbb{R}$, den natürlichen Parameterraum der Familie.

(iii) Die Poisson-Regression ist ein multiplikatives Modell. Für den Quotienten zweier Inzidenzraten (in der Epidemiologie: relatives Risiko, RR) bei zwei Kovariablenprofilen $\vec{x}_A$ und $\vec{x}_B$,

die sich in genau einer Ausprägung einer bestimmten Kovariable j unterscheiden, gilt nämlich

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

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

Speziell für dichotome Kovariable j: $RR = \exp(\beta_j)$

Dieses Konstanthalten aller anderen Kovariablen nennt man auch „Adjustieren“.

(iv) Die Parameterschätzung der Regressionskoeffizienten erfolgt vermittels des Maximum-Likelihood-Prinzips. Zwar existiert keine geschlossene Lösung, aber die Existenz eines eindeutig bestimmten Minimums der negativen Log-Likelihoodfunktion ist garantiert, da $-\ln(l(\beta, y))$ eine konvexe Funktion des Parametervektors β ist. Es können also numerische Routinen aus der konvexen Optimierung angewendet werden.

(v) Geschachtelte Modelle können mit einem Likelihood-Quotienten-Test verglichen werden.

Beispiel 3.59

Le (2003) berichtet eine Studie, in der Daten von $n = 44$ NotfallärztInnen in einem großen Krankenhaus erhoben worden sind. Die zu analysierende abhängige (Response-) Variable ist die Anzahl an Beschwerden für die jeweiligen ÄrztInnen. Der Zählrahmen pro Arzt/Ärztin ist die Anzahl an Patientenbesuchen, die vier zu berücksichtigenden Kovariablen sind Vergütung (in Dollar/Stunde) und Erfahrung (in Stunden) sowie Geschlecht und Facharztausbildung (ja/nein).
(siehe Präsentation mit R und Handout.)

Satz 3.60 (Multivariater Zentraler Grenzversatz)
Sei $\hat{\beta}(n)$ der MLE des Parametervektors $\beta = (\beta_1, \ldots, \beta_p)^t$ eines GLMs mit kanonischem Link zum Stichprobenumfang n. Falls alle p Kovariablen einen kompakten Wertebereich haben und für die Folge $(X_n)_{n \geq p}$ von Designmatrizen gilt, dass $(X_n^\top X_n)^{-1} \underset{n\to\infty}{\longrightarrow} 0$, so gilt:

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

Dieses asymptotische Ergebnis bleibt richtig, falls $F_n(\beta)$ durch $F_n(\hat{\beta}(n))$ ersetzt wird.

Beweis: Satz 2.2 in Kapitel 7 von Fahrmeir and Hamerle (1984), siehe auch Übungsaufgabe 37. ■

Bemerkung 3.61

(a) $Cov_n(Y) = diag\left(\left[Var\left(Y_i|\vec{X}_i = \vec{x}_i\right)\right]_{1\leq i\leq n}\right)$. Dabei hängt $Var\left(Y_i|\vec{X}_i = \vec{x}_i\right)$ von β ab.

(b) Im Falle der Poisson-Regression ist

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

Definition 3.62 (Saturiertes Modell)
Für ein gegebenes GLM basierend auf Responsevariablen $Y = (Y_1, \ldots, Y_n)^t$ und zugehörigen Beobachtungen $y = (y_1, \ldots, y_n)^t$ drücken wir die (gemeinsame) Likelihoodfunktion durch $\mu = (\mu_1, \ldots, \mu_n)^t$ aus und schreiben für sie $l(\mu, y)$.
Damit ist $\ln(l(\hat{\mu}, y))$ der (maximale) Wert der Loglikelihoodfunktion für das gegebene GLM.
Betrachten wir diesen Wert für alle in Frage kommenden Modelle, so existiert ein maximal erreichbarer Wert, der der optimalen Anpassungsgüte entspricht und durch $\ln(l(y, y))$ gegeben ist.
Das zugehörige „Modell", das jeder Beobachtung einen eigenen Parameter zuweist, heißt saturiertes Modell.

Bemerkung 3.63

(a) Das saturierte Modell ist kein Modell im eigentlichen Sinne, es kodiert die in der Stichprobe enthaltene Information lediglich um. Es ist auch nicht von eigenständigem Interesse, sondern ein Hilfsmittel, um zu einer verallgemeinerten Definition des Bestimmtheitsmaßes zu kommen.

(b) Für die Poisson-Regression gilt in der Nomenklatur von Definition 3.62:

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

(c) Das saturierte Modell kann auch anders kodiert sein (siehe Übung).

Definition 3.64 (Bestimmtheitsmaß eines GLM)
Sei unter der Nomenklatur von Definition 3.62 $\ln(l(y, y))$ der Loglikelihood-Wert des saturierten Modells, $\ln(l(\hat{\mu}, y))$ der eines fest vorgegebenen Modells (mit Designmatrix $X \in \mathbb{R}^{n \times p}$ mit $p < n$) und $\ln(l(\hat{\beta}_0, y))$ der des Null-Modells (das nur den Intercept enthält). Definiere

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

Dann ist das Bestimmtheitsmaß des Modells, welches $\hat{\mu}$ generiert hat, gegeben durch

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

Beschreibt das „$\hat{\mu}$-Modell“ die erhobenen Daten perfekt, so ist $D(\hat{\mu}) = 0 \Rightarrow R^2 = 1$.
Ist die Anpassungsgüte des „$\hat{\mu}$-Modell“ genau gleich der Anpassungsgüte des Null-Modells, so ist $D(\hat{\mu}) = D(\hat{\beta}_0) \Rightarrow R^2 = 0$.

Bemerkung 3.65 (Überdispersion)
Die Poissonverteilung ist einparametrisch mit $\mathbb{E}[\text{Poisson}(\lambda)] = \textit{Var}(\text{Poisson}(\lambda)) = \lambda$.
Daraus resultiert oftmals in der Praxis ein Problem für die Poisson-Regression, nämlich das der Überdispersion, d.h., dass $\textit{Var}(Y_i) > \lambda(\vec{x}_i)s_i$ ist.
Das ist insbesondere für das Prüfen der Regressionskoeffizienten auf statistische Signifikanz ein Problem, da unterschätzte Standardabweichungen zu große Teststatistiken (Z-scores) zur Folge haben.
Unter der Annahme, dass ein Skalierungsparameter ϕ existiert, so dass $\forall 1 \leq i \leq n : \textit{Var}(Y_i) = \phi\lambda(\vec{x}_i)s_i$ gilt, kann diesem Problem wir folgt begegnet werden:
Wir wissen aus der inferentiellen Likelihoodtheorie, dass (mit den Bezeichnungen von Definition 3.64) $D(\hat{\mu}) \underset{as.}{\sim} \chi^2_{n-p}$ gilt. Daraus folgt nach Pearson (vgl. Beispiel 3.12), dass

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

Nehmen wir an, dass $\hat{\mu}_i \approx \mu_i$ ist, so folgt, dass

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

denn $Q(Y)/\phi$ verhält sich dann wie die Summe von Quadraten von n standardisierten Zufallsvariablen bei p geschätzten Modellparametern, und $\mathbb{E}\left[\chi^2_\nu\right] = \nu$, also haben wir $\mathbb{E}\left[\frac{Q(Y)}{n-p}\right] \approx \phi$.
Aus dieser (heuristischen) Überlegung leiten sich zwei Schätzmethoden für den Überdispersionsparameter ϕ ab:

(a) Nehme ϕ als zusätzlichen Parameter in die (gemeinsame) Likelihoodfunktion auf und optimiere die (Log-)Likelihoodfunktion unter der Nebenbedingung $D(\hat{\mu}) = (n-p)$ (Quasi-Likelihood-Verfahren).

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

In beiden Fällen bleiben die Punktschätzungen der Regressionskoeffizienten unverändert, ihre geschätzten Standardabweichungen werden indes mit $\sqrt{\hat{\phi}}$ multipliziert.

3.6 Logistische Regression

Als zweites Beispiel für ein GLM betrachten wir die logistische Regression für Binärdaten als Response. Zur Einbettung in die allgemeine GLM-Theorie zunächst ein vorbereitendes Resultat.

Lemma 3.66
Die Familie der Bernoulliverteilungen mit Erfolgsparameter $p \in (0,1)$ bildet eine natürliche Exponentialfamilie im natürlichen Parameter $T(p) = \log(\frac{p}{1-p}) =: \operatorname{logit}(p)$.

Beweis: Für jedes $p \in (0,1)$ existiert eine Zähldichte (Likelihoodfunktion) der Form

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

Unter den Bezeichnungen von Definition 3.54 kann also $a(p) = 1-p$, $b(y) \equiv 1$, und $T(p) = \log(\frac{p}{1-p}) = \operatorname{logit}(p)$ gewählt werden. ■

Modell 3.67 (Logistische Regression mit Intercept)
Wir betrachten den Stichprobenraum $(\Omega, \mathcal{F}) = (\{0,1\}^n, 2^{\{0,1\}^n})$ und modellieren die Verteilung der stochastisch unabhängigen Responsevariablen $Y = (Y_1, \ldots, Y_n)^t \in \Omega$ als

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

wobei für die Kovariablen-abhängige Trefferwahrscheinlichkeit $p(\vec{x}_i)$ die strukturelle Annahme

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

getroffen wird. Dabei ist β_0 ein Intercept und $\vec{x}_i = (x_{i1}, \ldots, x_{ik})$ das Kovariablenprofil von Beobachtungseinheit $1 \le i \le n$.

Bemerkung 3.68

(a) Nach Lemma 3.66 ist das logistische Regressionsmodell ein GLM mit kanonischem Link.

(b) Für die (bedingten) Momente von Y_i gilt unter Modell 3.67

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

(c) Für $p \in (0,1)$ ist

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

Damit lässt sich die unter Modell 3.67 gemachte Strukturannahme auch schreiben als

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

Die (Umkehr-)Funktion $g^{-1} : \mathbb{R} \to (0,1),\ z \mapsto [1+\exp(-z)]^{-1}$ heißt die logistische Funktion.

(d) Das logistische Regressionsmodell ist ein multiplikatives Modell. Betrachten wir zwei Kovariablenprofile $\vec{x}_A$ und $\vec{x}_B$, die sich nur in den Ausprägungen einer festgelegten Kovariable $1 \leq j \leq k$ unterscheiden. Dann ist

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

die (bedingte) Chance (englisch: odds) für das Eintreten des Zielereignisses unter Kovariablenprofil $\vec{x}_A$. Folglich gilt für das logarithmische Chancenverhältnis (englisch: odds ratio, OR), dass

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

und folglich

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

Ist die Kovariable j eine dichotome Größe, so gilt insbesondere $\mathrm{OR} = \exp(\beta_j)$.

(e) Für die (Log-)Likelihoodfunktion des Gesamtexperimentes gilt unter Modell 3.67

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

da für jede Beobachtungseinheit $\forall 1 \leq i \leq n$ *gilt*

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

und

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

wobei

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

gilt, was den Einfluss von $\beta = (\beta_1, \ldots, \beta_k)^t$ *beschreibt.*

Anwendung 3.69 (Fall-Kontroll-Studien)

Die logistische Regression ist (zunächst) ein Modell, um prospektiv die bedingte Wahrscheinlichkeit für das Auftreten des Zielereignisses bei einer Beobachtungseinheit mit gegebenem Kovariablenprofil zu schätzen. Dazu dienen (prospektive) Kohortenstudien oder Querschnittsstudien. In der Epidemiologie werden indes auch (retrospektive) Fall-Kontroll-Studien durchgeführt, bei denen der (Krankheits-) Status $Y_i = y_i$ *der Response für alle Beobachtungseinheiten* $1 \leq i \leq n$ *schon zu Studienbeginn feststeht und retrospektiv das jeweilige Kovariablenprofil* $\vec{X}_i = \vec{x}_i$ *ermittelt wird. Unter einem solchen Studiendesign lässt sich an sich also nur* $\mathbb{P}_\beta\left[\vec{X}_i = \vec{x}_i | Y_i = y_i\right]$, $1 \leq i \leq n$, *schätzen.*

Durch Anwendung des Satzes von Bayes lässt sich indes auch in Fall-Kontroll-Studien ein logistisches Regressionsmodell aufstellen. Einziger Unterschied ist dabei, dass die Interpretierbarkeit des Intercepts verloren geht.

Notation:

Z_i : *Indikator für „Einschluss in die Fall-Kontroll-Studie“ (nein / ja),* $1 \leq i \leq n$.

$\pi_1 := \mathbb{P}(Z_i = 1 | Y_i = 1)$ *Einschlusswahrscheinlichkeit für Fälle, unabhängig von* $1 \leq i \leq n$.

$\pi_0 := \mathbb{P}(Z_i = 1 | Y_i = 0)$ *analog für Kontrollen*

Annahmen:

(1) Sampling-Wahrscheinlichkeiten hängen nur vom Wert des Response, nicht vom Kovariablenprofil ab, d.h., $\mathbb{P}(Z_i = 1 | Y_i = \ell, \vec{X}_i = \vec{x}_i) = \pi_\ell$, $\ell \in \{0, 1\}$.

(2) Logistisches Modell für die (bedingte) Verteilung der Responsevariablen:

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

Dann gilt nach dem Satz von Bayes

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

mit $\beta_0^* := \beta_0 + \log(\pi_1/\pi_0)$. Den Responsestatus-spezifischen Inklusionswahrscheinlichkeiten kann also durch geeignete Umdefinierung des Intercepts Rechnung getragen werden.
Um den Zusammenhang $p(\vec{x}_i)$ zu analysieren, kann also ein logistisches Regressionsmodell in Analogie zur Situation im Falle einer prospektiven Studie angepasst werden, wobei lediglich der Intercept seine (ursprüngliche) Interpretation verliert.
Können keine verlässlichen Angaben zu π_0 und π_1 gemacht werden, geht die Interpretationsfähigkeit des Intercepts völlig verloren; $p(\vec{x}_i)$ kann indes trotzdem noch mit den üblichen Verfahren inferiert werden.

Anwendung 3.70 (Receiver Operating Characteristic Kurve, englisch: ROC)
Hat man ein logistisches Regressionsmodell angepasst, so liegt es nahe, die geschätzten Regressionskoeffizienten und die daraus resultierenden geschätzten (bedingten) Wahrscheinlichkeiten für den Responsewert 0 zur Klassifizierung neuer Beobachtungseinheiten, von denen nur das Kovariablenprofil bekannt ist, zu verwenden (medizinisch: Diagnose). Dazu muss man also einen Schwellenwert p^* für $\hat{p}(\vec{x}_{neu})$ festlegen, ab welchem $\hat{y}_{neu} = 1$ diagnostiziert wird. Da die Schätzung der Regressionskoeffizienten stochastischen Schwankungen unterlegen ist, ist eine perfekte Diagnosefähigkeit des gefitteten Modells nicht zu erwarten.
Zur Festlegung von p^* kann eine sogenannte ROC-Analyse dienen. Dazu ordnen wir die aus der Stichprobe $((y_1, \vec{x}_1), \ldots, (y_n, \vec{x}_n))$ geschätzten Werte $(\hat{q}(\vec{x}_i))_{1 \leq i \leq n}$ mit

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

der Größe nach an und erhalten $\hat{q}_{1:n} \leq \hat{q}_{2:n} \leq \ldots \leq \hat{q}_{n:n}$. Wir bezeichnen zudem $n_0 := |\{1 \leq i \leq n : y_i = 0\}|$ und $n_1 := n - n_0 = |\{1 \leq i \leq n : y_i = 1\}|$.
Die ROC-Kurve ist nun der Graph einer zufälligen Irrfahrt mit n Schritten im Einheitsquadrat, startend in $(0, 0)$ und endend in $(1, 1)$. Die Irrfahrt wird dabei im Schritt $1 \leq \ell \leq n$ um einen Sprung der Breite $1/n_0$ nach rechts weitergeführt, falls $\hat{q}_{\ell:n}$ zu $y_{\ell:n} = 0$ gehört, und um einen Sprung der Höhe $1/n_1$ nach oben weitergeführt, falls $\hat{q}_{\ell:n}$ zu $y_{\ell:n} = 1$ gehört, wobei wir die Responsewerte gemäß der Ordnung der $(\hat{q}(\vec{x}_i))_{1 \leq i \leq n}$ permutieren. Es lässt sich leicht nachweisen, dass die Irrfahrt mit dieser Regel nach dem n-ten Schritt den Koordinatenpunkt $(1, 1)$ erreicht hat.
Beispiel:
Angenommen, es sind $n_0 = 2$ Indikatoren $y_i = 0$ und $n_1 = 3$ Indikatoren $y_i = 1$ beobachtet worden und $n = n_0 + n_1 = 5$. Wir nehmen an, dass $y_{1:5} = y_{2:5} = y_{4:5} = 1$ und $y_{3:5} = y_{5:5} = 0$.

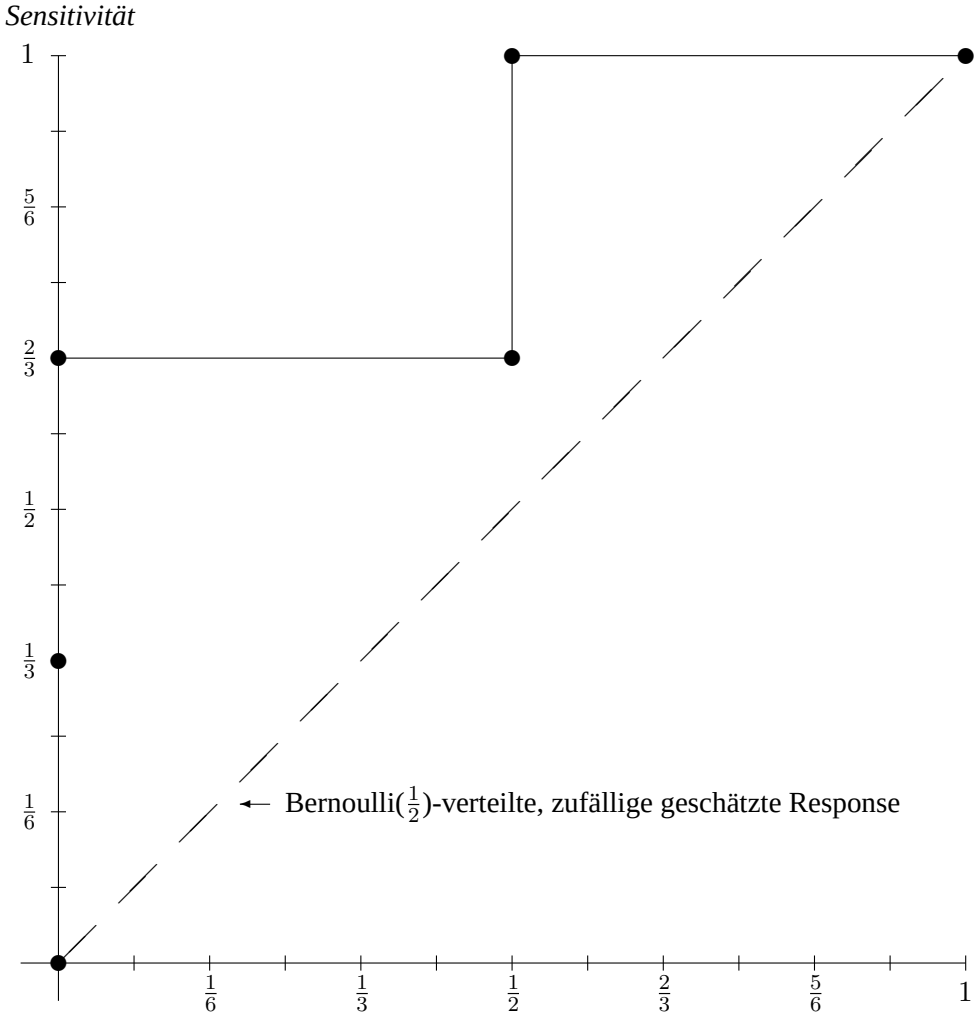


Abbildung 3.1: Beispiel für eine ROC-Kurve

Es wird nun der Schwellenwert $p^* \in \{1 - \hat{q}_{1:n}, \ldots, 1 - \hat{q}_{n:n}\}$ gewählt, der zu dem Schritt ℓ^* der Irrfahrt gehört, in welchem die Irrfahrt dem Koordinatenpunkt $(0, 1)$ am nächsten ist. Dieses Vorgehen minimiert die geschätzte (bzw. empirische) gewichtete Missklassifikationswahrscheinlichkeit. Im Beispiel würde $p^* = 1 - \hat{q}_{2:5}$ gewählt werden. Damit würden (empirisch in der erhobenen Stichprobe) zwei Drittel der „Fälle“ und alle „Kontrollen“ korrekt klassifiziert, denn

$$\hat{p}_{1:5} = 1 - \hat{q}_{1:5} > p^* \quad \Rightarrow \quad \hat{y}_{1:5} = 1 = y_{1:5},$$

$$\hat{p}_{2:5} = 1 - \hat{q}_{2:5} = p^* \quad \Rightarrow \quad \hat{y}_{2:5} = 1 = y_{2:5},$$

$$\hat{p}_{3:5} = 1 - \hat{q}_{3:5} < p^* \quad \Rightarrow \quad \hat{y}_{3:5} = 0 = y_{3:5},$$

$$\hat{p}_{4:5} = 1 - \hat{q}_{4:5} < p^* \quad \Rightarrow \quad \hat{y}_{4:5} = 0 \neq y_{4:5} = 1,$$

$$\hat{p}_{5:5} = 1 - \hat{q}_{5:5} < p^* \quad \Rightarrow \quad \hat{y}_{5:5} = 0 = y_{5:5}.$$

In der Epidemiologie (Theorie diagnostischer Tests) wird für binäres Y
$\mathbb{P}_{\text{Testverfahren}}(\hat{Y} = 1|Y = 1)$ *als Sensitivität,* $\mathbb{P}_{\text{Testverfahren}}(\hat{Y} = 0|Y = 0)$ *als Spezifizität und*
$\mathbb{P}_{\text{Testverfahren}}(\hat{Y} = 1|Y = 0) = 1 -$ *Spezifität als Falsch-Positiv-Rate,*
$\mathbb{P}_{\text{Testverfahren}}(\hat{Y} = 0|Y = 1) = 1 -$ *Sensitivität als Falsch-Negativ-Rate bezeichnet.*
Damit wird also in der ROC-Analyse die geschätzte $(1 -$ *Spezifizität*$)$ *bzw. Falsch-Positiv-Rate gegen die geschätzte Sensitivität (englisch auch: true positive rate) aufgetragen.*
Im Beispiel erhalten wir im optimalen Punkt eine geschätzte Falsch-Positiv-Rate von 0 *(alle „Kontrollen“ in der Stichprobe korrekt klassifiziert) und eine geschätzte Sensitivität von* $2/3$ *(zwei Drittel aller „Fälle“ in der Stichprobe korrekt diagnostiziert).*
Die Fläche unter der ROC-Kurve (englisch: area under the curve, ROC-AUC) ist ein zusammenfassendes Maß für die diagnostische Güte des Klassifikationsverfahrens. Als Vergleichswert kann $\mathrm{AUC}_{\mathrm{Raten}} = 1/2$ *herangezogen werden, was einer zufälligen (gleichverteilten) Zuordnung der geschätzten Response auf ein gegebenes Kovariablenprofil entspricht (Diagonale im Einheitsquadrat).*

Bemerkung 3.71 (Probit-Modell)
Obschon $g = \mathrm{logit}$ *der kanonische Link der logistischen Regression ist, existieren auch Alternativen zu dieser Link-Funktion. Zur Motivation beachten wir, dass* $G_{\mu,\tau} : \mathbb{R} \to (0, 1)$ *gegeben durch*

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

die Verteilungsfunktion der logistischen Verteilung mit Mittelwertsparameter μ *und Streuungsparameter* $\tau > 0$ *ist.*
Damit ist die logistische Funktion

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

also die Verteilungsfunktion einer standardisierten logistischen Verteilung mit $\mu = 0$ *und* $\tau = 1$ *und folglich ist die strukturelle Annahme der logistischen Regression gegeben durch*

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

Viele andere Familien von Verteilungen haben ebenfalls Lokations- und Skalenparameter und ihre Verteilungsfunktionen können demnach in analoger Weise als inverse Link-Funktionen zum Einsatz kommen. Wird speziell die Verteilungsfunktion einer Gauß’schen Normalverteilung benutzt, spricht man auch von einem Probit-Modell.

3.7 Cox-Regression, Überlebenszeitanalysen

Als letzte noch zu behandelnde Datenstruktur (der Response) beschäftigen wir uns in diesem Abschnitt mit Überlebenszeitdaten (Zeitspannen bis zum Eintritt eines festgelegten Zielereignisses).

Zunächst einige vorbereitende Begriffsbildungen aus der Überlebenszeitanalyse (englisch: Survival Analysis).

Definition 3.72 (Grundbegriffe der Survival Analysis)
Sei T eine nicht-negative, reellwertige Zufallsvariable über einem Wahrscheinlichkeitsraum mit Wahrscheinlichkeitsverteilung $\mathbb{P}$ und mit Verteilungsfunktion $F : [0, \infty) \to [0, 1]$, so dass $F(t) = \mathbb{P}(T \leq t)$, $t \in [0, \infty)$. Wir denken uns T als (zufällige) Zeitspanne bis zum Eintreten eines festgelegten Zielereignisses. Dann heißt

a) $S : [0, \infty) \to [0, 1]$*, gegeben durch* $S(t) = \mathbb{P}(T > t) = 1 - F(t)$*, Survivalfunktion von T.*

b) $\Lambda : [0, \infty) \to [0, \infty]$*, gegeben durch*

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

kumulative Hazardfunktion von T.

Nach Gill and Johansen (1990) existiert eine Folge von Unterteilungen $(t_i^{(n)})_{1 \leq i \leq k(n)}$ des Intervalls $(0, t]$, so dass mit $t_0^{(n)} \equiv 0$:

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

c) Ist die Verteilung von T stetig mit Lebesguedichte f und $t \mapsto S(t)$ differenzierbar, so heißt

$$\lambda(t) := \frac{d\Lambda(t)}{dt} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

Hazardfunktion oder Inzidenzdichte von T. Offenbar gilt dann für $t > 0$:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \,|T > t)}{\Delta t} \,.$$

Man nennt $\lambda(t)$ daher auch die instantane Ausfallrate oder instantanes Risiko.

d) Für zwei unterschiedliche Grundgesamtheiten A und B mit zugehörigen Inzidenzdichten λ_A und λ_B heißt die durch

$$\mathrm{RR}(t) := \frac{\lambda_A(t)}{\lambda_B(t)}$$

gegebene Funktion relatives Risiko.

Bemerkung 3.73

(i) Unter den Gegebenheiten von Definition 3.72.c) gilt

$$\begin{aligned}\Lambda(t) \quad &= \quad \int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1-F(s)}ds \\ &\underset{u:=F(s)}{=} \int_0^{F(t)} \frac{du}{1-u} = -\ln(1-u)\Big|_0^{F(t)} = -\ln(S(t)) \\ \iff S(t) &= \exp(-\Lambda(t)) \iff F(t) = 1-\exp(-\Lambda(t)).\end{aligned}$$

(ii) Ferner ist unter Definition 3.72.c) für $\Delta t > 0$ „klein“

$$\begin{aligned}\mathrm{R}(t,\Delta t\,) &:= \mathbb{P}(t < T \le t+\Delta t\,|T>t) \\ &= \frac{F(t+\Delta t\,)-F(t)}{1-F(t)} = \frac{S(t)-S(t+\Delta t\,)}{S(t)} \\ &\approx \Lambda(t+\Delta t\,)-\Lambda(t) \qquad \text{(*)}\end{aligned}$$

das Risiko, dass das Zielereignis in der Zeitspanne $(t, t+\Delta t\,]$ eintritt, gegeben, dass der Zeitpunkt t zielereignisfrei gewesen ist. Dabei gilt (*), denn

$$\frac{S(t)-S(t+\Delta t\,)}{S(t)} = 1-\exp(\Lambda(t)-\Lambda(t+\Delta t\,))$$

und $1-\exp(-x) \approx x$ für „kleines“ $x > 0$.

(iii) Ist T exponentialverteilt mit Intensitätsparameter $\vartheta > 0$, so ist

$$\lambda(t) = \frac{\vartheta\exp(-\vartheta t)}{\exp(-\vartheta t)} = \vartheta$$

die zeitlich konstante Inzidenzrate von T.

Eine erste Hauptaufgabe der Survival Analysis besteht in der Schätzung der Survivalfunktion. In der Praxis tritt dabei häufig das Problem zensierter Daten auf, d.h., dass nicht bei allen ursprünglich in die Studie eingeschlossenen Beobachtungseinheiten bis zum Ablauf der Studie entschieden werden kann, ob das Zielereignis eingetreten ist oder nicht. Dies verzerrt den zunächst naheliegenden, auf der empirischen Verteilungsfunktion basierenden Schätzer.

Ursachen für Zensierungen:

- Loss to follow-up (PatientIn zieht weg, etc.)
- Dropout (z.B.: unerwartete Nebenwirkungen treten bei einer Therapie auf)
- Studienende (z.B.: Finanzierung läuft aus)
- PatietIn stirbt durch eine Ursache, die nicht im Zusammenhang mit dem Zielereignis von Interesse steht.

Eine verfeinerte Schätzmethodik unter Einbeziehung von Zensierungen stellt der Kaplan-Meier-Schätzer (englisch auch: product-limit estimator) dar.

Definition 3.74 (Kaplan-Meier-Schätzer)
Gegeben sei eine iid. Stichprobe mit n Beobachtungseinheiten aus einer (homogenen) Grundgesamtheit mit (unbekannter) Survivalfunktion S bezüglich eines festgelegten Zielereignisses.

Seien $t_1 < t_1 < \ldots < t_k$ mit $k \leq n$ die geordneten, unterschiedlichen Beobachtungszeitpunkte in der Stichprobe und d_i, $1 \leq i \leq k$, die Anzahl an beobachteten Zielereignissen zum Zeitpunkt t_i. Dabei wird t_i in Bezug auf die Einschlusszeitpunkte der Beobachtungseinheiten in die Stichprobe ausgedrückt. Bezeichne ferner n_i, $1 \leq i \leq k$, die Anzahl an Beobachtungseinheiten in der Stichprobe, die unmittelbar vor dem Zeitpunkt t_i noch unter Risiko gestanden haben. Dann ist $\hat{S}$, gegeben durch

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), t \geq 0,$$

der Kaplan-Meier-Schätzer für S basierend auf der Stichprobe vom Umfang n.

Bemerkung 3.75

(i) Treten in der Stichprobe keinerlei Zensierungen auf, so gilt $\forall t \geq 0 : \hat{S}(t) = 1 - \hat{F}_n(t) = 1 - n^{-1} \sum_{i=1}^{n} \mathbf{1}_{[0,t]}(t_i)$, wobei $\hat{F}_n$ die empirische Verteilungsfunktion der n Ereigniszeitpunkte bezeichnet.

(ii) Eine erste heuristische Begründung für $\hat{S}$ liefert der Multiplikationssatz für (bedingt) stochastisch unabhängige Ereignisse.

Beispiel 3.76
In der folgenden Liste sind die Überlebenszeiten (in Wochen) von zehn Patienten mit einem superfizialen Harnblasenkarzinom nach Beginn einer (jeweiligen) Chemotherapie festgehalten. Dabei sind die zensierten Beobachtungen mit einem Kreuz („+“) gekennzeichnet:

$$63, 59, 57+, 37, 33, 21+, 11, 57, 44+, 37.$$

t_i	d_i	n_i	Faktor $= \frac{n_i - d_i}{n_i}$	$\hat{S}(t_i)$
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

Tabelle 3.3: Tabelle zur Berechnung des Kaplan-Meier-Schätzers

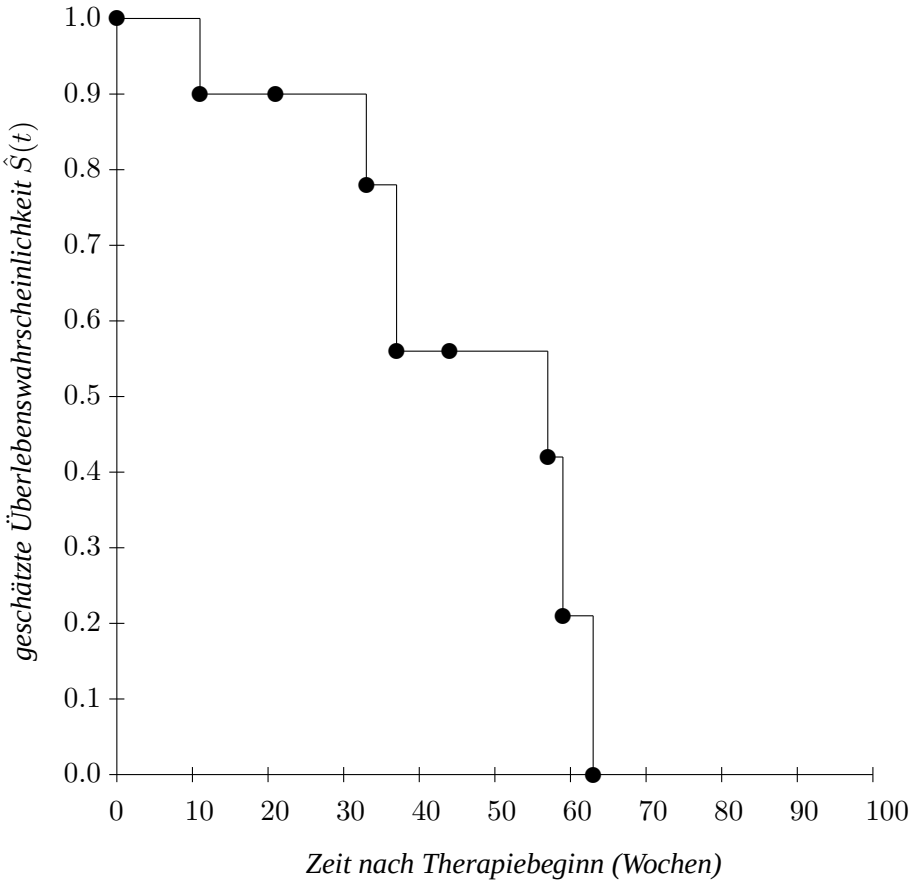


Abbildung 3.2: Beispielhafter Graph eines Kaplan-Meier-Schätzers

Definition 3.77 (Nichtparametrische Likelihoodfunktion)
Seien $X_1, \ldots, X_n$ reellwertige iid. Zufallsvariablen, $x = (x_1, \ldots, x_n)$ eine Realisierung von $(X_1, \ldots, X_n)$ und $\mathcal{M}$ die Menge aller Verteilungsfunktionen auf $\mathbb{R}$.

Dann heißt $L : \mathcal{M} \times \mathbb{R}^n \to [0,1]$

$$\begin{aligned}(F,x) \mapsto L(F,x) &:= \prod_{i=1}^{n} [F(x_i) - F(x_i-)] \\ &= \prod_{i=1}^{n} \mathbb{P}_F(\{x_i\})\end{aligned}$$

die nichtparametrische Likelihoodfunktion zu x.

Anmerkung: Offenbar kommen zur Maximierung der nichtparametrischen Likelihoodfunktion nur diskrete Verteilungen in Betracht, die ihre gesamte Masse auf die Punkte $(x_i)_{1\le i\le n}$ verteilen.

Satz 3.78 (Statistische Eigenschaften von $\hat{S}$)
Seien $T_1, \ldots, T_n$ *nicht-negative, reellwertige Zufallsvariablen mit* $\forall 1 \le i \le n : T_i \sim T$ *mit unbekannter Verteilungsfunktion* F *und zugehöriger Survivalfunktion* S.
Sei ferner $C = (C_1, \ldots, C_n)^t$ *ein Vektor ebenfalls nicht-negativer, reellwertiger Zufallsvariablen mit (gemeinsamer) Verteilung* $\mathcal{L}_C$*, die nicht von* F *abhängt.*
Wir nehmen an, dass, gegeben C*, die* $(T_i)_{1\le i\le n}$ *bedingt stochastisch unabhängig sind und dass wir* $Y = (Y_1, \ldots, Y_n)^t$ *beobachten können mit* $\forall 1 \le i \le n : Y_i = \min(T_i, C_i)$.
Ferner liege Zensierungsinformation vor durch Indikatoren $\delta_i := \mathbf{1}_{\{T_i \le C_i\}}$, $1 \le i \le n$.
Dann gilt:

(a) Mit den unter Definition 3.74 getroffenen Bezeichnungen ist $\hat{S}$ *nichtparametrischer Maximum-Likelihood-Schätzer (NPMLE) von* S.

(b) $\hat{S}$ *ist gleichmäßig konsistent auf Intervallen* $[0,t]$ *mit* $S(t) > 0$.

Für alle Punkte $t \ge 0$ *mit* $S(t) > 0$ *gilt:*

(c) Sind die $(C_i)_{1\le i\le n}$ *iid. mit Verteilungsfunktion* G *von* C_1*, so ist*

$$0 \le \mathbb{E}\left[\hat{S}(t)\right] - S(t) \le (1 - S(t)) \left\{1 - S(t)(1 - G(t))\right\}^n .$$

Insbesondere strebt der Bias von $\hat{S}(t)$ *für wachsenden Stichprobenumfang gegen Null. Es existieren weitere Bias-Abschätzungen und sogar exakte Formeln (siehe Stute (1994)).*

(d) Es gilt ein zentraler Grenzwertsatz für $\hat{S}(t) \equiv \hat{S}(n,t)$ *und die Varianz von* $\hat{S}(t)$ *kann approximiert werden durch die Greenwood-Formel:*

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)}.$$

Beweis: Sei $\vec{t} := (t_1, \ldots, t_n)^t$ und $\vec{c} := (c_1, \ldots, c_n)^t$. Zum Beweis von (a) beachten wir, dass

$$L(S, \mathcal{L}_c, \vec{t}, \vec{c}) = L(S, \mathcal{L}_c, \vec{c}) \cdot L(S, \vec{t} | C = \vec{c})$$

geschrieben werden kann. Dabei ist $L(S, \mathcal{L}_c, \vec{c}) = \mathcal{L}_C(\{\vec{c}\})$ und damit ohne Information über die interessierende Survivalfunktion S. Es genügt also, die bedingte nichtparametrische Likelihood-funktion $L(S, \vec{t}|C = \vec{c})$ zu optimieren. Wir erhalten

$$\begin{aligned} L(S, \vec{t}|C = \vec{c}) &= \prod_{i:\delta_i=1} \mathbb{P}^T(\{t_i\}) \prod_{i:\delta_i=0} S(c_i) \\ &= \prod_{i=1}^{n} \left[\mathbb{P}^T(\{y_i\})\right]^{\delta_i} \left[S(y_i)\right]^{1-\delta_i} \\ &= \prod_{j=1}^{k} \lambda_j^{d_j} (1-\lambda_j)^{n_j - d_j} \end{aligned}$$

nach Übungsaufgabe 42, wobei $k \leq n$ wie in Definition 3.74 und

$$\lambda_j = \mathbb{P}\big(T = t_{j:k} | T \geq t_{j:k}\big),\ 1 \leq j \leq k.$$

Einfache Algebra ergibt, dass folglich $\hat{\lambda}_j = d_j/n_j$, $1 \leq j \leq k$, geschätzt wird und es ergibt sich insgesamt

$$\hat{S}_{\mathrm{NPMLE}}(t) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j}$$

wie gewünscht, da nach Übungsaufgabe 42

$$\forall 0 \leq t \leq t_{k:k} : S(t) = \prod_{j:t_{j:k} \leq t} (1 - \lambda_j) \text{ gilt.}$$

Die Aussage unter Teil (b) findet sich als Theorem IV.3.1 in dem Buch von Andersen et al. (1993). Die asymptotische Normalität von $\hat{S}(t)$ unter Teil (d) ist eine Konsequenz aus Theorem IV.3.2 in Andersen et al. (1993) und die Greenwood-Formel folgt mit Hilfe der Delta-Methode (siehe Übungsaufgabe). ■

Sir David Cox hat ein Modell entwickelt, um Survival Analysis auch für heterogene Grundgesamtheiten (unter Einbeziehung von Kovariablen) betreiben zu können.

Modell 3.79 (Cox' proportional hazards-Modell)
Wir betrachten beobachtbare, stochastisch unabhängige Responsevariablen $Y_1, \ldots, Y_n$ mit $\forall 1 \leq i \leq n : Y_i = \min(T_i, C_i)$ wie in Satz 3.78.
Die Verteilung der interessierenden Ereigniszeiten $(T_i)_{1 \leq i \leq n}$ wird Kovariablen-abhängig modelliert und wir machen dabei die folgende Strukturannahme:

$$\forall 1 \leq i \leq n : \lambda(t|\vec{X}_i = \vec{x}_i) = \lambda_0(t) \exp(\eta_i) \tag{3.79.1}$$

mit dem linearen Prädiktor $\eta_i = \sum_{j=1}^{k} \beta_j x_{ij}$, $1 \leq i \leq n$, wobei $\vec{x}_i = (x_{i1}, \ldots, x_{ik})$ wie zuvor. Insbesondere gilt im Vergleich zur Baseline ($\vec{x}_{Baseline} \equiv \vec{0}$), dass $\mathrm{RR}(t|\vec{X}_i = \vec{x}_i)) \equiv \mathrm{RR}(\vec{x}_i) = \exp(\eta_i)$. Die Funktion λ_0 wird unspezifiziert gelassen und Baseline-Hazard genannt.
Ziel der statistischen Inferenz sind (vornehmlich) die Regressionskoeffizienten $\beta = (\beta_1, \ldots, \beta_k)^t$.

Bemerkung 3.80

(a) Das proportional hazards-Modell nach Cox (1972) ist ein semiparametrisches, multiplikatives Modell.

(b) Die Abbildung $t \mapsto \log(\lambda_0(t))$ kann als „Intercept-Funktion“ angesehen werden. Damit $\vec{x}_{Baseline} = \vec{0}$ sinnvoll zu interpretieren ist, sollten alle Kovariablen zunächst an ihren empirischen Mittelwerten zentriert werden („Standardisierung“).

(c) Kann eine plausible Annahme zu λ_0 gemacht werden (z.B. Weibull oder Gompertz), so kann Modell 3.79 zu einem parametrischen Modell modifiziert werden, auf welches dann die inferentielle Likelihood-Theorie angewendet werden kann.

(d) Die in (3.79.1) gemachte Proportionalitätsannahme kann und sollte durch Untersuchung der (nach Kaplan-Meier) geschätzten Survivalfunktionen überprüft werden (siehe 3.81). Bemerkenswerterweise ist Modell 3.79 in der Praxis oft ein akzeptables Modell.

(e) Die Schätzung der Regressionskoeffizienten im Falle einer unspezifizierten Baseline-Hazard erfolgt durch Maximierung der partiellen Likelihoodfunktion der Stichprobe, wobei auf die beobachteten Ereigniszeitpunkte bedingt wird.

Nehmen wir dazu der Einfachheit halber an, dass genau m Zielereignisse zu unterschiedlichen Zeitpunkten $t_{1:m} < t_{2:m} < \ldots < t_{m:m}$ beobachtet worden sind.

Nach Bemerkung 3.73.(ii) ist dann $\forall 1 \leq i \leq m$

$$\begin{aligned} \mathrm{R}(t_{i:m}, \Delta t \,|\vec{X}_i = \vec{x}_i) &= \Delta t \, \lambda(t_{i:m}|\vec{X}_i = \vec{x}_i) \\ &= \Delta t \, \lambda_0(t_{i:m}) \exp(\eta_{i:m}) \end{aligned}$$

für infinitesimal kleines Δt. Ist ferner R_i die Menge aller Beobachtungseinheiten, die unmittelbar vor dem Zeitpunkt $t_{i:m}$ unter Risiko stehen (bzw. gestanden haben), so ist für $1 \leq i \leq m$

$$\sum_{\ell \in R_i} \Delta t \, \lambda_0(t_{i:m}) \exp(\eta_\ell) = \Delta t \, \lambda_0(t_{i:m}) \sum_{\ell \in R_i} \exp(\eta_\ell)$$

die bedingte Wahrscheinlichkeit dafür, dass für irgendeine Beobachtungseinheit aus R_i das Zielereignis innerhalb der Zeitspanne $(t_{i:m}, t_{i:m} + \Delta t\,]$ eintritt. Als partielle Likelihoodfunktion wird dann das Produkt der auf die beobachteten Ereigniszeitpunkte bedingten Beobachtungswahrscheinlichkeiten aller Zielereignisse verwendet, also

$$L_{partiell}(\beta, y) = \prod_{i=1}^{m} \frac{\exp(\eta_{i:m})}{\sum_{\ell \in R_i} \exp(\eta_\ell)}$$

mit zugehöriger partieller Log-Likelihoodfunktion

$$\ln(L_{partiell}(\beta, y)) = \sum_{i=1}^{m} \eta_{i:m} - \ln\left(\sum_{\ell \in R_i} \exp(\eta_\ell)\right).$$

Mit Hilfe offensichtlicher (kombinatorischer) Modifikationen können auch Fälle behandelt werden, in denen Bindungen zwischen den $(t_{i:m})_{1 \le i \le m}$ auftreten.

Anwendung 3.81 (Tests auf proportionale Hazards)
Modellgleichung (3.79.1), zusammen mit der Beziehung $S(t) = \exp(-\Lambda(t))$ aus Bemerkung 3.73.(i) ergibt, dass unter den Gegebenheiten von Definition 3.72.(c) für ein gegebenes Kovariablenprofil $\vec{x}_A$ gilt:

$$\begin{aligned} S_A(t) &= \exp(-\Lambda_A(t)) = \exp\left(-\int_0^t \lambda_A(s)ds\right) \\ &= \exp\left(-\int_0^t \lambda_0(s)ds\, \exp(\eta_A)\right) = [S_0(t)]^{\exp(\eta_A)}. \end{aligned}$$

Damit ist

$$\log(-\log(S_A(t))) = \eta_A + \log(-\log(S_0(t)))$$

und für zwei unterschiedliche Kovariablenprofile $\vec{x}_A$ und $\vec{x}_B$ gilt folglich

$$\log(-\log(S_A(t))) - \log(-\log(S_B(t))) = \eta_A - \eta_B.$$

Für zwei unterschiedliche Strata von Kovariablenprofilen kann also eine visuelle Inspektion der log-minus log-transformierten geschätzten Survivalfunktionen als Modelldiagnosetechnik zum Einsatz kommen. Ein formaler Test auf proportionale Hazards kann konstruiert werden mit Hilfe der skalierten Schoenfeld-Residuen (siehe z.B. Grambsch and Therneau (1994)).

Beispiel 3.82
Häftlingkeitsdaten aus Rossi et al. (1980), siehe Präsentation mit R.

Definition 3.83 (Pseudo-Bestimmtheitsmaß bei partieller Likelihood)
Da zur Modellanpassung unter Bemerkung 3.80.(e) die partielle Likelihoodfunktion verwendet worden ist, ist die Schätzung der durch das Modell erklärten Streuung der Response hier komplizierter als im Falle der ANCOVA oder der GLMs. Eine Approximation liefert die einfache Formel von Maddala (Maddala, 1983, Seite 39). Bezeichne dazu in Anlehnung an Definition 3.64

$$D(\hat{\beta}) = 2\left[\log(L_{partiell}(\hat{\beta}, y)) - \log(L_{partiell}(\hat{\beta}_0, y))\right],$$

wobei $\hat{\beta}_0$ dem Null-Modell (nur Intercept) entspricht. Dann ist ein Pseudo-Bestimmtheitsmaß gegeben durch

$$\tilde{R}^2_{Maddala} := 1 - \exp\left(-\frac{D(\hat{\beta})}{n}\right).$$

Es existieren viele weitere Vorschläge in der Literatur, wobei manche wie im Falle der GLMs (siehe Definition 3.64) auch einen Vergleich mit dem saturierten Modell beinhalten.

Anwendung 3.84 (Zeitabhängige Kovariablen)
Neben statischen Kovariablen (wie zum Beispiel Geschlecht oder ethnische Zugehörigkeit) treten in der Survival Analysis oftmals auch dynamische Kovariablen auf, deren Ausprägungen sich selbst im Zeitverlauf ändern (können), z.B. kumulative Schadstoffexposition oder Therapieindikator (bei Crossover-Studien). Zur statistischen Analyse des Einflusses solcher zeitabhängigen Kovariablen müssen im Likelihood-Ansatz zu jedem Ereigniszeitpunkt $t_{i:m}$ für alle Beobachtungseinheiten unter Risiko ($\ell \in R_i$) die Zeit-aktuellen Werte der Kovariablen zur Verfügung stehen.
Eine modifizierte Form der partiellen Likelihoodfunktion ist dann unter den in Bemerkung 3.80.(e) getroffenen Annahmen gegeben durch

$$L_{partiell}(\beta, y) = \prod_{i=1}^{m} \frac{\exp(\eta_{i:m,i})}{\sum_{\ell \in R_i} \exp(\eta_{\ell,i})},$$

wobei $\eta_{\ell,i} = \sum_{j=1}^{k} \beta_j x_{\ell j,i}$ den (zeitabhängigen) Wert des linearen Prädiktors von Beobachtungseinheit ℓ zum Zeitpunkt $t_{i:m}$ bezeichnet und $\eta_{i:m,i}$ zu der Beobachtungseinheit gehört, bei der das Zielereignis zum Zeitpunkt $t_{i:m}$ eintritt (mit analogen Verallgemeinerungen bezüglich Bindungen wie unter Bemerkung 3.80.(e)). Da dieses Vorgehen unter Umständen zu einem erheblichen Berechnungsmehraufwand führen kann, ist ein vorgeschalteter Test auf Zeitabhängigkeit für solche Kovariablen empfehlenswert, für die Dynamik vermutet wird.
Sei X_1 eine solche Kovariable. Das ursprüngliche proportional hazards-Modell (der Einfachheit halber nur mit dieser einen Kovariablen) macht die Annahme

$$\lambda(t|X_{i1} = x_{i1}) = \lambda_0(t) \exp(\beta_1 x_{i1}),\ 1 \leq i \leq n.$$

Um nun eine Zeitabhängigkeit des Einflusses von X_1 auf die Response zu überprüfen, nehmen wir eine abgeleitete Variable mit Dynamik X_2 hinzu, etwa $X_2 = X_1 t$ oder $X_2 = X_1 \log(t)$, und betrachten das erweiterte Modell, das den Zusammenhang

$$\lambda(t|X_{i1} = x_{i1}) = \lambda_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2}),\ 1 \leq i \leq n,$$

annimmt. Durch Prüfen der Hypothese $H_0 : \beta_2 = 0$, welches mit Standardmethodik (Wald-Test, t-Test) möglich ist, lässt sich nun beurteilen, ob der Einfluss der interessierenden Kovariable X_1 eine Zeitabhängigkeit hat.

Beispiel 3.85 (Leukämie-Daten, Example 11.6 in Le (2003))
Im Rahmen einer kontrollierten Wirksamkeitsstudie für ein Medikament gegen Leukämie mit $n = 42$ PatientInnen wurde jede(r) Versuchsteilnehmende zufällig dem Therapiearm ($x_{i1} = 1$) oder dem Placeboarm ($x_{i1} = 0$) zugeordnet.

Das Zielereignis (Response) war die Remissionszeit bzw. Zensierung, falls bei Studienende noch kein Rückfall eingetreten war.
Es sollen die folgenden beiden Fragestellungen geprüft werden:

(1) Hat das Medikament eine Wirkung darauf, die Remissionszeit zu verlängern (ja/nein)?

(2) Hat das Medikament ein (zeitlich) kumulative Wirkung, d.h., spielt die Behandlungsdauer eine wichtige Rolle?

⇝ siehe Präsentation mit R und Handouts.

3.8 Bayesianische Behandlung linearer Modelle

In diesem Abschnitt geben wir einen kurzen Einblick in die Bayesianische Behandlung von (multiplen) linearen Regressionsmodellen und GLMs.
Wir beginnen mit dem klassischen ANCOVA-Modell aus 3.14, also $Y = X\beta + \varepsilon$ mit allen gemachten Zusatzannahmen. In Bayesianischer Notation haben wir:

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

wobei $\tilde{\beta}$ und $\tilde{\sigma}$ Zufallsgrößen sind.
Besonders einfach wird Bayesianische Inferenz bei Vorliegen von konjugierten Verteilungsklassen.

Definition 3.86 (Inverse Gamma-Verteilung)
Ist Z_1 eine nicht-negative Zufallsvariable mit $Z_1 \sim \mathrm{Gamma}(\alpha, r)$, so heißt $Z_2 := 1/Z_1$ invers gammaverteilt, in Zeichen $Z_2 \sim \mathrm{IG}(\alpha, r)$. Die Lebesguedichte von Z_2 ist gegeben durch

$$\begin{aligned} f_{Z_2}(z) &= \frac{\alpha^r}{\Gamma(r)} z^{-(r+1)} \exp(-\alpha/z) \mathbf{1}_{(0,\infty)}(z) \text{ und es gilt} \\ \mathbb{E}[Z_2] &= \frac{\alpha}{r-1} \text{ und } Var(Z_2) = \frac{\alpha^2}{(r-1)^2(r-2)}, \text{ falls } r > 2. \end{aligned}$$

Definition 3.87 (Normal-inverse Gammaverteilung)
Sei $\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2 \sim \mathcal{N}_p\left(m, \sigma^2 M\right)$ für Hyperparameter m und M, und sei zusätzlich $\tilde{\sigma}^2 \sim \mathrm{IG}(\alpha, r)$ mit Hyperparametern α und r. Dann heißt die gemeinsame Verteilung von $\tilde{\beta}$ und $\tilde{\sigma}^2$ Normal-inverse Gammaverteilung mit Parametern m, M, α und r. Als gemeinsame Dichte ergibt sich:

$$\begin{aligned} f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta, \sigma^2) &= f_{\tilde{\beta}|\tilde{\sigma}^2=\sigma^2}(\beta)\, f_{\tilde{\sigma}^2}(\sigma^2) \\ &= \left[(2\pi)^{\frac{p}{2}} \sigma^p |M|^{\frac{1}{2}}\right]^{-1} \exp\left(-\frac{1}{2\sigma^2}(\beta - m)^t M^{-1}(\beta - m)\right) \times \\ &\quad \frac{\alpha^r}{\Gamma(r)} (\sigma^2)^{-(r+1)} \exp\left(-\frac{\alpha}{\sigma^2}\right), \ \beta \in \mathbb{R}^p, \sigma^2 > 0. \end{aligned}$$

Wir schreiben $(\tilde{\beta}, \tilde{\sigma}^2) \sim \mathrm{NIG}(m, M, \alpha, r)$.

Lassen wir Faktoren in der gemeinsamen Dichte fort, die weder von β *noch von* σ^2 *abhängen, so erhalten wir*

$$f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta, \sigma^2) \propto \sigma^{-p} \exp\left(-\frac{1}{2\sigma^2}(\beta - m)^t M^{-1}(\beta - m)\right)(\sigma^2)^{-(r+1)} \exp\left(-\frac{\alpha}{\sigma^2}\right)$$

$$= (\sigma^2)^{-\frac{p}{2}-r-1} \exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}(\beta - m)^t M^{-1}(\beta - m) - \alpha\right]\right). \qquad (3.87.1)$$

Korollar 3.88

Nehmen wir $(\tilde{\beta}, \tilde{\sigma}^2) \sim \mathrm{NIG}(m, M, \alpha, r)$ *an, so gilt offenbar*

$$\mathbb{E}\big[\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2\big] = m,\ \mathit{Cov}\big(\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2\big) = \sigma^2 M,$$

$$\mathbb{E}\left[\tilde{\sigma}^2\right] = \frac{\alpha}{r-1} \quad \textit{falls } r > 1 \textit{ und}$$

$$\mathit{Var}\left(\tilde{\sigma}^2\right) = \frac{\alpha^2}{(r-1)^2(r-2)} \quad \textit{falls } r > 2.$$

Ferner gilt unbedingt für $r > 1$*:*

$$\begin{aligned}
\mathbb{E}\big[\tilde{\beta}\big] &= \mathbb{E}\left[\mathbb{E}\big[\tilde{\beta}|\tilde{\sigma}^2\big]\right] = m \text{ und} \\
\mathit{Cov}\big(\tilde{\beta}\big) &= \mathbb{E}\big[\mathit{Cov}\big(\tilde{\beta}|\tilde{\sigma}^2\big)\big] + \mathit{Cov}\big(\mathbb{E}\big[\tilde{\beta}|\tilde{\sigma}^2\big]\big) \\
&= \mathbb{E}\big[\tilde{\sigma}\big] M = \frac{\alpha}{r-1} M
\end{aligned}$$

nach Kovarianzzerlegungsformel, siehe z.B. (Ross, 2002, Seite 392).

Da ferner $f_{\tilde{\sigma}^2|\tilde{\beta}=\beta}(\sigma^2) \propto f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta, \sigma^2)$ *gilt, ist*

$$\tilde{\sigma}^2|\tilde{\beta} = \beta \sim \mathrm{IG}\left(\alpha + \frac{1}{2}(\beta - m)^t M^{-1}(\beta - m), r + \frac{p}{2}\right).$$

Schließlich nutzen wir zur Berechnung der unbedingten (Rand-)Verteilung von $\tilde{\beta}$ *aus, dass wegen der Normierungsbedingung der* $\mathrm{IG}\left(\alpha + \frac{1}{2}(\beta - m)^t M^{-1}(\beta - m), r + \frac{p}{2}\right)$*-Verteilung gilt:*

$$\int_0^\infty (\sigma^2)^{-\frac{p}{2}-r-1} \exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}(\beta - m)^t M^{-1}(\beta - m) - \alpha\right]\right) d\sigma^2$$

$$= \Gamma(r + \frac{p}{2}) \left\{\alpha + \frac{1}{2}(\beta - m)^t M^{-1}(\beta - m)\right\}^{-r-\frac{p}{2}}.$$

Damit folgt aus [3.87.1] (gemeinsame Dichte bezüglich σ^2 *ausintegriert), dass*

$$\begin{aligned}
f_{\tilde{\beta}}(\beta) &\propto \Gamma(r + \frac{p}{2}) \left\{\alpha + \frac{1}{2}(\beta - m)^t M^{-1}(\beta - m)\right\}^{-r-\frac{p}{2}} \\
&\propto \left\{1 + \frac{1}{2r}(\beta - m)^t \left[\frac{\alpha}{r} M\right]^{-1} (\beta - m)\right\}^{-r-\frac{p}{2}}.
\end{aligned}$$

Dies entspricht der λ^p*-Dichte einer multivariaten t-Verteilung mit* $2r$ *Freiheitsgraden, Lokationsparameter* m *und Dispersionsparameter* $\alpha M/r$*, also* $\tilde{\beta} \sim t(2r, m, \alpha/r \cdot M)$.

Anmerkung:

Die Varianz-Kovarianz-Matrix einer multivariaten t-Vereilung mit Dispersionsparameter Σ und ν Freiheitsgraden ist gegeben durch $\frac{\nu}{\nu-2}\Sigma$, falls $\nu > 2$.

Satz 3.89

Die Familie der Normal-inversen Gammaverteilungen für die Parameter $(\tilde{\beta}, \tilde{\sigma}^2)$ ist konjugiert zur Familie der Normalverteilungen für die Likelihood der Response Y im klassischen ANCOVA-Modell 3.14.

Genauer gilt:

Falls $Y|\tilde{\beta} = \beta, \tilde{\sigma}^2 = \sigma^2 \sim \mathcal{N}_n\left(X\beta, \sigma^2 I_n\right)$ *und* $(\tilde{\beta}, \tilde{\sigma}^2) \sim \mathrm{NIG}(m, M, \alpha, r)$*, so ist*

$$(\tilde{\beta}, \tilde{\sigma}^2)|Y = y \sim \mathrm{NIG}(m^*, M^*, \alpha^*, r^*) \text{ mit}$$

$$M^* = (X^tX + M^{-1})^{-1}, \qquad m^* = M^*(M^{-1}m + X^ty),$$

$$r^* = r + \frac{n}{2}, \qquad \alpha^* = \alpha + \frac{1}{2}\left(y^ty + m^tM^{-1}m - (m^*)^t(M^*)^{-1}m^*\right).$$

Beweis:

$$\begin{aligned} f_{(\tilde{\beta},\tilde{\sigma}^2)|Y=y}(\beta, \sigma^2) &\propto f_{(\tilde{\beta},\tilde{\sigma}^2)}(\beta, \sigma^2)\, l\left((\beta, \sigma^2), y\right) \\ &\propto (\sigma^2)^{-\frac{p}{2}-r-1} \exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}(\beta - m)^tM^{-1}(\beta - m) - \alpha\right]\right) \times \\ &\qquad \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta)\right). \end{aligned}$$

Dieser Ausdruck kann durch algebraische Manipulationen in die gewünschte Form gebracht werden (siehe Übungsaufgabe). ■

Korollar 3.90

Unter den Voraussetzungen von Satz 3.89 gilt:

(i) $\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2, Y = y \sim \mathcal{N}_p\left(\mu_\beta, \Sigma_\beta\right)$ *mit*

$$\begin{aligned} \Sigma_\beta &= \left(\frac{1}{\sigma^2}X^tX + \frac{1}{\sigma^2}M^{-1}\right)^{-1} \text{ und} \\ \mu_\beta &= \Sigma_\beta\left(\frac{1}{\sigma^2}X^ty + \frac{1}{\sigma^2}M^{-1}m\right). \end{aligned}$$

(ii) $\tilde{\sigma}^2|\tilde{\beta} = \beta, Y = y \sim \mathrm{IG}(\alpha', r')$ *mit*

$$\begin{aligned} \alpha' &= \alpha + \frac{1}{2}(y - X\beta)^t(y - X\beta) + \frac{1}{2}(\beta - m)^tM^{-1}(\beta - m) \text{ und} \\ r' &= r + \frac{n+p}{2}. \end{aligned}$$

Beweis: Aus Satz 3.89 ist bekannt, dass $(\tilde{\beta}, \tilde{\sigma}^2)|Y = y \sim \mathrm{NIG}(m^*, M^*, \alpha^*, r^*)$. Aus Definition 3.87 und Korollar 3.88 haben wir zudem die folgenden Charakterisierungen von $\mathrm{NIG}(m^*, M^*, \alpha^*, r^*)$ gewonnen:

(a) $\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2, Y = y \sim \mathcal{N}_p\left(m^*, \sigma^2 M^*\right)$

(b) $\tilde{\sigma}^2|\tilde{\beta} = \beta, Y = y \sim \mathrm{IG}\left(\alpha^* + \frac{1}{2}(\beta - m^*)^t (M^*)^{-1}(\beta - m^*), r^* + p/2\right)$

Es bleibt, die Parameter zu identifizieren:

(i)

$$\begin{aligned} \sigma^2 M^* = \sigma^2\left(X^t X + M^{-1}\right)^{-1} &= \left(\sigma^{-2} X^t X + \sigma^{-2} M^{-1}\right)^{-1} = \Sigma_\beta. \\ m^* = M^*\left(M^{-1} m + X^t y\right) &= \sigma^{-2}\Sigma_\beta\left(M^{-1} m + X^t y\right) \\ &= \Sigma_\beta\left(\sigma^{-2} M^{-1} m + \sigma^{-2} X^t y\right) = \mu_\beta. \end{aligned}$$

(ii) Leicht errechnen wir $r^* + p/2 = r + n/2 + p/2 = r + \frac{n+p}{2} = r'$. Ferner ergibt sich

$$\begin{aligned} \alpha^* + (\beta - m^*)^t (M^*)^{-1}(\beta - m^*)/2 &= \alpha + \left(y^t y + m^t M^{-1} m - (m^*)^t (M^*)^{-1} m^*\right)/2 \\ &\quad + \frac{1}{2}(\beta - m^*)^t (M^*)^{-1}(\beta - m^*) =: \alpha''. \end{aligned}$$

Wir vergleichen:
$2(\alpha'' - \alpha') = y^t y + m^t M^{-1} m - (m^*)^t (M^*)^{-1} m^* + (\beta - m^*)^t (M^*)^{-1}(\beta - m^*)$
$-(y - X\beta)^t (y - X\beta) - (\beta - m)^t M^{-1}(\beta - m) = 0.$

Beachte zu letzter Gleichheit:

$$\begin{aligned} (M^*)^{-1} m^* &= M^{-1} m + X^t y, \\ (m^*)^t (M^*)^{-1} &= \left(M^{-1} m + X^t y\right)^t, \end{aligned}$$

da $(M^*)^{-1}$ symmetrisch ist. ■

Bemerkung 3.91

Aus Korollar 1.9 ist bekannt, dass unter quadratischem Verlust die bedingte Erwartung $\mathbb{E}\left[\tilde{\beta}|Y\right]$ Bayes-optimaler (Punkt-)Schätzer für β ist. Wir erhalten

$$\hat{\beta}_{Bayes} = \mathbb{E}\left[\tilde{\beta}|Y\right] = (X^t X + M^{-1})^{-1}(M^{-1} m + X^t Y).$$

Definieren wir die Matrix A durch $A := \left(X^t X + M^{-1}\right)^{-1} X^t X$, so gilt damit
$\hat{\beta}_{Bayes} = (I_p - A)m + A\hat{\beta}$, wobei $\hat{\beta} = (X^t X)^{-1} X^t Y$ der Kleinste Quadrate Schätzer für β ist.

Die erarbeiteten Ansätze zur Bayesianischen Inferenz im klassischen ANCOVA-Modell lassen sich konzeptionell auch auf verallgemeinerte lineare Modelle übertragen.

Definition 3.92

Sei $l(\beta, y) = \prod_{i=1}^{n} l(\beta, y_i)$ die (gemeinsame) Likelihoodfunktion eines verallgemeinerten linearen Modells, wobei wir die Abhängigkeit von den Kovariablen zur notationellen Vereinfachung unterdrücken. Sei ferner $f_{\tilde{\beta}}$ eine a priori-Dichte bezüglich des Lebesguemaßes λ^p auf $\mathbb{R}^p$.
Dann heißt

(i) $f_{\tilde{\beta}|Y=y}$*, gegeben durch*

$$f_{\tilde{\beta}|Y=y}(\beta) = \frac{f_{\tilde{\beta}}(\beta) l(\beta, y)}{\int_{\mathbb{R}^p} f_{\tilde{\beta}}(\beta) l(\beta, y) d\beta} \propto f_{\tilde{\beta}}(\beta) l(\beta, y)$$

a posteriori-Dichte von $\tilde{\beta}$ (bezüglich λ^p) gegeben die beobachteten Daten $Y = y$.

(ii) $\mathbb{E}\big[\tilde{\beta}|Y = y\big] = \int_{\mathbb{R}^p} \beta f_{\tilde{\beta}|Y=y}(\beta) d\beta$ *a posteriori-Erwartungswert von $\tilde{\beta}$.*

(iii) $Cov\big(\tilde{\beta}|Y = y\big) = \int_{\mathbb{R}^p} \big(\beta - \mathbb{E}\big[\tilde{\beta}|Y = y\big]\big)\big(\beta - \mathbb{E}\big[\tilde{\beta}|Y = y\big]\big)^t f_{\tilde{\beta}|Y=y}(\beta) d\beta$ *a posteriori-Kovarianzmatrix von $\tilde{\beta}$ (gegeben $Y = y$).*

(iv) $\hat{\beta}_{post.} := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} f_{\tilde{\beta}|Y=y}(\beta) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} \Big\{\ln(f_{\tilde{\beta}}(\beta)) + \ln(l(\beta, y))\Big\}$ *a posteriori-Modus-Schätzer bzw. maximum a posteriori (MAP)-Schätzer für β.*

Anmerkung:
In der Praxis ist die Berechnung des a posteriori-Erwartungswertes $\mathbb{E}\big[\tilde{\beta}|Y = y\big]$ sowie der a posteriori-Kovarianzmatrix $\mathrm{Cov}\big(\tilde{\beta}|Y = y\big)$ nur in wenigen Spezialfällen analytisch möglich. Numerische Integrationen im $\mathbb{R}^p$ sind nur für kleine oder moderate Dimensionen p verlässlich bzw. stabil. Gerne zieht man sich daher auf MAP-Schätzer zurück.

Satz und Definition 3.93 (Ridge-Schätzer)
Wählen wir unter den Gegebenheiten von Definition 3.92 eine a priori-Normalverteilung, d.h., $\tilde{\beta} \sim \mathcal{N}_p(m, M)$, so erhalten wir

(a)

$$\begin{aligned} \hat{\beta}_{post.} &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} \left\{\ln(l(\beta, y)) - \frac{1}{2}(\beta - m)^t M^{-1}(\beta - m)\right\} \\ &=: \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} \ln(l_{post.}(\beta, y)). \end{aligned}$$

Damit kann die logarithmische a posteriori-Dichte $\ln(l_{post.}(\cdot, y))$ auch als penalisierte Loglikelihoodfunktion aufgefasst werden. Der Strafterm $(\beta - m)^t M^{-1}(\beta - m)$ penalisiert dabei Abweichungen vom a priori-Erwartungswert m.

(b) Wählen wir speziell $m = 0$ und $M = \tau^2 I_p$, so erhalten wir $\hat{\beta}_{post.}$ als den sogenannten Ridge-Schätzer mit Shrinkage-(Schrumpfungs-)Parameter $\lambda := [2\tau^2]^{-1}$ und der Strafterm vereinfacht sich zu

$$\lambda \cdot (\beta^t \beta) = \lambda \|\beta\|_2^2.$$

(c) Mit der penalisierten Fisher-Informationsmatrix $F_{post.}(\beta)$, gegeben durch

$$(F_{post.}(\beta))_{i,j} = -\mathbb{E}\left[\frac{\partial^2 \ln(l_{post.}(\beta, y))}{\partial\beta_i \partial\beta_j}\right],\ 1 \leq i, j \leq p,$$

gilt für $n \to \infty$:

$$\hat{\beta}_{post.}(n) \underset{as.}{\sim} \mathcal{N}_p\left(\beta, F^{-1}_{post.}(\hat{\beta}_{post.})\right),$$

wobei $F_{post.}$ (vermittels $l(\beta, y)$) ebenfalls vom Stichprobenumfang n abhängt.

Beweis: Abschnitt 4.6 in Fahrmeir et al. (2009). ■

Bemerkung 3.94

(i) Das Konzept der penalisierten Likelihood-Inferenz kann bedeutend verallgemeinert werden und stellt einen Schwerpunkt moderner Forschung im Bereich der mathematischen Statistik dar.

(ii) Ein allgemeiner Zugang zur Bayesianischen Inferenz in komplizierten Modellen ist gegeben durch sogenannte Markov Chain Monte Carlo (MCMC)-Verfahren, die Algorithmen liefern, um Pseudo-Stichproben aus den entsprechenden a posteriori-Verteilungen am Computer zu generieren.

Anwendung 3.95 (Markoffketten und Markov Chain Monte Carlo-Verfahren)
Siehe auch die entsprechende Präsentation (mit R-Beispielen).

Notizen zum Thema „Markoffketten auf endlichen Zustandsräumen“
Sei $(\Omega, 2^\Omega)$ ein Zustandsraum mit $|\Omega| = m$. Der Übergangskern $\mathcal{K}(x, \cdot)$ ist Wahrscheinlichkeitsmaß auf $\mathcal{F} = 2^\Omega$, d.h. $\mathcal{K}(x, \cdot) : \mathcal{F} \to [0, 1]$. Falls (X_n) zeithomogen ist, ist $\mathcal{K}(x, \cdot)$ bereits vollständig spezifiziert durch Angabe aller $\mathcal{K}(x, y)$, $x, y \in \Omega$. Also kann $\mathcal{K}$ beschrieben werden durch eine $(m \times m)$-Matrix P mit $P(x, y) = \mathbb{P}\big(X_1 = y | X_0 = y\big)$.

Beispiel: $m = 4$ *und*
$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

a) Ist P rekurrent? Ist P irreduzibel?
$(1,1) : (1) - (2) - (1)$ $\surd$
$(1,2) : (1) - (2)$ $\surd$
$(1,3) : (1) - (2) - (3)$ $\surd$
$(1,4) : (1) - (2) - (3) - (4)$ $\surd$

$(2,1): \surd$
$(2,2): (2)-(3)-(2)\ \surd$
$(2,3): \surd$
$(2,4): (2)-(3)-(4)\ \surd$
$(3,1): (3)-(2)-(1)\ \surd$
$(3,2):\ \surd$
$(3,3): (3)-(4)-(3)\ \surd$
$(3,4): \surd$
$(4,1): (4)-(3)-(2)-(1)\ \surd$
$(4,2): (4)-(3)-(2)\ \surd$
$(4,3): \surd$
$(4,4): (4)-(3)-(4)\ \surd$

b) Invariantes Maß (Startverteilung)

$$\mu \stackrel{!}{=} \mu P \Longleftrightarrow (\mu_1,\mu_2,\mu_3,\mu_4) = (\mu_1,\mu_2,\mu_3,\mu_4)\cdot \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Dies ist äquivalent zu dem linearen Gleichungssystem

$$\mu_1 = \frac{\mu_2}{3} \qquad [1]$$

$$\mu_2 = \mu_1 + \frac{2\mu_3}{3} \qquad [2]$$

$$\mu_3 = \frac{2\mu_2}{3} + \mu_4 \qquad [3]$$

$$\mu_4 = \frac{\mu_3}{3}. \qquad [4]$$

$$[1] \text{ in } [2]\text{: } \mu_2 = \frac{\mu_2}{3} + \frac{2\mu_3}{3} \iff \mu_2 = \mu_3. \qquad [5]$$

[5] *und* [4] *und* [1] $\Longrightarrow \mu_1 = \mu_4 = \frac{\mu_2}{3}$.

$$\begin{aligned} \sum_{i=1}^{4} \mu_i \equiv 1 \quad &\Longrightarrow \quad 2\mu_2 + \frac{2}{3}\mu_2 = 1 \\ &\Longleftrightarrow \quad \frac{8}{3}\mu_2 = 1 \iff \mu_2 = \frac{3}{8} \\ &\Longrightarrow \quad (\mu_1,\mu_2,\mu_3,\mu_4) = \left(\frac{1}{8},\frac{3}{8},\frac{3}{8},\frac{1}{8}\right). \end{aligned}$$

Probe:

$$\left(\frac{1}{8},\frac{3}{8},\frac{3}{8},\frac{1}{8}\right)\begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} = \left(\frac{1}{8},\frac{3}{8},\frac{3}{8},\frac{1}{8}\right) \quad \surd$$

Aber: Diese Kette ist periodisch (mit Periode 2). Daher findet Konvergenz gegen das invariante Maß nicht für jedes $\mu^{(0)}$ statt! Es ist dies eine Ehrenfest-Kette mit $N = 3$, allgemein gegeben durch

$$\begin{aligned} P(x,y) &= \frac{x-1}{N}, y = x-1, \\ P(x,y) &= \frac{N-x+1}{N}, y = x+1, \\ P(x,y) &= 0, \textit{ sonst.} \end{aligned}$$

Neue Kette ($m = 3$):

$$P = \begin{pmatrix} \frac{4}{10} & \frac{4}{10} & \frac{2}{10} \\ \frac{3}{10} & \frac{4}{10} & \frac{3}{10} \\ \frac{2}{10} & \frac{4}{10} & \frac{4}{10} \end{pmatrix}$$

P rekurrent und irreduzibel ($\surd$), P ist aperiodisch ($\surd$).

$$\mu^* = \left(\frac{3}{10},\frac{4}{10},\frac{3}{10}\right), \textit{ denn}$$

$$\left(\frac{3}{10},\frac{4}{10},\frac{3}{10}\right)\begin{pmatrix} \frac{4}{10} & \frac{4}{10} & \frac{2}{10} \\ \frac{3}{10} & \frac{4}{10} & \frac{3}{10} \\ \frac{2}{10} & \frac{4}{10} & \frac{4}{10} \end{pmatrix} = \left(\frac{3}{10},\frac{4}{10},\frac{3}{10}\right) \cdot \surd$$

Definition 3.96 (Doppelexponentialverteilung)
Die Doppelexponentialverteilung (auch: Laplace-Verteilung) mit Skalenparameter $\lambda > 0$ ist eine stetige Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit Lebesguedichte f_λ, gegeben durch

$$f_\lambda(t) = \frac{\lambda}{2}\exp(-\lambda|t|), t \in \mathbb{R}.$$

Satz 3.97
Wählen wir unter den Gegebenheiten von Definition 3.92 die a priori-Dichte als $f_{\tilde{\beta}}(\beta) = \prod_{j=1}^{p} f_\lambda(\beta_j)$, nehmen wir also a priori stochastisch unabhängige, identisch Laplace(λ)-verteilte Regressionskoeffizienten an, so ist der resultierende MAP-Schätzer für β ein L_1-Norm penalisierter Maximum-Likelihood-Schätzer.

Beweis: Für die a posteriori-Dichte von $\tilde{\beta}$ gilt

$$f_{\tilde{\beta}|Y=y}(\beta) \propto l(\beta, y) \cdot \prod_{j=1}^{p} \exp(-\lambda|\beta_j|).$$

$$\begin{aligned}
\text{Damit ist } \hat{\beta}_{post.} &= \arg\max_{\beta\in\mathbb{R}^p} f_{\tilde{\beta}|Y=y}(\beta) \\
&= \arg\max_{\beta\in\mathbb{R}^p} \{\ln\left(l(\beta, y)\right) - \lambda \sum_{j=1}^{p} |\beta_j|\} \\
&= \arg\max_{\beta\in\mathbb{R}^p} \{\ln\left(l(\beta, y)\right) - \lambda \cdot \|\beta\|_1\}.
\end{aligned}$$

■

Bemerkung 3.98

(a) Es gilt äquivalent

$$\hat{\beta}_{post.} = \arg\min_{\beta\in\mathbb{R}^p} \{-\ln\left(l(\beta, y)\right) + \lambda \|\beta\|_1\}.$$

Im Falle normalverteilter Responsevariablen ist $-\ln\left(l(\beta, y)\right)$ isotone Transformation der Fehlerquadratsumme.

(b) Im klassischen ANCOVA-Modell mit unbekannter Fehlervarianz σ^2 bietet es sich wiederum an, die bedingte a priori-Verteilung $\mathcal{L}(\tilde{\beta}|\tilde{\sigma}^2 = \sigma^2)$ als $[\text{Laplace}(\lambda/\sigma)]^p$ zu wählen, da dann für jede a priori-inverse Gammaverteilung von $\tilde{\sigma}^2$ sichergestellt ist, dass die a posteriori-Dichte unimodal ist (siehe Park and Casella (2008)).

(c) Der Skalierungs- bzw. Regularisierungsparameter λ kann entweder (z.B. durch Kreuzvalidierung oder marginale Likelihoodmaximierung) explizit gewählt werden oder es wird eine zusätzliche Hierarchiestufe in die Bayesianische Modellierung eingeführt, indem eine Hyper-a priori-Verteilung für $\tilde{\lambda}$ gewählt wird. Park and Casella (2008) empfehlen eine Gammaverteilung für $\tilde{\lambda}^2$.

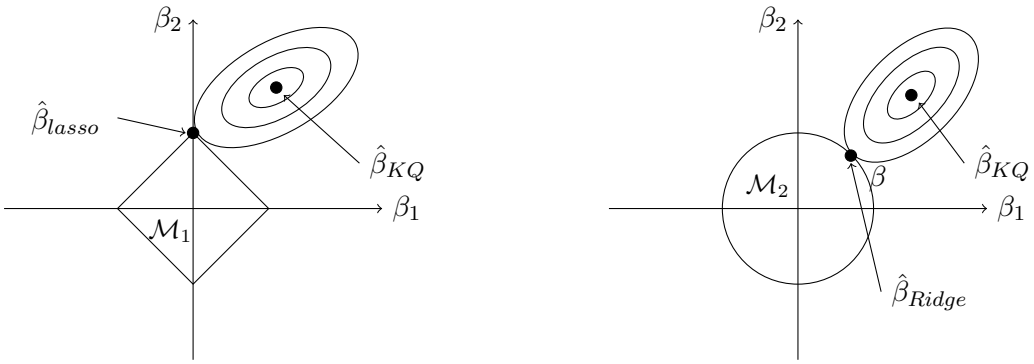
(d) Unter den Gegebenheiten von (b), also klassischer ANCOVA, entspricht $\hat{\beta}_{post.}$ aus Satz 3.97 dem lasso (least absolute shrinkage and selection operator)-Schätzer aus Tibshirani (1996).

(e) L_1-Regularisierung führt im Gegensatz zur L_2-Regularisierung oft implizit zu einer Variablenselektion. Dazu schreiben wir die Zielkriterien der beiden Verfahren um:

$$\begin{aligned}
\hat{\beta}_{lasso} &= \arg\min_{\mathcal{M}_1} \{-\ln\left(l(\beta, y)\right)\}, \quad \mathcal{M}_1 := \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq C_1\} \\
\hat{\beta}_{Ridge} &= \arg\min_{\mathcal{M}_2} \{-\ln\left(l(\beta, y)\right)\}, \quad \mathcal{M}_2 := \{\beta \in \mathbb{R}^p : \|\beta\|_2^2 \leq C_2\},
\end{aligned}$$

wobei $C_1 \equiv C_1(\lambda)$ und $C_2 \equiv C_2(\lambda)$ Transformationen des jeweiligen Regularisierungsparameters sind.

Im ANCOVA-Falle ist $-\ln(l(\beta, y))$ *äquivalent zur Fehlerquadratsumme* $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2$*, deren Konturlinien für* $p = 2$ *bezüglich* $\beta \in \mathbb{R}^2$ *Ellipsen sind.*
Wir erhalten die folgenden beiden Schaubilder (adaptiert nach Tibshirani (1996)).



(f) Es existieren eine Reihe verallgemeinerter Regularisierungstechniken, so zum Beispiel Bridge-Regression (L_q*-Norm Regularisierung mit allgemeinem* $q \geq 0$*) oder „elastic net“ (Straftterm ist gewichtetes Mittel aus* L_1*- und* L_2*-Norm Regularisierungstermen). Die Herleitung von (asymptotischen) Verteilungsaussagen für penalisierte ML- bzw. KQ-Schätzer ist Gegenstand aktueller Forschung.*

Kapitel 4

Das Statistik-Softwaresystem R

Siehe Präsentationen aus der ersten Übungswoche.

Tabellenverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

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.

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.

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.

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.

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.

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.

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.

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.