Statistik II

Universität Ulm
Abteilung Stochastik

Vorlesungsskript

Prof. Dr. Volker Schmidt

Stand: Wintersemester 2005/06

Ulm, im Februar 2006

Inhaltsverzeichnis

Literatur

[1] Büning, H., Trenkler, G. (1994)
Nichtparametrische statistische Methoden
de Gruyter, Berlin

[2] Cressie, N.A. (1993)
Statistics for Spatial Data
J. Wiley & Sons, New York

[3] Dobson, A.J. (2002)
An Introduction to Generalized Linear Models
Chapman & Hall, Boca Raton

[4] Falk, M., Marohn, F., Tewes, B. (2002)
Foundations of Statistical Analyses and Applications with SAS
Birkhäuser, Basel

[5] Hastie, T., Tibshirami, R., Friedman, J. (2001)
The Elements of Statistical Learnning
Springer, New York

[6] Koch, K.R. (1997)
Parameterschätzung und Hypothesentests in linearen Modellen
Dümmlers–Verlag, Bonn

[7] Lehmann, E.L. (1999)
Elements of Large–Sample Theory
Springer, New York

[8] Lehmann, E.L., Romano, J.P. (2005)
Testing Statistical Hypotheses
Springer, New York

[9] McCullagh, P., Nelder, J.A. (1989)
Genralized Linear Models
Chapman & Hall, London.

[10] Pruscha, H. (2000)
Vorlesungen über mathematische Statistik
Teubner–Verlag, Stuttgart.

[11] Van der Vaart, A., Wellner, J. (1996)
Weak Convergence and Empirical Processes
Springer–Verlag, New York

[12] Vapnik, V.N. (1998)
Statistical Learning Theory
J. Wiley & Sons, New York

1 Einleitung und Grundlagen

Die Vorlesung „Statistik II" ist für Studierende konzipiert, die bereits über Grundkenntnisse auf dem Gebiet der mathematischen Statistik verfügen. Die Schätz– und Testverfahren, die in „Statistik I" behandelt worden sind, werden dabei als bekannt vorausgesetzt.

Die Vorlesung „Statistik II" besteht aus den Teilen:

- multivariate Normalverteilung (reguläre und singuläre Normalverteilung, lineare und quadratische Formen)
- lineare Modelle (multiple Regression, normalverteilte Störgrößen, ein– und mehrfaktorielle Varianzanalyse)
- verallgemeinerte lineare Modelle (logistische Regression, Maximum–Likelihood–Gleichung, gewichteter KQ–Schätzer, Bewertung der Anpassungsgüte)
- Tests von Verteilungsannahmen (Kolmogorow–Smirnow–Test, χ^2–Anpassungstests von Pearson–Fisher)
- Nichtparametrische Lokalisationstests (Binomialtest, Iterationstests, lineare Rangtests)

Dabei werden wir insbesondere Begriffe und Ergebnisse nutzen, die in den Vorlesungen „Wahrscheinlichkeitsrechnung" bzw. „Statistik I" eingeführt worden sind, vgl. das Skript zur Vorlesung „Wahrscheinlichkeitsrechnung" im Wintersemester 2004/05:

http://www.mathematik.uni-ulm.de/stochastik/lehre/ws03_04/wr/skript/skript.html

bzw. das Skript zur Vorlesung „Statistik I" im Sommersemester 2005:

http://www.mathematik.uni-ulm.de/stochastik/lehre/ss04/statistik1/skript/skript.html

Verweise auf diese Vorlesungsmanuskripte werden wir mit dem Zusatz „WR" bzw. „I" vor der Nummer der zitierten Abschnitte, Lemmata, Theoreme, Korollare bzw. Formeln kennzeichnen.

1.1 Einige Grundbegriffe und Ergebnisse der Matrix–Algebra

Wir erinnern zunächst an einige grundlegende Begriffe und Ergebnisse der Matrix–Algebra, die im folgenden benötigt werden, vgl. hierzu auch das Skript zur Vorlesung „ Lineare Algebra" von W. Lütkebohmert (Universität Ulm, Wintersemester 2005/06)

1.1.1 Spur und Rang

- Die *Spur* $\mathrm{sp}(\mathbf{A})$ einer quadratischen $n \times n$ Matrix $\mathbf{A} = (a_{ij})$ ist gegeben durch

$$\mathrm{sp}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} \, . \tag{1}$$

- Sei $\mathbf{A}$ eine beliebige $n \times m$ Matrix. Der *Rang* $\mathrm{rg}(\mathbf{A})$ ist die maximale Anzahl der linear unabhängigen Zeilen (bzw. Spalten) von $\mathbf{A}$.
  - Dabei heißen die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_\ell \in \mathbb{R}^m$ *linear abhängig*, wenn es reelle Zahlen $c_1, \ldots, c_\ell \in \mathbb{R}$ gibt, die nicht alle gleich Null sind, so dass $c_1\mathbf{a}_1 + \ldots + c_\ell\mathbf{a}_\ell = \mathbf{o}$.
  - Anderenfalls heißen die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_\ell \in \mathbb{R}^m$ *linear unabhängig*.

Unmittelbar aus der Definitionsgleichung (1) der Matix–Spur und aus der Definition der Matrix–Multiplikation ergibt sich der folgende Hilfssatz.

Lemma 1.1 *Sei* $\mathbf{C}$ *eine beliebige* $n\times m$ *Matrix und* $\mathbf{D}$ *eine beliebige* $m\times n$ *Matrix. Dann gilt* $\mathrm{sp}(\mathbf{CD}) = \mathrm{sp}(\mathbf{DC})$.

Man kann zeigen, dass eine quadratische Matrix $\mathbf{A}$ genau dann invertierbar ist, wenn $\mathbf{A}$ vollen Rang hat bzw. wenn $\det \mathbf{A} \neq 0$ gilt. In diesem Zusammenhang ist auch das folgende Rsultat nützlich.

Lemma 1.2 *Sei* $\mathbf{A}$ *eine* $n \times m$ *Matrix mit* $n \geq m$ *und* $\mathrm{rg}(\mathbf{A}) = m$. *Dann gilt* $\mathrm{rg}(\mathbf{A}^\top \mathbf{A}) = m$.

Beweis

- Es ist klar, dass der Rang $\mathrm{rg}(\mathbf{A}^\top\mathbf{A})$ der $m \times m$ Matrix $\mathbf{A}^\top\mathbf{A}$ nicht größer als m sein kann.
- Wir nehmen nun an, dass $\mathrm{rg}(\mathbf{A}^\top\mathbf{A}) < m$. Dann gibt es einen Vektor $\mathbf{c} = (c_1, \ldots, c_m)^\top \in \mathbb{R}^m$, so dass $\mathbf{c} \neq \mathbf{o}$ und $\mathbf{A}^\top\mathbf{A}\mathbf{c} = \mathbf{o}$.
- Hieraus folgt, dass auch $\mathbf{c}^\top\mathbf{A}^\top\mathbf{A}\mathbf{c} = \mathbf{o}$ bzw. $(\mathbf{A}\mathbf{c})^\top(\mathbf{A}\mathbf{c}) = \mathbf{o}$, d.h. $\mathbf{A}\mathbf{c} = \mathbf{o}$.
- Dies ist jedoch ein Widerspruch zu der Voraussetzung, dass $\mathrm{rg}(\mathbf{A}) = m$. $\square$

Außerdem kann man zeigen, dass die beiden folgenden Eigenschaften von Spur bzw. Rang gelten.

Lemma 1.3 *Seien* $\mathbf{A}$ *und* $\mathbf{B}$ *beliebige* $n \times n$ *Matrizen. Dann gilt stets* $\mathrm{sp}(\mathbf{A} - \mathbf{B}) = \mathrm{sp}(\mathbf{A}) - \mathrm{sp}(\mathbf{B})$. *Wenn* $\mathbf{A}$ *idempotent und symmetrisch ist, d.h.,* $\mathbf{A} = \mathbf{A}^2$ *und* $\mathbf{A} = \mathbf{A}^\top$, *dann gilt außerdem* $\mathrm{sp}(\mathbf{A}) = \mathrm{rg}(\mathbf{A})$.

1.1.2 Eigenwerte und Eigenvektoren

Definition Sei $\mathbf{A}$ eine beliebige $n \times n$ Matrix. Jede (komplexe) Zahl $\lambda \in \mathbb{C}$, für die es einen Vektor $\mathbf{x} \in \mathbb{C}^n$ mit $\mathbf{x} \neq \mathbf{o}$ gibt, so dass

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{o}\,, \tag{2}$$

heißt *Eigenwert* der Matrix $\mathbf{A}$. Außerdem sagt man dann, dass $\mathbf{x}$ ein zu λ gehörender *Eigenvektor* ist.

Beachte

- Die Gleichung (2) hat nur für solche $\lambda \in \mathbb{C}$ eine Lösung $\mathbf{x} \in \mathbb{C}^n$ mit $\mathbf{x} \neq \mathbf{o}$, für die λ eine Lösung der so genannten *charakteristischen Polynomgleichung*

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \tag{3}$$

  ist, wobei die linke Seite $P(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$ von (3) das *charakteristische Polynom* der Matrix $\mathbf{A}$ genannt wird.
- Seien $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ die reellwertigen Lösungen von (3). Dann lässt sich das charakteristische Polynom $P(\lambda)$ in der Form

$$P(\lambda) = (-1)^n(\lambda - \lambda_1)^{a_1} \ldots (\lambda - \lambda_k)^{a_k} q(\lambda) \tag{4}$$

  darstellen, wobei $a_1, \ldots, a_k \in \mathbb{N}$ positive natürliche Zahlen sind, genannt die *algebraischen Vielfachheiten* von $\lambda_1, \ldots, \lambda_k$, und $q(\lambda)$ ein Polynom der Ordnung $n - \sum_{i=1}^k a_i$ ist, das keine reellen Lösungen besitzt.

Lemma 1.4 *Sei* $\mathbf{A} = (a_{ij})$ *eine symmetrische* $n\times n$ *Matrix mit reellwertigen Einträgen* a_{ij}. *Dann sind sämtliche Eigenwerte reell, und die zu verschiedenen Eigenwerten* $\lambda_i, \lambda_j \in \mathbb{R}$ *gehörenden Eigenvektoren* $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$ *sind zueinander orthogonal.*

Beweis

- Die Determinante $\det(\mathbf{A}-\lambda\mathbf{I})$ in (3) ist gegeben durch
$$\det(\mathbf{A}-\lambda\mathbf{I}) = \sum_{\boldsymbol{\pi}} (-1)^{r(\boldsymbol{\pi})} \prod_{i:\, i\neq\pi_i} a_{i\pi_i} \prod_{i:\, i=\pi_i} (a_{i\pi_i}-\lambda)\,, \tag{5}$$
wobei sich die Summation über alle $m!$ Permutationen $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_m)$ der natürlichen Zahlen $1,\ldots,m$ erstreckt und $r(\boldsymbol{\pi})$ die Anzahl der Zahlenpaare in $\boldsymbol{\pi}$ ist, die sich nicht in der natürlichen Ordnung befinden.
- Weil die Elemente von $\mathbf{A}$ reelle Zahlen sind, ist für jede Lösung $\lambda = a + \mathrm{i}\, b$ von (3) gleichzeitig auch $\overline{\lambda} = a - \mathrm{i}\, b$ eine Lösung von (3).
- Seien $\mathbf{x} = \mathbf{a} + \mathrm{i}\,\mathbf{b}$ und $\overline{\mathbf{x}} = \mathbf{a} - \mathrm{i}\,\mathbf{b}$ Eigenvektoren, die zu λ bzw. $\overline{\lambda}$ gehören. Dann gilt $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ und $\mathbf{A}\overline{\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}}$ bzw.
$$\overline{\mathbf{x}}^\top\mathbf{A}\mathbf{x} = \overline{\mathbf{x}}^\top\lambda\mathbf{x} = \lambda\overline{\mathbf{x}}^\top\mathbf{x}$$
und
$$\overline{\mathbf{x}}^\top\mathbf{A}\mathbf{x} = \big(\mathbf{A}^\top\overline{\mathbf{x}}\big)^\top\mathbf{x} = \big(\mathbf{A}\overline{\mathbf{x}}\big)^\top\mathbf{x} = \big(\overline{\lambda}\overline{\mathbf{x}}\big)^\top\mathbf{x} = \overline{\lambda}\overline{\mathbf{x}}^\top\mathbf{x}\,.$$
- Hieraus folgt, dass $\lambda\overline{\mathbf{x}}^\top\mathbf{x} = \overline{\lambda}\overline{\mathbf{x}}^\top\mathbf{x}$.
- Weil $\overline{\mathbf{x}}^\top\mathbf{x} = |\mathbf{a}|^2 + |\mathbf{b}|^2 > 0$, ergibt sich hieraus, dass $\lambda = \overline{\lambda}$, d.h., λ ist eine reelle Zahl.
- Auf ähnliche Weise lässt sich zeigen, dass es zu verschiedenen Eigenwerten $\lambda_i, \lambda_j \in \mathbb{R}$ gehörende Eigenvektoren $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$ mit reellwertigen Komponenten gibt, die zueinander orthogonal sind.
- Weil die Matrix $\mathbf{A} - \lambda_i\mathbf{I}$ nur reellwertige Eintragungen hat, sind mit $\mathbf{x}_i$ auch $\overline{\mathbf{x}}_i$ bzw. $\mathbf{x}_i + \overline{\mathbf{x}}_i \in \mathbb{R}^n$ zu λ_i gehörende Eigenvektoren.
- Wir können (und werden) deshalb o.B.d.A. annehmen, dass $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$. Aus der Gültigkeit von
$$(\mathbf{A}-\lambda_i\mathbf{I})\mathbf{x}_i = \mathbf{o} \qquad \text{und} \qquad (\mathbf{A}-\lambda_j\mathbf{I})\mathbf{x}_j = \mathbf{o}$$
ergibt sich außerdem, dass $\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i$ und $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j$ bzw.
$$\mathbf{x}_j^\top\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_j^\top\mathbf{x}_i \qquad \text{und} \qquad \mathbf{x}_i^\top\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_i^\top\mathbf{x}_j\,.$$
- Andererseits gilt offenbar $\mathbf{x}_j^\top\mathbf{x}_i = \mathbf{x}_i^\top\mathbf{x}_j$, und aus der Symmetrie von $\mathbf{A} = (a_{ij})$ ergibt sich die Identität $\mathbf{x}_j^\top\mathbf{A}\mathbf{x}_i = \mathbf{x}_i^\top\mathbf{A}\mathbf{x}_j$, denn es gilt
$$\mathbf{x}_j^\top\mathbf{A}\mathbf{x}_i = \sum_{m=1}^n\sum_{\ell=1}^n x_{\ell j}a_{\ell m}x_{mi} = \sum_{\ell=1}^n\sum_{m=1}^n x_{mi}a_{m\ell}x_{\ell j} = \mathbf{x}_i^\top\mathbf{A}\mathbf{x}_j\,.$$
- Insgesamt ergibt sich somit, dass $\lambda_i\mathbf{x}_j^\top\mathbf{x}_i = \lambda_j\mathbf{x}_i^\top\mathbf{x}_j$ bzw. $(\lambda_i-\lambda_j)\mathbf{x}_j^\top\mathbf{x}_i = 0$.
- Wegen $\lambda_i - \lambda_j \neq 0$ folgt hieraus, dass $\mathbf{x}_j^\top\mathbf{x}_i = 0$. □

1.1.3 Diagonalisierungsverfahren

- Sei nun $\mathbf{A}$ eine invertierbare symmetrische $n\times n$ Matrix.
- In Lemma 1.4 haben wir gezeigt, dass dann sämtliche Eigenwerte $\lambda_1,\ldots,\lambda_n$ von $\mathbf{A}$ reelle Zahlen sind (wobei in dieser Folge gegebenenfalls einunddieselbe Zahl mehrfach auftreten kann).
- Wegen $\det\mathbf{A}\neq 0$ ist $\lambda = 0$ keine Lösung von (3), d.h., sämtliche Eigenwerte $\lambda_1,\ldots,\lambda_n$ von $\mathbf{A}$ sind von Null verschieden.

- Außerdem kann man zeigen, dass es *orthonormale* (Basis-) Vektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ gibt, d.h.
$$\mathbf{v}_i^\top \mathbf{v}_i = 1\,, \qquad \mathbf{v}_i^\top \mathbf{v}_j = 0\,, \qquad \forall\, i,j \in \{1,\ldots,n\} \text{ mit } i \neq j\,, \tag{6}$$
so dass $\mathbf{v}_i$ ein zu λ_i gehörender Eigenvektor ist; $i = 1, \ldots, n$.
- Wenn sämtliche Eigenwerte $\lambda_1, \ldots, \lambda_n$ voneinander verschieden sind, dann folgt dies unmittelbar aus Teilaussage 2 von Lemma 1.4.
- Hieraus resultiert das folgende *Diagonalisierungsverfahren* für invertierbare symmetrische Matrizen.

Lemma 1.5

- *Sei $\mathbf{A}$ eine invertierbare symmetrische $n \times n$ Matrix, und sei $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ die $n \times n$ Matrix, die aus den orthonormalen Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n$ besteht.*
- *Dann gilt*
$$\mathbf{V}^\top \mathbf{A}\mathbf{V} = \mathbf{\Lambda}\,, \tag{7}$$
wobei $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ die $n \times n$ Diagonalmatrix bezeichnet, die aus den Eigenwerten $\lambda_1, \ldots, \lambda_n$ gebildet wird.

Beweis

- Aus der Defintionsgleichung (2) von Eigenwerten bzw. -vektoren ergibt sich, dass $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ für jedes $i = 1, \ldots, n$.
- Hieraus folgt, dass $\mathbf{A}\mathbf{V} = (\lambda_1 \mathbf{v}_1, \ldots, \lambda_n \mathbf{v}_n)$ bzw. $\mathbf{V}^\top \mathbf{A}\mathbf{V} = \mathbf{V}^\top(\lambda_1 \mathbf{v}_1, \ldots, \lambda_n \mathbf{v}_n) = \mathbf{\Lambda}$, wobei sich die letzte Gleichheit aus (6) ergibt. □

1.1.4 Symmetrie und Definitheit; Faktorisierung

Lemma 1.6 *Sei $\mathbf{A}$ eine symmetrische und positiv definite $n \times n$ Matrix, d.h., es gelte $\mathbf{A} = \mathbf{A}^\top$ und $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$ für jeden Vektor $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ mit $\mathbf{x} \neq \mathbf{o}$. Dann ist $\mathbf{A}$ invertierbar, und es gibt es eine invertierbare $n \times n$ Matrix $\mathbf{H}$, so dass*
$$\mathbf{A} = \mathbf{H}\mathbf{H}^\top\,. \tag{8}$$

Beweis Wir zeigen nur die Gültigkeit der zweiten Teilaussage.

- Aus Lemma 1.5 ergibt sich, dass $\mathbf{V}^\top \mathbf{A}\mathbf{V} = \mathbf{\Lambda}$ bzw.
$$\mathbf{A} = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda} \mathbf{V}^{-1}\,, \tag{9}$$
  - wobei $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ die $n \times n$ Matrix ist, die aus den orthonormalen Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n$ besteht,
  - und $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ die $n \times n$ Diagonalmatrix bezeichnet, die aus den (positiven) Eigenwerten $\lambda_1, \ldots, \lambda_n$ gebildet wird.
- Sei nun $\mathbf{\Lambda}^{1/2}$ die $n \times n$ Diagonalmatrix $\mathbf{\Lambda}^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$, und sei
$$\mathbf{H} = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top\,. \tag{10}$$
- Es ist klar, dass die in (10) gegebene Matrix $\mathbf{H}$ invertierbar ist. Wegen $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$ gilt außerdem
$$\begin{aligned} \mathbf{H}\mathbf{H}^\top &= (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top \Big((\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top \Big)^\top = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top \mathbf{V} \mathbf{\Lambda}^{1/2} \mathbf{V}^{-1} \\ &= (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{\Lambda}^{1/2} \mathbf{V}^{-1} = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda} \mathbf{V}^{-1} = \mathbf{A}\,, \end{aligned}$$
wobei sich die letzte Gleichheit aus (9) ergibt. □

Beachte

- Jede invertierbare $n \times n$ Matrix $\mathbf{H}$ mit $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$ wird *Quadratwurzel* von $\mathbf{A}$ genannt und mit $\mathbf{A}^{1/2}$ bezeichnet.
- Mit Hilfe der *Cholesky–Zerlegung* für symmetrische und positiv definite Matrizen kann man zeigen, dass es eine (eindeutig bestimmte) untere Dreiecksmatrix $\mathbf{H}$ mit $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$ gibt; vgl. die Vorlesung „Numerik 1a" von K. Urban (Universítät Ulm, Sommersemester 2005).

Die folgende Eigenschaft symmetrischer Matrizen ist eine Verallgemeinerung von Lemma 1.6.

Lemma 1.7 *Sei* $\mathbf{A}$ *eine symmetrische und nichtnegativ definite* $n \times n$ *Matrix, d.h., es gelte* $\mathbf{A} = \mathbf{A}^\top$ *und* $\mathbf{x}^\top\mathbf{A}\mathbf{x} \ge 0$ *für jeden Vektor* $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$*. Sei nun* $\operatorname{rg}(\mathbf{A}) = r\ \big(\le n\big)$*. Dann gibt es eine* $n \times r$ *Matrix* $\mathbf{H}$ *mit* $\operatorname{rg}(\mathbf{H}) = r$*, so dass* $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$.

Der *Beweis* von Lemma 1.7 verläuft ähnlich wie der Beweis von Lemma 1.6.

Lemma 1.8

- *Seien* $m, r \in \mathbb{N}$ *beliebige natürliche Zahlen mit* $1 \le r \le m$*. Sei* $\mathbf{A}$ *eine symmetrische und positiv definite* $m \times m$ *Matrix, und sei* $\mathbf{B}$ *eine* $r \times m$ *Matrix mit vollem Rang* $\operatorname{rg}(\mathbf{B}) = r$.
- *Dann sind auch die Matrizen* $\mathbf{B}\mathbf{A}\mathbf{B}^\top$ *und* $\mathbf{A}^{-1}$ *positiv definit.*

Beweis

- Wegen des vollen Ranges von $\mathbf{B}^\top$ gilt $\mathbf{B}^\top\mathbf{x} \ne \mathbf{o}$ für jedes $\mathbf{x} \in \mathbb{R}^r$ mit $\mathbf{x} \ne \mathbf{o}$.
- Weil $\mathbf{A}$ positiv definit ist, gilt damit auch
$$\mathbf{x}^\top\big(\mathbf{B}\mathbf{A}\mathbf{B}^\top\big)\mathbf{x} = (\mathbf{B}^\top\mathbf{x})^\top\mathbf{A}(\mathbf{B}^\top\mathbf{x}) > 0$$
für jedes $\mathbf{x} \in \mathbb{R}^r$ mit $\mathbf{x} \ne \mathbf{o}$, d.h., $\mathbf{B}\mathbf{A}\mathbf{B}^\top$ ist positiv definit.
- Für $\mathbf{B} = \mathbf{A}^{-1}$ ergibt sich hieraus insbesondere, dass
$$\mathbf{A}^{-1} = \mathbf{A}^{-1}\big(\mathbf{A}\mathbf{A}^{-1}\big) = \mathbf{A}^{-1}\mathbf{A}\big(\mathbf{A}^{-1}\big)^\top$$
positiv definit ist. □

1.2 Multivariate Normalverteilung

In diesem Abschnitt erinnern wir an den Begriff der multivariaten Normalverteilung und diskutieren einige grundlegende Eigenschaften dieser Verteilungsfamilie.

1.2.1 Definition und grundlegende Eigenschaften

- Seien $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ unabhängige und (identisch) normalverteilte Zufallsvariablen, d.h. insbesondere, dass
$$X_i \sim \mathrm{N}(\mu, \sigma^2)\,, \qquad \forall i = 1, \ldots, n\,, \tag{11}$$
wobei $\mu \in \mathbb{R}$ und $\sigma^2 > 0$.

- In Vektor–Schreibweise bedeutet die Normalverteilungseigenschaft (11) und die Unabhängigkeit der Stichprobenvariablen, dass die Verteilung der Zufallsstichprobe $\mathbf{X} = (X_1, \ldots, X_n)^\top$ gegeben ist durch
$$\mathbf{X} \sim \mathrm{N}\big(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\big)\,, \tag{12}$$
wobei $\boldsymbol{\mu} = (\mu, \ldots, \mu)^\top$ und $\mathrm{N}\big(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\big)$ die n–dimensionale Normalverteilung mit Erwartungswertvektor $\boldsymbol{\mu}$ und Kovarianzmatrix $\sigma^2 \mathbf{I}_n$ bezeichnet.

- *Zur Erinnerung* (vgl. Abschnitt WR-4.3.4): Allgemein wird die n–dimensionale Normalverteilung wie folgt definiert.
  - Sei $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ ein beliebiger Vektor, und sei $\mathbf{K}$ eine symmetrische und positiv definite $n \times n$-Matrix.
  - Sei $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ ein absolutstetiger Zufallsvektor, wobei die gemeinsame Dichte von $\mathbf{Z}$ gegeben sei durch
  $$f(\mathbf{z}) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \, \frac{1}{\sqrt{\det \mathbf{K}}} \, \exp\Big(-\frac{1}{2}\,(\mathbf{z}-\boldsymbol{\mu})^\top \mathbf{K}^{-1} (\mathbf{z}-\boldsymbol{\mu})\Big) \tag{13}$$
  für jedes $\mathbf{z} = (z_1, \ldots, z_n)^\top \in \mathbb{R}^n$.
  - Dann sagt man, dass der Zufallsvektor $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ (regulär) *normalverteilt* ist.
  - Schreibweise: $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$

Wir zeigen nun, dass die in (13) gegebene Funktion eine (n–dimensionale) Wahrscheinlichkeitsdichte ist.

Theorem 1.1 *Sei $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ ein beliebiger Vektor, und sei $\mathbf{K}$ eine symmetrische und positiv definite $n \times n$-Matrix. Dann gilt*
$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}\,(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1} (\mathbf{x}-\boldsymbol{\mu})\Big)\, dx_1 \ldots dx_n = (2\pi)^{n/2}\, (\det \mathbf{K})^{1/2}\,. \tag{14}$$

Beweis

- Weil $\mathbf{K}$ symmetrisch und positiv definit (und damit auch invertierbar) ist, gibt es wegen Lemma 1.5 eine $n \times n$ Matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, die aus den orthonormalen Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_n$ von $\mathbf{K}$ besteht, so dass
$$\mathbf{V}^\top \mathbf{K} \mathbf{V} = \boldsymbol{\Lambda}\,, \tag{15}$$
wobei $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ die $n \times n$ Diagonalmatrix bezeichnet, die aus den Eigenwerten $\lambda_1, \ldots, \lambda_n$ von $\mathbf{K}$ gebildet wird.
- Außerdem ergibt sich aus der positiven Definitheit von $\mathbf{K}$, dass $\mathbf{v}_i^\top \mathbf{K} \mathbf{v}_i = \lambda_i > 0$ für jedes $i = 1, \ldots, n$, d.h., sämtliche Eigenwerte $\lambda_1, \ldots, \lambda_n$ von $\mathbf{K}$ sind positiv.
- Wegen $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$ gilt auch $\mathbf{V}^\top = \mathbf{V}^{-1}$ bzw. $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$.
- Weil außerdem $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ gilt, ergibt sich hieraus und aus (15), dass
$$\big(\mathbf{V}^\top \mathbf{K} \mathbf{V}\big)^{-1} = \mathbf{V}^\top \mathbf{K}^{-1} \mathbf{V} = \mathrm{diag}\big(\lambda_1^{-1}, \ldots, \lambda_n^{-1}\big)\,.$$
- Die Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ mit $\mathbf{y} = \varphi(\mathbf{x}) = \mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu})$, d.h. $\mathbf{x} - \boldsymbol{\mu} = \mathbf{V}\mathbf{y}$, bildet den $\mathbb{R}^n$ bijektiv auf sich selbst ab, und für die Jacobi-Determinante der Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ gilt
$$\det\Big(\frac{\partial \varphi_i}{\partial x_j}\,(x_1, \ldots, x_n)\Big) = \det \mathbf{V} = \pm 1\,,$$
wobei sich die letzte Gleichheit aus der Tatsache ergibt, dass $1 = \det(\mathbf{V}^\top \mathbf{V}) = (\det \mathbf{V})^2$.

- Für das Integral auf der linken Seite von (14) gilt somit, dass

$$\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\Bigl(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Bigr)\,dx_1\ldots dx_n$$

$$=\int_{\mathbb{R}^n}\exp\Bigl(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Bigr)\,d(x_1,\ldots,x_n)=\int_{\mathbb{R}^n}\exp\left(-\frac{1}{2}\sum_{i=1}^n\frac{y_i^2}{\lambda_i}\right)d(y_1,\ldots,y_n)$$

$$=\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\sum_{i=1}^n\frac{y_i^2}{\lambda_i}\right)dy_1\ldots dy_n=\prod_{i=1}^n(2\pi\lambda_i)^{1/2}\,.$$

- Hieraus ergibt sich die Behauptung, weil

$$\prod_{i=1}^n\lambda_i=\det\boldsymbol{\Lambda}=\det\bigl(\mathbf{V}^\top\mathbf{K}\mathbf{V}\bigr)=\det\bigl(\mathbf{V}^\top\mathbf{V}\bigr)\det\mathbf{K}=\det\mathbf{K}\,.$$ $\square$

1.2.2 Charakteristiken der multivariaten Normalverteilung

- Sei $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^\top\in\mathbb{R}^n$ ein beliebiger Vektor, und sei $\mathbf{K}=(k_{ij})$ eine symmetrische und positiv definite $n\times n$ Matrix.
- Wir bestimmen zunächst die charakteristische Funktion von normalverteilten Zufallvektoren.
- *Zur Erinnerung*: Die *charakteristische Funktion* $\varphi:\mathbb{R}^n\to\mathbb{C}$ eines beliebigen n–dimensionalen Zufallsvektors $\mathbf{X}=(X_1,\ldots,X_n)^\top:\Omega\to\mathbb{R}^n$ ist gegeben durch

$$\varphi(\mathbf{t})=\mathbb{E}\,\exp\bigl(\mathrm{i}\,\mathbf{t}^\top\mathbf{X}\bigr)=\mathbb{E}\,\exp\left(\mathrm{i}\sum_{\ell=1}^n t_\ell X_\ell\right),\qquad\forall\,\mathbf{t}=(t_1,\ldots,t_n)^\top\in\mathbb{R}^n\,.\tag{16}$$

Theorem 1.2

- *Der Zufallsvektor* $\mathbf{X}=(X_1,\ldots,X_n)^\top:\Omega\to\mathbb{R}^n$ *sei normalverteilt mit* $\mathbf{X}\sim\mathrm{N}(\boldsymbol{\mu},\mathbf{K})$.
- *Dann gilt für die charakteristische Funktion* $\varphi:\mathbb{R}^n\to\mathbb{C}$ *von* $\mathbf{X}$, *dass*

$$\varphi(\mathbf{t})=\exp\Bigl(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}-\frac{1}{2}\,\mathbf{t}^\top\mathbf{K}\mathbf{t}\Bigr),\qquad\forall\,\mathbf{t}\in\mathbb{R}^n\,.\tag{17}$$

Beweis

- Aus (13) und (16) folgt, dass

$$\begin{aligned}\varphi(\mathbf{t})&=\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\left(\mathrm{i}\sum_{\ell=1}^n t_\ell x_\ell\right)f(x_1,\ldots,x_n)\,dx_1\ldots dx_n\\&=\frac{1}{(2\pi)^{n/2}(\det\mathbf{K})^{1/2}}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\Bigl(\mathrm{i}\,\mathbf{t}^\top\mathbf{x}-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Bigr)\,dx_1\ldots dx_n\\&=\frac{\exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})}{(2\pi)^{n/2}(\det\mathbf{K})^{1/2}}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\Bigl(\mathrm{i}\,\mathbf{t}^\top\mathbf{y}-\frac{1}{2}\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y}\Bigr)\,dy_1\ldots dy_n\,,\end{aligned}$$

wobei sich die letzte Gleichheit mit Hilfe der Substitution $\mathbf{y}=\mathbf{x}-\boldsymbol{\mu}$ ergibt, für die die Matrix der partiellen Ableitungen die Einheitsmatrix und somit die Jacobi-Determinante gleich 1 ist.

- Auf ähnliche Weise wie im Beweis von Theorem 1.1 ergibt sich nun hieraus mit Hilfe der Substitutionen $\mathbf{y} = \mathbf{V}\mathbf{x}$ und $\mathbf{t} = \mathbf{V}\mathbf{s}$, dass

$$\begin{aligned}\varphi(\mathbf{t}) &= \frac{\exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})}{(2\pi)^{n/2}\,(\det\mathbf{K})^{1/2}}\int\limits_{-\infty}^{\infty}\ldots\int\limits_{-\infty}^{\infty}\exp\Bigl(\mathrm{i}\;\mathbf{s}^\top\mathbf{x} - \frac{1}{2}\mathbf{x}^\top\mathbf{V}^\top\mathbf{K}^{-1}\mathbf{V}\mathbf{x}\Bigr)\,dx_1\ldots dx_n\\ &= \frac{\exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})}{(2\pi)^{n/2}\,(\det\mathbf{K})^{1/2}}\int\limits_{-\infty}^{\infty}\ldots\int\limits_{-\infty}^{\infty}\exp\Bigl(\sum_{\ell=1}^{n}\Bigl(\mathrm{i}\;s_\ell x_\ell - \frac{x_\ell^2}{2\lambda_\ell}\Bigr)\Bigr)\,dx_1\ldots dx_n\end{aligned}$$

  und somit

$$\begin{aligned}\varphi(\mathbf{t}) &= \frac{\exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})}{(2\pi)^{n/2}\,(\det\mathbf{K})^{1/2}}\prod_{\ell=1}^{n}\int\limits_{-\infty}^{\infty}\exp\Bigl(\mathrm{i}\;s_\ell x_\ell - \frac{x_\ell^2}{2\lambda_\ell}\Bigr)\,dx_\ell\\ &= \exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})\prod_{\ell=1}^{n}\frac{1}{\sqrt{2\pi\lambda_\ell}}\int\limits_{-\infty}^{\infty}\exp\Bigl(\mathrm{i}\;s_\ell x_\ell - \frac{x_\ell^2}{2\lambda_\ell}\Bigr)\,dx_\ell\,,\end{aligned}$$

  wobei die Matrix $\mathbf{V}$ aus den orthonormalen Eigenvektoren von $\mathbf{K}$ besteht und $\lambda_1,\ldots,\lambda_n > 0$ die Eigenwerte von $\mathbf{K}$ sind mit $\det\mathbf{K} = \lambda_1\cdot\ldots\cdot\lambda_n$.
- Nun genügt es zu beachten, dass $\varphi_\ell : \mathbb{R}\to\mathbb{C}$ mit

$$\varphi_\ell(s) = \int\limits_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\lambda_\ell}}\;\exp\Bigl(\mathrm{i}\;sx - \frac{x^2}{2\lambda_\ell}\Bigr)\,dx$$

  die charakteristische Funktion der (eindimensionalen) $\mathrm{N}(0,\lambda_\ell)$–Verteilung ist.
- Für diese Funktion hatten wir in Abschnitt WR-5.3.3 gezeigt, dass $\varphi_\ell(s) = \exp\bigl(-\lambda_\ell s^2/2\bigr)$.
- Es gilt somit

$$\begin{aligned}\varphi(\mathbf{t}) &= \exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})\prod_{\ell=1}^{n}\exp\Bigl(-\frac{\lambda_\ell s_\ell^2}{2}\Bigr) = \exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})\;\exp\Bigl(-\frac{\sum_{\ell=1}^{n}\lambda_\ell s_\ell^2}{2}\Bigr)\\ &= \exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})\;\exp\Bigl(-\frac{\mathbf{t}^\top\mathbf{K}\mathbf{t}}{2}\Bigr)\,.\end{aligned}$$

□

Mit Hilfe der in Theorem 1.2 hergeleiteten Formel (17) für die charakteristische Funktion lassen sich nun der Erwartungswert und die Kovarianzmatrix von normalverteilten Zufallsvektoren bestimmen.

Korollar 1.1 *Wenn $\mathbf{X} = (X_1,\ldots,X_n)^\top \sim \mathrm{N}(\boldsymbol{\mu},\mathbf{K})$, dann gilt für beliebige $i,j = 1,\ldots,n$*

$$\mathbb{E}\,X_i = \mu_i\,,\qquad \textit{und}\qquad \mathrm{Cov}\,(X_i,X_j) = k_{ij}\,. \tag{18}$$

Beweis

- Aus (17) folgt, dass

$$\frac{\partial\,\varphi(\mathbf{t})}{\partial t_i} = \Bigl(\mathrm{i}\,\mu_i - \sum_{\ell=1}^{n}k_{i\ell}t_\ell\Bigr)\varphi(\mathbf{t}) \tag{19}$$

  und

$$\frac{\partial^2\varphi(\mathbf{t})}{\partial t_i\partial t_j} = -\,k_{ij}\varphi(\mathbf{t}) + \Bigl(\mathrm{i}\,\mu_i - \sum_{\ell=1}^{n}k_{i\ell}t_\ell\Bigr)\Bigl(\mathrm{i}\,\mu_j - \sum_{\ell=1}^{n}k_{j\ell}t_\ell\Bigr)\varphi(\mathbf{t})\,. \tag{20}$$

- Man kann sich leicht überlegen, dass
$$\mathbb{E}\, X_i = \mathrm{i}^{-1}\, \frac{\partial\, \varphi(\mathbf{t})}{\partial t_i}\,\Big|_{\mathbf{t}=\mathbf{o}}\,.$$
Wegen $\varphi(\mathbf{o}) = 1$ ergibt sich nun hieraus und aus (19), dass $\mathbb{E}\, X_i = \mu_i$.
- Außerdem gilt (vgl. Übungsaufgabe 2.3)
$$\mathbb{E}\,(X_i X_j) = -\,\frac{\partial^2 \varphi(\mathbf{t})}{\partial t_i \partial t_j}\,\Big|_{\mathbf{t}=\mathbf{o}}\,.$$
Hieraus und aus (20) ergibt sich, dass $\mathrm{Cov}\,(X_i, X_j) = k_{ij}$. □

Beachte

- In Theorem WR-4.14 hatten wir gezeigt, dass die Kovarianzmatrix $\mathbf{K} = \mathbf{K}_{\mathbf{X}}$ eines beliebigen Zufallsvektors $\mathbf{X} = (X_1, \ldots, X_n)^\top$ stets symmetrisch und nichtnegativ definit ist.
- In der Definitionsgleichung (13) der Dichte der regulären multivariaten Normalverteilung wird zusätzlich vorausgesetzt, dass die Kovarianzmatrix $\mathbf{K}$ positiv definit ist.
- Dabei ist die positive Definitheit von $\mathbf{K}$ nicht nur hinreichend, sondern auch notwendig dafür, dass $\det \mathbf{K} \neq 0$, d.h., dass $\mathbf{K}$ invertierbar ist bzw. vollen Rang hat.

1.2.3 Randverteilungen und Unabhängigkeit von Teilvektoren; Faltungsstabilität

- In diesem Abschnitt zeigen wir, wie weitere interessante Eigenschaften der multivariaten Normalverteilung mit Hilfe von Theorem 1.2 hergeleitet werden können.
- Hierfür benötigen wir eine *vektorielle Version* des Eindeutigkeitssatzes für charakteristische Funktionen (vgl. Korollar WR-5.5), die wir ohne Beweis angeben.

Lemma 1.9 *Seien $\mathbf{X}, \mathbf{Y} : \Omega \to \mathbb{R}^n$ beliebige Zufallsvektoren; $\mathbf{X} = (X_1, \ldots, X_n)^\top$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$. Dann gilt*
$$\mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{Y} \qquad \textit{genau dann, wenn} \qquad \varphi_{\mathbf{X}}(\mathbf{t}) = \varphi_{\mathbf{Y}}(\mathbf{t}) \qquad \forall\, \mathbf{t} = (t_1, \ldots, t_n)^\top \in \mathbb{R}^n\,, \tag{21}$$
wobei
$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\, \exp\Bigl(\mathrm{i} \sum_{j=1}^n t_j X_j\Bigr)\,, \qquad \varphi_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}\, \exp\Bigl(\mathrm{i} \sum_{j=1}^n t_j Y_j\Bigr)$$
die charakteristischen Funktionen von $\mathbf{X}$ bzw. $\mathbf{Y}$ sind.

Zunächst zeigen wir, dass beliebige Teilvektoren von normalverteilten Zufallsvektoren erneut normalverteilt sind.

- Dabei setzen wir so wie bisher voraus, dass $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ ein beliebiger Vektor und $\mathbf{K} = (k_{ij})$ eine symmetrische und positiv definite $n \times n$-Matrix ist.
- Es ist klar, dass der Zufallsvektor $(X_{\pi_1}, \ldots, X_{\pi_n})^\top$ für jede Permutation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ der natürlichen Zahlen $1, \ldots, n$ normalverteilt ist, wenn $\mathbf{X} = (X_1, \ldots, X_n)^\top$ normalverteilt ist.
- Bei der Untersuchung der Verteilung von Teilvektoren normalverteilter Zufallsvektoren können wir uns somit o.B.d.A. auf die Betrachtung der ersten Komponenten beschränken.

Korollar 1.2 *Sei $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$, wobei $\mathbf{K}$ positiv definit sei. Dann gilt*

$$(X_1, \ldots, X_m)^\top \sim \mathrm{N}(\boldsymbol{\mu}_m, \mathbf{K}_m) \qquad \forall\, m = 1, \ldots, n\,,$$

wobei $\boldsymbol{\mu}_m = (\mu_1, \ldots, \mu_m)^\top$ und $\mathbf{K}_m$ diejenige $m \times m$ Matrix bezeichnet, die aus den ersten m Zeilen bzw. Spalten von $\mathbf{K}$ gebildet wird.

Beweis

- Sei $\varphi : \mathbb{R}^n \to \mathbb{C}$ die charakteristische Funktion von $(X_1, \ldots, X_n))^\top$.
- Für die charakteristische Funktion $\varphi_m : \mathbb{R}^m \to \mathbb{C}$ von $(X_1, \ldots, X_m))^\top$ gilt dann

$$\varphi_m(\mathbf{t}_m) = \varphi\Bigl((\mathbf{t}_m, \underbrace{0, \ldots, 0}_{n-m})\Bigr)\,, \qquad \forall\, \mathbf{t}_m = (t_1, \ldots, t_m))^\top \in \mathbb{R}^m\,.$$

- Hieraus und aus (17) ergibt sich, dass

$$\varphi_m(\mathbf{t}_m) = \exp\Bigl(\mathrm{i}\,\mathbf{t}_m^\top \boldsymbol{\mu}_m - \frac{1}{2}\,\mathbf{t}_m^\top \mathbf{K}_m \mathbf{t}_m\Bigr)\,, \qquad \forall\, \mathbf{t}_m \in \mathbb{R}^m\,.$$

- Weil mit $\mathbf{K}$ auch die $m \times m$ Matrix $\mathbf{K}_m$ symmetrisch und positiv definit ist, bedeutet dies wegen Theorem 1.2, dass die charakteristische Funktion des Teilvektors $(X_1, \ldots, X_m))^\top$ mit der charakteristischen Funktion der $\mathrm{N}(\boldsymbol{\mu}_m, \mathbf{K}_m)$–Verteilung übereinstimmt.
- Wegen des eineindeutigen Zusammenhanges zwischen der charakteristischen Funktion und der Verteilung von Zufallsvektoren (vgl. Lemma 1.9) ergibt sich hieraus die Behauptung. $\square$

Bei der Zerlegung des normalverteilten Zufallsvektors $\mathbf{X} = (X_1, \ldots, X_n)^\top$ in die zwei Teilvektoren $(X_1, \ldots, X_m)^\top$ und $(X_{m+1}, \ldots, X_n)^\top$, wobei $1 < m < n$, lässt sich ein einfaches Kriterium dafür angeben, dass $(X_1, \ldots, X_m)^\top$ und $(X_{m+1}, \ldots, X_n)^\top$ unabhängig sind.

Korollar 1.3 *Sei $\mathbf{X} = (X_1, \ldots, X_n)^\top$ ein normalverteilter Zufallsvektor mit $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$; $\mathbf{K} = (k_{ij})$. Die Teilvektoren $(X_1, \ldots, X_m)^\top$ und $(X_{m+1}, \ldots, X_n)^\top$ sind genau dann unabhängig, wenn $k_{ij} = 0$ für beliebige $i \in \{1, \ldots, m\}$ und $j \in \{m+1, \ldots, n\}$.*

Beweis

- Wenn die Teilvektoren $(X_1, \ldots, X_m)^\top$ und $(X_{m+1}, \ldots, X_n)^\top$ unabhängig sind, dann sind auch die (eindimensionalen) Zufallsvariablen X_i und X_j für beliebige $i \in \{1, \ldots, m\}$ und $j \in \{m+1, \ldots, n\}$ unabhängig.
- Damit gilt insbesondere $\mathrm{Cov}\,(X_i, X_j) = 0$, und aus Korollar 1.1 folgt, dass $k_{ij} = 0$.
- Wir nehmen nun umgekehrt an, dass $k_{ij} = 0$ für beliebige $i \in \{1, \ldots, m\}$ und $j \in \{m+1, \ldots, n\}$.
- Dann ergibt sich aus Theorem 1.2, dass sich die charakteristische Funktion $\varphi(\mathbf{t})$ von $\mathbf{X} = (X_1, \ldots, X_n)^\top$ wie folgt faktorisieren lässt.
- Für jedes $\mathbf{t} = (t_1, \ldots, t_n)^\top \in \mathbb{R}^n$ gilt

$$\begin{aligned}
\varphi(\mathbf{t}) = \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\,\mathbf{t}^\top \mathbf{K} \mathbf{t}\Bigr) &= \exp\Bigl(\mathrm{i}\sum_{i=1}^n t_i \mu_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n t_i k_{ij} t_j\Bigr) \\
&= \exp\Bigl(\mathrm{i}\sum_{i=1}^m t_i \mu_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m t_i k_{ij} t_j\Bigr)\,\exp\Bigl(\mathrm{i}\sum_{i=m+1}^n t_i \mu_i - \frac{1}{2}\sum_{i=m+1}^n \sum_{j=m+1}^n t_i k_{ij} t_j\Bigr)\,,
\end{aligned}$$

wobei die Faktoren des letzten Ausdruckes die charakteristischen Funktionen von $(X_1, \ldots, X_m)^\top$ und $(X_{m+1}, \ldots, X_n)^\top$ sind.

- Die Behauptung ergibt sich nun aus dem eineindeutigen Zusammenhang zwischen der Verteilung und der charakteristischen Funktion von Zufallsvektoren, vgl. Lemma 1.9. □

Beachte

- Schließlich diskutieren wir noch die Faltungsstabilität der multivariaten Normalverteilung und verallgemeinern dabei Korollar WR-3.2, wo wir diese Eigenschaft für die eindimensionale Normalverteilung bewiesen hatten.
- In diesem Zusammenhang ist die folgende Formel für die charakteristische Funktion von Summen unabhängiger Zufallsvektoren nützlich, die sich genauso wie die in Theorem WR-5.18 für den eindimensionalen Fall hergeleitete Formel beweisen lässt.

Lemma 1.10 *Seien $\mathbf{Z}_1, \mathbf{Z}_2 : \Omega \to \mathbb{R}^n$ unabhängige Zufallsvektoren. Für die charakteristische Funktion $\varphi_{\mathbf{Z}_1+\mathbf{Z}_2} : \mathbb{R}^n \to \mathbb{C}$ der Summe $\mathbf{Z}_1 + \mathbf{Z}_2$ gilt dann*

$$\varphi_{\mathbf{Z}_1+\mathbf{Z}_2}(\mathbf{t}) = \varphi_{\mathbf{Z}_1}(\mathbf{t})\, \varphi_{\mathbf{Z}_2}(\mathbf{t}), \qquad \forall\, \mathbf{t} \in \mathbb{R}^n, \tag{22}$$

wobei $\varphi_{\mathbf{Z}_i}$ die charakteristische Funktion von $\mathbf{Z}_i$ bezeichnet; $i = 1, 2$.

Die folgende Aussage wird *Faltungsstabilität* der multivariaten Normalverteilung genannt.

Korollar 1.4 *Seien $\mathbf{Z}_1, \mathbf{Z}_2 : \Omega \to \mathbb{R}^n$ unabhängige Zufallsvektoren mit $\mathbf{Z}_i \sim \mathrm{N}(\boldsymbol{\mu}_i, \mathbf{K}_i)$ für $i = 1, 2$. Dann gilt $\mathbf{Z}_1 + \mathbf{Z}_2 \sim \mathrm{N}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \mathbf{K}_1 + \mathbf{K}_2)$.*

Beweis

- Aus (17) und (22) ergibt sich, dass

$$\begin{aligned}
\varphi_{\mathbf{Z}_1+\mathbf{Z}_2}(\mathbf{t}) &= \varphi_{\mathbf{Z}_1}(\mathbf{t})\, \varphi_{\mathbf{Z}_2}(\mathbf{t}) \\
&= \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu}_1 - \frac{1}{2}\,\mathbf{t}^\top \mathbf{K}_1 \mathbf{t}\Bigr)\, \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu}_2 - \frac{1}{2}\,\mathbf{t}^\top \mathbf{K}_2 \mathbf{t}\Bigr) \\
&= \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{1}{2}\,\mathbf{t}^\top (\mathbf{K}_1 + \mathbf{K}_2)\mathbf{t}\Bigr).
\end{aligned}$$

- Hieraus und aus dem Eindeutigkeitssatz für charakteristische Funktionen von Zufallsvektoren (vgl. Lemma 1.9) ergibt sich die Behauptung. □

1.2.4 Lineare Transformation von normalverteilten Zufallsvektoren

Wir zeigen nun, dass die Lineartransformation normalverteilter Zufallsvektoren erneut zu normalverteilten Zufallsvektoren führt.

Theorem 1.3

- *Sei $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ ein n–dimensionaler normalverteilter Zufallsvektor mit Erwartungswertvektor $\boldsymbol{\mu} \in \mathbb{R}^n$ und mit (positiv definiter) Kovarianzmatrix $\mathbf{K}$.*
- *Außerdem gelte $m \le n$, und $\mathbf{A}$ sei eine beliebige $m \times n$ Matrix mit vollem Rang $\mathrm{rg}(\mathbf{A}) = m$ bzw. $\mathbf{c} \in \mathbb{R}^m$ ein beliebiger m–dimensionaler Vektor.*
- *Dann ist $\mathbf{Z} = \mathbf{A}\mathbf{Y} + \mathbf{c}$ ein (m–dimensionaler) normalverteilter Zufallsvektor mit*

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{c},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top). \tag{23}$$

Beweis

- Für jedes $\mathbf{a} \in \mathbb{R}^m$ gilt
$$\varphi_{\mathbf{Z}}(\mathbf{t}) = \exp(\mathrm{i}\,\mathbf{t}^\top \mathbf{a})\varphi_{\mathbf{Z}-\mathbf{a}}(\mathbf{t}), \qquad \forall\, \mathbf{t} \in \mathbb{R}^m .$$
- Aus der in Theorem 1.2 hergeleiteten Formel (17) und aus dem Eindeutigkeitssatz für die charakteristische Funktion von normalverteilten Zufallsvektoren folgt somit, dass
$$\mathbf{Z} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{c},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top) \quad \text{genau dann, wenn} \quad \mathbf{Z} - (\mathbf{A}\boldsymbol{\mu} + \mathbf{c}) \sim \mathrm{N}(\mathbf{o},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top).$$
- O.B.d.A. können (und werden) wir deshalb annehmen, dass $\mathbf{Y} \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$ und $\mathbf{c} = \mathbf{o}$.
- Für die charakteristische Funktion $\varphi_{\mathbf{Z}}(\mathbf{t})$ von $\mathbf{Z} = \mathbf{A}\mathbf{Y}$ ergibt sich dann, dass für jedes $\mathbf{t} \in \mathbb{R}^m$
$$\begin{aligned} \varphi_{\mathbf{Z}}(\mathbf{t}) &= \mathbb{E}\, e^{\mathrm{i}\,\mathbf{t}^\top \mathbf{Z}} \\ &= \mathbb{E}\, e^{\mathrm{i}\,\mathbf{t}^\top \mathbf{A}\mathbf{Y}} = \mathbb{E}\, e^{\mathrm{i}\,(\mathbf{A}^\top \mathbf{t})^\top \mathbf{Y}} \\ &= \varphi_{\mathbf{Y}}(\mathbf{A}^\top \mathbf{t}), \end{aligned}$$
  wobei $\varphi_{\mathbf{Y}}(\mathbf{A}^\top \mathbf{t})$ den Wert der charakteristischen Funktion des normalverteilten Zufallsvektors $\mathbf{Y}$ an der Stelle $\mathbf{A}^\top \mathbf{t} \in \mathbb{R}^n$ bezeichnet.
- Aus der Darstellungsformel (17) für die charakteristische Funktion normalverteilter Zufallsvektoren ergibt sich nun, dass
$$\begin{aligned} \varphi_{\mathbf{Z}}(\mathbf{t}) &= \varphi_{\mathbf{Y}}(\mathbf{A}^\top \mathbf{t}) \\ &= \exp\Bigl(-\frac{1}{2}\,(\mathbf{A}^\top \mathbf{t})^\top \mathbf{K}(\mathbf{A}^\top \mathbf{t})\Bigr) \\ &= \exp\Bigl(-\frac{1}{2}\,\mathbf{t}^\top (\mathbf{A}\mathbf{K}\mathbf{A}^\top)\mathbf{t}\Bigr). \end{aligned}$$
- Mit anderen Worten: Die charakteristische Funktion von $\mathbf{Z}$ stimmt mit der charakteristischen Funktion der $\mathrm{N}(\mathbf{o},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top)$–Verteilung überein.
- Aus dem Eindeutigkeitssatz für die charakteristische Funktion von Zufallsvektoren folgt somit, dass $\mathbf{Z} \sim \mathrm{N}(\mathbf{o},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top)$. □

Aus Theorem 1.3 ergibt sich insbesondere, dass sich normalverteilte Zufallsvektoren durch Lineartransformation von Vektoren konstruieren lassen, deren Komponenten unabhängige und $\mathrm{N}(0,1)$-verteilte Zufallsvariablen sind.

Korollar 1.5

- *Seien $Y_1, \ldots, Y_n : \Omega \to \mathbb{R}$ unabhängige Zufallsvariablen mit $Y_i \sim \mathrm{N}(0,1)$ für jedes $i = 1, \ldots, n$, d.h. $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \sim \mathrm{N}(\mathbf{o}, \mathbf{I})$.*
- *Sei $\mathbf{K}$ eine symmetrische und positiv definite $n \times n$ Matrix, und sei $\boldsymbol{\mu} \in \mathbb{R}^n$.*
- *Für den Zufallsvektor $\mathbf{Z} = \mathbf{K}^{1/2}\mathbf{Y} + \boldsymbol{\mu}$ gilt dann $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu},\, \mathbf{K})$, wobei $\mathbf{K}^{1/2}$ die Quadratwurzel von $\mathbf{K}$ ist.*

Beweis

- Aus Theorem 1.3 ergibt sich, dass
$$\mathbf{Z} \sim \mathrm{N}\bigl(\boldsymbol{\mu},\, \mathbf{K}^{1/2}(\mathbf{K}^{1/2})^\top\bigr).$$
- Hieraus und aus Lemma 1.6 folgt die Behauptung. □

1.2.5 Singuläre multivariate Normalverteilung

Der in Abschnitt 1.2.1 eingeführten Begriff der (regulären) multivariaten Normalverteilung lässt sich wie folgt verallgemeinern.

- Hierfür ist eine Faktorisierungseigenschaft von Kovarianzmatrizen nützlich, die wir bereits in Lemma 1.7 erwähnt hatten.
- *Zur Erinnerung*: Sei $\mathbf{K}$ eine symmetrische und nichtnegativ definite $n \times n$ Matrix mit $\operatorname{rg}(\mathbf{K}) = r \le n$. Dann gibt es eine $n \times r$ Matrix $\mathbf{B}$ mit $\operatorname{rg}(\mathbf{B}) = r$, so dass

$$\mathbf{K} = \mathbf{B}\mathbf{B}^\top . \tag{24}$$

Definition

- Sei $\mathbf{Y}$ ein n–dimensionaler Zufallsvektor mit Erwartungswertvektor $\boldsymbol{\mu} = \mathbb{E}\,\mathbf{Y}$ und Kovarianzmatrix $\mathbf{K} = \operatorname{Cov}(\mathbf{Y})$, so dass $\operatorname{rg}(\mathbf{K}) = r$ mit $r \le n$.
- Dann heißt $\mathbf{Y}$ normalverteilt, wenn $\mathbf{Y} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \mathbf{B}\mathbf{Z}$, wobei $\mathbf{B}$ eine $n \times r$ Matrix mit $\operatorname{rg}(\mathbf{B}) = r$ ist, die der Gleichung (24) genügt, und $\mathbf{Z}$ ein r–dimensionaler Zufallsvektor mit $\mathbf{Z} \sim \mathrm{N}(\mathbf{o}, \mathbf{I}_r)$ ist.
- Wir sagen, dass $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ *singulär normalverteilt* ist, wenn $\operatorname{rg}(\mathbf{K}) < n$.
(Schreibweise: $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$)

Beachte

- Wenn $\operatorname{rg}(\mathbf{K}) = r < n$, dann ist der Zufallsvektor $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ *nicht* absolutstetig,
  - denn die Werte von $\mathbf{Y} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \mathbf{B}\mathbf{Z}$ liegen mit Wahrscheinlichkeit 1 in der r–dimensionalen Teilmenge $\{\boldsymbol{\mu} + \mathbf{B}\mathbf{x} : \mathbf{x} \in \mathbb{R}^r\}$ des $\mathbb{R}^n$,
  - d.h., die Verteilung von $\mathbf{Y}$ besitzt keine Dichte bezüglich des n–dimensionalen Lebesgue-Maßes.
  - Ein Beispiel hierfür ist der Zufallsvektor $\mathbf{Y} = (Z, Z)^\top = \mathbf{B}Z$ mit $Z \sim \mathrm{N}(0, \sigma^2)$ und $\mathbf{B} = (1,1)^\top$, der nur Werte auf der Diagonalen $\{(z_1, z_2) \in \mathbb{R}^2 : z_1 = z_2\}$ annimmt.
- Die Verteilung des Zufallsvektors $\boldsymbol{\mu} + \mathbf{B}\mathbf{Z}$ hängt *nicht* von der Wahl der Matrix $\mathbf{B}$ in der Faktorisierungsgleichung (24) ab.
- Dies ergibt sich unmittelbar aus den folgenden beiden Kriterien für das Vorliegen von (singulären bzw. regulären) multivariaten Normalverteilungen.

Theorem 1.4

- *Sei* $\mathbf{Y}$ *ein* n*–dimensionaler Zufallsvektor mit Erwartungswertvektor* $\boldsymbol{\mu} = \mathbb{E}\,\mathbf{Y}$ *und Kovarianzmatrix* $\mathbf{K} = \operatorname{Cov}(\mathbf{Y})$*, so dass* $\operatorname{rg}(\mathbf{K}) = r$ *mit* $r \le n$.
- *Der Zufallsvektor* $\mathbf{Y}$ *ist genau dann normalverteilt, wenn eine der beiden folgenden Bedingungen erfüllt ist:*
  1. *Die charakteristische Funktion* $\varphi(\mathbf{t}) = \mathbb{E}\,\exp\Bigl(\mathrm{i}\sum_{j=1}^n t_j Y_j\Bigr)$ *von* $\mathbf{Y}$ *ist gegeben durch*

$$\varphi(\mathbf{t}) = \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\,\mathbf{t}^\top\mathbf{K}\mathbf{t}\Bigr), \qquad \forall\, \mathbf{t} = (t_1, \ldots, t_n)^\top \in \mathbb{R}^n . \tag{25}$$

  2. *Die lineare Funktion* $\mathbf{c}^\top\mathbf{Y}$ *von* $\mathbf{Y}$ *ist für jedes* $\mathbf{c} \in \mathbb{R}^n$ *mit* $\mathbf{c} \ne \mathbf{o}$ *normalverteilt mit*

$$\mathbf{c}^\top\mathbf{Y} \sim \mathrm{N}(\mathbf{c}^\top\boldsymbol{\mu}, \mathbf{c}^\top\mathbf{K}\mathbf{c}) .$$

Der *Beweis* von Theorem 1.4 wird in den Übungen diskutiert (vgl. Übungsaufgabe 2.2). Er wird deshalb hier weggelassen.

1.3 Lineare und quadratische Formen normalverteilter Zufallsvektoren

1.3.1 Definition, Erwartungswert und Kovarianz

Definition

- Seien $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ und $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ beliebige n–dimensionale Zufallsvektoren, und sei $\mathbf{A}$ eine symmetrische $n \times n$ Matrix mit reellwertigen Eintragungen.
- Dann heißt die (reellwertige) Zufallsvariable $\mathbf{Y}^\top \mathbf{A} \mathbf{Y} : \Omega \to \mathbb{R}$ *quadratische Form* von $\mathbf{Y}$ bezüglich $\mathbf{A}$.
- Die Zufallsvariable $\mathbf{Y}^\top \mathbf{A} \mathbf{Z} : \Omega \to \mathbb{R}$ heißt *bilineare Form* von $\mathbf{Y}$ und $\mathbf{Z}$ bezüglich $\mathbf{A}$.

Zunächst bestimmen wir den Erwartungswert von quadratischen bzw. bilinearen Formen.

Theorem 1.5 *Seien $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ und $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ beliebige n–dimensionale Zufallsvektoren, und sei $\mathbf{A}$ eine symmetrische $n \times n$ Matrix mit reellwertigen Eintragungen. Die Erwartungswertvektoren $\boldsymbol{\mu}_\mathbf{Y} = \mathbb{E}\,\mathbf{Y}$ und $\boldsymbol{\mu}_\mathbf{Z} = \mathbb{E}\,\mathbf{Z}$ sowie die Kovarianzmatrizen $\mathbf{K}_{\mathbf{YY}} = \bigl(\mathrm{Cov}\,(Y_i, Y_j)\bigr)$ und $\mathbf{K}_{\mathbf{ZY}} = \bigl(\mathrm{Cov}\,(Z_i, Y_j)\bigr)$ seien wohldefiniert. Dann gilt*

$$\mathbb{E}\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Y}\bigr) = \mathrm{sp}(\mathbf{A}\mathbf{K}_{\mathbf{YY}}) + \boldsymbol{\mu}_\mathbf{Y}^\top \mathbf{A} \boldsymbol{\mu}_\mathbf{Y} \qquad \text{und} \qquad \mathbb{E}\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Z}\bigr) = \mathrm{sp}(\mathbf{A}\mathbf{K}_{\mathbf{ZY}}) + \boldsymbol{\mu}_\mathbf{Y}^\top \mathbf{A} \boldsymbol{\mu}_\mathbf{Z}\,. \tag{26}$$

Beweis

- Wir beweisen nur die zweite Formel in (26), denn die erste Formel ergibt sich hieraus als Spezialfall für $\mathbf{Z} = \mathbf{Y}$.
- Offenbar gilt $\mathbf{Y}^\top \mathbf{A} \mathbf{Z} = \mathrm{sp}\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Z}\bigr)$. Außerdem folgt aus Lemma 1.1, dass $\mathrm{sp}\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Z}\bigr) = \mathrm{sp}\bigl(\mathbf{A}\mathbf{Z}\mathbf{Y}^\top\bigr)$.
- Insgesamt ergibt sich also, dass

$$\begin{aligned} \mathbb{E}\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Z}\bigr) &= \mathbb{E}\,\mathrm{sp}\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Z}\bigr) = \mathbb{E}\,\mathrm{sp}\bigl(\mathbf{A}\mathbf{Z}\mathbf{Y}^\top\bigr) = \mathrm{sp}\bigl(\mathbf{A}\mathbb{E}\,(\mathbf{Z}\mathbf{Y}^\top)\bigr) \\ &= \mathrm{sp}\bigl(\mathbf{A}(\mathbf{K}_{\mathbf{ZY}} + \boldsymbol{\mu}_\mathbf{Z}\boldsymbol{\mu}_\mathbf{Y}^\top)\bigr) = \mathrm{sp}(\mathbf{A}\mathbf{K}_{\mathbf{ZY}}) + \boldsymbol{\mu}_\mathbf{Y}^\top \mathbf{A} \boldsymbol{\mu}_\mathbf{Z}\,. \end{aligned}$$

□

Auf ähnliche Weise lässt sich eine Formel für die Kovarianz von quadratischen Formen normalverteilter Zufallsvektoren herleiten. Dabei sind die folgenden Formeln für die dritten bzw. vierten gemischten Momente der Komponenten von zentrierten normalverteilten Zufallsvektoren nützlich.

Lemma 1.11 *Sei $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$ ein normalverteilter Zufallsvektor mit Erwartungswertvektor $\boldsymbol{\mu} = \mathbf{o}$ und mit beliebiger Kovarianzmatrix $\mathbf{K} = (k_{ij})$. Dann gilt*

$$\mathbb{E}\,(Z_i Z_j Z_\ell) = 0 \quad \text{und} \quad \mathbb{E}\,(Z_i Z_j Z_\ell Z_m) = k_{ij}k_{\ell m} + k_{i\ell}k_{jm} + k_{j\ell}k_{im} \qquad \forall\, i, j, \ell, m \in \{1, \ldots, n\}\,. \tag{27}$$

Der *Beweis* von Lemma 1.11 wird hier weggelassen und in den Übungen diskutiert, vgl. Übungsaufgabe 2.3.

Theorem 1.6

- *Sei $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ ein n–dimensionaler Zufallsvektor mit $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$, und seien $\mathbf{A} = (a_{ij})$, $\mathbf{B} = (b_{ij})$ beliebige symmetrische $n \times n$ Matrizen.*
- *Dann gilt*

$$\mathrm{Cov}\,\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Y},\, \mathbf{Y}^\top \mathbf{B} \mathbf{Y}\bigr) = 2\,\mathrm{sp}(\mathbf{A}\mathbf{K}\mathbf{B}\mathbf{K}) + 4\boldsymbol{\mu}^\top \mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu}\,. \tag{28}$$

- *Insbesondere gilt* $\mathrm{Var}\,\bigl(\mathbf{Y}^\top \mathbf{A} \mathbf{Y}\bigr) = 2\,\mathrm{sp}\bigl((\mathbf{A}\mathbf{K})^2\bigr) + 4\boldsymbol{\mu}^\top \mathbf{A}\mathbf{K}\mathbf{A}\boldsymbol{\mu}$.

Beweis

- Aus der Definition der Kovarianz und aus Theorem 1.5 ergibt sich, dass
$$\begin{aligned}\operatorname{Cov}\left(\mathbf{Y}^\top\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right) &= \mathbb{E}\left((\mathbf{Y}^\top\mathbf{A}\mathbf{Y}-\mathbb{E}\,(\mathbf{Y}^\top\mathbf{A}\mathbf{Y}))(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}-\mathbb{E}\,(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}))\right)\\ &= \mathbb{E}\left((\mathbf{Y}^\top\mathbf{A}\mathbf{Y}-\operatorname{sp}(\mathbf{A}\mathbf{K})-\boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}-\operatorname{sp}(\mathbf{B}\mathbf{K})-\boldsymbol{\mu}^\top\mathbf{B}\boldsymbol{\mu})\right).\end{aligned}$$
- Mit der Substitution $\mathbf{Z}=\mathbf{Y}-\boldsymbol{\mu}$ bzw. $\mathbf{Y}=\mathbf{Z}+\boldsymbol{\mu}$ ergibt sich hieraus, dass
$$\begin{aligned}\operatorname{Cov}\left(\mathbf{Y}^\top\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right) &= \mathbb{E}\left((\mathbf{Z}^\top\mathbf{A}\mathbf{Z}+2\boldsymbol{\mu}^\top\mathbf{A}\mathbf{Z}-\operatorname{sp}(\mathbf{A}\mathbf{K}))(\mathbf{Z}^\top\mathbf{B}\mathbf{Z}+2\boldsymbol{\mu}^\top\mathbf{B}\mathbf{Z}-\operatorname{sp}(\mathbf{B}\mathbf{K}))\right)\\ &= \mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)+2\boldsymbol{\mu}^\top\mathbf{A}\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)+2\boldsymbol{\mu}^\top\mathbf{B}\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\right)\\ &\quad -\mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\right)\operatorname{sp}(\mathbf{B}\mathbf{K})-\mathbb{E}\left(\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)\operatorname{sp}(\mathbf{A}\mathbf{K})\\ &\quad +4\boldsymbol{\mu}^\top\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu}+\operatorname{sp}(\mathbf{A}\mathbf{K})\operatorname{sp}(\mathbf{B}\mathbf{K})\\ &= \mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)+2\boldsymbol{\mu}^\top\mathbf{A}\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)+2\boldsymbol{\mu}^\top\mathbf{B}\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\right)\\ &\quad +4\boldsymbol{\mu}^\top\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu}-\operatorname{sp}(\mathbf{A}\mathbf{K})\operatorname{sp}(\mathbf{B}\mathbf{K}),\end{aligned}$$
wobei sich die letzte Gleichheit aus Theorem 1.5 ergibt, weil $Z\sim \mathrm{N}(\mathbf{o},\mathbf{K})$ und somit $\mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\right)=\operatorname{sp}(\mathbf{A}\mathbf{K})$ gilt.
- Weil die Matrizen $\mathbf{A}$, $\mathbf{B}$ und $\mathbf{K}$ symmetrisch sind, ergibt sich aus Lemma 1.11, dass
$$\begin{aligned}\mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right) &= \mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\cdot\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)\\ &= \sum_{i=1}^n\sum_{j=1}^n\sum_{\ell=1}^n\sum_{m=1}^n a_{ij}b_{\ell m}\mathbb{E}\left(Z_iZ_jZ_\ell Z_m\right)\\ &= \sum_{i=1}^n\sum_{j=1}^n\sum_{\ell=1}^n\sum_{m=1}^n \left(a_{ij}k_{ji}b_{\ell m}k_{m\ell}+a_{ji}k_{i\ell}b_{\ell m}k_{mj}+a_{ij}k_{j\ell}b_{\ell m}k_{mi}\right)\\ &= \operatorname{sp}(\mathbf{A}\mathbf{K})\operatorname{sp}(\mathbf{B}\mathbf{K})+2\operatorname{sp}(\mathbf{A}\mathbf{K}\mathbf{B}\mathbf{K}).\end{aligned}$$
- Außerdem ergibt sich aus Lemma 1.11, dass
$$\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\right)=\left(\sum_{i=1}^n\sum_{j=1}^n a_{ij}\mathbb{E}\left(Z_iZ_jZ_\ell\right)\right)_\ell=\mathbf{o} \tag{29}$$
und entsprechend $\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)=\mathbf{o}$.
- Zusammen mit dem oben hergeleiteten Ausdruck für $\operatorname{Cov}\left(\mathbf{Y}^\top\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right)$ ergibt sich nun hieraus die Behauptung. □

Wir leiten nun noch die folgende Formel für den Kovarianzvektor von linearen bzw. quadratischen Formen normalverteilter Zufallsvektoren her.

Theorem 1.7 *Sei $\mathbf{Y}=(Y_1,\ldots,Y_n)^\top$ ein n-dimensionaler Zufallsvektor mit $\mathbf{Y}\sim\mathrm{N}(\boldsymbol{\mu},\mathbf{K})$, und seien $\mathbf{A}=(a_{ij})$, $\mathbf{B}=(b_{ij})$ beliebige symmetrische $n\times n$ Matrizen. Dann gilt*
$$\operatorname{Cov}\left(\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right)=2\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu}. \tag{30}$$

Beweis

- Weil $\mathbb{E}\,(\mathbf{A}\mathbf{Y})=\mathbf{A}\boldsymbol{\mu}$ und weil in Theorem 1.5 gezeigt wurde, dass
$$\mathbb{E}\left(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right)=\operatorname{sp}(\mathbf{B}\mathbf{K})+\boldsymbol{\mu}^\top\mathbf{B}\boldsymbol{\mu},$$
ergibt sich, dass
$$\begin{aligned}\operatorname{Cov}\left(\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right) &= \mathbb{E}\left((\mathbf{A}\mathbf{Y}-\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}-\boldsymbol{\mu}^\top\mathbf{B}\boldsymbol{\mu}-\operatorname{sp}(\mathbf{B}\mathbf{K}))\right)\\ &= \mathbb{E}\left((\mathbf{A}\mathbf{Y}-\mathbf{A}\boldsymbol{\mu})((\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu})+2(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}\boldsymbol{\mu}-\operatorname{sp}(\mathbf{B}\mathbf{K}))\right).\end{aligned}$$

- Außerdem gilt $\mathbb{E}(\mathbf{AY}-\mathbf{A}\boldsymbol{\mu})=\mathbf{o}$, und aus (29) folgt mit $\mathbf{Z}=\mathbf{Y}-\boldsymbol{\mu}$, dass
$$\mathbb{E}\left((\mathbf{AY}-\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu})\right)=\mathbf{A}\mathbb{E}\left((\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu})\right)=\mathbf{o}\,.$$
- Somit ergibt sich, dass
$$\begin{aligned}\operatorname{Cov}\left(\mathbf{AY},\,\mathbf{Y}^\top\mathbf{BY}\right) &= 2\mathbb{E}\left((\mathbf{AY}-\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}\boldsymbol{\mu}\right)\\ &= 2\mathbf{A}\mathbb{E}\left((\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\right)\mathbf{B}\boldsymbol{\mu}\\ &= 2\mathbf{AKB}\boldsymbol{\mu}\,.\end{aligned}$$
□

1.3.2 Nichtzentrale χ^2–Verteilung

Um die Verteilung von quadratischen Formen normalverteilter Zufallsvektoren zu bestimmen, führen wir die (parametrische) Familie der nichtzentralen χ^2–Verteilungen ein.

Definition Sei $\boldsymbol{\mu}\in\mathbb{R}^n$ und $(X_1,\ldots,X_n)^\top\sim\mathrm{N}(\boldsymbol{\mu},\mathbf{I})$. Dann sagt man, dass die Zufallsvariable
$$Z=(X_1,\ldots,X_n)(X_1,\ldots,X_n)^\top=\sum_{i=1}^n X_i^2$$
eine *nichtzentrale χ^2–Verteilung* mit n Freiheitsgraden und dem *Nichtzentralitätsparameter* $\lambda=\boldsymbol{\mu}^\top\boldsymbol{\mu}$ hat. (Schreibweise: $Z\sim\chi^2_{n,\lambda}$)

Beachte

- Für $\boldsymbol{\mu}=\mathbf{o}$ ergibt sich als Spezialfall die bereits in Abschnitt I–1.3.1 eingeführte (zentrale) χ^2–Verteilung χ^2_n mit n Freiheitsgraden.
- Um eine Formel für die Dichte der nichtzentralen χ^2–Verteilung herzuleiten, betrachten wir (neben der charakteristischen Funktion) noch eine weitere *Integraltransformation* von Wahrscheinlichkeitsdichten.

Definition

- Sei $f:\mathbb{R}\to[0,\infty)$ die Dichte einer reellwertigen Zufallsvariable, so dass das Integral $\int_{-\infty}^{\infty}e^{tx}f(x)\,dx$ wohldefiniert ist für jedes $t\in(a,b)$ aus einem gewissen Intervall (a,b) mit $a<b$.
- Dann heißt die Abbildung $\psi:(a,b)\to\mathbb{R}$ mit
$$\psi(t)=\int_{-\infty}^{\infty}e^{tx}f(x)\,dx\,,\qquad\forall\,t\in(a,b)\tag{31}$$
die *momenterzeugende Funktion* der Dichte f.

Es gilt der folgende *Eindeutigkeitssatz* für momenterzeugende Funktionen, den wir hier ohne Beweis angeben.

Lemma 1.12

- *Seien $f,f':\mathbb{R}\to[0,\infty)$ die Dichten von reellwertigen Zufallsvariablen, und seien die zugehörigen momenterzeugenden Funktionen $\psi:(a,b)\to\mathbb{R}$ bzw. $\psi':(a,b)\to\mathbb{R}$ auf einem (gemeinsamen) Intervall (a,b) mit $a<b$ wohldefiniert.*
- *Es gilt $\psi(t)=\psi'(t)$ für jedes $t\in(a,b)$ genau dann, wenn $f(x)=f'(x)$ für fast jedes $x\in\mathbb{R}$.*

Mit Hilfe von Lemma 1.12 können wir nun die Dichte der nichtzentralen χ^2–Verteilung bestimmen.

Theorem 1.8

- *Sei $Z_{n,\lambda}:\Omega\to\mathbb{R}$ eine $\chi^2_{n,\lambda}$–verteilte Zufallsvariable mit n Freiheitsgraden und Nichtzentralitätsparameter λ.*
- *Dann ist die Dichte von $Z_{n,\lambda}$ gegeben durch*

$$f_{Z_{n,\lambda}}(z)=\begin{cases}\exp\Bigl(-\dfrac{\lambda+z}{2}\Bigr)\displaystyle\sum_{j=0}^{\infty}\dfrac{\Bigl(\dfrac{\lambda}{2}\Bigr)^j z^{\frac{n}{2}+j-1}}{j!\,2^{\frac{n}{2}+j}\,\Gamma\Bigl(\dfrac{n}{2}+j\Bigr)}\,, & \textit{wenn } z>0,\\[2ex] 0 & \textit{sonst.}\end{cases}\tag{32}$$

Beweis

- Sei $\boldsymbol{\mu}\in\mathbb{R}^n$ und $(X_1,\ldots,X_n)^\top\sim\mathrm{N}(\boldsymbol{\mu},\mathbf{I})$.
- Die momenterzeugende Funktion $\psi_Z(t)$ von $Z=(X_1,\ldots,X_n)(X_1,\ldots,X_n)^\top=\sum_{j=1}^n X_j^2$ ist im Intervall $(-\infty,1/2)$ wohldefiniert, und es gilt für jedes $t<1/2$, dass

$$\begin{aligned}\psi_Z(t)&=\mathbb{E}\exp\Bigl(t\sum_{j=1}^n X_j^2\Bigr)=\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\Bigl(t\sum_{j=1}^n x_j^2\Bigr)\prod_{j=1}^n\frac{1}{\sqrt{2\pi}}\exp\Bigl(\frac{1}{2}\,(x_j-\mu_j)^2\Bigr)\,dx_1\ldots dx_n\\&=\Bigl(\frac{1}{2\pi}\Bigr)^{n/2}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\Bigl(t\sum_{j=1}^n x_j^2-\frac{1}{2}\sum_{j=1}^n(x_j-\mu_j)^2\Bigr)\,dx_1\ldots dx_n\\&=\prod_{j=1}^n\int_{-\infty}^{\infty}(2\pi)^{-1/2}\exp\Bigl(tx_j^2-\frac{1}{2}(x_j-\mu_j)^2\Bigr)\,dx_j\,.\end{aligned}$$

- Dabei lässt sich der Exponent des letzten Ausdruckes wie folgt umformen:

$$\begin{aligned}tx_j^2-\frac{1}{2}(x_j-\mu_j)^2&=-\frac{1}{2}\,(-2tx_j^2+x_j^2-2x_j\mu_j+\mu_j^2)\\&=-\frac{1}{2}\Bigl(x_j^2(1-2t)-2x_j\mu_j+\mu_j^2(1-2t)^{-1}+\mu_j^2-\mu_j^2(1-2t)^{-1}\Bigr)\\&=-\frac{1}{2}\Bigl((x_j-\mu_j(1-2t)^{-1})^2(1-2t)+\mu_j^2(1-(1-2t)^{-1})\Bigr)\,.\end{aligned}$$

- Somit gilt

$$\begin{aligned}\psi_Z(t)&=\exp\Bigl(-\frac{1}{2}\,(1-(1-2t)^{-1})\sum_{j=1}^n\mu_j^2\Bigr)\prod_{j=1}^n\int_{-\infty}^{\infty}(2\pi)^{-1/2}\exp\Bigl(-\frac{(x_j-\mu_j(1-2t)^{-1})^2}{2(1-2t)^{-1}}\Bigr)\,dx_j\\&=(1-2t)^{-n/2}\exp\Bigl(-\frac{\lambda}{2}\,(1-(1-2t)^{-1})\Bigr)\,,\end{aligned}$$

  weil unter dem Integral die Dichte der eindimensionalen Normalverteilung (bis auf den konstanten Faktor $(1-2t)^{1/2}$) steht; $\lambda=\boldsymbol{\mu}^\top\boldsymbol{\mu}$.
- Andererseits ergibt sich für die momenterzeugende Funktion $\psi(t)$ der in (32) gegebenen Dichte $f_{Z_{n,\lambda}}(z)$, dass

$$\psi(t)=\sum_{j=0}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^j}{j!}\int_0^{\infty}e^{tz}\,\frac{z^{n/2+j-1}e^{-z/2}}{2^{\frac{n}{2}+j}\,\Gamma\bigl(\frac{n}{2}+j\bigr)}\,dz\,,$$

  wobei das Integral die momenterzeugende Funktion der (zentralen) χ^2–Verteilung χ^2_{n+2j} mit $n+2j$ Freiheitsgraden ist.

- Ähnlich wie die charakteristische Funktion (vgl. Theorem I–1.5) ist die momenterzeugende Funktion dieser Verteilung gegeben durch

$$\psi_{\chi^2_{n+2j}}(t) = \frac{1}{(1-2t)^{n/2+j}}\,.$$

- Somit gilt

$$\int\limits_0^\infty e^{tz}\,\frac{z^{n/2+j-1}e^{-z/2}}{2^{\frac{n}{2}+j}\,\Gamma\Bigl(\frac{n}{2}+j\Bigr)}\,dz = \frac{1}{(1-2t)^{n/2+j}}\,,$$

  bzw.

$$\begin{aligned}\psi(t) &= e^{-\lambda/2}(1-2t)^{-n/2}\sum_{j=0}^\infty \frac{1}{j!}\,\Bigl(\frac{\lambda}{2}(1-2t)^{-1}\Bigr)^j\\ &= (1-2t)^{-n/2}\exp\Bigl(-\,\frac{\lambda}{2}\,(1-(1-2t)^{-1})\Bigr)\,.\end{aligned}$$

- Somit gilt $\psi(t) = \psi_Z(t)$ für jedes $t < 1/2$, und die Behauptung folgt aus Lemma 1.12. □

1.3.3 Verteilungs– und Unabhängigkeitseigenschaften linearer und quadratischer Formen

- *Zur Erinnerung*: Bei der Definition der nichtzentralen χ^2–Verteilung in Abschnitt 1.3.2 wurde die Quadratsumme der Komponenten von $\mathrm{N}(\boldsymbol{\mu},\mathbf{I})$-verteilten Zufallsvektoren betrachtet.
- Man kann nun zeigen, dass die (entsprechend modifizierte) Quadratsumme auch dann eine nichtzentrale χ^2–Verteilung besitzt, wenn der betrachtete normalverteilte Zufallsvektor eine *beliebige* positiv definite Kovarianzmatrix hat.
- Und zwar sei $\boldsymbol{\mu}\in\mathbb{R}^n$, und sei $\mathbf{K}$ eine symmetrische und positiv definite $n\times n$ Matrix.
- Wenn $\mathbf{Z} = (Z_1,\ldots,Z_n)^\top \sim \mathrm{N}(\boldsymbol{\mu},\mathbf{K})$, dann ergibt sich aus Theorem 1.3, dass

$$\mathbf{K}^{-1/2}\mathbf{Z}\sim \mathrm{N}(\mathbf{K}^{-1/2}\boldsymbol{\mu},\mathbf{I})\,.$$

- Aus der Definition der nichtzentralen χ^2– Verteilung folgt somit, dass

$$\mathbf{Z}^\top\mathbf{K}^{-1}\mathbf{Z} = \bigl(\mathbf{K}^{-1/2}\mathbf{Z}\bigr)^\top\mathbf{K}^{-1/2}\mathbf{Z}\sim\chi^2_{n,\lambda}\,, \tag{33}$$

  wobei $\lambda = (\mathbf{K}^{-1/2}\boldsymbol{\mu})^\top\mathbf{K}^{-1/2}\boldsymbol{\mu} = \boldsymbol{\mu}^\top\mathbf{K}^{-1}\boldsymbol{\mu}$.

Die Verteilungseigenschaft (33) für quadratische Formen von normalverteilten Zufallsvektoren lässt sich wie folgt verallgemeinern. Dabei ist Lemma 1.7 über die Faktorisierung symmetrischer und nichtnegativ definiter Matrizen nützlich.

Theorem 1.9

- *Sei* $\mathbf{Z} = (Z_1,\ldots,Z_n)^\top \sim \mathrm{N}(\boldsymbol{\mu},\mathbf{K})$, *wobei die Kovarianzmatrix* $\mathbf{K}$ *positiv definit sei. Außerdem sei* $\mathbf{A}$ *eine symmetrische* $n\times n$ *Matrix mit* $\mathrm{rg}(\mathbf{A}) = r\le n$.
- *Wenn die Matrix* $\mathbf{AK}$ *idempotent ist, d.h., wenn* $\mathbf{AK} = (\mathbf{AK})^2$, *dann gilt* $\mathbf{Z}^\top\mathbf{AZ}\sim\chi^2_{r,\lambda}$, *wobei* $\lambda = \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}$.

Beweis

- Die Matrix $\mathbf{AK}$ sei idempotent. Dann gilt
$$\mathbf{AK} = \mathbf{AKAK}\,.$$
- Weil $\mathbf{K}$ regulär ist, kann man beide Seiten dieser Gleichung von rechts mit $\mathbf{K}^{-1}$ multiplizieren. Dabei ergibt sich, dass
$$\mathbf{A} = \mathbf{AKA} \tag{34}$$
bzw. für jedes $\mathbf{x} \in \mathbb{R}^n$
$$\mathbf{x}^\top\mathbf{Ax} = \mathbf{x}^\top\mathbf{AKAx} = (\mathbf{Ax})^\top\mathbf{K}(\mathbf{Ax}) \ge 0\,,$$
d.h., $\mathbf{A}$ ist nichtnegativ definit.
- Gemäß Lemma 1.7 gibt es somit eine Zerlegung
$$\mathbf{A} = \mathbf{HH}^\top\,, \tag{35}$$
so dass die $n \times r$ Matrix $\mathbf{H}$ den vollen Spaltenrang r hat.
- Wegen Lemma 1.2 bedeutet dies, dass die inverse Matrix $(\mathbf{H}^\top\mathbf{H})^{-1}$ existiert.
- Aus Theorem 1.3 über die lineare Transformation von normalverteilten Zufallvektoren ergibt sich nun für den r–dimensionalen Vektor $\mathbf{Z}' = \mathbf{H}^\top\mathbf{Z}$, dass
$$\mathbf{Z}' \sim \mathrm{N}(\mathbf{H}^\top\boldsymbol{\mu}, \mathbf{I}_r)\,, \tag{36}$$
weil
$$\begin{aligned}\mathbf{H}^\top\mathbf{KH} &= (\mathbf{H}^\top\mathbf{H})^{-1}(\mathbf{H}^\top\mathbf{H})(\mathbf{H}^\top\mathbf{KH})(\mathbf{H}^\top\mathbf{H})(\mathbf{H}^\top\mathbf{H})^{-1}\\ &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top(\mathbf{AKA})\mathbf{H}(\mathbf{H}^\top\mathbf{H})^{-1}\\ &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{AH}(\mathbf{H}^\top\mathbf{H})^{-1} = \mathbf{I}_r\,,\end{aligned}$$
wobei sich die letzten drei Gleichheiten aus (34) bzw. (35) ergeben.
- Weil andererseits
$$\mathbf{Z}^\top\mathbf{AZ} \;=\; \mathbf{Z}^\top\mathbf{HH}^\top\mathbf{Z} \;=\; \bigl(\mathbf{H}^\top\mathbf{Z}\bigr)^\top\mathbf{H}^\top\mathbf{Z} \;=\; (\mathbf{Z}')^\top\mathbf{Z}'$$
und weil
$$\bigl(\mathbf{H}^\top\boldsymbol{\mu}\bigr)^\top\mathbf{H}^\top\boldsymbol{\mu} = \boldsymbol{\mu}^\top\mathbf{HH}^\top\boldsymbol{\mu} = \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}\,,$$
ergibt sich die Behauptung nun aus (36) und aus der Definition der nichtzentralen χ^2–Verteilung. $\square$

Außerdem ist das folgende Kriterium für die Unabhängigkeit von linearen bzw. quadratischen Formen normalverteilter Zufallsvektoren nützlich. Es kann als (vektorielle) Verallgemeinerung von Lemma 5.3 im Skript zur Vorlesung „Statistik I" aufgefasst werden.

Theorem 1.10

- *Sei* $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$, *wobei* $\mathbf{K}$ *eine beliebige (symmetrische, nichtnegativ definite) Kovarianzmatrix sei.*
- *Außerdem seien* $\mathbf{A}$, $\mathbf{B}$ *beliebige* $r_1 \times n$ *bzw.* $r_2 \times n$ *Matrizen mit* $r_1, r_2 \le n$, *und sei* $\mathbf{C}$ *eine symmetrische und nichtnegativ definite* $n \times n$ *Matrix.*
- *Wenn zusätzlich die Bedingung*
$$\mathbf{AKB}^\top = \mathbf{0} \qquad \textit{bzw.} \qquad \mathbf{AKC} = \mathbf{0} \tag{37}$$
erfüllt ist, dann sind die Zufallsvariablen $\mathbf{AZ}$ *und* $\mathbf{BZ}$ *bzw.* $\mathbf{AZ}$ *und* $\mathbf{Z}^\top\mathbf{CZ}$ *unabhängig.*

Beweis

- Wir zeigen zunächst, dass aus (37) die Unabhängigkeit der linearen Formen $\mathbf{AZ}$ und $\mathbf{BZ}$ folgt.
- Wegen des Eindeutigkeitssatzes für charakteristische Funktionen von Zufallsvektoren (vgl. Lemma 1.9) genügt es zu zeigen, dass für beliebige $\mathbf{t}_1 \in \mathbb{R}^{r_1}$, $\mathbf{t}_2 \in \mathbb{R}^{r_2}$

$$\mathbb{E}\,\exp\bigl(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{AZ} + \mathbf{t}_2^\top\mathbf{BZ})\bigr) = \mathbb{E}\,\exp\bigl(\mathrm{i}\,\mathbf{t}_1^\top\mathbf{AZ}\bigr)\mathbb{E}\,\exp\bigl(\mathrm{i}\,\mathbf{t}_2^\top\mathbf{BZ}\bigr)\,.$$

- Aus (37) folgt, dass

$$\mathbf{BKA}^\top = \Bigl((\mathbf{BKA}^\top)^\top\Bigr)^\top = \Bigl(\mathbf{AKB}^\top\Bigr)^\top = \mathbf{0}$$

  und somit auch, dass für beliebige $\mathbf{t}_1 \in \mathbb{R}^{r_1}$, $\mathbf{t}_2 \in \mathbb{R}^{r_2}$

$$(\mathbf{t}_1^\top\mathbf{A})\mathbf{K}(\mathbf{t}_2^\top\mathbf{B})^\top = \mathbf{t}_1^\top\mathbf{AKB}^\top\mathbf{t}_2 = \mathbf{0}\,, \qquad (\mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A})^\top = \mathbf{t}_2^\top\mathbf{BKA}^\top\mathbf{t}_1 = \mathbf{0}\,. \tag{38}$$

- Aus der in Theorem 1.4 hergeleiteten Darstellungsformel (25) für die charakteristische Funktion von normalverteilten Zufallsvektoren und aus (38) ergibt sich dann, dass

$$\begin{aligned}
&\mathbb{E}\,\exp\bigl(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{AZ} + \mathbf{t}_2^\top\mathbf{BZ})\bigr) = \mathbb{E}\,\exp\bigl(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\mathbf{Z}\bigr)\\
&= \exp\Bigl(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})^\top\Bigr)\\
&= \exp\Bigl(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_1^\top\mathbf{A})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A})^\top - \frac{1}{2}\,(\mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_2^\top\mathbf{B})^\top\Bigr)\\
&= \exp\Bigl(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_1^\top\mathbf{A})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A})^\top\Bigr)\exp\Bigl(\mathrm{i}\,(\mathbf{t}_2^\top\mathbf{B})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_2^\top\mathbf{B})^\top\Bigr)\\
&= \mathbb{E}\,\exp\bigl(\mathrm{i}\,\mathbf{t}_1^\top\mathbf{AZ}\bigr)\mathbb{E}\,\exp\bigl(\mathrm{i}\,\mathbf{t}_2^\top\mathbf{BZ}\bigr)\,.
\end{aligned}$$

- Wir zeigen nun noch, dass die Unabhängigkeit von $\mathbf{AZ}$ und $\mathbf{Z}^\top\mathbf{CZ}$ aus der zweiten Bedingung in (37) folgt.
- Sei $\mathrm{rg}(\mathbf{C}) = r \le n$. Gemäß Lemma 1.7 gibt es dann eine $n \times r$ Matrix $\mathbf{H}$ mit $\mathrm{rg}(\mathbf{H}) = r$, so dass $\mathbf{C} = \mathbf{HH}^\top$.
- Aus (37) ergibt sich dann, dass $\mathbf{AKHH}^\top = \mathbf{0}$ bzw. $\mathbf{AKHH}^\top\mathbf{H} = \mathbf{0}$.
- Hieraus folgt schließlich, dass $\mathbf{AKH} = \mathbf{0}$, weil die $r \times r$ Matrix $\mathbf{H}^\top\mathbf{H}$ wegen Lemma 1.2 den (vollen) Rang $\mathrm{rg}(\mathbf{H}) = r$ hat und deshalb invertierbar ist.
- Aus dem ersten Teil des Beweises ergibt sich somit, dass die linearen Formen $\mathbf{AZ}$ und $\mathbf{H}^\top\mathbf{Z}$ unabhängig sind.
- Wegen

$$\mathbf{Z}^\top\mathbf{CZ} = \mathbf{Z}^\top\mathbf{HH}^\top\mathbf{Z} = (\mathbf{H}^\top\mathbf{Z})^\top\mathbf{H}^\top\mathbf{Z}$$

  ergibt sich nun aus dem Transformationssatz für unabhängige Zufallsvektoren (vgl. Theorem I–1.8), dass auch $\mathbf{AZ}$ und $\mathbf{Z}^\top\mathbf{CZ}$ unabhängig sind. □

2 Lineare Modelle; Designmatrix mit vollem Rang

Zur Erinnerung (vgl. Kapitel 5 der Vorlesung „Statistik I"):

- Bei der *einfachen linearen Regression* wird von zwei Datensätzen $(x_1,\ldots,x_n)\in\mathbb{R}^n$ und $(y_1,\ldots,y_n)\in\mathbb{R}^n$ ausgegangen, die stochastisch modelliert werden sollen.
- Dabei fassen wir die Vektoren $(x_1,y_1),\ldots,(x_n,y_n)$ als Realisierungen von n Zufallsvektoren $(X_1,Y_1),\ldots,(X_n,Y_n)$ auf, die typischerweise *nicht* identisch verteilt sind.
- Wir deuten die Zufallsvariablen $Y_1,\ldots,Y_n$ als *Zielvariablen* und nehmen an, dass sie auf die folgende Weise von den *Ausgangsvariablen* $X_1,\ldots,X_n$ abhängen:
$$Y_i=\varphi(X_i)+\varepsilon_i\,,\qquad \forall\, i=1,\ldots,n\,,\tag{1}$$
wobei
  - $\varphi:\mathbb{R}\to\mathbb{R}$ eine beliebige (Borel–messbare) Funktion, die so genannte *Regressionsfunktion* ist und
  - $\varepsilon_1,\ldots,\varepsilon_n:\Omega\to\mathbb{R}$ Zufallsvariablen, so genannte *Störgrößen* sind, durch die beispielsweise zufällige Messfehler modelliert werden können.
- Ein wichtiger Spezialfall liegt vor, wenn die Regressionsfunktion $\varphi:\mathbb{R}\to\mathbb{R}$ eine lineare Funktion ist, die so genannte *Regressionsgerade*, d.h., wenn es reelle Zahlen $\beta_1,\beta_2\in\mathbb{R}$ gibt mit
$$\varphi(x)=\beta_1+\beta_2x\,,\qquad \forall\, x\in\mathbb{R}\,,\tag{2}$$
wobei β_1 die *Regressionskonstante* und β_2 der *Regressionskoeffizient* genannt wird.
- Die Größen $\beta_1,\beta_2\in\mathbb{R}$ sind unbekannte Modellparameter, die aus den beobachteten Daten $(x_1,\ldots,x_n)\in\mathbb{R}^n$ und $(y_1,\ldots,y_n)\in\mathbb{R}^n$ geschätzt werden sollen.

Wir betrachten nun die folgende *multivariate* Verallgemeinerung des einfachen linearen Regressionsmodells, wobei $m,n\ge 2$ beliebige natürliche Zahlen seien, so dass $m\le n$.

- Wir nehmen an, dass die Zielvariablen $Y_1,\ldots,Y_n$ von *vektoriellen* m–dimensionalen Ausgangsvariablen $(X_{11},\ldots,X_{1m})^\top,\ldots,(X_{n1},\ldots,X_{nm})^\top$ abhängen, d.h., es gelte
$$Y_i=\varphi(X_{i1},\ldots,X_{im})+\varepsilon_i\,,\qquad \forall\, i=1,\ldots,n\,,\tag{3}$$
wobei
  - die Regressionsfunktion $\varphi:\mathbb{R}^m\to\mathbb{R}$ gegeben ist durch
$$\varphi(x_1,\ldots,x_m)=\beta_1x_1+\ldots+\beta_mx_m\,,\qquad \forall\,(x_1,\ldots,x_m)^\top\in\mathbb{R}^m\tag{4}$$
mit (unbekannten) Regressionskoeffizienten $\beta_1,\ldots,\beta_m\in\mathbb{R}$ und
  - die zufälligen Störgrößen $\varepsilon_1,\ldots,\varepsilon_n:\Omega\to\mathbb{R}$ den folgenden Bedingungen genügen:
$$\mathbb{E}\,\varepsilon_i=0\,,\qquad \operatorname{Var}\varepsilon_i=\sigma^2\,,\qquad \operatorname{Cov}(\varepsilon_i,\varepsilon_j)=0\,,\qquad \forall\, i,j=1,\ldots,n \text{ mit } i\ne j\tag{5}$$
für eine gewisse (unbekannte) Zahl $\sigma^2>0$.
- Dabei betrachten wir hier nur den Fall, dass die Ausgangsvariablen $(X_{11},\ldots,X_{1m})^\top,\ldots,(X_{n1},\ldots,X_{nm})^\top$ deterministisch sind, d.h., es gelte
$$(X_{11},\ldots,X_{1m})^\top=(x_{11},\ldots,x_{1m})^\top,\ldots,(X_{n1},\ldots,X_{nm})^\top=(x_{n1},\ldots,x_{nm})^\top$$
für gewisse Vektoren $(x_{11},\ldots,x_{1m})^\top,\ldots,(x_{n1},\ldots,x_{nm})^\top\in\mathbb{R}^m$.

Beachte

- In Matrixschreibweise lässt sich dann das in (3) und (4) gegebene Modell wie folgt formulieren:
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{6}$$
wobei
$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{7}$$
- Dabei wird $\mathbf{X}$ die *Designmatrix* des Regressionsmodells genannt.

2.1 Methode der kleinsten Quadrate

Das Ziel dieses Abschnittes besteht darin, die unbekannten Modellparameter $\beta_1, \ldots, \beta_m$ und σ^2 aus den beobachteten Daten $(x_{11}, \ldots, x_{1m})^\top, \ldots, (x_{n1}, \ldots, x_{nm})^\top \in \mathbb{R}^m$ und $(y_1 \ldots, y_n)^\top \in \mathbb{R}^n$ zu schätzen.

- Ähnlich wie in Abschnitt I–5.1 betrachten wir hierfür die *Methode der kleinsten Quadrate* zur Bestimmung von Schätzern $\widehat{\beta}_1, \ldots, \widehat{\beta}_m$ für die unbekannten Regressionskoeffizienten $\beta_1, \ldots, \beta_m$.
- Und zwar soll ein Zufallsvektor $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)^\top$ bestimmt werden, so dass der *mittlere quadratische Fehler*
$$e(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \bigl(Y_i - (\beta_1 x_{i1} + \ldots + \beta_m x_{im})\bigr)^2 \tag{8}$$
für $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ minimal wird, vgl. auch Kapitel 3 der Vorlesung „Numerik 1a" von K. Urban (Universität Ulm, Sommersemester 2005).

Beachte Außer den in (5) gemachten Modellannahmen werden zunächst keine zusätzlichen Voraussetzungen über die Verteilung der zufälligen Störgrößen $\varepsilon_1, \ldots, \varepsilon_n : \Omega \to \mathbb{R}$ benötigt.

2.1.1 Normalengleichung

Man kann leicht zeigen, dass die in (8) betrachtete Funktion $e(\boldsymbol{\beta})$ ein eindeutig bestimmtes Minimum hat, wenn die Designmatrix $\mathbf{X}$ vollen (Spalten–) Rang hat, d.h., wenn $\operatorname{rg}(\mathbf{X}) = m$ gilt.

Theorem 2.1 *Sei* $\operatorname{rg}(\mathbf{X}) = m$.

- *Der mittlere quadratische Fehler* $e(\boldsymbol{\beta})$ *in* (8) *ist genau dann minimal, wenn* $\boldsymbol{\beta}$ *Lösung der folgenden Normalengleichung ist:*
$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y}\,. \tag{9}$$
- *Dabei hat* (9) *die eindeutig bestimmte Lösung*
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\,. \tag{10}$$

Beweis

- Die in (8) gegebene Funktion $e(\boldsymbol{\beta})$ ist differenzierbar, wobei

$$e'(\boldsymbol{\beta}) = \Bigl(\frac{\partial e(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial e(\boldsymbol{\beta})}{\partial \beta_m}\Bigr)^\top = \frac{2}{n}\,\Bigl(\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top\mathbf{Y}\Bigr)$$

  und

$$e''(\boldsymbol{\beta}) = \Bigl(\frac{\partial^2 e(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}\Bigr) = \frac{2}{n}\,\mathbf{X}^\top\mathbf{X}\,.$$

- Aus $e'(\boldsymbol{\beta}) = \mathbf{o}$ ergibt sich die Normalengleichung (9).
- Außerdem folgt aus Lemma 1.2, dass $\operatorname{rg}(\mathbf{X}^\top\mathbf{X}) = m$.
  - Die $m \times m$ Matrix $\mathbf{X}^\top\mathbf{X}$ (und somit auch $e''(\boldsymbol{\beta})$) ist deshalb invertierbar und positiv definit.
  - Folglich ist $e(\boldsymbol{\beta})$ ist genau dann minimal, wenn $\boldsymbol{\beta}$ Lösung von (9) ist.
- Weil die $m \times m$ Matrix $\mathbf{X}^\top\mathbf{X}$ invertierbar ist, besitzt (9) eine eindeutig bestimmte Lösung $\widehat{\boldsymbol{\beta}}$, die durch (10) gegeben ist. □

Beachte Der Schätzer $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ für $\boldsymbol{\beta}$ ist eine Lineartransformation der Zufallsstichprobe $\mathbf{Y}$, d.h., $\widehat{\boldsymbol{\beta}}$ ist ein *linearer Schätzer.*

Beispiele (einfaches und multiples lineares Regressionsmodell)

- Für $m = 2$ und

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{11}$$

  ergibt sich das bereits in Abschnitt I–5.1 betrachtete *einfache lineare Regressionsmodell* als Spezialfall.
- Die Designmatrix $\mathbf{X}$ in (11) hat genau dann vollen Rang $\operatorname{rg}(\mathbf{X}) = 2$, wenn nicht alle $x_1, \ldots, x_n$ gleich sind.
- Der in (10) betrachtete Schätzer $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2)$ für die Regressionskonstante β_1 bzw. den Regressionskoeffizient β_2 hat dann die Form (vgl. auch Theorem I–5.1)

$$\widehat{\beta}_2 = \frac{s^2_{xy}}{s^2_{xx}}\,, \qquad \widehat{\beta}_1 = \overline{y}_n - \widehat{\beta}_2 \overline{x}_n\,, \tag{12}$$

  wobei $\overline{x}_n$, $\overline{y}_n$ die Stichprobenmittel bezeichnen, d.h.

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i\,, \qquad \overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i\,,$$

  und die Stichprobenvarianzen s^2_{xx}, s^2_{yy} bzw. die Stichprobenkovarianz s^2_{xy} gegeben sind durch

$$s^2_{xx} = \frac{1}{n-1}\sum_{i=1}^n \bigl(x_i - \overline{x}_n\bigr)^2\,, \quad s^2_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)\,, \quad s^2_{yy} = \frac{1}{n-1}\sum_{i=1}^n \bigl(y_i - \overline{y}_n\bigr)^2\,.$$

- Für $m > 2$ und

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \ldots & x_{nm} \end{pmatrix} \tag{13}$$

  ergibt sich das so genannte *multiple lineare Regressionsmodell.*

2.1.2 Güteeigenschaften des KQ–Schätzers $\widehat{\boldsymbol\beta}$

Wir setzen von jetzt an in Abschnitt 2.1 stets voraus, dass die Designmatrix $\mathbf{X}$ vollen (Spalten–) Rang hat und leiten drei verschiedene *Güteeigenschaften* des in (10) gegebenen KQ–Schätzers $\widehat{\boldsymbol\beta} = (\widehat\beta_1, \ldots, \widehat\beta_m)^\top$ her.

Theorem 2.2 *Der Schätzer $\widehat{\boldsymbol\beta}$ ist erwartungstreu für $\boldsymbol\beta$, d.h., es gilt $\mathbb{E}\,\widehat{\boldsymbol\beta} = \boldsymbol\beta$ für jedes $\boldsymbol\beta \in \mathbb{R}^m$.*

Beweis Wegen $\mathbb{E}\,\boldsymbol\varepsilon = \mathbf{o}$ ergibt sich aus (6) und (10), dass

$$\begin{aligned}\mathbb{E}\,\widehat{\boldsymbol\beta} &\overset{(10)}{=} \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\right) \overset{(6)}{=} \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol\beta+\boldsymbol\varepsilon)\right) = \boldsymbol\beta + \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol\varepsilon\right)\\ &= \boldsymbol\beta + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\,\boldsymbol\varepsilon = \boldsymbol\beta\,.\end{aligned}$$

□

Der KQ–Schätzer $\widehat{\boldsymbol\beta}$ besitzt außerdem die folgende Eigenschaft der *Varianzminimalität*. Dabei bezeichne $\mathcal{L}$ die Familie aller erwartungstreuen linearen Schätzer $\widetilde{\boldsymbol\beta} = \mathbf{A}\mathbf{Y} + \mathbf{a}$ für $\boldsymbol\beta$, wobei $\mathbf{A}$ eine $(m \times n)$–dimensionale Matrix ist und $\mathbf{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$.

Theorem 2.3 *Für jedes $\widetilde{\boldsymbol\beta} = (\widetilde\beta_1, \ldots, \widetilde\beta_m) \in \mathcal{L}$ gilt*

$$\operatorname{Var}\widehat\beta_i \le \operatorname{Var}\widetilde\beta_i\,, \qquad \forall\, i = 1, \ldots, m\,, \tag{14}$$

wobei die Gleichheit in (14) *genau dann für jedes $i = 1, \ldots, m$ gilt, wenn $\widetilde{\boldsymbol\beta} = \widehat{\boldsymbol\beta}$.*

Beweis

- Weil vorausgesetzt wird, dass der lineare Schätzer $\widetilde{\boldsymbol\beta} = \mathbf{A}\mathbf{Y} + \mathbf{a}$ erwartungstreu für $\boldsymbol\beta$ ist, gilt

  $$\boldsymbol\beta = \mathbb{E}\,\widetilde{\boldsymbol\beta} = \mathbb{E}\,(\mathbf{A}\mathbf{Y}+\mathbf{a}) \overset{(6)}{=} \mathbb{E}\left(\mathbf{A}(\mathbf{X}\boldsymbol\beta+\boldsymbol\varepsilon)\right) + \mathbf{a} = \mathbf{A}\mathbf{X}\boldsymbol\beta + \mathbf{A}\mathbb{E}\,\boldsymbol\varepsilon + \mathbf{a} = \mathbf{A}\mathbf{X}\boldsymbol\beta + \mathbf{a}$$

  für jedes $\boldsymbol\beta \in \mathbb{R}^m$, wobei sich die letzte Gleichheit aus $\mathbb{E}\,\boldsymbol\varepsilon = \mathbf{o}$ ergibt.
- Hieraus folgt, dass

  $$\mathbf{A}\mathbf{X} = \mathbf{I} \qquad \text{und} \qquad \mathbf{a} = \mathbf{o}\,. \tag{15}$$

- Somit gilt

  $$\widetilde{\boldsymbol\beta} = \mathbf{A}\mathbf{Y} = \mathbf{A}(\mathbf{X}\boldsymbol\beta+\boldsymbol\varepsilon) = \mathbf{A}\mathbf{X}\boldsymbol\beta + \mathbf{A}\boldsymbol\varepsilon = \boldsymbol\beta + \mathbf{A}\boldsymbol\varepsilon\,,$$

  d.h., jeder lineare erwartungstreue Schätzer $\widetilde{\boldsymbol\beta}$ für $\boldsymbol\beta$ hat die Form

  $$\widetilde{\boldsymbol\beta} = \boldsymbol\beta + \mathbf{A}\boldsymbol\varepsilon\,. \tag{16}$$

- Für die Kovarianzmatrix $\operatorname{Cov}(\widetilde{\boldsymbol\beta})$ des Zufallsvektors $\widetilde{\boldsymbol\beta}$ gilt also

  $$\operatorname{Cov}(\widetilde{\boldsymbol\beta}) = \mathbb{E}\left((\widetilde{\boldsymbol\beta}-\boldsymbol\beta)(\widetilde{\boldsymbol\beta}-\boldsymbol\beta)^\top\right) = \mathbb{E}\left((\mathbf{A}\boldsymbol\varepsilon)(\mathbf{A}\boldsymbol\varepsilon)^\top\right) = \mathbf{A}\mathbb{E}\,(\boldsymbol\varepsilon\boldsymbol\varepsilon^\top)\mathbf{A}^\top = \sigma^2\mathbf{A}\mathbf{A}^\top\,,$$

  d.h.

  $$\operatorname{Cov}(\widetilde{\boldsymbol\beta}) = \sigma^2\mathbf{A}\mathbf{A}^\top\,. \tag{17}$$

- Außerdem ergibt sich aus (17) ergibt sich mit $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, dass die Kovarianzmatrix $\operatorname{Cov}(\widehat{\boldsymbol{\beta}})$ des KQ–Schätzers $\widehat{\boldsymbol{\beta}}$ gegeben ist durch
$$\operatorname{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\,, \tag{18}$$
denn es gilt
$$\begin{aligned}\operatorname{Cov}(\widehat{\boldsymbol{\beta}}) &= \sigma^2\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)^\top\\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\,.\end{aligned}$$
- Um die Gültigkeit von (14) zu beweisen, ist somit zu zeigen, dass
$$\big((\mathbf{X}^\top\mathbf{X})^{-1}\big)_{ii} \le \big(\mathbf{A}\mathbf{A}^\top\big)_{ii}\,, \qquad \forall\, i = 1,\ldots,m\,. \tag{19}$$
- Mit $\mathbf{D} = \mathbf{A} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ gilt
$$\begin{aligned}\mathbf{A}\mathbf{A}^\top &= \big(\mathbf{D} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\big(\mathbf{D} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)^\top\\ &= \mathbf{D}\mathbf{D}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top + \mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}\\ &= \mathbf{D}\mathbf{D}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}\,,\end{aligned}$$
denn wegen (15) gilt
$$\mathbf{D}\mathbf{X} = \big(\mathbf{A} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\mathbf{X} = \mathbf{A}\mathbf{X} - \mathbf{I} = \mathbf{I} - \mathbf{I} = \mathbf{0}\,,$$
wobei $\mathbf{0}$ die Nullmatrix bezeichnet.
- Weil mit $\mathbf{D} = (d_{ij})$ die Ungleichung $\big(\mathbf{D}\mathbf{D}^\top\big)_{ii} = \sum_{j=1}^m d_{ij}^2 \ge 0$ gilt, ergibt sich hieraus die Gültigkeit von (19).
- Außerdem wird klar, dass die Gleichheit in (19) für jedes $i = 1,\ldots,m$ genau dann gilt, wenn $\mathbf{D} = \mathbf{0}$, d.h. $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. □

Beachte

- Aus den Theoremen 2.1 und 2.2 folgt, dass $\widehat{\boldsymbol{\beta}} \in \mathcal{L}$. Aus Theorem 2.3 ergibt sich außerdem, dass $\widehat{\boldsymbol{\beta}}$ im Sinne von (14) *bester erwartungstreuer linearer Schätzer* für $\boldsymbol{\beta}$ ist.
- Wir leiten nun noch eine hinreichende Bedingung dafür her, dass $\widehat{\boldsymbol{\beta}}$ ein schwach konsistenter Schätzer für $\boldsymbol{\beta}$ ist, wobei der Stichprobenumfang n, d.h. die Anzahl der Zeilen der Designmatrix $\mathbf{X} = \mathbf{X}_n$ gegen ∞ strebt.
- *Zur Erinnerung*: Ein Schätzer $\widetilde{\boldsymbol{\beta}}_n = \widetilde{\boldsymbol{\beta}}(Y_1,\ldots,Y_n)$ für $\boldsymbol{\beta}$ heißt *schwach konsistent*, wenn
$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}}(|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| > \varepsilon) = 0\,, \qquad \forall\, \varepsilon > 0,\ \boldsymbol{\beta} \in \mathbb{R}^m\,.$$
- Unter ähnlichen Bedingungen kann man auch zeigen, dass $\widehat{\boldsymbol{\beta}}_n$ asymptotisch normalverteilt ist, wenn $n \to \infty$ (vgl. Abschnitt III.3.2 in Pruscha (2000)).

Theorem 2.4 *Sei $f : \mathbb{N} \to \mathbb{R} \setminus \{0\}$ eine Funktion mit $\lim_{n\to\infty} f(n) = 0$, so dass der Grenzwert*
$$\mathbf{Q} = \lim_{n\to\infty}\big(f(n)\mathbf{X}_n^\top\mathbf{X}_n\big) \tag{20}$$
existiert und die $m \times m$ Matrix $\mathbf{Q}$ invertierbar ist. Dann ist $\widehat{\boldsymbol{\beta}}_n$ ein schwach konsistenter Schätzer für $\boldsymbol{\beta}$.

Beweis

- Weil $\widehat{\boldsymbol{\beta}}_n$ erwartungstreu ist (vgl. Theorem 2.2), gilt für jedes $n \ge m$

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\beta}}(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| > \varepsilon) &= \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}|^2 > \varepsilon^2) \;=\; \mathbb{P}_{\boldsymbol{\beta}}\Bigl(\sum_{i=1}^m \bigl(\widehat{\beta}_{in} - \beta_i\bigr)^2 > \varepsilon^2\Bigr)\\
&\le \mathbb{P}_{\boldsymbol{\beta}}\Bigl(\bigcup_{i=1}^m \{(\widehat{\beta}_{in} - \beta_i)^2 > \frac{\varepsilon^2}{m}\}\Bigr) \;\le\; \sum_{i=1}^m \mathbb{P}_{\boldsymbol{\beta}}\Bigl(\bigl(\widehat{\beta}_{in} - \beta_i\bigr)^2 > \frac{\varepsilon^2}{m}\Bigr)\\
&\le \frac{m}{\varepsilon^2}\sum_{i=1}^m \operatorname{Var}\widehat{\beta}_{in}\,,
\end{aligned}$$

  wobei sich die letzte Abschätzung aus der Tschebyschev–Ungleichung ergibt (vgl. Theorem WR-4.18).
- Es genügt somit zu zeigen, dass

$$\lim_{n\to\infty} \operatorname{Var}\widehat{\beta}_{in} = 0\,, \qquad \forall\, i = 1, \ldots, m\,. \tag{21}$$

- Die Matrix $\mathbf{Q}^{-1}$ ist wohldefiniert, weil vorausgesetzt wird, dass die (Grenz–) Matrix $\mathbf{Q}$ invertierbar ist. Außerdem ergibt sich aus (20), dass

$$\mathbf{Q}^{-1} = \lim_{n\to\infty}\bigl(f(n)\mathbf{X}_n^\top\mathbf{X}_n\bigr)^{-1}.$$

- Aus der in (18) hergeleiteten Formel für die Kovarianzmatrix des Zufallsvektors $\widehat{\beta}_n$ ergibt sich nun, dass

$$\lim_{n\to\infty}\operatorname{Cov}(\widehat{\beta}_n) = \sigma^2 \lim_{n\to\infty}\bigl(\mathbf{X}_n^\top\mathbf{X}_n\bigr)^{-1} = \sigma^2 \lim_{n\to\infty} f(n) \lim_{n\to\infty}\bigl(f(n)\mathbf{X}_n^\top\mathbf{X}_n\bigr)^{-1} = \Bigl(\sigma^2 \lim_{n\to\infty} f(n)\Bigr)\mathbf{Q}^{-1} = \mathbf{0}\,.$$

- Hieraus ergibt sich insbesondere die Gültigkeit von (21). □

2.1.3 Erwartungstreue Schätzung der Varianz σ^2 der Störgrößen

- Neben den in (5) formulierten Bedingungen an die Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ gelte nun $n > m$, wobei wir erneut voraussetzen, dass die Designmatrix $\mathbf{X}$ vollen Rang hat, d.h. $\operatorname{rg}(\mathbf{X}) = m$.
- In Verallgemeinerung des Ansatzes, den wir in Abschnitt I–5.1.3 bei der Schätzung von σ^2 im einfachen linearen Regressionsmodell betrachtet hatten, setzen wir nun

$$S^2 = \frac{1}{n-m}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,. \tag{22}$$

- Bei normalverteilten Störgrößen kann S^2 als eine modifizierte Version eines Maximum–Likelihood–Schätzers für σ^2 aufgefasst werden; vgl. Abschnitt 2.2.

Wir zeigen, dass durch (22) ein erwartungstreuer Schätzer für σ^2 gegeben ist. Hierbei sind die folgenden Hilfssätze nützlich.

Lemma 2.1 *Die $n \times n$ Matrix*

$$\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \tag{23}$$

ist idempotent und symmetrisch, d.h., es gilt

$$\mathbf{G} = \mathbf{G}^2 \qquad \text{und} \qquad \mathbf{G} = \mathbf{G}^\top\,. \tag{24}$$

Beweis

- Die zweite Teilaussage in (24) ergibt sich unmittelbar aus der Definition von $\mathbf{G}$ und den Rechenregeln für transponierte Matrizen, denn es gilt
$$\mathbf{G}^\top = \Bigl(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Bigr)^\top = \mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{G}\,.$$
- Außerdem gilt
$$\begin{aligned}\mathbf{G}^2 &= \Bigl(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Bigr)\Bigl(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Bigr)\\ &= \mathbf{I}-2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top+\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\\ &= \mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top=\mathbf{G}\,.\end{aligned}$$ □

Lemma 2.2 *Für die in* (23) *gegebene $n\times n$ Matrix $\mathbf{G}$ gilt* $\mathrm{sp}(\mathbf{G})=n-m$.

Beweis

- Man kann sich leicht überlegen, dass
  - $\mathrm{sp}(\mathbf{A}-\mathbf{B})=\mathrm{sp}(\mathbf{A})-\mathrm{sp}(\mathbf{B})$ für beliebige $n\times n$ Matrizen $\mathbf{A}$ und $\mathbf{B}$,
  - $\mathrm{sp}(\mathbf{CD})=\mathrm{sp}(\mathbf{DC})$ für beliebige $n\times m$ Matrizen $\mathbf{C}$ und beliebige $m\times n$ Matrizen $\mathbf{D}$.
- Hieraus und aus der Definitionsgleichung (23) von $\mathbf{G}$ ergibt sich, dass
$$\begin{aligned}\mathrm{sp}(\mathbf{G}) &= \mathrm{sp}\Bigl(\mathbf{I}_n-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Bigr) = \mathrm{sp}(\mathbf{I}_n)-\mathrm{sp}\Bigl(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Bigr)\\ &= \mathrm{sp}(\mathbf{I}_n)-\mathrm{sp}\Bigl(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Bigr) = \mathrm{sp}(\mathbf{I}_n)-\mathrm{sp}(\mathbf{I}_m) = n-m\,,\end{aligned}$$
wobei $\mathbf{I}_\ell$ die $(\ell\times\ell)$–dimensionale Einheitsmatrix bezeichnet. □

Theorem 2.5 *Es gilt $\mathbb{E}\,S^2=\sigma^2$ für jedes $\sigma^2>0$, d.h., S^2 ist ein erwartungstreuer Schätzer für σ^2.*

Beweis

- Offenbar gilt
$$\mathbf{GX}=\bigl(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\bigr)\mathbf{X}=\mathbf{0}\,. \tag{25}$$
- Hieraus ergibt sich mit Hilfe von (10) und (23), dass
$$\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol\beta} \overset{(10)}{=} \mathbf{Y}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \overset{(23)}{=} \mathbf{GY} = \mathbf{GX}\boldsymbol\beta+\mathbf{G}\boldsymbol\varepsilon \overset{(25)}{=} \mathbf{G}\boldsymbol\varepsilon\,.$$
- Für den in (22) eingeführten Schätzer S^2 gilt somit wegen $\mathbf{G}^\top\mathbf{G}=\mathbf{G}^2=\mathbf{G}$ (vgl. Lemma 2.1), dass
$$\begin{aligned}S^2 &= \frac{1}{n-m}\,(\mathbf{G}\boldsymbol\varepsilon)^\top(\mathbf{G}\boldsymbol\varepsilon)=\frac{1}{n-m}\,\boldsymbol\varepsilon^\top\mathbf{G}^\top\mathbf{G}\boldsymbol\varepsilon=\frac{1}{n-m}\,\boldsymbol\varepsilon^\top\mathbf{G}\boldsymbol\varepsilon\\ &= \frac{1}{n-m}\,\mathrm{sp}\bigl(\boldsymbol\varepsilon^\top\mathbf{G}\boldsymbol\varepsilon\bigr)=\frac{1}{n-m}\,\mathrm{sp}\bigl(\mathbf{G}\boldsymbol\varepsilon\boldsymbol\varepsilon^\top\bigr)\,.\end{aligned}$$
- Wegen $\mathbb{E}\,(\boldsymbol\varepsilon\boldsymbol\varepsilon^\top)=\sigma^2\mathbf{I}_n$ ergibt sich hieraus, dass
$$\mathbb{E}\,S^2=\frac{1}{n-m}\,\mathrm{sp}\bigl(\mathbf{G}\mathbb{E}\,(\boldsymbol\varepsilon\boldsymbol\varepsilon^\top)\bigr)=\frac{1}{n-m}\,\mathrm{sp}\bigl(\mathbf{G}\sigma^2\mathbf{I}_n\bigr)=\frac{\sigma^2}{n-m}\,\mathrm{sp}\bigl(\mathbf{G}\bigr)=\sigma^2\,,$$
wobei sich die letzte Gleichheit aus Lemma 2.2 ergibt. □

2.2 Normalverteilte Störgrößen

- Zusätzlich zu den Modellannahmen, die am Anfang von Kapitel 2 formuliert worden sind, setzen wir in diesem Abschnitt noch voraus, dass die zufälligen Störgrößen $\varepsilon_1, \ldots, \varepsilon_n : \Omega \to \mathbb{R}$ unabhängig und normalverteilt sind, d.h. $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ für jedes $i = 1, \ldots, n$.
- Darüber hinaus gelte auch $\mathrm{rg}(\mathbf{X}) = m$ und $n > m$.
- Gemäß Theorem 1.3 ist dann die Verteilung des Vektors $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ der Zielvariablen bzw. des KQ–Schätzers $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ gegeben durch

$$\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}) \tag{26}$$

und

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\bigl(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\bigr)\,. \tag{27}$$

2.2.1 Maximum–Likelihood–Schätzer

- Durch (26) ist ein *parametrisches Modell* für die Verteilung des Vektors $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ der Stichprobenvariablen $Y_1, \ldots, Y_n$ gegeben.
- Außer der Methode der kleinsten Quadrate, die in Abschnitt 2.1 diskutiert worden ist, kann nun auch die Maximum–Likelihood–Methode zur Gewinnung von Schätzern für die unbekannten Modellparameter $\boldsymbol{\beta}$ und σ^2 verwendet werden.
- Aus (1.13) und (26) ergibt sich, dass

$$f_{\mathbf{Y}}(\mathbf{y}) = \Bigl(\frac{1}{\sigma\sqrt{2\pi}}\Bigr)^n \exp\Bigl(-\frac{1}{2\sigma^2}\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Bigr) \tag{28}$$

für jedes $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$.

- Wir betrachten also die *Likelihood-Funktion*

$$L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \Bigl(\frac{1}{\sigma\sqrt{2\pi}}\Bigr)^n \exp\Bigl(-\frac{1}{2\sigma^2}\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Bigr) \tag{29}$$

bzw. die *Loglikelihood-Funktion*

$$\log L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\,\log(2\pi) - \frac{n}{2}\,\log(\sigma^2) - \frac{1}{2\sigma^2}\,|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2\,. \tag{30}$$

- Wir suchen Schätzer $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$ für $\boldsymbol{\beta}$, σ^2, so dass mit Wahrscheinlichkeit 1

$$L(\mathbf{Y}; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \sup_{\boldsymbol{\beta}\in\mathbb{R}^m,\, \sigma^2>0} L(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2) \tag{31}$$

bzw. äquivalent hierzu

$$\log L(\mathbf{Y}; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \sup_{\boldsymbol{\beta}\in\mathbb{R}^m,\, \sigma^2>0} \log L(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2)\,. \tag{32}$$

Beachte Die Maximierung in (31) bzw. (32) kann in zwei Schritten erfolgen: zuerst bezüglich $\boldsymbol{\beta}$ und dann bezüglich σ^2. Wegen (30) ist der erste Schritt identisch mit dem in Abschnitt 2.1.1 betrachteten Minimierungsverfahren.

Theorem 2.6 *Die Lösung des Maximierungsproblems* (31) *bzw.* (32) *ist eindeutig bestimmt und gegeben durch*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \tag{33}$$

bzw.

$$\widehat{\sigma}^2 = \frac{1}{n}\,\bigl(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr)^\top\bigl(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr)\,. \tag{34}$$

Beweis

- Für beliebige, jedoch fest vorgegebene $\mathbf{y} \in \mathbb{R}^n$ und $\sigma^2 > 0$ betrachten wir zunächst die Abbildung
$$\mathbb{R}^m \ni \boldsymbol{\beta} \mapsto \log L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) \,. \tag{35}$$
- In Theorem 2.1 hatten wir gezeigt, dass die in (35) gegebene Abbildung das eindeutig bestimmte globale Maximum $\widehat{\boldsymbol{\beta}}(\mathbf{y}) = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ besitzt, das nicht von σ^2 abhängt.
- Für jedes (fest vorgegebene) $\mathbf{y} \in \mathbb{R}^n$ betrachten wir nun die Abbildung
$$(0,\infty) \ni \sigma^2 \mapsto \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2) \,. \tag{36}$$
- Diese Abbildung ist stetig, und es gilt offenbar
$$\lim_{\sigma^2 \to \infty} \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2) = -\infty \,.$$
- Weil $n > m$ vorausgesetzt wird, nimmt der n–dimensionale absolutstetige Zufallsvektor $\mathbf{Y}$ nur mit Wahrscheinlichkeit 0 Werte in der m–dimensionalen Teilmenge $\{\mathbf{X}y : y \in \mathbb{R}^m\}$ des $\mathbb{R}^n$ an.
- Deshalb gilt $|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}|^2 > 0$ mit Wahrscheinlichkeit 1 und somit
$$\lim_{\sigma^2 \to 0} \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2) = -\infty$$
für fast jedes $\mathbf{y} \in \mathbb{R}^n$.
- Für fast jedes $\mathbf{y} \in \mathbb{R}^n$ besitzt also die in (36) gegebene Abbildung mindestens ein globales Maximum in $(0,\infty)$.
- Für jedes dieser Maxima gilt
$$\frac{\partial \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2)}{\partial \sigma^2} = -\,\frac{n}{2\sigma^2} \;+\; \frac{1}{2\sigma^4}\,(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y})) = 0 \,.$$
- Die (eindeutig bestimmte) Lösung dieser Gleichung ist
$$\widehat{\sigma}^2(\mathbf{y}) = \frac{1}{n}\,(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y})) \,.$$
□

Beachte

- Der in Theorem 2.6 hergeleitete ML–Schätzer $\widehat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ stimmt mit dem in Theorem 2.1 hergeleiteten KQ–Schätzer überein.
- Der ML–Schätzer $\widehat{\sigma}^2$ für σ^2 unterscheidet sich dagegen von dem in Abschnitt 2.1.3 betrachteten (erwartungstreuen) Schätzer S^2 für σ^2 um einen konstanten Proportionalitätsfaktor, denn es gilt
$$\widehat{\sigma}^2 = \frac{n-m}{n}\, S^2 \,.$$

2.2.2 Verteilungs– und Unabhängigkeitseigenschaften von $\widehat{\boldsymbol{\beta}}$ und S^2

- Außer der bereits in (27) erwähnten Normalverteilungseigenschaft $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\bigl(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\bigr)$ lässt sich auch die Verteilung des Schätzers
$$S^2 = \frac{1}{n-m}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \,. \tag{37}$$
für die Varianz σ^2 der Störgrößen bestimmen.

- Hierfür benutzen wir die Darstellungsformel
$$\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{G}\boldsymbol{\varepsilon}\,, \tag{38}$$
die wir im Beweis von Theorem 2.5 gezeigt hatten, wobei $\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

Dabei benötigen wir die folgende *Rangformel*, die wir hier ohne Beweis angeben; vgl. Theorem 2.2.2 im Skript zur Vorlesung „ Lineare Algebra" von W. Lütkebohmert (Universität Ulm, Wintersemester 2005/06).

Lemma 2.3 *Sei* $\mathbf{A}$ *eine beliebige* $n \times n$ *Matrix. Dann gilt*
$$\mathrm{rg}(\mathbf{A}) = n - \dim \mathrm{Ker}(\mathbf{A})\,, \tag{39}$$
wobei $\mathrm{Ker}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{o}\}$ *und* $\dim \mathrm{Ker}(\mathbf{A})$ *die Dimension von* $\mathrm{Ker}(\mathbf{A}) \subset \mathbb{R}^n$ *bezeichnet.*

Aus der in Theorem 1.9 hergeleiteten Bedingung für die χ^2–Verteiltheit von quadratischen Formen normalverteilter Zufallsvektoren ergibt sich nun das folgende Resultat.

Theorem 2.7 *Es gilt*
$$\frac{(n-m)S^2}{\sigma^2} \sim \chi^2_{n-m}\,, \tag{40}$$
d.h., die Zufallsvariable $(n-m)S^2/\sigma^2$ *hat eine (zentrale)* χ^2*–Verteilung mit* $n-m$ *Freiheitsgraden.*

Beweis

- In Lemma 2.1 hatten wir gezeigt, dass die Matrix $\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ idempotent und symmetrisch ist.
- Hieraus und aus (38) ergibt sich, dass
$$\begin{aligned}\frac{(n-m)S^2}{\sigma^2} &= \frac{1}{\sigma^2}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \frac{1}{\sigma^2}\,(\mathbf{G}\boldsymbol{\varepsilon})^\top\mathbf{G}\boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\,\boldsymbol{\varepsilon}^\top\mathbf{G}^\top\mathbf{G}\boldsymbol{\varepsilon}\\ &= \bigl(\sigma^{-1}\boldsymbol{\varepsilon}\bigr)^\top\mathbf{G}\bigl(\sigma^{-1}\boldsymbol{\varepsilon}\bigr)\,.\end{aligned}$$
- Weil $\sigma^{-1}\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \mathbf{I})$ und weil die Matrix $\mathbf{G}\mathbf{I} = \mathbf{G}$ idempotent ist, genügt es wegen Theorem 1.9 noch zu zeigen, dass $\mathrm{rg}(\mathbf{G}) = n - m$.
- Im Beweis von Theorem 2.5 hatten wir gezeigt, dass $\mathbf{G}\mathbf{X} = \mathbf{0}$, d.h.
$$\{\mathbf{X}\mathbf{y} : \mathbf{y} \in \mathbb{R}^m\} \subset \mathrm{Ker}(\mathbf{G})\,.$$
- Andererseits gilt für jedes $\mathbf{x} \in \mathrm{Ker}(\mathbf{G})$, dass
$$\bigl(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\bigr)\mathbf{x} = \mathbf{o} \qquad \text{bzw.} \qquad \mathbf{x} = \mathbf{X}\,\underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{x}}_{=\mathbf{y}\in\mathbb{R}^m}$$
und somit $\mathrm{Ker}(\mathbf{G}) = \{\mathbf{X}\mathbf{y} : \mathbf{y} \in \mathbb{R}^m\}$.
- Hieraus folgt, dass $\dim \mathrm{Ker}(\mathbf{G}) = \mathrm{rg}(\mathbf{X}) = m$ bzw. $\mathrm{rg}(\mathbf{G}) = n - m$ wegen Lemma 2.3. □

Außerdem nutzen wir das in Theorem 1.10 hergeleitete Kriterium für die Unabhängigkeit von linearen bzw. quadratischen Formen normalverteilter Zufallsvektoren, um das folgende Resultat zu zeigen.

Theorem 2.8 *Die Schätzer* $\widehat{\boldsymbol{\beta}}$ *und* S^2 *für* $\boldsymbol{\beta}$ *bzw.* σ^2 *sind unabhängig.*

Beweis

- Aus $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ und $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ergibt sich, dass
$$\widehat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \;=\; (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} + \boldsymbol{\beta}\,.$$
- Außerdem hatten wir im Beweis von Theorem 2.7 gezeigt, dass sich der Schätzer
$$S^2 = \frac{1}{n-m}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$
als quadratische Form von $\boldsymbol{\varepsilon}$ darstellen lässt:
$$S^2 = \frac{1}{n-m}\,\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}\,, \qquad \text{wobei }\; \mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$
- Weil $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I})$ und weil
$$\Bigl((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Bigr)\Bigl(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Bigr) = \mathbf{0}\,,$$
ergibt sich aus Theorem 1.10, dass die lineare Form $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$ und die quadratische Form $\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}$ unabhängig sind.
- Damit sind auch die Zufallsvariablen $\widehat{\boldsymbol{\beta}}$ und S^2 unabhängig. □

2.2.3 Tests für die Regressionskoeffizienten; Quadratsummenzerlegung

- Mit Hilfe der Verteilungs– und Unabhängigkeitseigenschaften von linearen bzw. quadratischen Formen normalverteilter Zufallsvektoren, die in den Abschnitten 1.3.3 bzw. 2.2.2 hergeleitet wurden, kann man t–Tests bzw. F–Tests zur Verifizierung von Hypothesen über die Regressionskoeffizienten $\beta_1, \ldots, \beta_m$ konstruieren.
- Dabei verwenden wir so wie bisher die (unabhängigen) Schätzer $\widehat{\boldsymbol{\beta}}$ und S^2 für $\boldsymbol{\beta}$ bzw. σ^2, wobei
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \sim \mathrm{N}\bigl(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\bigr) \tag{41}$$
bzw.
$$\frac{(n-m)S^2}{\sigma^2} = \frac{1}{\sigma^2}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \sim \chi^2_{n-m}\,. \tag{42}$$

Zunächst diskutieren wir den folgenden F–Test, der auch *Test auf Gesamtzusammenhang* bzw. *Test auf Signifikanz des Modells* genannt wird.

- Hierbei wird die Null-Hypothese $H_0 : \beta_1 = \ldots = \beta_m = 0$ (gegen die Alternative $H_1 : \beta_j \neq 0$ für ein $j \in \{1, \ldots, m\}$) getestet.
- Die Wahl der Testgröße ist durch die folgende *Quadratsummenzerlegung* motiviert.

Theorem 2.9 *Mit der Schreibweise $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ gilt*
$$\mathbf{Y}^\top\mathbf{Y} = \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}} + \bigl(\mathbf{Y} - \widehat{\mathbf{Y}}\bigr)^\top\bigl(\mathbf{Y} - \widehat{\mathbf{Y}}\bigr)\,. \tag{43}$$

Beweis

- Es gilt

$$\begin{aligned}\mathbf{Y}^\top\mathbf{Y} &= \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n\Bigl((Y_i-\widehat{Y}_i)+\widehat{Y}_i\Bigr)^2 = \sum_{i=1}^n \widehat{Y}_i^2 + \sum_{i=1}^n (Y_i-\widehat{Y}_i)^2 + 2\underbrace{\sum_{i=1}^n (Y_i-\widehat{Y}_i)\widehat{Y}_i}_{=0}\\ &= \sum_{i=1}^n \widehat{Y}_i^2 + \sum_{i=1}^n (Y_i-\widehat{Y}_i)^2 = \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}} + \bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)^\top\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr).\end{aligned}$$

- Dabei ergibt sich die vorletzte Gleichheit aus der folgenden Überlegung: Es gilt

$$\begin{aligned}\sum_{i=1}^n (Y_i-\widehat{Y}_i)\widehat{Y}_i &= (\mathbf{Y}-\widehat{\mathbf{Y}})^\top\widehat{\mathbf{Y}} = (\mathbf{Y}^\top - \widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top)\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Y}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\top \underbrace{\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}}_{=\mathbf{X}^\top\mathbf{Y}}\\ &= \mathbf{Y}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{Y} = \bigl(\mathbf{Y}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}\bigr)^\top - \widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{Y} = 0.\end{aligned}$$

□

Beachte

- Der erste Summand $\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}$ auf der rechten Seite von (43) ist die quadrierte Länge des Vektors $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ der geschätzten Zielwerte $\widehat{Y}_1,\ldots,\widehat{Y}_n$.
- Die zweite Komponente der Quadratsummenzerlegung (43), d.h. die Summe der Abweichungsquadrate $\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)^\top\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)$, wird *Reststreuung* genannt.
- Manchmal wird auch die so genannte *Bestimmtheitsmaßzahl* R^2 betrachtet, die gegeben ist durch

$$R^2 = 1 - \frac{\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)^\top\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)}{\sum\limits_{i=1}^n (Y_i-\overline{Y})^2}, \qquad \text{wobei} \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i.$$

Aus unserer Modellannahme, dass die Designmatrix $\mathbf{X}$ vollen Rang hat, d.h. $\mathrm{rg}(\mathbf{X}) = m$, ergibt sich die Ungleichung $(\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta} > 0$, wenn die Hypothese $H_0 : \beta_1 = \ldots = \beta_m = 0$ falsch ist.

- Deshalb ist es naheliegend, die Hypothese H_0 abzulehnen, wenn die quadrierte Länge $\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}$ des Zufallsvektors $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ hinreichend groß ist.
- Dabei wird auch die Variabilität σ^2 der Daten bei der Entscheidung berücksichtigt, was „hinreichend groß" ist.
  - Unter $H_0 : \boldsymbol{\beta} = \mathbf{o}$ gilt

$$\mathbb{E}\bigl(\mathbf{Y}^\top\mathbf{Y}\bigr) = \mathbb{E}\bigl(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\bigr) = \mathbb{E}\Bigl(\sum_{i=1}^n \varepsilon_i^2\Bigr) = \sum_{i=1}^n \mathbb{E}\,\varepsilon_i^2 = n\sigma^2.$$

  - Aus Theorem 2.9 folgt in diesem Fall, dass

$$n\sigma^2 = \mathbb{E}\bigl(\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}\bigr) + \mathbb{E}\bigl(\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)^\top\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)\bigr),$$

  - weshalb bei der Überprüfung der Hypothese $H_0 : \boldsymbol{\beta} = \mathbf{o}$ der Quotient von $\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}$ und der Summe der Abweichungsquadrate $\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)^\top\bigl(\mathbf{Y}-\widehat{\mathbf{Y}}\bigr)$ betrachtet wird.

- Genauer gesagt: Wir betrachten die folgende Testgröße

$$T_{\mathrm{mod}} = \frac{\widehat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X}) \widehat{\boldsymbol{\beta}}}{m S^2} . \tag{44}$$

Um einen auf T_{mod} basierenden Test der Hypothese $H_0 : \beta_1 = \ldots = \beta_m = 0$ konstruieren zu können, muss die Verteilung der Testgröße T_{mod} bestimmt werden.

Theorem 2.10 *Unter $H_0 : \beta_1 = \ldots = \beta_m = 0$ gilt*

$$T_{\mathrm{mod}} \sim \mathrm{F}_{m,n-m} , \tag{45}$$

d.h., die in (44) *gegebene Testgröße T_{mod} ist* F*–verteilt mit $(m, n-m)$ Freiheitsgraden.*

Beweis

- Unter $H_0 : \beta_1 = \ldots = \beta_m = 0$ gilt $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$ mit $\mathbf{K} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.
- Hieraus folgt, dass $(\sigma^{-1}\mathbf{X})^\top (\sigma^{-1}\mathbf{X})\mathbf{K} = (\sigma^{-1}\mathbf{X})^\top (\sigma^{-1}\mathbf{X})\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{I}$, d.h. insbesondere, dass die Matrix $(\sigma^{-1}\mathbf{X})^\top (\sigma^{-1}\mathbf{X})\mathbf{K}$ idempotent ist.
- Aus Theorem 1.9 ergibt sich nun, dass die quadratische Form $\sigma^{-2} \widehat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X}) \widehat{\boldsymbol{\beta}}$ eine (zentrale) χ^2–Verteilung mit m Freiheitsgraden hat.
- Außerdem hatten wir in Theorem 2.7 gezeigt, dass die Zufallsvariable $(n-m)S^2/\sigma^2$ eine (zentrale) χ^2–Verteilung mit $n-m$ Freiheitsgraden hat.
- In Theorem 2.8 hatten wir gezeigt, dass $\widehat{\boldsymbol{\beta}}$ und S^2 unabhängig sind.
- Aus dem Transformationssatz für unabhängige Zufallsvektoren (vgl. Theorem I–1.8) folgt somit, dass auch die Zufallsvariablen $\sigma^{-2} \widehat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X}) \widehat{\boldsymbol{\beta}}$ und $(n-m)S^2/\sigma^2$ unabhängig sind.
- Die Behauptung ergibt sich nun aus der Definition der F–Verteilung, vgl. Abschnitt I–3.1.3. □

Beachte

- Beim Test der Hypothese $H_0 : \beta_1 = \ldots = \beta_m = 0$ zum Niveau $\alpha \in (0,1)$ (gegen die Alternative $H_1 : \beta_j \neq 0$ für ein $j \in \{1, \ldots, m\}$) wird die Nullhypothese H_0 abgelehnt, wenn

$$T_{\mathrm{mod}} > \mathrm{F}_{m,n-m,1-\alpha} , \tag{46}$$

  wobei $\mathrm{F}_{m,n-m,1-\alpha}$ das $1-\alpha$–Quantil der F–Verteilung mit $(m, n-m)$ Freiheitsgraden bezeichnet.
- Auf ähnliche Weise lässt sich ein F–Test zur Verifizierung der Hypothese $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ zum Niveau $\alpha \in (0,1)$ (gegen die Alternative $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$) für einen beliebigen hypothetischen Parametervektor $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})$ konstruieren.
- So wie im Beweis von Theorem 2.10 vorgehend kann man zeigen, dass unter $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ die Testgröße

$$T_{\boldsymbol{\beta}_0} = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{m S^2} \tag{47}$$

  F–verteilt ist mit $(m, n-m)$ Freiheitsgraden.
- Die Nullhypothese $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ wird somit abgelehnt, wenn

$$T_{\boldsymbol{\beta}_0} > \mathrm{F}_{m,n-m,1-\alpha} . \tag{48}$$

Zur Verifizierung von Hypothesen über *einzelne* Komponenten von $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ werden dagegen t–Tests verwendet.

- Sei $j \in \{1, \ldots, m\}$. Um einen hypothetischen Wert $\beta_{0,j}$ der j–ten Komponente β_j des Parametervektors $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ zu testen, betrachten wir die Testgröße

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{S\sqrt{x^{jj}}}\,, \tag{49}$$

wobei x^{ij} die (i,j)–te Eintragung der (inversen) Matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ bezeichnet.

- Aus (41) – (42) und aus der Unabhängigkeit von $\widehat{\boldsymbol{\beta}}$ und S^2 ergibt sich, dass $T_j \sim \mathrm{t}_{n-m}$.
- Beim Test der Hypothese $H_0 : \beta_j = \beta_{0,j}$ zum Niveau $\alpha \in (0,1)$ (gegen die Alternative $H_1 : \beta_j \neq \beta_{0,j}$) wird die Nullhypothese H_0 abgelehnt, wenn

$$\frac{\bigl|\widehat{\beta}_j - \beta_{0,j}\bigr|}{S\sqrt{x^{jj}}} > \mathrm{t}_{n-m,1-\alpha/2}\,, \tag{50}$$

wobei $\mathrm{t}_{n-m,1-\alpha/2}$ das $(1-\alpha/2)$–Quantil der t–Verteilung mit $n-m$ Freiheitsgraden bezeichnet.

Beachte

- Der Test der Hypothese $H_0 : \beta_j = 0$ (gegen die Alternative $H_1 : \beta_j \neq 0$) ist von besonderem Interesse, weil damit verifiziert werden kann, inwieweit die Zielvariablen $Y_1, \ldots, Y_n$ überhaupt von dem j–ten Einflussfaktor abhängen.
- Bei diesem Test auf Signifikanz des j–ten Einflussfaktors wird die Nullhypothese $H_0 : \beta_j = 0$ abgelehnt, wenn

$$\frac{\bigl|\widehat{\beta}_j\bigr|}{S\sqrt{x^{jj}}} > \mathrm{t}_{n-m,1-\alpha/2}\,. \tag{51}$$

Die bisher in diesem Abschnitt betrachteten Tests sind Spezialfälle des folgenden *universellen Tests*. Dabei wird ein *beliebiger* Teil der Komponenten des Parametervektors $\boldsymbol{\beta}$ getestet.

- Für $\ell \in \{1, \ldots, m\}$ und $\beta_{0,\ell}, \ldots, \beta_{0m} \in \mathbb{R}$ soll die Hypothese

$$H_0 : \beta_\ell = \beta_{0,\ell}, \ldots, \beta_m = \beta_{0m} \qquad \text{versus} \qquad H_1 : \beta_j \neq \beta_{0j} \text{ für ein } j \in \{\ell, \ldots, m\} \tag{52}$$

getestet werden.

- Hierfür betrachten wir die folgende $(m-\ell+1) \times (m-\ell+1)$–dimensionale Teilmatrix $\mathbf{K}_{\mathrm{uni}}$ der Matrix $(\mathbf{X}^\top\mathbf{X})^{-1} = (x^{ij})$ mit

$$\mathbf{K}_{\mathrm{uni}} = \begin{pmatrix} x^{\ell,\ell} & \ldots & x^{\ell,m} \\ \vdots & & \vdots \\ x^{m,\ell} & \ldots & x^{mm} \end{pmatrix}.$$

  - Man kann zeigen, dass die inverse Matrix $\mathbf{K}_{\mathrm{uni}}^{-1}$ wohldefiniert ist, denn es gilt $\mathbf{K}_{\mathrm{uni}} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top$, wobei $\mathbf{H} = (\mathbf{0}, \mathbf{I})$, die Nullmatrix $\mathbf{0}$ die Dimension $(m-\ell+1) \times (\ell-1)$ und die Einheitsmatrix $\mathbf{I}$ die Dimension $(m-\ell+1) \times (m-\ell+1)$ hat.
  - Hieraus und aus Lemma 1.8 folgt, dass die Matrix $\mathbf{K}_{\mathrm{uni}}$ positiv definit und damit invertierbar ist.

- Ein Ansatz zur Lösung des Testproblems (52) ist dann durch die Testgröße

$$T_{\mathrm{uni}} = \frac{\bigl(\widehat{\boldsymbol{\beta}}_{\mathrm{uni}} - \boldsymbol{\beta}_{\mathrm{uni}}\bigr)^\top \mathbf{K}_{\mathrm{uni}}^{-1} \bigl(\widehat{\boldsymbol{\beta}}_{\mathrm{uni}} - \boldsymbol{\beta}_{\mathrm{uni}}\bigr)}{(m-\ell+1)S^2} \tag{53}$$

gegeben, wobei $\widehat{\boldsymbol{\beta}}_{\mathrm{uni}} = (\widehat{\beta}_\ell, \ldots, \widehat{\beta}_m)$ und $\boldsymbol{\beta}_{\mathrm{uni}} = (\beta_{0\ell}, \ldots, \beta_{0m})$.

- Denn aus dem folgenden Theorem 2.11 ergibt sich, dass unter der in (52) formulierten Nullhypothese H_0

$$T_{\rm uni} \sim \mathrm{F}_{m-\ell+1,n-m}\,. \tag{54}$$

- Die Hypothese $H_0 : \beta_\ell = \beta_{0,\ell}, \ldots, \beta_m = \beta_{0m}$ wird somit abgelehnt, wenn

$$T_{\rm uni} > \mathrm{F}_{m-\ell+1,n-m,1-\alpha}\,. \tag{55}$$

Wir diskutieren nun noch einen allgemeinen *Test für Linearformen* des Parametervektors $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$.

- Sei $r \in \{1, \ldots, m\}$, sei $\mathbf{H}$ eine $r \times m$ Matrix mit vollem Rang $\operatorname{rg}(\mathbf{H}) = r$, und sei $\mathbf{c} \in \mathbb{R}^r$.
- Getestet werden soll die Hypothese

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c} \qquad \text{versus} \qquad H_1 : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}\,, \tag{56}$$

  wobei die folgende Testgröße $T_{\mathbf{H}}$ betrachtet wird:

$$T_{\mathbf{H}} = \frac{\bigl(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{c}\bigr)^\top \bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{c}\bigr)}{rS^2}\,. \tag{57}$$

Theorem 2.11 *Unter $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ gilt*

$$T_{\mathbf{H}} \sim \mathrm{F}_{r,n-m}\,, \tag{58}$$

d.h., die in (57) *gegebene Testgröße $T_{\mathbf{H}}$ ist* F*-verteilt mit $(r, n-m)$ Freiheitsgraden.*

Beweis

- Weil die Designmatrix $\mathbf{X}$ vollen Rang hat, ist die symmetrische Matrix $\mathbf{X}^\top\mathbf{X}$ positiv definit.
  - Gemäß Lemma 1.8 sind damit auch die Matrizen $(\mathbf{X}^\top\mathbf{X})^{-1}$ bzw. $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top$ positiv definit,
  - d.h. insbesondere, dass die Matrix $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top$ vollen Rang besitzt und deshalb invertierbar ist.
- Die in (57) betrachtete quadratische Form $\mathbf{Z}^\top(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top)^{-1}\mathbf{Z}$ ist somit wohldefiniert, wobei

$$\mathbf{Z} = \mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{c} \qquad \text{mit } \widehat{\boldsymbol{\beta}} \sim \mathrm{N}\bigl(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\bigr).$$

- Aus Theorem 1.3 ergibt sich, dass unter $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$

$$\mathbf{Z} \sim \mathrm{N}\bigl(\mathbf{o}, \sigma^2\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\bigr)\,.$$

- Außerdem ist die $r \times r$ Matrix $\mathbf{A} = \bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\bigr)^{-1}$ symmetrisch, denn es gilt

$$\begin{aligned}
\mathbf{A}^\top &= \Bigl(\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\bigr)^{-1}\Bigr)^\top = \Bigl(\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\bigr)^\top\Bigr)^{-1} = \Bigl(\mathbf{H}\bigl((\mathbf{X}^\top\mathbf{X})^{-1}\bigr)^\top\mathbf{H}^\top\Bigr)^{-1}\\
&= \Bigl(\mathbf{H}\bigl((\mathbf{X}^\top\mathbf{X})^\top\bigr)^{-1}\mathbf{H}^\top\Bigr)^{-1} = \bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\bigr)^{-1} = \mathbf{A}\,.
\end{aligned}$$

- Weil die Matrix $\bigl(\sigma^{-2}\mathbf{A}\bigr)\bigl(\sigma^2\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\bigr) = \mathbf{I}$ offenbar idempotent ist, ergibt sich aus Theorem 1.9, dass $\sigma^{-2}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}$ eine χ^2_r-verteilte Zufallsvariable ist.
- Der Rest des Beweises verläuft genauso wie der Beweis von Theorem 2.10. □

Beachte Die Nullhypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ wird abgelehnt, wenn $T_{\mathbf{H}} > \mathrm{F}_{r,n-m,1-\alpha}$, wobei $T_{\mathbf{H}}$ die in (57) gegebene Testgröße ist.

2.2.4 Konfidenzbereiche; Prognose von Zielvariablen

- *Zur Erinnerung:* In Abschnitt 2.2.3 hatten wir die Testgröße $T_j = (\widehat{\beta}_j - \beta_j)/(S\sqrt{x^{jj}})$ betrachtet, wobei x^{ij} die (i,j)-te Eintragung der (inversen) Matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ bezeichnet.
- Dabei hatten wir gezeigt, dass $T_j \sim \mathrm{t}_{n-m}$ für jedes $j \in \{1,\ldots,m\}$ gilt.
- Hieraus ergeben sich die folgenden Konfidenzintervalle zum Niveau $1-\alpha \in (0,1)$ für jeden einzelnen Regressionskoeffizienten β_j.
- Und zwar gilt jeweils mit Wahrscheinlichkeit $1-\alpha$

$$\widehat{\beta}_j - \mathrm{t}_{n-m,1-\alpha/2} S\sqrt{x^{jj}} < \beta_j < \widehat{\beta}_j + \mathrm{t}_{n-m,1-\alpha/2} S\sqrt{x^{jj}}\,. \tag{59}$$

Beachte

- Auf die gleiche Weise wie im Beweis von Theorem I–5.8 ergibt sich mit Hilfe der Bonferroni-Ungleichung (vgl. Lemma I–5.4) ein *gemeinsamer Konfidenzbereich* zum Niveau $1-\alpha \in (0,1)$ für sämtliche m Regressionskoeffizienten $\beta_1,\ldots,\beta_m$.
- Und zwar ist die Wahrscheinlichkeit, dass

$$\widehat{\beta}_j - \mathrm{t}_{n-m,1-\alpha/2m} S\sqrt{x^{jj}} < \beta_j < \widehat{\beta}_j + \mathrm{t}_{n-m,1-\alpha/2m} S\sqrt{x^{jj}} \tag{60}$$

  gleichzeitig für jedes $j = 1,\ldots,m$ gilt, mindestens gleich $1-\alpha$.
- Aus Theorem 2.10 ergibt sich außerdem ein *exakter* gemeinsamer Konfidenzbereich zum Niveau $1-\alpha$ für sämtliche m Regressionskoeffizienten $\beta_1,\ldots,\beta_m$.
  - Denn es gilt (vgl. (47) – (48)), dass

$$\mathbb{P}_{\boldsymbol{\beta}}\Bigl(\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{mS^2} < \mathrm{F}_{m,n-m,1-\alpha}\Bigr) = 1-\alpha\,.$$

  - Dabei stellt der Konfidenzbereich E mit

$$E = \Bigl\{\boldsymbol{\beta} = (\beta_1,\ldots,\beta_m):\ \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{mS^2} < \mathrm{F}_{m,n-m,1-\alpha}\Bigr\}$$

    einen (zufälligen) *Ellipsoid* dar mit dem Mittelpunkt $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1,\ldots,\widehat{\beta}_m)$.
- Man kann zeigen, dass sich der Ellipsoid E in einen m–dimensionalen achsenparallelen Quader $E' \supset E$ einbetten lässt, wobei

$$E' = \prod_{j=1}^m \Bigl(\widehat{\beta}_j - S\sqrt{mx^{jj}\mathrm{F}_{m,n-m,1-\alpha}},\ \widehat{\beta}_j + S\sqrt{mx^{jj}\mathrm{F}_{m,n-m,1-\alpha}}\Bigr)\,.$$

- Der Konfidenzbereich E' hat eine einfachere Gestalt als E. Wegen $E' \supset E$ ist E' jedoch eine ungenauere Schätzung als E.

Auf ähnliche Weise ergibt sich ein *Konfidenzintervall für den erwarteten Zielwert*

$$\varphi(x_{0,1},\ldots,x_{0,m}) = \beta_1 x_{0,1} + \ldots + \beta_m x_{0,m}\,,$$

der einem vorgegebenen Vektor $\mathbf{x}_0 = (x_{0,1},\ldots,x_{0,m})^\top \in \mathbb{R}^m$ von Werten $x_{0,1},\ldots,x_{0,m}$ der m Einflussfaktoren entspricht.

- Hierfür betrachten wir die $1 \times m$ Matrix $\mathbf{H} = (x_{0,1}, \ldots, x_{0,m})$ $(= \mathbf{x}_0^\top)$.
- Dann ergibt sich aus Theorem 2.11, dass die Zufallsvariable

$$\sqrt{T_{\mathbf{H}}} = \frac{\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - \varphi(\mathbf{x}_0)}{S\sqrt{\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}}$$

t–verteilt ist mit $n - m$ Freiheitsgraden.

- Mit Wahrscheinlichkeit $1 - \alpha$ gilt also

$$\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - Z_0 < \varphi(\mathbf{x}_0) < \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 + Z_0\,, \tag{61}$$

wobei

$$Z_0 = \mathrm{t}_{n-m,1-\alpha/2} S \sqrt{\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}\,.$$

Beachte

- Völlig analog ergibt sich ein *Prognoseintervall* für die Zielvariable $Y_0 = \beta_1 x_{0,1} + \ldots + \beta_m x_{0,m} + \varepsilon_0$, wobei die Störgröße ε_0 normalverteilt und unabhängig von den Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ ist; $\varepsilon_0 \sim \mathrm{N}(0, \sigma^2)$.
- Und zwar gilt $\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - Y_0 \sim \mathrm{N}\bigl(0, 1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0\bigr)$ und somit mit Wahrscheinlichkeit $1 - \alpha$

$$\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - Z_0' < Y_0 < \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 + Z_0'\,, \tag{62}$$

wobei $Z_0' = \mathrm{t}_{n-m,1-\alpha/2} S \sqrt{1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}$.

2.2.5 Konfidenzband

In diesem Abschnitt nehmen wir an, dass die Designmatrix $\mathbf{X}$ die Form

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \ldots & x_{nm} \end{pmatrix} \tag{63}$$

hat, d.h., wir betrachten das (multiple) lineare Regressionsmodell.

- In der Definitionsgleichung (4) für die Regressionsfunktion $\varphi(x_1, \ldots, x_m) = \beta_1 x_1 + \ldots + \beta_m x_m$ setzen wir nun $x_1 = 1$ und bestimmen ein *Konfidenzband für die Regressionshyperebene*

$$y = \varphi(1, x_2, \ldots, x_m) = \beta_1 + \beta_2 x_2 + \ldots + \beta_m x_m\,, \qquad \forall\, x_2, \ldots, x_m \in \mathbb{R}\,.$$

- Dabei ist eine Zahl $a_\gamma > 0$ gesucht, so dass mit der vorgegebenen (Überdeckungs–) Wahrscheinlichkeit $\gamma = 1 - \alpha \in (0, 1)$

$$\widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_m x_m - a_\gamma Z_{\mathbf{x}} < \varphi(1, x_2, \ldots, x_m) < \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_m x_m + a_\gamma Z_{\mathbf{x}}\,, \tag{64}$$

gleichzeitig für jedes $\mathbf{x} = (1, x_2, \ldots, x_m) \in \mathbb{R}^m$ gilt, wobei

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \qquad \text{und} \qquad Z_{\mathbf{x}} = S\sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}}\,.$$

Bei der Lösung dieser Fragestellung ist das folgende Hilfsergebnis nützlich.

Lemma 2.4 *Mit Wahrscheinlichkeit* 1 *gilt*

$$\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\bigl((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\bigr)^2}{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} = (\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})\,, \tag{65}$$

wobei $\mathbb{R}_1^{m-1}$ *die Menge aller derjenigen Vektoren* $\mathbf{x}\in\mathbb{R}^m$ *mit* $\mathbf{x}=(1,x_2,\ldots,x_m)^\top$ *bezeichnet.*

Beweis

- Aus den Lemmata 1.6 und 1.8 folgt, dass $(\mathbf{X}^\top\mathbf{X})^{-1}=\mathbf{H}\mathbf{H}^\top$ für eine invertierbare $m\times m$ Matrix $\mathbf{H}$.
- Somit kann der Ausdruck
$$(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x} = ((\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon})^\top\mathbf{H}^\top\mathbf{x}$$
als Skalarprodukt der m–dimensionalen Vektoren $(\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon}$ und $\mathbf{H}^\top\mathbf{x}$ aufgefasst werden.
- Analog gilt
$$(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} = ((\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon} \qquad\text{und}\qquad \mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x} = (\mathbf{H}^\top\mathbf{x})^\top\mathbf{H}^\top\mathbf{x}\,.$$
- Hieraus und aus der Ungleichung
$$|\mathbf{y}^\top\mathbf{z}| \le \sqrt{\mathbf{y}^\top\mathbf{y}}\,\sqrt{\mathbf{z}^\top\mathbf{z}} \qquad \forall\,\mathbf{y},\mathbf{z}\in\mathbb{R}^m \tag{66}$$
ergibt sich mit $\mathbf{y}=(\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon}$ und $\mathbf{z}=\mathbf{H}^\top\mathbf{x}$, dass
$$\bigl|(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\bigr| \le \sqrt{(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})}\,\sqrt{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}$$
bzw.
$$\frac{\bigl((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\bigr)^2}{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} \le (\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})\,. \tag{67}$$
- Weil der Zufallsvektor $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^\top$ unabhängige absolutstetige Komponenten hat, gilt $\sum_{i=1}^n\varepsilon_i\ne 0$ mit Wahrscheinlichkeit 1.
- Sei nun $\sum_{i=1}^n\varepsilon_i\ne 0$. Dann ergibt sich aus der in (63) betrachteten Form der Designmatrix $\mathbf{X}$, dass der Vektor $\mathbf{x}=\mathbf{X}^\top\boldsymbol{\varepsilon}/\sum_{i=1}^n\varepsilon_i$ zu $\mathbb{R}_1^{m-1}$ gehört und dass dann in (67) die Gleichheit gilt. □

Aus dem folgenden Resultat, das eine vektorielle Verallgemeinerung von Theorem I–5.9 ist, ergibt sich das gesuchte Konfidenzband.

Theorem 2.12 *Sei* $a_\gamma=\sqrt{m\,\mathrm{F}_{m,n-m,\gamma}}$. *Dann gilt*

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\bigl(\widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\varphi(\mathbf{x})\bigr)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} \le a_\gamma^2\right) = \gamma\,. \tag{68}$$

Beweis

- Für jedes $\mathbf{x}\in\mathbb{R}_1^{m-1}$ gilt
$$\begin{aligned}\widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\varphi(\mathbf{x}) &= \widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\boldsymbol{\beta}^\top\mathbf{x} = \bigl((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\bigr)^\top\mathbf{x}-\boldsymbol{\beta}^\top\mathbf{x} = \bigl(\boldsymbol{\beta}+(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\bigr)^\top\mathbf{x}-\boldsymbol{\beta}^\top\mathbf{x}\\ &= (\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\end{aligned}$$
und somit
$$\begin{aligned}\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\bigl(\widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\varphi(\mathbf{x})\bigr)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} &= \max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\bigl((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\bigr)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} = \frac{1}{S^2}\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\bigl((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\bigr)^2}{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}\\ &= \frac{(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})}{S^2}\,,\end{aligned}$$
wobei sich die letzte Gleichheit aus Lemma 2.4 ergibt.

- Es gilt also
$$\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\big(\widehat{\boldsymbol\beta}^\top \mathbf{x}-\varphi(\mathbf{x})\big)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} = \frac{(\mathbf{X}^\top\boldsymbol\varepsilon)^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol\varepsilon)}{S^2}\,. \tag{69}$$
- Weil $\boldsymbol\varepsilon\sim \mathrm{N}(\mathbf{o},\sigma^2\mathbf{I})$ und weil
$$\mathbf{X}^\top\Big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)=\mathbf{0}\,,$$
ergibt sich aus Theorem 1.10, dass $\mathbf{X}^\top\boldsymbol\varepsilon$ und $\boldsymbol\varepsilon^\top\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\boldsymbol\varepsilon$ unabhängig sind.
- Aus der bereits im Beweis von Theorem 2.7 hergeleiteten Darstellungsformel
$$S^2=\frac{1}{n-m}\,\boldsymbol\varepsilon^\top\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\boldsymbol\varepsilon$$
ergibt sich somit, dass auch die Zufallsvariablen $(\mathbf{X}^\top\boldsymbol\varepsilon)^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol\varepsilon)$ und S^2 unabhängig sind.
- In Theorem 2.7 hatten wir gezeigt, dass
$$(n-m)S^2/\sigma^2\sim\chi^2_{n-m}\,.$$
- Außerdem ergibt sich aus Theorem 1.9, dass
$$(\mathbf{X}^\top\boldsymbol\varepsilon)^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol\varepsilon)/\sigma^2\sim\chi^2_m\,,$$
weil die $m\times m$ (Kovarianz-) Matrix $\mathbf{X}^\top\mathbf{X}$ des normalverteilten Zufallsvektors $\mathbf{X}^\top\boldsymbol\varepsilon$ vollen Rang hat und weil die Matrix $(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})=\mathbf{I}$ idempotent ist.
- Wegen (69) haben wir also insgesamt gezeigt, dass
$$\frac{1}{m}\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\big(\widehat{\boldsymbol\beta}^\top \mathbf{x}-\varphi(\mathbf{x})\big)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}\sim \mathrm{F}_{m,n-m}\,.$$
- Für den in (64) bzw. (68) betrachteten Schwellenwert ergibt sich deshalb $a_\gamma=\sqrt{m\,\mathrm{F}_{m,n-m,\gamma}}$. □

3 Beliebige Designmatrix; verallgemeinerte Inverse

- Wir betrachten nun die folgende Verallgemeinerung des in Kapitel 2 behandelten linearen Modells

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{1}$$

für das wir bisher stets vorausgestzt hatten, dass die Designmatrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix} \tag{2}$$

eine $(n \times m)$–dimensionale Matrix mit vollem (Spalten–) Rang $\operatorname{rg}(\mathbf{X}) = m$ ist, wobei $n \geq m$.

- In diesem Kapitel werden wir dagegen den Fall $\operatorname{rg}(\mathbf{X}) \leq m$ betrachten, d.h., wir lassen zu, dass $\mathbf{X}$ keinen vollen Rang besitzt.
- So wie in Abschnitt 2.1 setzen wir zunächst über den Zufallsvektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ lediglich voraus, dass

$$\mathbb{E}\,\varepsilon_i = 0\,, \qquad \operatorname{Var}\varepsilon_i = \sigma^2\,, \qquad \operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0\,, \qquad \forall\, i, j = 1, \ldots, n \text{ mit } i \neq j \tag{3}$$

für eine gewisse (unbekannte) Zahl $\sigma^2 > 0$.

3.1 Varianzanalyse als lineares Modell

Wir diskutieren zunächst zwei Beispiele von Fragestellungen, die zu linearen Modellen führen, deren Designmatrix keinen vollen Rang hat, vgl. auch Abschnitt 3.4.

Der Begriff „Varianzanalyse” bedeutet dabei nicht, dass Varianzen von Zufallsvariablen untersucht werden, sondern es handelt sich um die Analyse der *Variablitität von Erwartungswerten*. In der englischsprachigen Literatur ist die Abkürzung ANOVA üblich (ANOVA= analysis of variance).

3.1.1 Einfaktorielle Varianzanalyse; ANOVA–Nullhypothese

- Bei der einfaktoriellen Varianzanalyse nehmen wir an, dass sich die Zufallsstichprobe $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ in k Klassen von Teilstichproben $(Y_{ij},\ j = 1, \ldots, n_i)$ zerlegen läßt,
  - wobei $n_i > 1$ für jedes $i = 1, \ldots, k$ und $\sum_{i=1}^k n_i = n$
  - und die Stichprobenvariablen, die zu einundderselben Klasse gehören, jeweils den gleichen Erwartungswert θ_i haben mögen.
- Mit anderen Worten: Wir nehmen an, dass

$$Y_{ij} = \theta_i + \varepsilon_{ij}\,, \qquad \forall\, i = 1, \ldots, k,\ j = 1, \ldots, n_i\,, \tag{4}$$

wobei $\theta_1, \ldots, \theta_k \in \mathbb{R}$ (unbekannte) Parameter sind und die Störgrößen $\varepsilon_{ij} : \Omega \to \mathbb{R}$ unkorreliert sind mit

$$\mathbb{E}\,\varepsilon_{ij} = 0\,, \qquad \operatorname{Var}\varepsilon_{ij} = \sigma^2\,, \qquad \forall\, i = 1, \ldots, k,\ j = 1, \ldots, n_i\,. \tag{5}$$

Beachte

- Die Nummern $i = 1, \ldots, k$ der Klassen $(Y_{ij},\ j = 1, \ldots, n_i)$ werden als *Stufen eines Einflussfaktors* gedeutet.

- Die oben gemachten Modellannahmen bedeuten insbesondere, dass die beobachteten Werte $y_1, \ldots, y_n$ der Zielvariablen $Y_1, \ldots, Y_n$ wie folgt tabellarisch strukturiert werden können:

Stufe	1	2	3	$\ldots$	k
	y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
	y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	y_{k3}
			y_{3n_3}		$\vdots$
	y_{1n_1}				
		y_{2n_2}			y_{kn_k}

Wir zeigen, dass die klassische ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ mit Hilfe von so genannten Kontrasten ausgedrückt werden kann.

- Hierfür betrachten wir folgende Menge $\mathcal{A} \subset \mathbb{R}^k$ mit

$$\mathcal{A} = \Bigl\{ \mathbf{a} = (a_1, \ldots, a_k)^\top : \mathbf{a} \neq \mathbf{o}, \sum_{i=1}^k a_i = 0 \Bigr\} .$$

- Sei $\mathbf{t} = (t_1, \ldots, t_k)^\top \in \mathbb{R}^k$ ein beliebiger Vektor von Variablen, und sei $\mathbf{a} = (a_1, \ldots, a_k)^\top \in \mathcal{A}$ ein Vektor von (bekannten) Konstanten. Die Abbildung $\mathbf{t} \to \sum_{i=1}^k a_i t_i$ heißt dann *Kontrast.*

Lemma 3.1 *Seien $\theta_1, \ldots, \theta_k \in \mathbb{R}$ beliebige relle Zahlen. Für die Gültigkeit von $\theta_1 = \ldots = \theta_k$ ist dann notwendig und hinreichend, dass*

$$\sum_{i=1}^k a_i \theta_i = 0 \qquad \forall\, \mathbf{a} \in \mathcal{A} . \tag{6}$$

Beweis

- Wenn $\theta_1 = \ldots = \theta_k = \theta$, dann gilt für jedes $\mathbf{a} \in \mathcal{A}$

$$\sum_{i=1}^k a_i \theta_i = \theta \sum_{i=1}^k a_i = 0 .$$

- Um die Hinlänglichkeit der Bedingung zu beweisen, betrachten wir die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_{k-1} \in \mathcal{A}$ mit

$$\mathbf{a}_1 = (1, -1, 0, \ldots, 0)^\top , \quad \mathbf{a}_2 = (0, 1, -1, 0, \ldots, 0)^\top , \quad \ldots , \quad \mathbf{a}_{k-1} = (0, \ldots, 0, 1, -1)^\top .$$

- Für jedes $i \in \{1, \ldots, k-1\}$ ergibt sich aus der Gültigkeit der Bedingung (6) für $\mathbf{a}_i$, dass $-\theta_i + \theta_{i+1} = 0$ bzw. $\theta_i = \theta_{i+1}$. Hieraus folgt, dass $\theta_1 = \ldots \theta_k$. $\square$

Beachte

- Wegen Lemma 3.1 ist die klassische ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ äquivalent ist mit der Hypothese $H_0 : \sum_{i=1}^k a_i \theta_i = 0$ für jedes $\mathbf{a} = (a_1, \ldots, a_k)^\top \in \mathcal{A}$.

- Außerdem ist klar,
  - dass $\sum_{i=1}^k a_i\overline{Y}_{i\,\cdot}$ mit $\overline{Y}_{i\,\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$ für jedes $\mathbf{a} \in \mathcal{A}$ ein erwartungstreuer Schätzer für $\sum_{i=1}^k a_i\theta_i = 0$ ist,
  - dass die Varianz von $\sum_{i=1}^k a_i\overline{Y}_{i\,\cdot}$ gegeben ist durch $\operatorname{Var}\sum_{i=1}^k a_i\overline{Y}_{i\,\cdot} = \sigma^2\ \sum_{i=1}^k a_i^2/n_i$
  - und dass
$$S_p^2 = \frac{1}{n-k}\ \sum_{i=1}^k\sum_{j=1}^{n_i}\bigl(Y_{ij} - \overline{Y}_{i\,\cdot}\bigr)^2 \tag{7}$$
    ein Schätzer für σ^2 ist, die so genannte *gepoolte Stichprobenvarianz.*
- Es ist somit naheliegend, $H_0 : \theta_1 = \ldots = \theta_k$ abzulehnen, wenn das Supremum über alle $\mathbf{a} \in \mathcal{A}$ der (geeignet normierten) Beträge von $\sum_{i=1}^k a_i\overline{Y}_{i\,\cdot}$ einen gewissen Schwellenwert überschreitet, wobei die Testgröße $\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2$ betrachtet wird mit $T_{\mathbf{a}} = \Bigl(\sum_{i=1}^k a_i\overline{Y}_{i\,\cdot}\Bigr)\Big/\Bigl(S_p^2\ \sum_{i=1}^k a_i^2/n_i\Bigr)$.
- Ähnlich wie im Beweis von Lemma 2.4 kann man zeigen, dass unter $H_0 : \theta_1 = \ldots = \theta_k$
$$\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2 = \frac{\sum\limits_{i=1}^k n_i\bigl(\overline{Y}_{i\,\cdot} - \overline{Y}_{\cdot\cdot}\bigr)^2}{S_p^2}\ , \tag{8}$$
  wobei $\overline{Y}_{\cdot\cdot} = \sum_{i=1}^k n_i\overline{Y}_{i\,\cdot}\Big/\sum_{i=1}^k n_i$.

Durch die folgende *Quadratsummenzerlegung* ergibt sich eine anschauliche Deutung von Zähler und Nenner der in (8) betrachteten Testgröße $\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2$, vgl. auch Theorem 2.9.

Theorem 3.1 *Es gilt*
$$\sum_{i=1}^k\sum_{j=1}^{n_i}\bigl(Y_{ij} - \overline{Y}_{\cdot\cdot}\bigr)^2 = \sum_{i=1}^k n_i\bigl(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\bigr)^2 + \sum_{i=1}^k\sum_{j=1}^{n_i}\bigl(Y_{ij} - \overline{Y}_{i\cdot}\bigr)^2 . \tag{9}$$

Beweis Durch Ausmultiplizieren der linken Seite von (9) ergibt sich, dass
$$\begin{aligned}
\sum_{i=1}^k\sum_{j=1}^{n_i}\bigl(Y_{ij} - \overline{Y}_{\cdot\cdot}\bigr)^2 &= \sum_{i=1}^k\sum_{j=1}^{n_i}\bigl((Y_{ij} - \overline{Y}_{i\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\bigr)^2\\
&= \sum_{i=1}^k\sum_{j=1}^{n_i}\bigl((Y_{ij} - \overline{Y}_{i\cdot})^2 + 2(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2\bigr)\\
&= \sum_{i=1}^k\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2 + 2\sum_{i=1}^k(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\underbrace{\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})}_{=0} + \sum_{i=1}^k n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 .
\end{aligned}$$
□

Beachte

- Die Doppelsumme auf der linken Seite von (9) kann als eine Maßzahl für die (Gesamt-) *Variabilität der Stichprobenvariablen* $\{Y_{ij},\ i = 1,\ldots,k,\ j = 1,\ldots,n_i\}$ aufgefasst werden.
- Die erste Summe auf der rechten Seite von (9) ist eine Maßzahl für die Variabilität *zwischen* den Stufen des Einflussfaktors, während die Doppelsumme auf der rechten Seite von (9) eine Maßzahl für die Variabilität *innerhalb* der Stufen des Einflussfaktors ist.

- Wegen der in (7) gegebenen Definition von S_p^2 ist die in (8) betrachtete Testgröße also proportional zu dem Quotienten, der aus der Variabilität zwischen den Stufen des Einflussfaktors und der Variabilität innerhalb der Stufen gebildet wird.
- Die ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ wird somit abgelehnt, wenn die Variabilität zwischen den Stufen signifikant größer als die Variabilität innerhalb der Stufen des Einflussfaktors ist.

3.1.2 Reparametrisierung der Erwartungswerte

Das in Abschnitt 3.1.1 betrachtete Modell der einfaktoriellen Varianzanalyse kann auf zwei verschiedene Weisen als lineares Modell dargestellt werden.

- In beiden Fällen wird die Zufallsstichprobe $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ „strukturiert", d.h., wir verwenden die Schreibweise $\mathbf{Y} = \bigl(Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, \ldots, Y_{k1}, \ldots, Y_{kn_k}\bigr)^\top$, wobei $n_1 + \ldots + n_k = n$.
- Der Zufallsvektor $\mathbf{Y}$ wird in der Form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ dargestellt, wobei die Designmatrix $\mathbf{X}$ und der Parametervektor $\boldsymbol{\beta}$ jeweils unterschiedlich gewählt werden.
  - Dabei hat $\mathbf{X}$ im ersten Fall vollen Rang, im zweiten Fall jedoch keinen vollen Rang.
  - Die zweite (reparametrisierte) Darstellung ist auf die Anwendung der allgemeinen Schätz– und Testverfahren ausgerichtet, die in den Abschnitten 3.2 und 3.3 behandelt werden.
  - Bei normalverteilten Störgrößen lässt sich auf diese Weise unter $H_0 : \theta_1 = \ldots = \theta_k$ die Verteilung der in (8) betrachteten Testgröße $\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2$ bestimmen, vgl. die Formel (88) in Abshcnitt 3.4.1.

Fall 1

- In diesem Fall ist die Designmatrix $\mathbf{X}$ gegeben durch die $n \times k$ Matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \tag{10}$$

  und der Parametervektor $\boldsymbol{\beta}$ ist gegeben durch $\boldsymbol{\beta} = (\theta_1, \ldots, \theta_k)^\top$.

Fall 2

- Wir betrachten die folgende *Reparametrisierung* der Erwartungswerte $\theta_1, \ldots, \theta_k$, die den Stufen des Einflussfaktors entsprechen.
- Und zwar seien $\mu \in \mathbb{R}$ und $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ reelle Zahlen, so dass

$$\theta_i = \mu + \alpha_i\,, \qquad \forall\, i = 1, \ldots, k \tag{11}$$

  und

$$\sum_{i=1}^k n_i\alpha_i = 0\,. \tag{12}$$

- Dann lässt sich die Zufallsstichprobe $\mathbf{Y}$ des einfaktoriellen Varianzanalyse-Modells ebenfalls in der Form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ darstellen, wobei die Designmatrix $\mathbf{X}$ jetzt allerdings gegeben ist durch die $n \times (k+1)$ Matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & 0 & \ldots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \tag{13}$$

  und der Parametervektor $\boldsymbol{\beta}$ ist gegeben durch $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^\top$.

Beachte

- Die lineare Nebenbedingung (12) an die Komponenten $\alpha_1, \ldots, \alpha_k$ des Parametervektors $\boldsymbol{\beta}$ bewirkt, dass die Darstellung (11) – (12) der Erwartungswerte $\theta_1, \ldots, \theta_k$ eindeutig ist.
- Aus (11) und (12) ergibt sich außerdem, dass

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mathbb{E}\, Y_{ij} = \mu\,,$$

  wobei
  - der Parameter μ als *allgemeines Mittel* der Erwartungswerte $\mathbb{E}\, Y_{ij}$ der Stichprobenvariablen Y_{ij} aufgefasst werden kann und
  - der (Abweichungs-) Parameter α_i der *Effekt* der i-ten Stufe des Einflussfaktors genannt wird.
- Für die in (13) gegebene Designmatrix $\mathbf{X}$ gilt $\operatorname{rg}(\mathbf{X}) = k$, d.h., die $n \times (k+1)$–dimensionale Matrix $\mathbf{X}$ hat keinen vollen Spaltenrang.

Theorem 3.2 *Es gilt*

$$\mathbb{E}\,\overline{Y}_{\cdot\cdot} = \mu \qquad \textit{und} \qquad \mathbb{E}\left(\overline{Y}_{i\,\cdot} - \overline{Y}_{\cdot\cdot}\right) = \alpha_i \tag{14}$$

für jedes $i = 1, \ldots, k$, d.h., durch $\overline{Y}_{\cdot\cdot}$ und $\overline{Y}_{i\,\cdot} - \overline{Y}_{\cdot\cdot}$ sind erwartungstreue Schätzer für die Modellparameter μ bzw. α_i gegeben .

Beweis Aus der Definitionsgleichung von $\overline{Y}_{\cdot\cdot}$ ergibt sich, dass

$$\mathbb{E}\,\overline{Y}_{\cdot\cdot\cdot} = \frac{1}{\sum_{i=1}^k n_i}\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mathbb{E}\, Y_{ij} = \frac{1}{\sum_{i=1}^k n_i}\sum_{i=1}^{k} n_i\theta_i = \mu + \frac{1}{\sum_{i=1}^k n_i}\sum_{i=1}^{k} n_i\alpha_i = = \mu\,,$$

wobei sich die letzte Gleichheit aus der Reparametrisierungsbedingung (12) ergibt. Die zweite Teilaussage in (14) lässt sich auf analoge Weise beweisen. □

3.1.3 Zweifaktorielle Varianzanalyse

- Wir modifizieren nun das in Abschnitt 3.1.1 eingeführte Modell der einfaktoriellen Varianzanalyse und nehmen an, dass die Zielvariablen $Y_1, \ldots, Y_n$ von *zwei* Einflussfaktoren abhängen.
- Dabei zerlegen wir die Zufallsstichprobe $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ in $k_1 \cdot k_2$ Teilstichproben $(Y_{i_1 i_2 j},\ j = 1, \ldots, n_{i_1 i_2})$, wobei $n_{i_1 i_2} > 1$ für alle $i_1 = 1, \ldots, k_1$ bzw. $i_2 = 1, \ldots, k_2$ und
$$\sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} n_{i_1 i_2} = n\,,$$
- Wir nehmen an, dass die Stichprobenvariablen, die zu einundderselben Klasse gehören, jeweils den gleichen Erwartungswert $\theta_{i_1 i_2}$ haben.
- Mit anderen Worten: Wir nehmen an, dass
$$Y_{i_1 i_2 j} = \theta_{i_1 i_2} + \varepsilon_{i_1 i_2 j}\,, \qquad \forall\, i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2,\ j = 1, \ldots, n_{i_1 i_2}\,, \tag{15}$$
wobei $\theta_{i_1 i_2} \in \mathbb{R}$ (unbekannte) Parameter und die Störgrößen $\varepsilon_{i_1 i_2 j} : \Omega \to \mathbb{R}$ unkorreliert sind mit
$$\mathbb{E}\,\varepsilon_{i_1 i_2 j} = 0\,, \qquad \operatorname{Var} \varepsilon_{i_1 i_2 j} = \sigma^2\,, \qquad \forall\, i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2,\ j = 1, \ldots, n_{i_1 i_2}\,. \tag{16}$$

Beachte

- Die Darstellung (15) der Stichprobenvariablen $Y_{i_1 i_2 j}$ führt zu der gleichen Art eines linearen Modells, wie es in Fall 1 von Abschnitt 3.1.2 betrachtet wurde.
- Die Nummern $i_1 = 1, \ldots, k_1$ bzw. $i_2 = 1, \ldots, k_2$ der Klassen $(Y_{i_1 i_2 j},\ j = 1, \ldots, n_{i_1 i_2})$ werden erneut als *Stufen* des jeweiligen Einflussfaktors gedeutet.
- Die Designmatrix $\mathbf{X}$ hat dabei die Dimension $n \times (k_1 \cdot k_2)$ und den vollen Spaltenrang $k_1 \cdot k_2$.

Außerdem betrachten wir eine ähnliche Reparametrisierung der Erwartungswerte $\theta_{i_1 i_2}$ wie in Abschnitt 3.1.2.

- Dabei diskutieren wir hier lediglich den so genannten *balancierten Fall*, d.h.,
  - wir setzen zusätzlich voraus, dass sämtliche $k_1 \cdot k_2$ Teilstichproben $(Y_{i_1 i_2 j},\ j = 1, \ldots, n_{i_1 i_2})$ identische Stichprobenumfänge besitzen.
  - Es gelte also $n_{i_1 i_2} = r$ für alle $i_1 = 1, \ldots, k_1$ und $i_2 = 1, \ldots, k_2$, wobei $r = n/(k_1 k_2)$.
- Sei $\mu \in \mathbb{R}$, und für alle $i_1 \in \{1, \ldots, k_1\}$ und $i_2 \in \{1, \ldots, k_2\}$ seien $\alpha^{(1)}_{i_1} \in \mathbb{R}$, $\alpha^{(2)}_{i_2} \in \mathbb{R}$ und $\alpha_{i_1 i_2} \in \mathbb{R}$ reelle Zahlen, so dass
$$\theta_{i_1 i_2} = \mu + \alpha^{(1)}_{i_1} + \alpha^{(2)}_{i_2} + \alpha_{i_1 i_2}\,, \qquad \forall\, i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2 \tag{17}$$
und
$$\sum_{i_1=1}^{k_1} \alpha^{(1)}_{i_1} = \sum_{i_2=1}^{k_2} \alpha^{(2)}_{i_2} = \sum_{i_1=1}^{k_1} \alpha_{i_1 i_2} = \sum_{i_2=1}^{k_2} \alpha_{i_1 i_2} = 0\,. \tag{18}$$
- Dann lässt sich die Zufallsstichprobe $\mathbf{Y}$ in der Form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ darstellen, wobei
  - die Designmatrix $\mathbf{X}$ gegeben ist durch eine Matrix der Dimension $n \times (1 + k_1 + k_2 + k_1 k_2)$, deren Eintragungen nur aus Nullen und Einsen bestehen und die keinen vollen Rang hat.
  - Der Parametervektor $\boldsymbol{\beta}$ hat somit die folgende Form:
$$\boldsymbol{\beta} = \bigl(\mu, \alpha^{(1)}_1, \ldots, \alpha^{(1)}_{k_1}, \alpha^{(2)}_1, \ldots, \alpha^{(2)}_{k_2}, \alpha_{11}, \ldots, \alpha_{k_1 k_2}\bigr)^\top\,.$$

Beachte

- Die linearen Nebenbedingungen (18) an die Komponenten des Parametervektors $\boldsymbol{\beta}$ bewirken, ähnlich wie bei dem in Abschnitt 3.1.2 betrachteten Modell der einfaktoriellen Varianzanalyse, dass die Darstellung (17) – (18) der Erwartungswerte $\theta_{11},\ldots,\theta_{k_1k_2}$ eindeutig ist.
- Dabei kann
  - μ als *allgemeines Mittel* der Erwartungswerte $\mathbb{E}\,Y_{i_1i_2j}$ der Stichprobenvariablen $Y_{i_1i_2j}$ aufgefasst werden,
  - $\alpha^{(1)}_{i_1}$ wird *Haupteffekt* der i_1-ten Stufe des ersten Einflussfaktors genannt,
  - $\alpha^{(2)}_{i_2}$ heißt *Haupteffekt* der i_2-ten Stufe des zweiten Einflussfaktors, und
  - $\alpha_{i_1i_2}$ heißt *Wechselwirkung* zwischen den Stufen i_1 und i_2 der Stufenkombination (i_1,i_2).

Zur Konstruktion von Schätzern für die Modellparameter μ, $\alpha^{(1)}_{i_1}$, $\alpha^{(2)}_{i_2}$ bzw. $\alpha_{i_1i_2}$ verwenden wir die folgende Notation: Sei

$$Y_{i_1\cdot\cdot}=\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}Y_{i_1i_2j}\,,\qquad Y_{\cdot\,i_2\cdot}=\sum_{i_1=1}^{k_1}\sum_{j=1}^{r}Y_{i_1i_2j}\,,\qquad Y_{i_1i_2\cdot}=\sum_{j=1}^{r}Y_{i_1i_2j}\tag{19}$$

bzw.

$$\overline{Y}_{i_1\cdot\cdot}=\frac{1}{rk_2}\,Y_{i_1\cdot\cdot}\,,\qquad \overline{Y}_{\cdot\,i_2\cdot}=\frac{1}{rk_1}\,Y_{\cdot\,i_2\cdot}\,,\qquad \overline{Y}_{i_1i_2\cdot}=\frac{1}{r}\,Y_{i_1i_2\cdot}\,,\qquad \overline{Y}_{\cdots}=\frac{1}{rk_1k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}Y_{i_1i_2j}\,.\tag{20}$$

Theorem 3.3 *Es gilt*

$$\mathbb{E}\,\overline{Y}_{\cdots}=\mu\,,\quad \mathbb{E}\big(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdots}\big)=\alpha^{(1)}_{i_1}\,,\quad \mathbb{E}\big(\overline{Y}_{\cdot\,i_2\cdot}-\overline{Y}_{\cdots}\big)=\alpha^{(2)}_{i_2}\,,\quad \mathbb{E}\big(\overline{Y}_{\cdots}+\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\,i_2\cdot}\big)=\alpha_{i_1i_2}\tag{21}$$

für beliebige $i_1=1,\ldots,k_1$, $i_2=1,\ldots,k_2$, *d.h., durch* $\overline{Y}_{\cdots}$, $\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdots}$, $\overline{Y}_{\cdot\,i_2\cdot}-\overline{Y}_{\cdots}$ *und* $\overline{Y}_{\cdots}+\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\,i_2\cdot}$ *sind erwartungstreue Schätzer für die Modellparameter* μ, $\alpha^{(1)}_{i_1}$, $\alpha^{(2)}_{i_2}$ *bzw.* $\alpha_{i_1i_2}$ *gegeben .*

Beweis Aus der Definitionsgleichung von $\overline{Y}_{\cdots}$ in (20) ergibt sich, dass

$$\begin{aligned}\mathbb{E}\,\overline{Y}_{\cdots} &= \frac{1}{rk_1k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\mathbb{E}\,Y_{i_1i_2j} \;=\; \frac{1}{k_1k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\theta_{i_1i_2}\\ &= \mu+\frac{1}{k_1k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\big(\alpha^{(1)}_{i_1}+\alpha^{(2)}_{i_2}+\alpha_{i_1i_2}\big)=\mu\,,\end{aligned}$$

wobei sich die letzte Gleichheit aus den Reparametrisierungsbedingungen (18) ergibt. Die anderen drei Teilaussagen in (21) lassen sich auf analoge Weise beweisen. □

Beachte

- Die Bedingungen (18), d.h. die Annahme, dass der Parametervektor $\boldsymbol{\beta}$ zu einem linearen Unterraum des $\mathbb{R}^{1+k_1+k_2+k_1k_2}$ gehört, spielen eine wesentliche Rolle im Beweis von Theorem 3.3.
- Dabei können die Aussagen von Theorem 3.3 als Erwartungstreue der betrachteten Schätzer bezüglich dieses eingeschränkten Parameterraumes interpretiert werden.
- Wenn jedoch zugelassen wird, dass $\boldsymbol{\beta}$ ein *beliebiger* Vektor der Dimension $1+k_1+k_2+k_1k_2$ ist, dann gibt es *keinen* KQ–Schätzer für $\boldsymbol{\beta}$, der gleichzeitig erwartungstreu ist, vgl. die Diskussion am Ende von Abschnitt 3.2.1.

Das folgende Resultat enthält eine *Quadratsummenzerlegung*, vgl. auch die Theoreme 2.9 und 3.1.

Theorem 3.4 *Es gilt*

$$\begin{aligned}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j}-\overline{Y}_{\cdot\cdot\cdot}\bigr)^2 &= rk_2\sum_{i_1=1}^{k_1}\bigl(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)^2+rk_1\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{\cdot\,i_2\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)^2+\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\bigr)^2\\ &\quad+r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\,i_2\cdot}+\overline{Y}_{\cdot\cdot\cdot}\bigr)^2.\end{aligned}\tag{22}$$

Beweis Mit der in (19) bzw. (20) eingeführten Notation gilt

$$\begin{aligned}&\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Bigl(Y_{i_1i_2j}-\overline{Y}_{\cdot\cdot\cdot}\Bigr)^2\\ &=\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Bigl(\bigl(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)+\bigl(\overline{Y}_{\cdot\,i_2\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)+\bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\bigr)+\bigl(\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\,i_2\cdot}+\overline{Y}_{\cdot\cdot\cdot}\bigr)\Bigr)^2\\ &=\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Bigl(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\Bigr)^2+\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Bigl(\overline{Y}_{\cdot\,i_2\cdot}-\overline{Y}_{\cdot\cdot\cdot}\Bigr)^2+\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\Bigr)^2\\ &\quad+\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Bigl(\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\,i_2\cdot}+\overline{Y}_{\cdot\cdot\cdot}\Bigr)^2+R\,,\end{aligned}$$

wobei ähnlich wie im Beweis von Theorem 3.1 gezeigt werden kann, dass die Summe R der gemischten Produkte gleich Null ist. □

Beachte

- Die Quadratsumme auf der linken Seite von (22) kann als eine Maßzahl für die (Gesamt-) *Variabilität der Stichprobenvariablen* $\{Y_{i_1i_2j},\ i_1=1,\ldots,k_1,\ i_2=1,\ldots,k_2,\ j=1,\ldots,r\}$ aufgefasst werden.
- Die ersten beiden Quadratsummen auf der rechten Seite von (22) sind Maßzahlen für die Variabilität *zwischen* den Stufen des ersten bzw. zweiten Einflussfaktors, während die dritte Quadratsumme auf der rechten Seite von (22) eine Maßzahl für die Variabilität *innerhalb* der Stufenpaare (i_1,i_2) der beiden Einflussfaktoren ist, die so genannte *Reststreuung.*
- Die vierte Quadratsumme auf der rechten Seite von (22) ist eine Maßzahl für die Wechselwirkungen *zwischen* den Komponenten der Stufenpaare (i_1,i_2) der beiden Einflussfaktoren.
- Mit ähnlichen Überlegungen wie im Beweis von Theorem 2.5 kann man zeigen, dass eine geeignet normierte Version der Reststreuung ein erwartungstreuer Schätzer der Varianz σ^2 der Störgrößen ist.
- Und zwar gilt $\mathbb{E}\,S^2=\sigma^2$, wobei

$$S^2=\frac{1}{k_1k_2(r-1)}\;\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\Bigr)^2.$$

3.2 Schätzung der Modellparameter

Wir kehren nun zur Untersuchung des in (1) – (3) gegebenen linearen Modells mit allgemeiner Designmatrix $\mathbf{X}$ zurück, wobei wir in diesem Abschnitt annehmen, dass

- $\mathrm{rg}(\mathbf{X})=r<m$, d.h., $\mathbf{X}$ hat *keinen vollen Spaltenrang*, und dass
- $\boldsymbol{\beta}\in\mathbb{R}^m$ ein *beliebiger* m–dimensionaler Vektor ist, d.h., es werden zunächst *keine Nebenbedingungen* vom Typ (12) bzw. (18) betrachtet.

3.2.1 KQ–Schätzer für β

Die folgende Eigenschaft des Ranges von Matrixprodukten ist nützlich, die sich unmittelbar aus der in Lemma 2.3 betrachteten Rangformel ergibt.

Lemma 3.2 *Seien $m, n, r \in \mathbb{N}$ beliebige natürliche Zahlen, und seien $\mathbf{A}, \mathbf{B}$ beliebige $m \times n$ bzw. $n \times r$ Matrizen. Dann gilt*

$$\mathrm{rg}(\mathbf{AB}) \le \min\{\,\mathrm{rg}(\mathbf{A}),\, \mathrm{rg}(\mathbf{B})\}\,. \tag{23}$$

Beachte

- Weil wir jetzt annehmen, dass die Designmatrix $\mathbf{X}$ keinen vollen Rang besitzt, ist die $m \times m$ Matrix $\mathbf{X}^\top\mathbf{X}$ *nicht* invertierbar, denn gemäß Lemma 3.2 gilt $\mathrm{rg}\big(\mathbf{X}^\top\mathbf{X}\big) \le \mathrm{rg}(\mathbf{X}) < m$.
- Die Normalengleichung (2.9), d.h.,
$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}\,, \tag{24}$$
besitzt deshalb keine eindeutig bestimmte Lösung.
- Um die Lösungsmenge der Gleichung (24) zu beschreiben, benötigen wir den Begriff der verallgemeinerten inversen Matrix, vgl. auch Abschnitt 3.4 der Vorlesung „Numerik 1a" von K. Urban (Universität Ulm, Sommersemester 2005).

Definition Eine $m \times n$ Matrix $\mathbf{A}^-$ heißt *verallgemeinerte Inverse* der $n \times m$ Matrix $\mathbf{A}$, wenn

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}\,. \tag{25}$$

Um zu zeigen, dass es immer eine Lösung $\mathbf{A}^-$ der Definitionsgleichung (25) gibt, benutzen wir die folgende allgemeine Matrix-Darstellungsformel, die wir hier ohne Beweis angeben.

Lemma 3.3 *Sei $\mathbf{A}$ eine $n \times m$ Matrix mit $n \ge m$ und $\mathrm{rg}(\mathbf{A}) = r \le m$. Dann gibt es invertierbare $n \times n$ bzw. $m \times m$ Matrizen $\mathbf{P}$ bzw. $\mathbf{Q}$, so dass*

$$\mathbf{PAQ} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \qquad \textit{bzw.} \qquad \mathbf{A} = \mathbf{P}^{-1}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1}\,. \tag{26}$$

Mit Hilfe von Lemma 3.3 kann man zeigen, wie man zu Lösungen $\mathbf{A}^-$ von (25) gelangen kann.

- Seien $\mathbf{P}$ und $\mathbf{Q}$ Matrizen mit den in Lemma 3.3 betrachteten Eigenschaften, und sei $\mathbf{B}$ eine beliebige $m \times n$ Matrix mit
$$\mathbf{B} = \mathbf{Q}\begin{pmatrix} \mathbf{I}_r & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{pmatrix}\mathbf{P}\,, \tag{27}$$
wobei $\mathbf{R}$, $\mathbf{S}$, $\mathbf{T}$ beliebige Matrizen sind mit den Dimensionen $r \times (n-r)$, $(m-r) \times r$ bzw. $(m-r) \times (n-r)$.
- Dann ergibt sich aus (26) und (27), dass
$$\begin{aligned} \mathbf{ABA} &= \mathbf{P}^{-1}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1}\mathbf{Q}\begin{pmatrix} \mathbf{I}_r & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{pmatrix}\mathbf{P}\mathbf{P}^{-1}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} \\ &= \mathbf{P}^{-1}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{I}_r & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{pmatrix}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} \\ &= \mathbf{A}\,, \end{aligned}$$
d.h., die in (27) gegebene Matrix $\mathbf{B}$ ist eine verallgemeinerte Inverse von $\mathbf{A}$.

- Sei $k \in \{r, \ldots, m\}$ eine beliebige natürliche Zahl. Setzt man nun beispielsweise

$$\mathbf{R} = \mathbf{0} \qquad \text{und} \qquad \mathbf{S} = \mathbf{0} \qquad \text{und} \qquad \mathbf{T} = \begin{pmatrix} \mathbf{I}_{k-r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{28}$$

  dann gilt $\operatorname{rg}(\mathbf{B}) = k$.

Insgesamt erhalten wir somit das folgende Ergebnis.

Lemma 3.4 *Sei $\mathbf{A}$ eine $n \times m$ Matrix mit $n \geq m$ und $\operatorname{rg}(\mathbf{A}) = r \leq m$. Außerdem sei $\mathbf{B}$ für jedes $k \in \{r, \ldots, m\}$ die in (27) – (28) gegebene $m \times n$ Matrix. Dann gilt $\operatorname{rg}(\mathbf{B}) = k$ und $\mathbf{A}^- = \mathbf{B}$ ist eine Lösung der Gleichung (25).*

Außerdem sind die folgenden Eigenschaften der verallgemeinerten Inversen nützlich.

Lemma 3.5

- *Sei $\mathbf{A}$ eine beliebige $n \times m$ Matrix mit $n \geq m$, und sei $(\mathbf{A}^\top \mathbf{A})^-$ eine verallgemeinerte Inverse der symmetrischen $m \times m$ Matrix $\mathbf{A}^\top \mathbf{A}$.*
- *Dann ist auch die transponierte Matrix $\left((\mathbf{A}^\top \mathbf{A})^-\right)^\top$ eine verallgemeinerte Inverse von $\mathbf{A}^\top \mathbf{A}$.*
- *Außerdem gilt*

$$\mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top = \mathbf{A}^\top. \tag{29}$$

Beweis

- Definitionsgemäß gilt für die verallgemeinerte Inverse, dass $\mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top \mathbf{A} = \mathbf{A}^\top \mathbf{A}$.
- Hieraus und aus der Symmetrie der Matrix $\mathbf{A}^\top \mathbf{A}$ ergibt sich, dass

$$\mathbf{A}^\top \mathbf{A} = (\mathbf{A}^\top \mathbf{A})^\top = \left(\mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top \mathbf{A}\right)^\top = \mathbf{A}^\top \mathbf{A} \left((\mathbf{A}^\top \mathbf{A})^-\right)^\top \mathbf{A}^\top \mathbf{A},$$

  d.h. die transponierte Matrix $\left((\mathbf{A}^\top \mathbf{A})^-\right)^\top$ ist ebenfalls eine verallgemeinerte Inverse von $\mathbf{A}^\top \mathbf{A}$.
- Um die zweite Teilaussage (29) zu beweisen, betrachten wir die Matrix

$$\mathbf{B} = \mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top - \mathbf{A}^\top.$$

- Dann gilt

$$\begin{aligned} \mathbf{B}\mathbf{B}^\top &= \left(\mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top - \mathbf{A}^\top\right)\left(\mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top - \mathbf{A}^\top\right)^\top \\ &= \mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top \mathbf{A} \left((\mathbf{A}^\top \mathbf{A})^-\right)^\top \mathbf{A}^\top \mathbf{A} \\ &\quad - \mathbf{A}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{A})^- \mathbf{A}^\top \mathbf{A} - \mathbf{A}^\top \mathbf{A} \left((\mathbf{A}^\top \mathbf{A})^-\right)^\top \mathbf{A}^\top \mathbf{A} + \mathbf{A}^\top \mathbf{A} \\ &= \mathbf{A}^\top \mathbf{A} - \mathbf{A}^\top \mathbf{A} - \mathbf{A}^\top \mathbf{A} + \mathbf{A}^\top \mathbf{A} = \mathbf{0}. \end{aligned}$$

- Hieraus folgt, dass $\mathbf{B} = \mathbf{0}$. $\square$

Mit Hilfe der verallgemeinerten Inversen $(\mathbf{X}^\top \mathbf{X})^-$ von $\mathbf{X}^\top \mathbf{X}$ und ihrer (in Lemma 3.5 betrachteten) Eigenschaften lässt sich die Lösungsmenge der Normalengleichung (24) beschreiben.

Theorem 3.5 *Die allgemeine Lösung* $\boldsymbol{\beta}$ *der Normalengleichung* $\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}$ *hat die Form*

$$\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \bigl(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigr)\mathbf{z}\,, \tag{30}$$

wobei $(\mathbf{X}^\top\mathbf{X})^-$ *eine beliebige Lösung der Gleichung*

$$\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{X}^\top\mathbf{X} \tag{31}$$

und $\mathbf{z} \in \mathbb{R}^m$ *ein beliebiger* m*–dimensionaler Vektor ist.*

Beweis

- Durch Einsetzen von (30) in die linke Seite der Normalengleichung (24) erkennt man,
  - dass für jedes $\mathbf{z} \in \mathbb{R}^m$ durch (30) eine Lösung von (24) gegeben ist,
  - denn es gilt

$$\begin{aligned}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^\top\mathbf{X}\Bigl((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \bigl(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigr)\mathbf{z}\Bigr)\\ &= \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \;=\; \mathbf{X}^\top\mathbf{Y}\,,\end{aligned}$$

    wobei sich die letzte Gleichheit aus Lemma 3.5 ergibt.
- Sei nun $\widetilde{\boldsymbol{\beta}}$ bzw. $\boldsymbol{\beta}$ eine beliebige Lösung bzw. eine durch den Ansatz (30) gegebene Lösung von (24).
  - Dann ergibt sich durch seitenweise Subtraktion von (24), dass

$$\mathbf{X}^\top\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{o}\,. \tag{32}$$

  - Für ein $\mathbf{z} \in \mathbb{R}^m$ gilt also

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} &= \boldsymbol{\beta} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \stackrel{(30)}{=} (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \bigl(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigr)\mathbf{z} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})\\ &= (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \bigl(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigr)(\mathbf{z} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})) + (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\\ &\stackrel{(32)}{=} (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \bigl(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigr)(\mathbf{z} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}))\,.\end{aligned}$$

  - Hieraus folgt, dass $\widetilde{\boldsymbol{\beta}}$ ebenfalls eine durch den Ansatz (30) gegebene Lösung von (24) ist. □

Beispiel (einfaktorielle Varianzanalyse)

- *Zur Erinnerung*: Im reparametrisierten Modell der einfaktoriellen Varianzanalyse (vgl. Fall 2 des in Abschnitt 3.1.2 betrachteten Beispiels) ist die Designmatrix gegeben durch die $n \times (k+1)$ Matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\ 1 & 1 & 0 & 0 & \ldots & 0 & 0\\ 1 & 0 & 1 & 0 & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\ 1 & 0 & 1 & 0 & \ldots & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & 0 & 0 & 0 & \ldots & 0 & 1\\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\ 1 & 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \tag{33}$$

  und der Parametervektor $\boldsymbol{\beta}$ ist gegeben durch $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^\top$.

- Man kann sich leicht überlegen, dass dann

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} n & n_1 & n_2 & n_3 & \ldots & n_{k-1} & n_k \\ n_1 & n_1 & 0 & 0 & \ldots & 0 & 0 \\ n_2 & 0 & n_2 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ n_k & 0 & 0 & 0 & \ldots & 0 & n_k \end{pmatrix} \tag{34}$$

  und dass eine verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$ gegeben ist durch

$$\left(\mathbf{X}^\top\mathbf{X}\right)^- = \begin{pmatrix} \frac{1}{n} & 0 & 0 & 0 & \ldots & 0 & 0 \\ -\frac{1}{n} & \frac{1}{n_1} & 0 & 0 & \ldots & 0 & 0 \\ -\frac{1}{n} & 0 & \frac{1}{n_2} & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -\frac{1}{n} & 0 & 0 & 0 & \ldots & 0 & \frac{1}{n_k} \end{pmatrix}. \tag{35}$$

- Die Normalengleichung (24), d.h. $\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}$, besitzt somit die folgende Gestalt:

$$n\mu + \sum_{i=1}^k n_i\alpha_i \;=\; Y_{\cdot\cdot}\,, \qquad n_i\mu + n_i\alpha_i \;=\; Y_{i\cdot}\,, \qquad \forall\, i = 1,\ldots,k\,.$$

- Wenn wir die Lösungen dieses Gleichungssystems lediglich in dem *eingeschränkten Parameterraum* $\Theta \subset \mathbb{R}^{k+1}$ suchen, wobei

$$\Theta = \Bigl\{\boldsymbol{\beta} = (\mu,\alpha_1,\ldots,\alpha_k) : \sum_{i=1}^k n_i\alpha_i = 0\Bigr\}\,, \tag{36}$$

  dann ergibt sich die (eindeutig bestimmte) Lösung $\widehat{\boldsymbol{\beta}} = (\widehat{\mu},\widehat{\alpha}_1,\ldots,\widehat{\alpha}_k)$ mit

$$\widehat{\mu} \;=\; \overline{Y}_{\cdot\cdot}\,, \qquad \widehat{\alpha}_i \;=\; \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\,, \quad \forall\, i = 1,\ldots,k\,. \tag{37}$$

- Man kann sich leicht überlegen, dass die in (37) gegebene Lösung $\widehat{\boldsymbol{\beta}}$ der Normalengleichung (24)
  - die Gestalt $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ hat, wobei die verallgemeinerte Inverse $(\mathbf{X}^\top\mathbf{X})^-$ durch (35) gegeben ist, und
  - ein erwartungstreuer Schätzer für $\boldsymbol{\beta} = (\mu,\alpha_1,\ldots,\alpha_k)$ bezüglich des eingeschränkten Parameterraumes Θ ist, der die in (36) gegebene Form hat.
- Ohne die in (36) betrachtete Nebenbedingung gibt es jedoch keinen KQ–Schätzer für $\boldsymbol{\beta}$, der gleichzeitig erwartungstreu ist, vgl. Theorem 3.8.

Wir betrachten jetzt erneut das in (1) – (3) gegebene lineare Modell mit allgemeiner Designmatrix $\mathbf{X}$. Insbesondere betrachten wir die in Theorem 3.5 diskutierten Lösungen der Normalengleichung (24) und zeigen, dass für $\mathbf{z} = \mathbf{o}$ der in (2.8) gegebene mittlere quadratische Fehler $e(\boldsymbol{\beta})$ minimiert wird.

Theorem 3.6 *Sei $(\mathbf{X}^\top\mathbf{X})^-$ eine beliebige verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$. Dann minimiert die Stichprobenfunktion*

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \tag{38}$$

den mittleren quadratischen Fehler $e(\boldsymbol{\beta})$, d.h., $\overline{\boldsymbol{\beta}}$ ist ein KQ–Schätzer für $\boldsymbol{\beta}$.

Beweis

- Für jeden m–dimensionalen Vektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ gilt

$$\begin{aligned} n\, e(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \;=\; \bigl(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}} + \mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\bigr)^\top \bigl(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}} + \mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\bigr) \\ &= \bigl(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\bigr)^\top \bigl(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\bigr) + (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\geq\; \bigl(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\bigr)^\top \bigl(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\bigr) = n\, e(\overline{\boldsymbol{\beta}})\,, \end{aligned}$$

  weil

$$(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \Bigl(\mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\Bigr)^\top \Bigl(\mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\Bigr) \geq 0$$

  und

$$(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \bigl(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\bigr) = (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \bigl(\mathbf{X}^\top - \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)\mathbf{Y} = 0\,,$$

  wobei sich die letzte Gleichheit aus Lemma 3.5 ergibt. □

3.2.2 Erwartungswertvektor und Kovarianzmatrix des KQ–Schätzers $\overline{\boldsymbol{\beta}}$

Aus den Modellannahmen (3) über die Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ und aus den allgemeinen Rechenregeln für den Erwartungswert bzw. die Kovarianz von reellwertigen Zufallsvariablen ergibt sich, dass Erwartungswertvektor und Kovarianzmatrix des KQ–Schätzers $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{Y}$ die folgende Form haben.

Theorem 3.7 *Es gilt*

$$\mathbb{E}\, \overline{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \tag{39}$$

und

$$\operatorname{Cov} \overline{\boldsymbol{\beta}} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} \bigl((\mathbf{X}^\top \mathbf{X})^-\bigr)^\top . \tag{40}$$

Beweis

- Aus $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ und $\mathbb{E}\, \boldsymbol{\varepsilon} = \mathbf{o}$ ergibt sich, dass

$$\mathbb{E}\, \overline{\boldsymbol{\beta}} = \mathbb{E}\Bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{Y}\Bigr) = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbb{E}\, \mathbf{Y} = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}\,.$$

- Außerdem gilt für beliebige $i, j \in \{1, \ldots, m\}$

$$\begin{aligned} \operatorname{Cov}\bigl(\overline{\beta}_i, \overline{\beta}_j\bigr) &= \operatorname{Cov}\Bigl(\sum_{\ell=1}^n \bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)_{i\ell} Y_\ell,\, \sum_{r=1}^n \bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)_{jr} Y_r\Bigr) \\ &= \sum_{\ell=1}^n \sum_{r=1}^n \bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)_{i\ell} \bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)_{jr} \operatorname{Cov}(Y_\ell, Y_r) \\ &= \sigma^2 \sum_{\ell=1}^n \bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)_{i\ell} \bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)_{j\ell} \\ &= \sigma^2 \sum_{\ell=1}^n \bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\bigr)_{i\ell} \Bigl(\mathbf{X}\bigl((\mathbf{X}^\top \mathbf{X})^-\bigr)^\top\Bigr)_{\ell j} \\ &= \sigma^2 \Bigl((\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} \bigl((\mathbf{X}^\top \mathbf{X})^-\bigr)^\top\Bigr)_{ij}\,. \end{aligned}$$

□

Aus den Theoremen 3.5 und 3.7 ergibt sich mit Hilfe von Lemma 3.2, dass es *keinen* KQ–Schätzer für $\boldsymbol{\beta}$ gibt, der gleichzeitig erwartungstreu ist. Insbesondere ist der in (38) gegebene KQ–Schätzer $\overline{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ nicht erwartungstreu.

Theorem 3.8 *Wenn* $\mathrm{rg}(\mathbf{X}) < m$*, dann gibt es keinen erwartungstreuen KQ–Schätzer für* $\boldsymbol{\beta}$*.*

Beweis

- Wegen $\mathrm{rg}(\mathbf{X}) < m$ ergibt sich aus Lemma 3.2, dass auch $\mathrm{rg}(\mathbf{X}^\top\mathbf{X}) < m$ bzw.
$$\mathrm{rg}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big) < m\,.$$
  - Es gibt also ein $\boldsymbol{\beta} \neq \mathbf{o}$ mit $(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{o}$, d.h., die Gleichung
$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{41}$$
gilt *nicht* für jedes $\boldsymbol{\beta} \in \mathbb{R}^m$.
  - Wegen (39) ist somit der in (38) gegebene KQ–Schätzer $\overline{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ nicht erwartungstreu.
- Weil (41) nicht für jedes $\boldsymbol{\beta} \in \mathbb{R}^m$ gilt, ergibt sich darüber hinaus, dass auch für jedes beliebige, jedoch fest vorgegebene $\mathbf{z} \in \mathbb{R}^m$ die Gleichung
$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta} - \mathbf{z}) = \boldsymbol{\beta} - \mathbf{z}$$
bzw.
$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \big(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)\mathbf{z} = \boldsymbol{\beta}$$
nicht für jedes $\boldsymbol{\beta} \in \mathbb{R}^m$ gilt.
- Wegen Theorem 3.5 bedeutet dies, dass es keinen KQ–Schätzer für $\boldsymbol{\beta}$ gibt, der gleichzeitig erwartungstreu ist. □

3.2.3 Schätzbare Funktionen

- In Abschnitt 3.2.2 hatten wir gezeigt, dass es im linearen Modell ohne Nebenbedingungen keinen erwartungstreuen KQ–Schätzer für $\boldsymbol{\beta}$ gibt, wenn die Designmatrix $\mathbf{X}$ keinen vollen Rang besitzt.
- Anstelle des Vektors $\boldsymbol{\beta}$ betrachtet man deshalb eine Klasse von (reellwertigen) linearen Funktionen $\mathbf{a}^\top\boldsymbol{\beta}$ des Parametervektors $\boldsymbol{\beta}$, für die erwartungstreue KQ–Schätzer konstruiert werden können.
- Mit anderen Worten: Anstelle der (vektoriellen) Lineartransformation $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ der Zufallsstichprobe $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ betrachtet man eine Klasse von (reellwertigen) linearen Funktionen $\mathbf{c}^\top\mathbf{Y}$ von $\mathbf{Y}$, die als Schätzer von $\mathbf{a}^\top\boldsymbol{\beta}$ aufgefasst werden.
- Dies führt zu der folgenden Begriffsbildung.

Definition

- Sei $\mathbf{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$ ein beliebiger m–dimensionaler Vektor.
- Die lineare Funktion $\mathbf{a}^\top\boldsymbol{\beta}$ des Parametervektors $\boldsymbol{\beta}$ heißt *erwartungstreu schätzbar* bzw. *schätzbare Funktion*, wenn es einen n–dimensionalen Vektor $\mathbf{c} = (c_1, \ldots, c_n)^\top$ gibt, so dass
$$\mathbb{E}\left(\mathbf{c}^\top\mathbf{Y}\right) = \mathbf{a}^\top\boldsymbol{\beta}\,, \qquad \forall\,\boldsymbol{\beta} \in \mathbb{R}^m\,. \tag{42}$$

Beispiel (einfaktorielle Varianzanalyse)

- Für das reparametrisierte Modell der einfaktoriellen Varianzanalyse mit dem Parametervektor $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^\top \in \mathbb{R}^{k+1}$ kann man zeigen, dass beispielsweise $\alpha_1 - \alpha_2$ eine schätzbare Funktion im Sinne der Definitionsgleichung (42) ist.

- Denn mit
$$\mathbf{a}^\top = (0,1,-1,0,\ldots,0) \qquad \text{und} \qquad \mathbf{c}^\top = (\underbrace{0,\ldots,0}_{n_1-1},1,-1,0,\ldots,0)$$
gilt
$$\mathbb{E}\left(\mathbf{c}^\top \mathbf{Y}\right) = \mathbb{E}\left(Y_{1n_1} - Y_{21}\right) = (\mu+\alpha_1) - (\mu+\alpha_2) = \alpha_1 - \alpha_2 = \mathbf{a}^\top \boldsymbol{\beta}$$
für jedes $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^\top \in \mathbb{R}^{k+1}$.
- Auf ähnliche Weise kann man zeigen, dass auch $\mu + \alpha_i$ für $i = 1,\ldots,k$ bzw. $\alpha_i - \alpha_{i'}$ für $i, i' = 1,\ldots,k$ mit $i \neq i'$ schätzbare Funktionen von $\boldsymbol{\beta}$ sind.

Beispiel (zweifaktorielle Varianzanalyse mit balancierten Teilstichproben)

- Für das in Abschnitt 3.1.3 eingeführte Modell der zweifaktoriellen Varianzanalyse mit balancierten Teilstichproben besitzt die Normalengleichung (24) die folgende Gestalt:
$$\begin{aligned}
rk_1k_2\mu + rk_2 \sum_{i_1=1}^{k_1} \alpha_{i_1}^{(1)} + rk_1 \sum_{i_2=1}^{k_2} \alpha_{i_2}^{(2)} + r \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} \alpha_{i_1i_2} &= Y_{\cdots} \\
rk_2\mu + rk_2\alpha_{i_1}^{(1)} + r\sum_{i_2=1}^{k_2} \alpha_{i_2}^{(2)} + r\sum_{i_2=1}^{k_2} \alpha_{i_1i_2} &= Y_{i_1\cdot\cdot} \qquad \forall\, i_1 = 1,\ldots,k_1 \\
rk_1\mu + r\sum_{i_1=1}^{k_1} \alpha_{i_1}^{(1)} + rk_1\alpha_{i_2}^{(2)} + r\sum_{i_1=1}^{k_1} \alpha_{i_1i_2} &= Y_{\cdot i_2\cdot} \qquad \forall\, i_2 = 1,\ldots,k_2 \\
r\mu + r\alpha_{i_1}^{(1)} + r\alpha_{i_2}^{(2)} + r\alpha_{i_1i_2} &= Y_{i_1i_2\cdot} \qquad \forall\, i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2
\end{aligned}$$
- Unter Berücksichtigung der Nebenbedingung (18) ist dieses Gleichungssystem eindeutig lösbar. Mit anderen Worten: Wenn nur Parametervektoren
$$\boldsymbol{\beta} = \left(\mu, \alpha_1^{(1)}, \ldots, \alpha_{k_1}^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_{k_2}^{(2)}, \alpha_{11}, \ldots, \alpha_{k_1k_2}\right)^\top$$
aus dem eingeschränkten Parameterraum
$$\Theta = \Big\{\boldsymbol{\beta} : \sum_{i_1=1}^{k_1} \alpha_{i_1}^{(1)} = \sum_{i_2=1}^{k_2} \alpha_{i_2}^{(2)} = \sum_{i_1=1}^{k_1} \alpha_{i_1i_2} = \sum_{i_2=1}^{k_2} \alpha_{i_1i_2} = 0\Big\}$$
betrachtet werden, dann ergibt sich die eindeutig bestimmte Lösung
$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\mu}, \widehat{\alpha}_1^{(1)}, \ldots, \widehat{\alpha}_{k_1}^{(1)}, \widehat{\alpha}_1^{(2)}, \ldots, \widehat{\alpha}_{k_2}^{(2)}, \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{k_1k_2}\right)^\top \tag{43}$$
der Normalengleichung, wobei
$$\widehat{\mu} = \overline{Y}_{\cdots}\,, \quad \widehat{\alpha}_{i_1}^{(1)} = \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdots}\,, \quad \widehat{\alpha}_{i_2}^{(2)} = \overline{Y}_{\cdot i_2\cdot} - \overline{Y}_{\cdots}\,, \quad \widehat{\alpha}_{i_1i_2} = \overline{Y}_{\cdots} + \overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdot i_2\cdot} \tag{44}$$
für beliebige $i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2$.
- Man kann zeigen, dass die in (43) – (44) gegebene Lösung $\widehat{\boldsymbol{\beta}}$ der Normalengleichung die Gestalt
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$$
hat, wobei $(\mathbf{X}^\top\mathbf{X})^-$ eine verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$ und $\mathbf{X}$ die Designmatrix des zweifaktoriellen Varianzanalyse-Modells mit balancierten Teilstichproben ist.
- *Beachte*: Die in (43) – (44) gegebene Stichprobenfunktion $\widehat{\boldsymbol{\beta}}$ wurde bereits in Theorem 3.3 diskutiert, wobei gezeigt wurde, dass $\widehat{\boldsymbol{\beta}}$ ein erwartungstreuer Schätzer für $\boldsymbol{\beta}$ bezüglich des eingeschränkten Parameterraumes Θ ist.

- Außerdem kann man zeigen, dass die linearen Funktionen $\mu + \alpha^{(1)}_{i_1} + \alpha^{(2)}_{i_2} + \alpha_{i_1 i_2}$ des Parametervektors $\boldsymbol{\beta}$ für beliebige $i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2$ im Sinne der Definitionsgleichung (42) (d.h. ohne Berücksichtigung der Nebenbedingungen (18)) erwartungstreu schätzbar sind.
- Im Modell der zweifaktoriellen Varianzanalyse ohne Wechselwirkungen, d.h. $\alpha_{i_1 i_2} = 0$ für beliebige $i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2$, sind auch $\alpha^{(1)}_{i_1} - \alpha^{(1)}_{i'_1}$ für beliebige $i_1, i'_1 = 1,\ldots,k_1$ mit $i_1 \neq i'_1$ bzw. $\alpha^{(2)}_{i_2} - \alpha^{(2)}_{i'_2}$ für beliebige $i_2, i'_2 = 1,\ldots,k_2$ mit $i_2 \neq i'_2$ erwartungstreu schätzbar.

Das folgende Hilfsergebnis, das eine Ergänzung von Lemma 3.5 ist, benötigen wir, um zwei allgemeine *Kriterien für die erwartungstreue Schätzbarkeit* von linearen Funktionen $\mathbf{a}^\top\boldsymbol{\beta}$ des Parametervektors $\boldsymbol{\beta}$ herzuleiten.

Lemma 3.6 *Sei* $\big(\mathbf{X}^\top\mathbf{X}\big)^-$ *eine verallgemeinerte Inverse von* $\mathbf{X}^\top\mathbf{X}$. *Dann gilt*

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{X}\,. \tag{45}$$

Beweis In Lemma 3.5 hatten wir gezeigt, dass

- die transponierte Matrix $\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top$ ebenfalls eine verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$ ist und dass $\mathbf{X}^\top\mathbf{X}\big(\mathbf{X}^\top\mathbf{X}\big)^-\mathbf{X}^\top = \mathbf{X}^\top$.
- Damit gilt auch $\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\mathbf{X}^\top = \mathbf{X}^\top$.
- Hieraus ergibt sich (45) durch Vertauschen von Spalten und Zeilen. □

Theorem 3.9 *Sei* $\mathbf{a} = (a_1,\ldots,a_m)^\top \in \mathbb{R}^m$ *ein beliebiger Vektor. Die lineare Funktion* $\mathbf{a}^\top\boldsymbol{\beta}$ *des Parametervektors* $\boldsymbol{\beta}$ *ist genau dann erwartungstreu schätzbar, wenn eine der folgenden beiden Bedingungen erfüllt ist:*

1. *Es gibt ein* $\mathbf{c} \in \mathbb{R}^n$, *so dass*
$$\mathbf{a}^\top = \mathbf{c}^\top\mathbf{X}\,. \tag{46}$$
2. *Der Vektor* $\mathbf{a}$ *genügt dem folgenden Gleichungssystem:*
$$\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{a}^\top\,. \tag{47}$$

Beweis

- Sei $\mathbf{a}^\top\boldsymbol{\beta}$ eine schätzbare Funktion des Parametervektors $\boldsymbol{\beta}$.
  - Dann ergibt sich aus (42), dass für jedes $\boldsymbol{\beta} \in \mathbb{R}^m$
  $$\mathbf{a}^\top\boldsymbol{\beta} = \mathbb{E}\left(\mathbf{c}^\top\mathbf{Y}\right) = \mathbf{c}^\top\mathbb{E}\,\mathbf{Y} = \mathbf{c}^\top\mathbf{X}\boldsymbol{\beta} \qquad \text{bzw.} \qquad \Big(\mathbf{c}^\top\mathbf{X} - \mathbf{a}^\top\Big)\boldsymbol{\beta} = 0\,.$$
  - Hieraus folgt, dass $\mathbf{a}^\top = \mathbf{c}^\top\mathbf{X}$.
- Umgekehrt sei $\mathbf{c}$ ein Vektor, der der Bedingung (46) genügt.
  - Hieraus folgt, dass
  $$\mathbb{E}\left(\mathbf{c}^\top\mathbf{Y}\right) = \mathbf{c}^\top\mathbb{E}\,\mathbf{Y} = \mathbf{c}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^\top\boldsymbol{\beta}\,, \qquad \forall\,\boldsymbol{\beta} \in \mathbb{R}^m\,.$$
  - Damit ist auch die Hinlänglichkeit der Bedingung (46) bewiesen.
- Um die Notwendigkeit der Bedingung (47) zu zeigen, benutzen wir das Ergebnis von Lemma 3.6.
  - Sei $\mathbf{a}^\top\boldsymbol{\beta}$ eine schätzbare Funktion des Parametervektors $\boldsymbol{\beta}$.
  - Dann ergibt sich aus (46) und (45), dass
  $$\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{c}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{c}^\top\mathbf{X} = \mathbf{a}^\top\,.$$

- Um die Hinlänglichkeit der Bedingung (47) zu zeigen, genügt es zu beachten,
  - dass sich aus (47) die Gültigkeit von $\mathbf{a}^\top = \mathbf{c}^\top \mathbf{X}$ für $\mathbf{c}^\top = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top$ ergibt.
  - Die erwartungstreue Schätzbarkeit von $\mathbf{a}^\top \boldsymbol{\beta}$ ergibt sich nun aus der ersten Teilaussage. □

Beachte

- Wenn die Designmatrix $\mathbf{X}$ vollen Rang hat, d.h. $(\mathbf{X}^\top \mathbf{X})^- = (\mathbf{X}^\top \mathbf{X})^{-1}$, dann ist die Bedingung (47) offenbar für jedes $\mathbf{a} \in \mathbb{R}^m$ erfüllt.
- In diesem Fall ist somit *jede* lineare Funktion des Parametervektors $\boldsymbol{\beta}$ erwartungstreu schätzbar, was bereits in Theorem 2.2 gezeigt wurde.

Für den Fall, dass die Designmatrix $\mathbf{X} = (x_{ij})$ keinen vollen Rang hat, zeigen wir,

- wie sich aus der zweiten Teilaussage von Theorem 3.9 die Schätzbarkeit der folgenden linearen Funktionen $\mathbf{a}^\top \boldsymbol{\beta}$ von $\boldsymbol{\beta}$ ergibt.
- Der (Gewichts-) Vektor $\mathbf{c}$ des linearen erwartungstreuen Schätzers $\mathbf{c}^\top \mathbf{Y}$ für $\mathbf{a}^\top \boldsymbol{\beta}$ kann dabei jeweils so wie im Beweis von Theorem 3.9 gewählt werden, d.h.,

$$\mathbf{c}^\top = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top . \tag{48}$$

Theorem 3.10 *Die folgenden linearen Funktionen des Parametervektors $\boldsymbol{\beta}$ sind erwartungstreu schätzbar:*

1. *die Komponenten $\sum_{j=1}^m x_{1j}\beta_j, \ldots, \sum_{j=1}^m x_{nj}\beta_j$ des Erwartungswertvektors $\mathbb{E}\, \mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$,*
2. *jede lineare Funktion von schätzbaren Funktionen,*
3. *die Komponenten $\beta'_1, \ldots, \beta'_m$ des so genannten projizierten Parametervektors $\boldsymbol{\beta}' = (\beta'_1, \ldots, \beta'_m)^\top$, wobei*

$$\boldsymbol{\beta}' = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} . \tag{49}$$

Beweis

- Sei $\mathbf{a}_i = (x_{i1}, \ldots, x_{im})^\top$ für jedes $i \in \{1, \ldots, n\}$. Dann gilt

$$\begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top \end{pmatrix} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{X} = \begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top \end{pmatrix},$$

  wobei sich die vorletzte Gleichheit aus Lemma 3.6 ergibt. Aus Theorem 3.9 folgt nun, dass jede der Linearkombinationen $\sum_{j=1}^m x_{1j}\beta_j, \ldots, \sum_{j=1}^m x_{nj}\beta_j$ eine schätzbare Funktion ist.
- Um die zweite Teilaussage zu beweisen, betrachten wir eine beliebige (endliche) Familie $\mathbf{a}_1^\top \boldsymbol{\beta}, \ldots, \mathbf{a}_s^\top \boldsymbol{\beta}$ von s schätzbaren Funktionen, die wir in der Form $\mathbf{A}\boldsymbol{\beta}$ darstellen, wobei $\mathbf{A} = \bigl(\mathbf{a}_1, \ldots, \mathbf{a}_s\bigr)^\top$ eine $s \times m$–dimensionale Matrix und $s \in \mathbb{N}$ eine beliebige natürliche Zahl ist.
- Aus Theorem 3.9 ergibt sich, dass

$$\mathbf{A}(\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{A} .$$

  - Für jeden s–dimensionalen Vektor $\mathbf{b} = (b_1, \ldots, b_s)^\top \in \mathbb{R}^s$ gilt damit auch

$$\mathbf{b}^\top \mathbf{A}(\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{b}^\top \mathbf{A} .$$

  - Hieraus ergibt sich mit Hilfe von Theorem 3.9, dass die lineare Funktion $\mathbf{b}^\top\mathbf{A}\boldsymbol{\beta}$ der schätzbaren Funktionen $\mathbf{A}\boldsymbol{\beta}$ selbst eine schätzbare Funktion ist.

- In der dritten Teilaussage wird die Familie $\mathbf{A}\boldsymbol{\beta}$ von linearen Funktionen des Parametervektors $\boldsymbol{\beta}$ betrachtet, wobei die $m \times m$ Matrix $\mathbf{A}$ gegeben ist durch $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}$.
  - Hieraus folgt, dass

$$\mathbf{A}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{A}\,,$$

    wobei sich die vorletzte Gleichheit aus der Definitionsgleichung (25) der verallgemeinerten Inversen ergibt.
  - Aus Theorem 3.9 folgt nun, dass die Komponenten $\beta_1', \ldots, \beta_m'$ des projizierten Parametervektors

$$\boldsymbol{\beta}' = \mathbf{A}\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$$

    schätzbare Funktionen sind. □

3.2.4 Beste lineare erwartungstreue Schätzer; Gauß–Markow–Theorem

- In diesem Abschnitt zeigen wir, wie BLUE–Schätzer für schätzbare Funktionen des Parametervektors $\boldsymbol{\beta}$ konstruiert werden können.
- *Zur Erinnerung*: Ein linearer erwartungstreuer Schätzer wird BLUE–Schätzer genannt, wenn es keinen linearen erwartungstreuen Schätzer gibt, dessen Varianz kleiner ist (BLUE = best linear unbiased estimator).
- In der Theorie linearer Modelle wird das folgende Resultat das *Gauß–Markow–Theorem* genannt.

Theorem 3.11

- *Sei $\mathbf{a}^\top\boldsymbol{\beta}$ eine schätzbare Funktion des Parametervektors $\boldsymbol{\beta}$, sei $(\mathbf{X}^\top\mathbf{X})^-$ eine beliebige verallgemeinerte Inverse der $m \times m$ Matrix $\mathbf{X}^\top\mathbf{X}$, und sei $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$.*
- *Dann ist $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ ein BLUE–Schätzer für $\mathbf{a}^\top\boldsymbol{\beta}$, wobei*

$$\operatorname{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) = \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a}\,. \tag{50}$$

Beweis

- Wir zeigen zuerst, dass $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ ein linearer erwartungstreuer Schätzer für $\mathbf{a}^\top\boldsymbol{\beta}$ ist.
  - Es ist klar, dass

$$\mathbf{a}^\top\overline{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$$

    eine lineare Funktion der Zufallsstichprobe $\mathbf{Y} = (Y_1, \ldots, Y_n)$ ist.
  - Weil vorausgesetzt wird, dass $\mathbf{a}^\top\boldsymbol{\beta}$ eine schätzbare Funktion des Parametervektors $\boldsymbol{\beta}$ ist, folgt aus Theorem 3.9, dass es ein $\mathbf{c} \in \mathbb{R}^n$ gibt, so dass

$$\mathbf{a}^\top = \mathbf{c}^\top\mathbf{X}\,. \tag{51}$$

  - Somit gilt für jedes $\boldsymbol{\beta} \in \mathbb{R}^m$, dass

$$\mathbb{E}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) = \mathbf{c}^\top\mathbf{X}\mathbb{E}\,\overline{\boldsymbol{\beta}} = \mathbf{c}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbb{E}\,\mathbf{Y} = \mathbf{c}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{c}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^\top\boldsymbol{\beta}\,,$$

    wobei sich die vorletzte Gleichheit aus Lemma 3.6 ergibt.
  - Damit ist gezeigt, dass $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ ein linearer erwartungstreuer Schätzer für $\mathbf{a}^\top\boldsymbol{\beta}$ ist.

- Aus den Rechenregeln für die Varianz (vgl. Theorem WR-4.13) ergibt sich, dass
$$\mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) = \mathrm{Var}\Bigl(\sum_{i=1}^m a_i\overline{\beta}_i\Bigr) = \sum_{i=1}^m\sum_{j=1}^m a_ia_j\;\mathrm{Cov}\,(\overline{\beta}_i,\overline{\beta}_j)\,.$$
  - Außerdem hatten wir in Theorem 3.7 gezeigt, dass
$$\mathrm{Cov}\,(\overline{\beta}_i,\overline{\beta}_j) = \sigma^2\bigl((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigl((\mathbf{X}^\top\mathbf{X})^-\bigr)^\top\bigr)_{ij}\,.$$
  - Hieraus folgt, dass
$$\begin{aligned}\mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) &= \sigma^2\sum_{i=1}^m\sum_{j=1}^m a_ia_j\bigl((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigl((\mathbf{X}^\top\mathbf{X})^-\bigr)^\top\bigr)_{ij}\\ &= \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigl((\mathbf{X}^\top\mathbf{X})^-\bigr)^\top\mathbf{a}\\ &= \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\bigl((\mathbf{X}^\top\mathbf{X})^-\bigr)^\top\mathbf{X}^\top\mathbf{c}\\ &= \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{c}\,,\end{aligned}$$
    wobei sich die vorletzte Gleichheit aus (51) und die letzte Gleichheit aus Lemma 3.5 ergibt.
  - Die erneute Anwendung von (51) liefert die Varianzformel (50).
- Es ist noch zu zeigen, dass der Schätzer $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ die kleinste Varianz in der Klasse aller linearen erwartungstreuen Schätzer für $\mathbf{a}^\top\boldsymbol{\beta}$ hat.
  - Sei $\mathbf{b}\in\mathbb{R}^n$, so dass $\mathbf{b}^\top\mathbf{Y}$ ein linearer erwartungstreuer Schätzer für $\mathbf{a}^\top\boldsymbol{\beta}$ ist. Dann gilt
$$\mathbf{a}^\top\boldsymbol{\beta} = \mathbb{E}\left(\mathbf{b}^\top\mathbf{Y}\right) = \mathbf{b}^\top\mathbf{X}\boldsymbol{\beta}\qquad\forall\,\boldsymbol{\beta}\in\mathbb{R}^m$$
    und somit auch
$$\mathbf{b}^\top\mathbf{X} = \mathbf{a}^\top\,. \tag{52}$$
  - Für die Kovarianz von $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ und $\mathbf{b}^\top\mathbf{Y}$ gilt
$$\mathrm{Cov}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}},\,\mathbf{b}^\top\mathbf{Y}\right) = \mathrm{Cov}\left(\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y},\,\mathbf{b}^\top\mathbf{Y}\right) = \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{b} = \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a}\,,$$
    wobei sich die letzte Gleichheit aus (52) ergibt.
  - Hieraus und aus der Varianzformel (50) folgt, dass
$$\begin{aligned}0 &\le \mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}-\mathbf{b}^\top\mathbf{Y}\right) = \mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right)+\mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right)-2\,\mathrm{Cov}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}},\,\mathbf{b}^\top\mathbf{Y}\right)\\ &= \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a}+\mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right)-2\,\sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a}\\ &= \mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right)-\sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a} = \mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right)-\mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right).\end{aligned}$$
□

Beachte

- Im Beweis von Theorem 3.11 wurde nirgendwo explizit genutzt, dass $\mathrm{rg}(\mathbf{X})<m$.
- Mit anderen Worten: Wenn die Designmatrix $\mathbf{X}$ vollen Rang hat, d.h. $\mathrm{rg}(\mathbf{X})=m$, dann ist $\mathbf{a}^\top\boldsymbol{\beta}$ für *jeden* m–dimensionalen Vektor $\mathbf{a}^\top\in\mathbb{R}^m$ erwartungstreu schätzbar, und $\mathbf{a}^\top\widehat{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ ist ein BLUE–Schätzer für $\mathbf{a}^\top\boldsymbol{\beta}$.

Aus der folgenden *Invarianzeigenschaft* der verallgemeinerten Inversen $(\mathbf{X}^\top\mathbf{X})^-$ von $\mathbf{X}^\top\mathbf{X}$ ergibt sich, dass der in Theorem 3.11 betrachtete BLUE–Schätzer $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ *nicht* von der spezifischen Wahl von $(\mathbf{X}^\top\mathbf{X})^-$ abhängt.

Lemma 3.7 *Seien* $\mathbf{A}$ *und* $\mathbf{A}'$ *beliebige verallgemeinerte Inverse der Matrix* $\mathbf{X}^\top\mathbf{X}$. *Dann gilt*
$$\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}\mathbf{A}'\mathbf{X}^\top\,. \tag{53}$$

Beweis

- In Lemma 3.6 hatten wir gezeigt, dass
$$\mathbf{X}\mathbf{A}\mathbf{X}^\top\mathbf{X} = \mathbf{X} = \mathbf{X}\mathbf{A}'\mathbf{X}^\top\mathbf{X} \tag{54}$$
für beliebige verallgemeinerte Inverse $\mathbf{A}$ und $\mathbf{A}'$ der Matrix $\mathbf{X}^\top\mathbf{X}$.
- Wenn diese Gleichungskette von rechts mit $\mathbf{A}\mathbf{X}^\top$ multipliziert wird, dann ergibt sich
$$\mathbf{X}\mathbf{A}\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}\mathbf{A}'\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top\,. \tag{55}$$
- Aus (54) ergibt sich für die linke Seite der letzten Gleichheit, dass $\mathbf{X}\mathbf{A}\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}\mathbf{A}\mathbf{X}^\top$.
- Außerdem hatten wir in Lemma 3.5 gezeigt, dass $\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}^\top$ für jede verallgemeinerte Inverse $\mathbf{A}$ von $\mathbf{X}^\top\mathbf{X}$ gilt.
- Wenn dies in die rechte Seite von (55) eingesetzt wird, dann ergibt sich (53). □

Mit Hilfe von Lemma 3.7 kann nun die oben erwähnte Invarianzeigenschaft des in Theorem 3.11 betrachteten BLUE–Schätzers $\mathbf{a}^\top\overline{\boldsymbol\beta}$ bewiesen werden.

Theorem 3.12 *Sei $\mathbf{a}^\top\boldsymbol\beta$ eine schätzbare Funktion des Parametervektors $\boldsymbol\beta$. Dann hängt der BLUE–Schätzer $\mathbf{a}^\top\overline{\boldsymbol\beta} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ nicht von der Wahl der verallgemeinerten Inversen $(\mathbf{X}^\top\mathbf{X})^-$ ab.*

Beweis

- Zur Erinnerung: Für jede schätzbare Funktion $\mathbf{a}^\top\boldsymbol\beta$ von $\boldsymbol\beta$ ergibt sich aus Theorem 3.9, dass $\mathbf{a}^\top = \mathbf{c}^\top\mathbf{X}$ für ein $\mathbf{c}\in\mathbb{R}^n$.
- Hieraus und aus Lemma 3.7 folgt, dass
$$\mathbf{a}^\top\overline{\boldsymbol\beta} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = \mathbf{c}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$$
nicht von der Wahl der verallgemeinerten Inversen $(\mathbf{X}^\top\mathbf{X})^-$ abhängt. □

Beispiel (einfaktorielle Varianzanalyse)

- Für das reparametrisierte Modell der einfaktoriellen Varianzanalyse sind $\mu+\alpha_i$ für $i=1,\ldots,k$ bzw. $\alpha_i-\alpha_{i'}$ für $i,i'=1,\ldots,k$ mit $i\neq i'$ schätzbare Funktionen von $\boldsymbol\beta$, vgl. Theorem 3.10.
- Aus Theorem 3.11 ergibt sich, dass $\widehat\mu+\widehat\alpha_i$ bzw. $\widehat\alpha_i-\widehat\alpha_{i'}$ BLUE–Schätzer für $\mu+\alpha_i$ bzw. $\alpha_i-\alpha_{i'}$ sind, wobei
$$\overline{\boldsymbol\beta} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = (\widehat\mu,\widehat\alpha_1,\ldots,\widehat\alpha_k)^\top$$
mit $\widehat\mu = \overline{Y}_{\cdot\cdot}$ bzw. $\widehat\alpha_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$ die Lösung der Normalengleichung (24) ist, die bereits in Abschnitt 3.2.1 betrachtet wurde.

Beispiel (zweifaktorielle Varianzanalyse mit balancierten Teilstichproben)

- Für das in Abschnitt 3.1.3 eingeführte balancierte Modell der zweifaktoriellen Varianzanalyse mit balancierten Teilstichproben sind die linearen Funktionen $\mu+\alpha^{(1)}_{i_1}+\alpha^{(2)}_{i_2}+\alpha_{i_1i_2}$ des Parametervektors $\boldsymbol\beta = \bigl(\mu,\alpha^{(1)}_1,\ldots,\alpha^{(1)}_{k_1},\alpha^{(2)}_1,\ldots,\alpha^{(2)}_{k_2},\alpha_{11},\ldots,\alpha_{k_1k_2}\bigr)$ für beliebige $i_1=1,\ldots,k_1,\ i_2=1,\ldots,k_2$ erwartungstreu schätzbar, vgl. Theorem 3.10.

- Aus Theorem 3.11 ergibt sich, dass $\widehat{\mu}+\widehat{\alpha}_{i_1}^{(1)}+\widehat{\alpha}_{i_2}^{(2)}+\widehat{\alpha}_{i_1i_2}$ ein BLUE–Schätzer für $\mu+\alpha_{i_1}^{(1)}+\alpha_{i_2}^{(2)}+\alpha_{i_1i_2}$ ist, wobei
$$\overline{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}=\bigl(\widehat{\mu},\widehat{\alpha}_1^{(1)},\ldots,\widehat{\alpha}_{k_1}^{(1)},\widehat{\alpha}_1^{(2)},\ldots,\widehat{\alpha}_{k_2}^{(2)},\widehat{\alpha}_{11},\ldots,\widehat{\alpha}_{k_1k_2}\bigr)$$
mit
$$\widehat{\mu}=\overline{Y}_{\cdots}\,,\qquad \widehat{\alpha}_{i_1}^{(1)}=\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdots}\,,\qquad \widehat{\alpha}_{i_2}^{(2)}=\overline{Y}_{\cdot\, i_2\cdot}-\overline{Y}_{\cdots}\,,\qquad \widehat{\alpha}_{i_1i_2}=\overline{Y}_{\cdots}+\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\, i_2\cdot}$$
die bereits in Abschnitt 3.2.3 betrachtete Lösung der Normalengleichung (24) ist.
- Darüber hinaus ergibt sich aus Theorem 3.10, dass im Modell der zweifaktoriellen Varianzanalyse ohne Wechselwirkungen, d.h. $\alpha_{i_1i_2}=0$ für beliebige $i_1=1,\ldots,k_1,\ i_2=1,\ldots,k_2$, auch $\alpha_{i_1}^{(1)}-\alpha_{i_1'}^{(1)}$ für beliebige $i_1,i_1'=1,\ldots,k_1$ mit $i_1\neq i_1'$ bzw. $\alpha_{i_2}^{(2)}-\alpha_{i_2'}^{(2)}$ für beliebige $i_2,i_2'=1,\ldots,k_2$ mit $i_2\neq i_2'$ erwartungstreu schätzbar sind.
- Aus Theorem 3.11 folgt somit, dass $\widehat{\alpha}_{i_1}^{(1)}-\widehat{\alpha}_{i_1'}^{(1)}$ bzw. $\widehat{\alpha}_{i_2}^{(2)}-\widehat{\alpha}_{i_2'}^{(2)}$ BLUE–Schätzer für $\alpha_{i_1}^{(1)}-\alpha_{i_1'}^{(1)}$ bzw. $\alpha_{i_2}^{(2)}-\alpha_{i_2'}^{(2)}$ sind.

3.3 Normalverteilte Störgrößen

Zusätzlich zu den Modellannahmen, die am Anfang von Kapitel 3 formuliert worden sind, setzen wir jetzt erneut voraus, dass $n>m$ und dass die zufälligen Störgrößen $\varepsilon_1,\ldots,\varepsilon_n:\Omega\to\mathbb{R}$ unabhängig und (identisch) normalverteilt sind, d.h. insbesondere, dass $\varepsilon_i\sim\mathrm{N}(0,\sigma^2)$ für jedes $i=1,\ldots,n$.

3.3.1 Maximum–Likelihood–Schätzer

- Genauso wie im Fall von Designmatrizen mit vollem Spaltenrang, der in Abschnitt 2.2.1 diskutiert wurde, ergibt sich aus Theorem 1.3, dass der Vektor $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ der Zielvariablen normalverteilt ist mit
$$\mathbf{Y}\sim\mathrm{N}(\mathbf{X}\boldsymbol{\beta},\,\sigma^2\mathbf{I})\,. \tag{56}$$
- Mit anderen Worten: Der Zufallsvektor $\mathbf{Y}$ ist absolutstetig mit der Dichte
$$f_{\mathbf{Y}}(\mathbf{y})=\Bigl(\frac{1}{\sigma\sqrt{2\pi}}\Bigr)^n\,\exp\Bigl(-\,\frac{1}{2\sigma^2}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\Bigr) \tag{57}$$
für jedes $\mathbf{y}=(y_1,\ldots,y_n)^\top\in\mathbb{R}^n$.
- Die Loglikelihood-Funktion $\log L(\mathbf{y};\boldsymbol{\beta},\sigma^2)=\log f_{\mathbf{Y}}(\mathbf{y})$ hat somit die Gestalt
$$\log L(\mathbf{y};\boldsymbol{\beta},\sigma^2)=-\,\frac{n}{2}\,\log(2\pi)-\,\frac{n}{2}\,\log(\sigma^2)-\,\frac{1}{2\sigma^2}\,|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|^2\,. \tag{58}$$
- Um einen Maximum–Likelihood–Schätzer für den Parametervektor $(\boldsymbol{\beta},\sigma^2)$ zu bestimmen, betrachten wir zunächst (genauso wie im Beweis von Theorem 2.6) für beliebige, jedoch fest vorgegebene $\mathbf{y}\in\mathbb{R}^n$ und $\sigma^2>0$ die Abbildung
$$\mathbb{R}^m\ni\boldsymbol{\beta}\mapsto\log L(\mathbf{y};\boldsymbol{\beta},\sigma^2)\,. \tag{59}$$
- Aus (58) und (59) ergibt sich, dass dabei der folgende Ausdruck $e(\boldsymbol{\beta})$ minimiert werden muss, wobei
$$e(\boldsymbol{\beta})=\frac{1}{n}\,|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|^2=\frac{1}{n}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,.$$

- Aus Theorem 3.6 folgt somit, dass der KQ–Schätzer $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ gleichzeitig ein ML–Schätzer für $\boldsymbol{\beta}$ ist (der nicht von σ^2 abhängt).
- Außerdem ergibt sich genauso wie im Beweis von Theorem 2.6, dass durch $(\overline{\boldsymbol{\beta}}, \overline{\sigma}^2)$ ein ML–Schätzer für $(\boldsymbol{\beta}, \sigma^2)$ gegeben ist, wobei

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \qquad \text{und} \qquad \overline{\sigma}^2 = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\right)^\top\left(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\right). \tag{60}$$

Beachte

- In Abschnitt 3.2.2 hatten wir gezeigt, dass $\overline{\boldsymbol{\beta}}$ im allgemeinen kein erwartungstreuer Schätzer für $\boldsymbol{\beta}$ ist.
- Ebenso ist $\overline{\sigma}^2$ kein erwartungstreuer Schätzer für σ^2, wobei sich jedoch durch eine einfache Modifikation von $\overline{\sigma}^2$ ein erwartungstreuer Schätzer für σ^2 ergibt.
- Um dies zu zeigen, sind die folgenden Eigenschaften der Matrix

$$\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top \tag{61}$$

  nützlich, die als Verallgemeinerung der entsprechenden Matrix-Eigenschaften aufgefasst werden können, die in Lemma 2.1 bzw. in Lemma 2.2 für den Fall von Designmatrizen $\mathbf{X}$ mit vollem Spaltenrang hergeleitet worden sind.

Lemma 3.8 *Sei* $\operatorname{rg}(\mathbf{X}) = r \le m$. *Für die in* (61) *gegebene Matrix* $\mathbf{G}$ *gilt dann:*

1. $\mathbf{G}$ *ist idempotent und symmetrisch,*
2. $\mathbf{G}\mathbf{X} = \mathbf{0}$ *und*
3. $\operatorname{sp}(\mathbf{G}) = n - r$.

Beweis

- Aus Lemma 3.5 ergibt sich, dass

$$\begin{aligned}
\mathbf{G}^2 &= \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\right)\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\right)\\
&= \mathbf{I} - 2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top}_{=\mathbf{X}^\top}\\
&= \mathbf{I} - 2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top = \mathbf{G}\,.
\end{aligned}$$

- In Lemma 3.5 hatten wir gezeigt, dass $\left((\mathbf{X}^\top\mathbf{X})^-\right)^\top$ eine verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$ ist. Somit ergibt sich aus Lemma 3.7, dass

$$\mathbf{G}^\top = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\right)^\top = \mathbf{I} - \mathbf{X}\left((\mathbf{X}^\top\mathbf{X})^-\right)^\top\mathbf{X}^\top = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top = \mathbf{G}\,.$$

- Damit ist die erste Teilaussage bewiesen. Um die zweite Teilaussage zu beweisen, genügt es zu beachten, dass

$$\mathbf{G}\mathbf{X} = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\right)\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}\,,$$

  wobei sich die vorletzte Gleichheit aus Lemma 3.6 ergibt.
- Die dritte Teilaussage lässt sich wie folgt zeigen:
  - Aus Lemma 3.2 und aus Lemma 3.6 ergibt sich, dass

$$r = \operatorname{rg}(\mathbf{X}) = \operatorname{rg}\left(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\right) \le \operatorname{rg}\left(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\right) \le \operatorname{rg}(\mathbf{X}) = r\,,$$

    d.h.,

$$\operatorname{rg}\left(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\right) = r\,. \tag{62}$$

– Aus Lemma 3.5 ergibt sich, dass die Matrix $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ idempotent ist, denn es gilt

$$\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big) = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^- \underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top}_{=\mathbf{X}^\top} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top .$$

– Weil $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ außerdem symmetrisch ist, folgt aus Lemma 1.3 mit Hilfe von (62), dass

$$\begin{aligned} \mathrm{sp}(\mathbf{G}) &= \mathrm{sp}\big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big) \\ &= \mathrm{sp}(\mathbf{I}_n) - \mathrm{sp}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big) \\ &= n - \mathrm{rg}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big) = n - r\,. \end{aligned}$$

□

Aus Lemma 3.8 ergibt sich nun eine Formel für den Erwartungswert des in (60) betrachteten ML–Schätzers $\overline{\sigma}^2$.

Theorem 3.13 *Es gilt*

$$\mathbb{E}\,\overline{\sigma}^2 = \frac{n-r}{n}\,\sigma^2\,. \tag{63}$$

Beweis Für den in (60) betrachteten ML–Schätzer $\overline{\sigma}^2 = \big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol\beta}\big)^\top\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol\beta}\big)/n$ ergibt sich mit den in Lemma 3.8 hergeleiteten Eigenschaften der Matrix $\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$, dass

$$\begin{aligned} \mathbb{E}\,\overline{\sigma}^2 &= \frac{1}{n}\,\mathbb{E}\Big(\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol\beta}\big)^\top\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol\beta}\big)\Big) \\ &= \frac{1}{n}\,\mathbb{E}\Big(\mathbf{Y}^\top\big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)^\top\big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\mathbf{Y}\Big) \\ &= \frac{1}{n}\,\mathbb{E}\Big(\mathbf{Y}^\top\mathbf{G}^\top\mathbf{G}\mathbf{Y}\Big) = \frac{1}{n}\,\mathbb{E}\left(|\mathbf{G}\mathbf{Y}|^2\right) = \frac{1}{n}\,\mathbb{E}\left(|\mathbf{G}(\mathbf{X}\boldsymbol\beta + \boldsymbol\varepsilon)|^2\right) \\ &= \frac{1}{n}\,\mathbb{E}\left(|\mathbf{G}\boldsymbol\varepsilon|^2\right) = \frac{1}{n}\,\mathbb{E}\Big(\boldsymbol\varepsilon^\top\mathbf{G}^\top\mathbf{G}\boldsymbol\varepsilon\Big) = \frac{1}{n}\,\mathbb{E}\Big(\boldsymbol\varepsilon^\top\mathbf{G}\boldsymbol\varepsilon\Big) \\ &= \frac{\sigma^2}{n}\,\mathrm{sp}(\mathbf{G}) = \frac{n-r}{n}\,\sigma^2\,. \end{aligned}$$

□

Beachte Mit der Schreibweise

$$S^2 = \frac{1}{n-r}\,\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol\beta}\big)^\top\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol\beta}\big) \qquad \text{bzw.} \qquad S^2 = \frac{1}{n-r}\,\boldsymbol\varepsilon^\top\mathbf{G}\boldsymbol\varepsilon \tag{64}$$

ergibt sich aus Theorem 3.13, dass $\mathbb{E}\,S^2 = \sigma^2$, d.h., S^2 ist ein erwartungstreuer Schätzer für σ^2.

Um die Verteilungen der Schätzer $\overline{\boldsymbol\beta}$ bzw. S^2 bestimmen zu können, benötigen wir den Begriff der singulären multivariaten Normalverteilung, vgl. Abschnitt 1.2.5.

Theorem 3.14 *Sei* $\mathrm{rg}(\mathbf{X}) = r \le m$. *Dann gilt*

$$\overline{\boldsymbol\beta} \sim \mathrm{N}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol\beta,\, \sigma^2(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\big) \tag{65}$$

und

$$\frac{(n-r)S^2}{\sigma^2} \sim \chi^2_{n-r}\,, \tag{66}$$

wobei die Zufallsvariablen $\overline{\boldsymbol\beta}$ *und* S^2 *unabhängig sind.*

Beweis

- Für den in (60) gegebenen Schätzer $\overline{\boldsymbol{\beta}}$ gilt
$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = \big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\big(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\big) = \boldsymbol{\mu}+\mathbf{B}\boldsymbol{\varepsilon}\,,$$
wobei
$$\boldsymbol{\mu} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\,, \qquad \mathbf{B} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\,, \qquad \boldsymbol{\varepsilon}\sim \mathrm{N}(\mathbf{o},\,\sigma^2\mathbf{I}_n)\,.$$
- Aus der Definition der (singulären) multivariaten Normalverteilung ergibt sich nun, dass $\overline{\boldsymbol{\beta}}\sim \mathrm{N}(\boldsymbol{\mu},\,\mathbf{K})$, wobei
$$\mathbf{K} = \sigma^2\mathbf{B}\mathbf{B}^\top = \sigma^2(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\,.$$
- Damit ist (65) bewiesen. Um die Gültigkeit von (66) zu zeigen, nutzen wir die im Beweis von Theorem 3.13 hergeleitete Identität
$$S^2 = \frac{1}{n-r}\,\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}\,, \qquad \text{wobei } \mathbf{G} = \mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top. \tag{67}$$
- Weil $\boldsymbol{\varepsilon}\sim \mathrm{N}(\mathbf{o},\,\sigma^2\mathbf{I}_n)$ und weil wir in Lemma 3.8 gezeigt hatten, dass
  - die Matrix $\mathbf{G}$ idempotent und symmetrisch ist
  - mit $\mathrm{rg}(\mathbf{G}) = n-r$,

  ergibt sich aus Theorem 1.9, dass die quadratische Form $(n-r)S^2/\sigma^2$ eine (zentrale) χ^2–Verteilung mit $n-r$ Freiheitsgraden hat, d.h., $(n-r)S^2/\sigma^2\sim\chi^2_{n-r}$.
- Weil jede idempotente und symmetrische Matrix gleichzeitig auch nichtnegativ definit ist und weil wegen Lemma 3.5
$$\mathbf{B}\mathbf{G} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big) = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top - (\mathbf{X}^\top\mathbf{X})^-\underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top}_{=\mathbf{X}^\top} = \mathbf{0}\,,$$
ergibt sich aus Theorem 1.10, dass die Zufallsvariablen $\mathbf{B}\boldsymbol{\varepsilon}$ und $\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}$ unabhängig sind. Damit sind auch die Zufallsvariablen $\overline{\boldsymbol{\beta}}$ und S^2 unabhängig. □

3.3.2 Tests linearer Hypothesen

In diesem Abschnitt diskutieren wir eine verallgemeinerte Version des Tests für Linearformen von $\boldsymbol{\beta}$, den wir in Abschnitt 2.2.3 für den Fall von Designmatrizen $\mathbf{X}$ mit vollem Spaltenrang betrachtet hatten, vgl. Theorem 2.11. Jetzt nehmen wir dagegen an, dass $\mathrm{rg}(\mathbf{X}) = r < m$.

- Sei $s\in\{1,\ldots,m\}$, sei $\mathbf{H}$ eine $s\times m$ Matrix mit vollem Rang $\mathrm{rg}(\mathbf{H}) = s$, und sei $\mathbf{d}\in\mathbb{R}^s$.
- Getestet werden soll die Hypothese
$$H_0:\mathbf{H}\boldsymbol{\beta} = \mathbf{d} \qquad \text{versus} \qquad H_1:\mathbf{H}\boldsymbol{\beta}\neq\mathbf{d}\,, \tag{68}$$
wobei angenommen wird, dass die Eintragungen der Matrix $\mathbf{H} = (\mathbf{h}_1,\ldots,\mathbf{h}_s)^\top$ und die Komponenten des Vektors $\mathbf{d} = (d_1,\ldots,d_s)^\top$ bekannt sind.
- Zur Verifizierung der in (68) betrachteten Nullhypothese $H_0:\mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ wird (ähnlich wie in Theorem 2.11) eine Testgröße konstruiert, deren Verteilung *nicht* von dem unbekannten Parametervektor $(\boldsymbol{\beta},\sigma^2)$ abhängt.
- Hierfür wird der folgende Begriff eingeführt, wobei vorausgesetzt wird , dass die Komponenten des Vektors $\mathbf{H}\boldsymbol{\beta}$ erwartungstreu schätzbar sind.

Definition Die Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ heißt *testbar*, wenn sämtliche Komponenten $\mathbf{h}_1^\top\boldsymbol{\beta}, \ldots, \mathbf{h}_s^\top\boldsymbol{\beta}$ des Vektors $\mathbf{H}\boldsymbol{\beta}$ schätzbare Funktionen des Parametervektors $\boldsymbol{\beta}$ sind.

Beachte Aus Theorem 3.9 folgt, dass die Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ genau dann testbar ist, wenn

- es eine $s \times n$ Matrix $\mathbf{C}$ gibt, so dass
$$\mathbf{H} = \mathbf{C}\mathbf{X}\,, \tag{69}$$

bzw.

- die Matrix $\mathbf{H}$ der folgenden Gleichung genügt:
$$\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{H}\,. \tag{70}$$

Bei der Konstruktion einer Testgröße zur Verifizierung der in (68) betrachteten Nullhypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ ist der folgende Hilfssatz nützlich.

Lemma 3.9

- *Sei $s \le m$, sei $\mathbf{H}$ eine $s \times m$ Matrix mit vollem Rang* $\mathrm{rg}(\mathbf{H}) = s$, *die der Gleichung* (69) *bzw.* (70) *genügt, und sei* $(\mathbf{X}^\top\mathbf{X})^-$ *eine beliebige verallgemeinerte Inverse von* $\mathbf{X}^\top\mathbf{X}$.
- *Dann ist die $s \times s$ Matrix* $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top$ *positiv definit (und damit invertierbar).*

Beweis

- Man kann zeigen, dass die symmetrische $m \times m$ Matrix $\mathbf{X}^\top\mathbf{X}$ mit $\mathrm{rg}(\mathbf{X}^\top\mathbf{X}) = r < m$ dargestellt werden kann in der Form:
$$\mathbf{X}^\top\mathbf{X} = \mathbf{P}^{-1}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^{-1}\,,$$
wobei die $m \times m$ Matrix $\mathbf{P}$ invertierbar und symmetrisch ist; vgl. auch die Aussage von Lemma 3.3.
- Im Beweis von Lemma 3.4 hatten wir gezeigt, dass dann durch
$$\mathbf{P}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-r} \end{pmatrix}\mathbf{P} = \mathbf{P}\mathbf{P}$$
eine verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$ gegeben ist, die offenbar positiv definit ist.
- Aus Lemma 1.8 folgt nun, dass auch die Matrix $\mathbf{H}\mathbf{P}\mathbf{P}\mathbf{H}^\top$ positiv definit ist.
- Hieraus folgt, dass $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top$ auch für jede beliebige verallgemeinerte Inverse $(\mathbf{X}^\top\mathbf{X})^-$ von $\mathbf{X}^\top\mathbf{X}$ positiv definit ist, denn aus (69) ergibt sich, dass
$$\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top = \mathbf{C}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{C}^\top = \mathbf{C}\mathbf{X}\mathbf{P}\mathbf{P}\mathbf{X}^\top\mathbf{C}^\top = \mathbf{H}\mathbf{P}\mathbf{P}\mathbf{H}^\top\,,$$
wobei sich die zweite Gleichheit aus Lemma 3.7 ergibt. □

Beachte

- Aus Lemma 3.9 ergibt sich, dass die Testgröße $T_\mathbf{H}$ mit
$$T_\mathbf{H} = \frac{\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\bigr)}{sS^2} \tag{71}$$
wohldefiniert ist, wobei $\overline{\boldsymbol{\beta}}$ und S^2 die in (60) bzw. (64) gegebenen Schätzer für $\boldsymbol{\beta}$ bzw. σ^2 sind.

- Diese Testgröße ist eine Verallgemeinerung der entsprechenden Testgröße $T_{\mathbf{H}}$, die in Abschnitt 2.2.3 für Designmatrizen mit vollem Rang betrachtet wurde. Die Verteilung der in (71) gegebenen Testgröße $T_{\mathbf{H}}$ lässt sich wie folgt bestimmen.

Theorem 3.15 *Die Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ sei testbar. Unter $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ gilt dann $T_{\mathbf{H}} \sim \mathrm{F}_{s,n-r}$, d.h., die in (71) gegebene Testgröße $T_{\mathbf{H}}$ ist F-verteilt mit $(s, n-r)$ Freiheitsgraden.*

Beweis

- Aus der Definitionsgleichung (60) von $\overline{\boldsymbol{\beta}}$ ergibt sich, dass

$$\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} - \mathbf{d} = \mathbf{H}\bigl((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\bigr)\bigl(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\bigr) - \mathbf{d} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\varepsilon}\,,$$

  wobei

$$\boldsymbol{\mu} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{d} = \mathbf{C}\underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}}_{=\mathbf{X}}\boldsymbol{\beta} - \mathbf{d} = \mathbf{H}\boldsymbol{\beta} - \mathbf{d} = \mathbf{o}$$

  und $\mathbf{B} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ bzw. $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I}_n)$.
- Für den Zähler $Z = \bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\bigr)$ der in (71) gegebenen Testgröße $T_{\mathbf{H}}$ gilt somit

$$\begin{aligned} Z &= \boldsymbol{\varepsilon}^\top\mathbf{B}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\mathbf{B}\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}^\top\mathbf{A}\boldsymbol{\varepsilon}\,, \end{aligned}$$

  - wobei die Matrix $\mathbf{A} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ idempotent ist, denn wegen (70) gilt

$$\begin{aligned} \mathbf{A}^2 &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1} \\ &\quad\times \underbrace{\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}}_{=\mathbf{H}}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top \\ &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top = \mathbf{A}\,. \end{aligned}$$

  - Weil $\mathbf{A}$ auch symmetrisch ist mit $\mathrm{rg}(\mathbf{A}) = s$, ergibt sich aus Theorem 1.9, dass die quadratische Form Z/σ^2 eine χ^2–Verteilung mit s Freiheitsgraden hat.
- In Theorem 3.14 wurde außerdem gezeigt, dass $(n-r)S^2/\sigma^2 \sim \chi^2_{n-r}$ und dass die Zufallsvariablen $\overline{\boldsymbol{\beta}}$ und S^2 unabhängig sind.
- Damit sind auch die Zufallsvariablen Z und S^2 unabhängig, und es gilt

$$T_{\mathbf{H}} = \frac{Z/s\sigma^2}{S^2/\sigma^2} \sim \mathrm{F}_{s,n-r}\,. \qquad \square$$

Beachte

- Die Wahl der Testgröße $T_{\mathbf{H}}$ in (71) kann wie folgt motiviert werden: Ähnlich wie im Beweis von Theorem 3.15 ergibt sich aus Theorem 1.9, dass die quadratische Form Z/σ^2 mit

$$Z = \bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\bigr)$$

  im allgemeinen (d.h., ohne die Gültigkeit von $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ vorauszusetzen) eine nichtzentrale χ^2–Verteilung $\chi^2_{s,\lambda}$ hat mit

$$\lambda = \frac{\bigl(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\bigr)}{\sigma^2}\,.$$

- Hieraus folgt, dass
$$\mathbb{E}\Big(\frac{Z}{\sigma^2}\Big) = \frac{d}{dt}\,\mathbb{E}\exp\Big(\frac{tZ}{\sigma^2}\Big)\Big|_{t=0} = s+\lambda\,,$$
wobei sich die letzte Gleichheit aus der Formel für die momenterzeugende Funktion der $\chi^2_{s,\lambda}$–Verteilung ergibt, die im Beweis von Theorem 1.8 hergeleitet wurde.
  - Mit anderen Worten: Es gilt
$$\mathbb{E}\Big(\frac{Z}{s}\Big) = \sigma^2 + \frac{\big(\mathbf{H}\boldsymbol{\beta}-\mathbf{d}\big)^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\boldsymbol{\beta}-\mathbf{d}\big)}{s}\,, \tag{72}$$
und aus Theorem 3.13 ergibt sich, dass $\mathbb{E}\,(S^2)=\sigma^2$.
  - Unter der Nullhypothese $H_0 : \mathbf{H}\boldsymbol{\beta}=\mathbf{d}$ sind somit die Erwartungswerte von Zähler und Nenner der Testgröße $T_\mathbf{H}$ gleich.
- Andererseits ergibt sich aus Lemma 1.8 und Lemma 3.9, dass die inverse Matrix $\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}$ positiv definit ist und dass somit
$$\big(\mathbf{H}\boldsymbol{\beta}-\mathbf{d}\big)^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\boldsymbol{\beta}-\mathbf{d}\big) > 0\,,$$
wenn die Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta}=\mathbf{d}$ falsch ist. Aus (72) folgt dann in diesem Fall, dass
$$\mathbb{E}\Big(\frac{Z}{s}\Big) > \sigma^2 = \mathbb{E}\,(S^2)\,. \tag{73}$$
- Allgemein gilt (wegen der Unabhängigkeit von Z und S^2), dass $\mathbb{E}\,T_\mathbf{H} = \mathbb{E}\,\big(Z/s\big)\mathbb{E}\,(1/S^2)$, und aus der Jensen-Ungleichung folgt, dass $\mathbb{E}\,(1/S^2) > 1/\mathbb{E}\,(S^2)$.
- Aus (73) ergibt sich somit, dass
$$\mathbb{E}\,T_\mathbf{H} > \frac{\mathbb{E}\Big(\dfrac{Z}{s}\Big)}{\mathbb{E}\,(S^2)} > 1\,,$$
wenn H_0 falsch ist.
- Es liegt deshalb nahe, die Nullhypothese $H_0 : \mathbf{H}\boldsymbol{\beta}=\mathbf{d}$ abzulehnen, wenn die Testgröße $T_\mathbf{H}$ Werte annimmt, die signifikant größer als 1 sind.
- Wegen der in Theorem 3.15 hergeleiteten Verteilungseigenschaft der Testgröße $T_\mathbf{H}$ wird somit H_0 abgelehnt, wenn $T_\mathbf{H} > \mathrm{F}_{s,n-r,1-\alpha}$.

In manchen Fällen ist es zweckmäßiger, eine *alternative Darstellung* der in (71) gegebenen Testgröße $T_\mathbf{H}$ zu betrachten. Hierfür definieren wir die folgenden beiden *Summen von quadratischen Abweichungen SSE* bzw. SSE_H (**S**ums of **S**quared **E**rrors) mit
$$SSE = (\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}})^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}})\,,\qquad \overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \tag{74}$$
und
$$SSE_H = (\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H)\,,\qquad \overline{\boldsymbol{\beta}}_H = \overline{\boldsymbol{\beta}} - (\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\big)\,. \tag{75}$$

Theorem 3.16 *Für die in* (71) *gegebene Testgröße $T_\mathbf{H}$ gilt*
$$T_\mathbf{H} = \frac{(SSE_H - SSE)/s}{SSE/(n-r)}\,. \tag{76}$$

Beweis

- Es gilt

$$\begin{aligned} SSE_H &= (\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H) \\ &= \bigl(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}+\mathbf{X}(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}_H)\bigr)^\top\bigl(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}+\mathbf{X}(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}_H)\bigr) \\ &= (\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}})^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}})+(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}_H)^\top\mathbf{X}^\top\mathbf{X}(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}_H)\,, \end{aligned}$$

  weil sich aus den Teilaussagen 1 und 2 von Lemma 3.8 ergibt, dass

$$\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}})=\mathbf{X}^\top\bigl(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\bigr)\mathbf{Y}=\underbrace{\mathbf{X}^\top\mathbf{G}}_{=\mathbf{0}}\mathbf{Y}=\mathbf{o}\,.$$

- Aus (69), d.h. $\mathbf{H}=\mathbf{C}\mathbf{X}$, und aus Lemma 3.7 ergibt sich somit, dass

$$\begin{aligned} SSE_H &= (\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}})^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}) \\ &\quad+\bigl(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\underbrace{\mathbf{H}\bigl((\mathbf{X}^\top\mathbf{X})^-\bigr)^\top\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top}_{=\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top} \\ &\quad\times\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\bigr) \\ &= SSE+\bigl(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\bigr)\,. \end{aligned}$$

□

Beachte

- Aus der Definitionsgleichung (75) von $\overline{\boldsymbol{\beta}}_H$ ergibt sich, dass

$$\mathbf{H}\overline{\boldsymbol{\beta}}_H=\mathbf{H}\Bigl(\overline{\boldsymbol{\beta}}-(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\bigr)\Bigr)=\mathbf{d}\,,$$

  d.h., der in (75) gegebene Zufallsvektor $\overline{\boldsymbol{\beta}}_H$ nimmt nur Werte in dem *eingeschränkten Parameterraum* $\Theta_H=\bigl\{\boldsymbol{\beta}\in\mathbb{R}^m:\ \mathbf{H}\boldsymbol{\beta}=\mathbf{d}\bigr\}$ an.
- Außerdem kann man leicht zeigen, dass $\overline{\boldsymbol{\beta}}_H$ den mittleren quadratischen Fehler $e(\boldsymbol{\beta})$ für alle $\boldsymbol{\beta}\in\Theta_H$ minimiert, wobei

$$e(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^n\bigl(Y_i-(\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_mx_{im})\bigr)^2\,.$$

3.3.3 Konfidenzbereiche

- Bei der Konstruktion von Konfidenzbereichen gehen wir ähnlich wie in Abschnitt 2.2.4 vor, wo der Fall betrachtet wurde, dass die Designmatrix $\mathbf{X}$ vollen Rang $\operatorname{rg}(\mathbf{X})=m$ hat. Dabei nehmen wir jetzt allerdings so wie in Abschnitt 3.3.2 an, dass $\operatorname{rg}(\mathbf{X})=r<m$.
- Sei $s\in\{1,\ldots,m\}$, und $\mathbf{H}$ sei eine $s\times m$ Matrix mit vollem Rang $\operatorname{rg}(\mathbf{H})=s$, deren Eintragungen bekannt seien, wobei $\mathbf{H}=(\mathbf{h}_1,\ldots,\mathbf{h}_s)^\top$.
- Dann ergibt sich unmittelbar aus Theorem 3.15 der folgende *Konfidenzbereich für den Vektor* $\mathbf{H}\boldsymbol{\beta}$ zum Niveau $1-\alpha\in(0,1)$.

Theorem 3.17 *Sämtliche Komponenten $\mathbf{h}_1^\top\boldsymbol{\beta},\ldots,\mathbf{h}_s^\top\boldsymbol{\beta}$ des Vektors $\mathbf{H}\boldsymbol{\beta}$ seien schätzbare Funktionen von $\boldsymbol{\beta}$. Dann ist der (zufällige) Ellipsoid*

$$E=\Bigl\{\mathbf{d}\in\mathbb{R}^s:\ \frac{\bigl(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}}-\mathbf{d}\bigr)}{sS^2}\le \mathrm{F}_{s,n-r,1-\alpha}\Bigr\} \tag{77}$$

ein Konfidenzbereich für $\mathbf{H}\boldsymbol{\beta}$ zum Niveau $1-\alpha\in(0,1)$, wobei $\overline{\boldsymbol{\beta}}$ und S^2 die in (60) bzw. (64) gegebenen Schätzer für $\boldsymbol{\beta}$ bzw. σ^2 sind.

Aus Theorem 3.17 ergibt sich insbesondere das folgende Resultat.

Korollar 3.1 *Für jedes $i \in \{1,\ldots,s\}$ ist durch*

$$(\underline{\theta},\overline{\theta}) = \Bigl(\mathbf{h}_i^\top\overline{\boldsymbol{\beta}} - \mathrm{t}_{n-r,1-\alpha/2}S\sqrt{\mathbf{h}_i^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}_i}\;,\; \mathbf{h}_i^\top\overline{\boldsymbol{\beta}} + \mathrm{t}_{n-r,1-\alpha/2}S\sqrt{\mathbf{h}_i^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}_i}\Bigr) \tag{78}$$

ein Konfidenzintervall $(\underline{\theta},\overline{\theta})$ für $\mathbf{h}_i^\top\boldsymbol{\beta}$ zum Niveau $1-\alpha \in (0,1)$ gegeben.

Beispiel

- Wir betrachten das folgende lineare Modell, vgl. N. Ravishanker und D.K. Dey (2002) *A First Course in Linear Model Theory*, Chapman & Hall/CRC, S. 235:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix},$$

  wobei $\boldsymbol{\varepsilon} = (\varepsilon_1,\varepsilon_2,\varepsilon_3)^\top \sim \mathrm{N}(\mathbf{o},\sigma^2\mathbf{I})$.
- Mit Hilfe von Korollar 3.1 soll ein Konfidenzintervall für $\beta_1+\beta_2/3+2\beta_3/3$ zum Niveau $1-\alpha = 0.95$ bestimmt werden.
- Weil $\mathrm{rg}(\mathbf{X}) = 2 < m = 3$, muss zunächst geprüft werden, ob

$$\mathbf{h}^\top\boldsymbol{\beta} = \beta_1+\beta_2/3+2\beta_3/3 \tag{79}$$

  mit $\mathbf{h}^\top = (1,1/3,2/3)$ eine erwartungstreu schätzbare Funktion von $\boldsymbol{\beta}^\top = (\beta_1,\beta_2,\beta_3)$ ist.
  - Gemäß Kriterium 1 in Theorem 3.9 ist dies genau dann der Fall, wenn es ein $\mathbf{c}^\top = (c_1,c_2,c_3) \in \mathbb{R}^3$ gibt, so dass $\mathbf{h}^\top = \mathbf{c}^\top\mathbf{X}$, d.h., wenn

$$\begin{aligned} 1 &= c_1+c_2+c_3 \\ 1/3 &= c_1+c_3 \\ 2/3 &= c_2\,. \end{aligned}$$

  - Weil dieses Gleichungssystem offenbar lösbar ist, ist somit $\mathbf{h}^\top\boldsymbol{\beta}$ schätzbar.
- Außerdem gilt

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

  und eine verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$ ist gegeben durch:

$$(\mathbf{X}^\top\mathbf{X})^- = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 3/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

- Hieraus folgt, dass $\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h} = 1/2$ und

$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & -1 & 1/2 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{bzw.} \qquad \overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} Y_2 \\ (Y_1-2Y_2+Y_3)/2 \\ 0 \end{pmatrix}.$$

- Somit ergibt sich, dass $(\mathbf{X}\overline{\boldsymbol{\beta}})^\top = \bigl((Y_1+Y_3)/2, Y_2, (Y_1+Y_3)/2\bigr)$ bzw.
$$\mathbf{h}^\top\overline{\boldsymbol{\beta}} = (Y_1+4Y_2+Y_3)/6 \qquad \text{und} \qquad S^2 = (Y_1-Y_3)^2/2\,.$$
- Das gesuchte Konfidenzintervall $(\underline{\theta}, \overline{\theta})$ für $\mathbf{h}^\top\boldsymbol{\beta}$ zum Niveau $1-\alpha = 0.95$ hat also die Form
$$(\underline{\theta}, \overline{\theta}) = \Bigl((Y_1+4Y_2+Y_3)/6 - Z\,,\,(Y_1+4Y_2+Y_3)/6 + Z\Bigr)$$
wobei $Z = t_{1,0.975}|Y_1 - Y_3|/2$.

In Verallgemeinerung von Theorem 2.12 leiten wir nun ein so genanntes *Scheffé–Konfidenzband* her, d.h. simultane Konfidenzintervalle für eine ganze Klasse von schätzbaren Funktionen des Parametervektors $\boldsymbol{\beta}$.

- Sei $s \in \{1, \ldots, m\}$, und $\mathbf{H}$ sei erneut eine $s\times m$ Matrix mit vollem Rang $\mathrm{rg}(\mathbf{H}) = s$, wobei $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_s)^\top$, so dass sämtliche Komponenten $\mathbf{h}_1^\top\boldsymbol{\beta}, \ldots, \mathbf{h}_s^\top\boldsymbol{\beta}$ des Vektors $\mathbf{H}\boldsymbol{\beta}$ schätzbare Funktionen von $\boldsymbol{\beta}$ sind.
- Weil $\mathbf{H}$ vollen (Zeilen-) Rang hat, sind die Vektoren $\mathbf{h}_1, \ldots, \mathbf{h}_s$ linear unabhängig und bilden die Basis eines s–dimensionalen linearen Unterraumes in $\mathbb{R}^m$, den wir mit $\mathcal{L} = \mathcal{L}(\mathbf{h}_1, \ldots, \mathbf{h}_s)$ bezeichnen.
- Wegen Theorem 3.10 ist $\mathbf{h}^\top\boldsymbol{\beta}$ für jedes $\mathbf{h} \in \mathcal{L}$ eine schätzbare Funktionen von $\boldsymbol{\beta}$.
- Gesucht ist eine Zahl $a_\gamma > 0$, so dass mit der (vorgegebenen) Wahrscheinlichkeit $\gamma \in (0,1)$
$$\mathbf{h}^\top\overline{\boldsymbol{\beta}} - a_\gamma Z_{\mathbf{h}} \le \mathbf{h}^\top\boldsymbol{\beta} \le \mathbf{h}^\top\overline{\boldsymbol{\beta}} + a_\gamma Z_{\mathbf{h}} \tag{80}$$
gleichzeitig für jedes $\mathbf{h} \in \mathcal{L}$ gilt, wobei $Z_{\mathbf{h}} = S\sqrt{\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}}$ und $\overline{\boldsymbol{\beta}}$ bzw. S^2 die in (60) bzw. (64) gegebenen Schätzer für $\boldsymbol{\beta}$ bzw. σ^2 sind.

Theorem 3.18 *Sei $a_\gamma = \sqrt{s\,\mathrm{F}_{s,n-r,\gamma}}$. Dann gilt*
$$\mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{h}\in\mathcal{L}} \frac{(\mathbf{h}^\top\overline{\boldsymbol{\beta}} - \mathbf{h}^\top\boldsymbol{\beta})^2}{S^2\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}} \le a_\gamma^2\right) = \gamma\,. \tag{81}$$

Beweis

- Ähnlich wie im Beweis von Theorem 2.12 ergibt sich aus der Ungleichung von Cauchy–Schwarz für Skalarprodukte, vgl. (66), dass
$$\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\bigr) = \max_{\mathbf{x}\neq\mathbf{o}} \frac{\Bigl(\mathbf{x}^\top\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\bigr)\Bigr)^2}{\mathbf{x}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)\mathbf{x}}\,,$$
wobei sich das Maximum über sämtliche Vektoren $\mathbf{x} \in \mathbb{R}^s$ mit $\mathbf{x} \neq \mathbf{o}$ erstreckt.
- Hieraus und aus Theorem 3.15 ergibt sich nun, dass
$$\begin{aligned}
\gamma &= \mathbb{P}_{\boldsymbol{\beta}}\Bigl(\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\bigr)^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)^{-1}\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\bigr) \le sS^2\,\mathrm{F}_{s,n-r,\gamma}\Bigr)\\
&= \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{x}\neq\mathbf{o}} \frac{\Bigl(\mathbf{x}^\top\bigl(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\bigr)\Bigr)^2}{\mathbf{x}^\top\bigl(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\bigr)\mathbf{x}} \le sS^2\,\mathrm{F}_{s,n-r,\gamma}\right)\\
&= \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{x}\neq\mathbf{o}} \frac{\Bigl((\mathbf{H}^\top\mathbf{x})^\top\overline{\boldsymbol{\beta}} - (\mathbf{H}^\top\mathbf{x})^\top\boldsymbol{\beta}\Bigr)^2}{(\mathbf{H}^\top\mathbf{x})^\top(\mathbf{X}^\top\mathbf{X})^-(\mathbf{H}^\top\mathbf{x})} \le sS^2\,\mathrm{F}_{s,n-r,\gamma}\right)\\
&= \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{h}\in\mathcal{L}} \frac{\Bigl(\mathbf{h}^\top\overline{\boldsymbol{\beta}} - \mathbf{h}^\top\boldsymbol{\beta}\Bigr)^2}{\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}} \le sS^2\,\mathrm{F}_{s,n-r,\gamma}\right).
\end{aligned}$$
$\square$

3.4 Beispiele

3.4.1 F–Test der ANOVA-Nullhypothese

- Wir betrachten das reparametrisierte Modell der *einfaktoriellen Varianzanalyse*, d.h., die Designmatrix $\mathbf{X}$ sei die in (13) gegebene $n \times (k+1)$ Matrix mit $\operatorname{rg}(\mathbf{X}) = k < m = k+1$, wobei

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & 0 & \ldots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \tag{82}$$

  und der Parametervektor $\boldsymbol{\beta}$ hat die Form $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^\top$.

- Getestet werden soll, ob die Stufen des Einflussfaktors signifikant sind, d.h., wir testen die ANOVA–Nullhypothese $H_0 : \alpha_1 = \ldots = \alpha_k$ (gegen die Alternative $H_1 : \alpha_i \neq \alpha_j$ für ein Paar $i, j \in \{1, \ldots, k\}$ mit $i \neq j$). Dabei nutzen wir den allgemeinen Testansatz von Theorem 3.15 bzw. Theorem 3.16.

- Eine äquivalente Formulierung der Nullhypothese $H_0 : \alpha_1 = \ldots = \alpha_k$ ist gegeben durch

$$H_0 : \alpha_1 - \alpha_2 = 0, \ldots, \alpha_1 - \alpha_k = 0 \qquad \text{bzw.} \qquad H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}\,, \tag{83}$$

  wobei $\mathbf{H}$ eine $(k-1) \times (k+1)$ Matrix ist mit

$$\mathbf{H} = \begin{pmatrix} 0 & 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \ldots & -1 & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & -1 \end{pmatrix} \tag{84}$$

- Es ist klar, dass $\mathbf{H}$ eine Matrix mit vollem Zeilenrang $\operatorname{rg}(\mathbf{H}) = k-1$ ist. Aus Theorem 3.10 ergibt sich außerdem, dass sämtliche Komponenten $\alpha_1 - \alpha_2, \ldots, \alpha_1 - \alpha_k$ des Vektors $\mathbf{H}\boldsymbol{\beta}$ schätzbare Funktionen von $\boldsymbol{\beta}$ sind.

- Mit anderen Worten: Die Matrix $\mathbf{H}$ genügt den Bedingungen von Theorem 3.15 bzw. Theorem 3.16, so dass zur Verifizierung der Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ die in Theorem 3.16 betrachtete Testgröße

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k-1)}{SSE/(n-k)}$$

  verwendet werden kann, wobei sich die in (74) bzw. (75) definierten Quadratsummen SSE und SSE_H wie folgt bestimmen lassen.

- *Zur Erinnerung*: In Abschnitt 3.2.1 hatten wir gezeigt, dass eine verallgemeinerte Inverse von $\mathbf{X}^\top\mathbf{X}$ gegeben ist durch (35), d.h.

$$\left(\mathbf{X}^\top\mathbf{X}\right)^- = \begin{pmatrix} \frac{1}{n} & 0 & 0 & 0 & \dots & 0 & 0 \\ -\frac{1}{n} & \frac{1}{n_1} & 0 & 0 & \dots & 0 & 0 \\ -\frac{1}{n} & 0 & \frac{1}{n_2} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -\frac{1}{n} & 0 & 0 & 0 & \dots & 0 & \frac{1}{n_k} \end{pmatrix}. \tag{85}$$

- Hieraus und aus (82) folgt, dass

$$\overline{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^-\mathbf{X}^\top\mathbf{Y} = \left(\overline{Y}_{\cdot\cdot}, \overline{Y}_{1\cdot} - \overline{Y}_{\cdot\cdot}, \ldots, \overline{Y}_{k\cdot} - \overline{Y}_{\cdot\cdot}\right)^\top$$

bzw.

$$\mathbf{X}\overline{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^-\mathbf{X}^\top\mathbf{Y} = \Bigl(\underbrace{\overline{Y}_{1\cdot}, \ldots, \overline{Y}_{1\cdot}}_{n_1}, \ldots, \underbrace{\overline{Y}_{k\cdot}, \ldots, \overline{Y}_{k\cdot}}_{n_k}\Bigr)^\top.$$

- Somit ergibt sich für die Quadratsumme $SSE = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})$, dass

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2. \tag{86}$$

Beachte

- Wegen der speziellen Gestalt (82) der Designmatrix $\mathbf{X}$ lässt sich die Formel (86) auch direkt aus der Tatsache herleiten, dass $\overline{\boldsymbol{\beta}}$ ein KQ–Schätzer ist. Und zwar gilt

$$\begin{aligned} SSE &= (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{k}\Bigl\{\min_{x\in\mathbb{R}}\sum_{j=1}^{n_i}\left(Y_{ij} - x\right)^2\Bigr\} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2. \end{aligned}$$

- Außerdem ergibt sich aus der Bemerkung am Ende von Abschnitt 3.3.2, dass

$$SSE_H = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H) = \min_{\boldsymbol{\beta}\in\Theta_H}\left|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right|^2,$$

wobei $\Theta_H = \{\boldsymbol{\beta} \in \mathbb{R}^{k+1} : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}\}$ und $\{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \Theta_H\} \subset \mathbb{R}^n$ die Menge derjenigen n–dimensionalen Vektoren ist, für die sämtliche Komponenten gleich sind.

- Somit gilt

$$SSE_H = \min_{x\in\mathbb{R}}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - x\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2, \tag{87}$$

weil der Mittelwert $\overline{Y}_{\cdot\cdot}$ die Quadratsumme $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - x\right)^2$ minimiert.

- Hieraus und aus (86) folgt, dass

$$SSE_H - SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\bigl(Y_{ij} - \overline{Y}_{\cdot\cdot}\bigr)^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Bigl(Y_{ij} - \overline{Y}_{i\cdot}\Bigr)^2 = \sum_{i=1}^{k} n_i\bigl(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\bigr)^2 ,$$

  wobei sich die letzte Gleichheit aus der Quadratsummenzerlegung

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\bigl(Y_{ij} - \overline{Y}_{\cdot\cdot}\bigr)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Bigl(Y_{ij} - \overline{Y}_{i\cdot}\Bigr)^2 + \sum_{i=1}^{k} n_i\bigl(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\bigr)^2$$

  ergibt, vgl. die Formel (9) in Theorem 3.1.

- Für die in Theorem 3.16 betrachtete Testgröße $T_{\mathbf{H}}$ gilt also, dass

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k-1)}{SSE/(n-k)} = \frac{(n-k)\sum\limits_{i=1}^{k} n_i\bigl(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\bigr)^2}{(k-1)\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\Bigl(Y_{ij} - \overline{Y}_{i\cdot}\Bigr)^2} \sim \mathrm{F}_{k-1,n-k} \,. \tag{88}$$

3.4.2 F–Tests für die zweifaktorielle Varianzanalyse

Wir konstruieren nun F–Tests für das in Abschnitt 3.1.3 eingeführte Modell der *zweifaktoriellen Varianzanalyse* mit balancierten Teilstichproben, d.h.,

- der Parametervektor $\boldsymbol{\beta}$ hat die Form

$$\boldsymbol{\beta} = \bigl(\mu, \alpha_1^{(1)}, \ldots, \alpha_{k_1}^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_{k_2}^{(2)}, \alpha_{11}, \ldots, \alpha_{k_1 k_2}\bigr)^\top ,$$

- die Designmatrix $\mathbf{X}$ hat die Dimension $n \times m$, wobei $n = rk_1k_2$ und $m = 1 + k_1 + k_2 + k_1k_2$,
- die Eintragungen von $\mathbf{X}$ bestehen nur aus Nullen und Einsen, und es gilt $\operatorname{rg}(\mathbf{X}) = k_1k_2 < m$.

Signifikanz der Einflussfaktoren

Wir konstruieren zunächst einen Test zur Untersuchung der Frage, ob die Stufen des ersten Einflussfaktors signifikant sind. Hierfür prüfen wir die Hypothese, ob die Effekte

$$\alpha_{i_1}^{(1)*} = \alpha_{i_1}^{(1)} + \frac{1}{k_2}\sum_{i_2=1}^{k_2}\alpha_{i_1 i_2}$$

des ersten Einflussfaktors, zuzüglich ihrer Wechselwirkungen gemittelt über sämtliche Stufen des zweiten Einflussfaktors, gleich sind. Mit anderen Worten: Wir testen die Hypothese

$$H_0 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} = 0 \quad \forall\, i_1 \in \{1, \ldots, k_1\} \quad \text{vs.} \quad H_1 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} \neq 0 \quad \text{für ein } i_1 \in \{1, \ldots, k_1\} , \tag{89}$$

wobei es eigentlich genügt, das Hypothesenpaar

$$H_0 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} = 0 \quad \forall\, i_1 \in \{2, \ldots, k_1\} \quad \text{vs.} \quad H_1 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} \neq 0 \quad \text{für ein } i_1 \in \{2, \ldots, k_1\}$$

zu betrachten.

- Man kann leicht zeigen, dass die Nullhypothese in (89) die Form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ hat,

– wobei

$$
\mathbf{H}=\begin{pmatrix}
0 & 1 & -1 & 0 & \dots & 0 & 0 & \dots & 0 & \frac{1}{k_2} & \dots & \frac{1}{k_2} & \frac{-1}{k_2} & \dots & \frac{-1}{k_2} & 0 & \dots & 0\\
0 & 1 & 0 & -1 & & 0 & 0 & \dots & 0 & \frac{1}{k_2} & \dots & \frac{1}{k_2} & 0 & \dots & 0 & 0 & \dots & 0\\
\vdots & \vdots & \vdots & & \ddots & & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \ddots & \vdots & \vdots\\
0 & 1 & 0 & & & -1 & 0 & \dots & 0 & \frac{1}{k_2} & \dots & \frac{1}{k_2} & 0 & \dots & 0 & \frac{-1}{k_2} & \dots & \frac{-1}{k_2}
\end{pmatrix}
$$

eine $(k_1-1)\times m$ Matrix mit vollem Zeilenrang $\mathrm{rg}(\mathbf{H})=k_1-1$ und mit Zeilenblöcken der Längen $1,\,1,\,k_1-1,\,k_2,\,k_2$ bzw. $(k_1-1)k_2$ ist

– und sämtliche Komponenten des Vektors $\mathbf{H}\boldsymbol{\beta}$ schätzbare Funktionen von $\boldsymbol{\beta}$ sind, denn es gilt

$$\alpha_1^{(1)*}-\alpha_{i_1}^{(1)*}=\frac{1}{k_2}\sum_{i_2=1}^{k_2}\Bigl(\theta_{1i_2}-\theta_{i_1i_2}\Bigr)\,.$$

- Zur Verifizierung der Hypothese $H_0:\mathbf{H}\boldsymbol{\beta}=\mathbf{o}$ kann somit erneut die in Theorem 3.16 betrachtete Testgröße $T_{\mathbf{H}}$ verwendet werden, wobei

$$T_{\mathbf{H}}=\frac{(SSE_H-SSE)/(k_1-1)}{SSE/(k_1k_2(r-1))}\sim \mathrm{F}_{k_1-1,\,k_1k_2(r-1)}$$

mit

$$SSE=\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\bigr)^2 \tag{90}$$

und

$$SSE_H-SSE=rk_2\sum_{i_1=1}^{k_1}\bigl(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)^2\,. \tag{91}$$

- Dabei lassen sich die Formeln (90) und (91) für die in (74) bzw. (75) definierten Quadratsummen SSE und SSE_H auf ähnliche Weise wie in Abschnitt 3.4.1 herleiten.
- Und zwar ergibt sich (90) mit der gleichen Minimierungstechnik, die bei der direkten Herleitung von (86) verwendet wurde. Darüber hinaus lässt sich die Quadratsumme SSE_H wie folgt bestimmen.
  – Sowie bisher ist $\Theta_H=\{\boldsymbol{\beta}\in\mathbb{R}^{1+k_1+k_2+k_1k_2}:\mathbf{H}\boldsymbol{\beta}=\mathbf{o}\}$ der eingeschränkte Parameterraum.
  – Wegen der speziellen Gestalt der Matrizen $\mathbf{X}$ und $\mathbf{H}$ kann bei der Minimierung in

$$SSE_H=(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H)=\min_{\boldsymbol{\beta}\in\Theta_H}\bigl|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\bigr|^2$$

die Menge $\{\mathbf{X}\boldsymbol{\beta}:\boldsymbol{\beta}\in\Theta_H\}\subset\mathbb{R}^{rk_1k_2}$ auf die folgende Weise (ähnlich wie in Formel (87)) durch eine Minimierung bezüglich der Menge $\mathbb{R}\times\mathbb{R}_H^{k_2}\times\mathbb{R}_H^{k_1k_2}\subset\mathbb{R}^{1+k_2+k_1k_2}$ derjenigen Vektoren $\mathbf{x}=(x,x_1,\ldots,x_{k_2},x_{11},\ldots,x_{k_1k_2})\in\mathbb{R}^{1+k_2+k_1k_2}$ ersetzt werden, die den folgenden Bedingungen genügen:

$$\sum_{i_2=1}^{k_2}x_{i_2}=\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}x_{i_1i_2}=0\,.$$

- Und zwar gilt

$$
\begin{aligned}
SSE_H &= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_2}\times\mathbb{R}_H^{k_1k_2}}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j}-(x+x_{i_2}+x_{i_1i_2})\bigr)^2\\
&= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_2}\times\mathbb{R}_H^{k_1k_2}}\Biggl\{\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\bigr)^2+k_1k_2r\bigl(\overline{Y}_{\cdot\cdot\cdot}-x\bigr)^2+k_2r\sum_{i_1=1}^{k_1}\bigl(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)^2\\
&\qquad +k_1r\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{\cdot i_2\cdot}-\overline{Y}_{\cdot\cdot\cdot}-x_{i_2}\bigr)^2+r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot i_2\cdot}+\overline{Y}_{\cdot\cdot\cdot}-x_{i_1i_2}\bigr)^2\Biggr\}\,,
\end{aligned}
$$

d.h.,

$$SSE_H = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\bigr)^2 + k_2r\sum_{i_1=1}^{k_1}\bigl(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)^2 .$$

- Hieraus und aus (90) ergibt sich (91).

Beachte

- Auf die gleiche Weise ergibt sich ein Test, um zu prüfen, ob die Stufen des zweiten Einflussfaktors signifikant sind. Dabei prüfen wir die Hypothese, ob die Effekte

$$\alpha_{i_2}^{(2)*} = \alpha_{i_2}^{(2)} + \frac{1}{k_1}\sum_{i_1=1}^{k_1}\alpha_{i_1i_2}$$

  des zweiten Einflussfaktors, zuzüglich ihrer Wechselwirkungen gemittelt über sämtliche Stufen des ersten Einflussfaktors, gleich sind.
- Wir testen also die Hypothese

$$H_0 : \alpha_1^{(2)*}-\alpha_2^{(2)*}=0\,,\ \ldots\,,\ \alpha_1^{(2)*}-\alpha_{k_2}^{(2)*}=0 \quad \text{versus} \quad H_1 : \alpha_1^{(2)*}-\alpha_{i_2}^{(2)*}\neq 0 \quad \text{für ein } i_2\in\{1,\ldots,k_2\}\,.$$

- Als Testgröße ergibt sich in diesem Fall:

$$T_{\mathbf{H}} = \frac{k_1k_2(r-1)rk_1\sum\limits_{i_2=1}^{k_2}\bigl(\overline{Y}_{\cdot\, i_2\cdot}-\overline{Y}_{\cdot\cdot\cdot}\bigr)^2}{(k_2-1)\sum\limits_{i_1=1}^{k_1}\sum\limits_{i_2=1}^{k_2}\sum\limits_{j=1}^{r}\bigl(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\bigr)^2} \sim \mathrm{F}_{k_2-1,\,k_1k_2(r-1)}$$

Wechselwirkungen zwischen den beiden Einflussfaktoren

Wir konstruieren nun einen Test, um zu prüfen, ob es signifikante Wechselwirkungen zwischen den beiden Einflussfaktoren gibt. Hierfür wird die Hypothese

$$H_0 : \alpha_{11}^* - \alpha_{i_1i_2}^* = 0 \quad \forall\,(i_1,i_2)\in\{1,\ldots,k_1\}\times\{1,\ldots,k_2\} \tag{92}$$

getestet, wobei

$$\alpha_{i_1i_2}^* = \alpha_{i_1i_2} - \overline{\alpha}_{i_1\,\cdot} - \overline{\alpha}_{\cdot\, i_2} + \overline{\alpha}_{\cdot\,\cdot}$$

und

$$\overline{\alpha}_{i_1\,\cdot} = \frac{1}{k_2}\sum_{i_2=1}^{k_2}\alpha_{i_1i_2}\,,\qquad \overline{\alpha}_{\cdot\, i_2} = \frac{1}{k_1}\sum_{i_1=1}^{k_1}\alpha_{i_1i_2}\,,\qquad \overline{\alpha}_{\cdot\,\cdot} = \frac{1}{k_1k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\alpha_{i_1i_2}\,.$$

- Auf ähnliche Weise wie bisher kann man zeigen, dass sich die in (92) betrachtete Hypothese in der Form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ schreiben lässt, wobei
  - $\mathbf{H}$ eine $(k_1k_2-1)\times m$ Matrix mit vollem Zeilenrang $\mathrm{rg}(\mathbf{H}) = k_1k_2-1$ ist und
  - sämtliche Komponenten des Vektors $\mathbf{H}\boldsymbol{\beta}$ schätzbare Funktionen von $\boldsymbol{\beta}$ sind, denn es gilt

$$\alpha_{i_1i_2}^* = \theta_{i_1i_2} - \frac{1}{k_2}\sum_{i_2=1}^{k_2}\theta_{i_1i_2} - \frac{1}{k_1}\sum_{i_1=1}^{k_1}\theta_{i_1i_2} + \frac{1}{k_1k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\theta_{i_1i_2}\,.$$

- Zur Verifizierung der Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ kann somit die in Theorem 3.16 betrachtete Testgröße

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k_1k_2-1)}{SSE/(k_1k_2(r-1))}$$

  verwendet werden, wobei SSE so wie bisher durch (90) gegeben ist, während sich die Quadratsumme SSE_H aus den folgenden Überlegungen ergibt.

- Wegen der speziellen Gestalt der Matrizen $\mathbf{X}$ und $\mathbf{H}$ kann bei der Minimierung in

$$SSE_H = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H) = \min_{\boldsymbol{\beta}\in\Theta_H} \left|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right|^2$$

die Menge $\{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \Theta_H\} \subset \mathbb{R}^{rk_1k_2}$ ähnlich wie bisher durch die Menge $\mathbb{R} \times \mathbb{R}_H^{k_1} \times \mathbb{R}_H^{k_2} \subset \mathbb{R}^{1+k_1+k_2}$ derjenigen Vektoren

$$\mathbf{x} = (x, x_1^{(1)}, \ldots, x_{k_1}^{(1)}, x_1^{(2)}, \ldots, x_{k_2}^{(2)}) \in \mathbb{R}^{1+k_1+k_2}$$

ersetzt werden, die den folgenden Bedingungen genügen:

$$\sum_{i_1=1}^{k_1} x_{i_1}^{(1)} = \sum_{i_2=1}^{k_2} x_{i_2}^{(2)} = 0\,.$$

- Und zwar gilt

$$SSE_H = \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_1}\times\mathbb{R}_H^{k_2}} \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j} - (x + x_{i_1}^{(1)} + x_{i_2}^{(2)})\bigr)^2$$

bzw.

$$\begin{aligned}
SSE_H &= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_1}\times\mathbb{R}_H^{k_2}} \Biggl\{ \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\bigr)^2 \\
&\quad + k_1k_2r\bigl(\overline{Y}_{\cdots} - x\bigr)^2 + k_2r\sum_{i_1=1}^{k_1}\bigl(\overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdots} - x_{i_1}^{(1)}\bigr)^2 + k_1r\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{\cdot\, i_2\cdot} - \overline{Y}_{\cdots} - x_{i_2}^{(2)}\bigr)^2 \\
&\quad + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdot\, i_2\cdot} + \overline{Y}_{\cdots}\bigr)^2 \Biggr\} \\
&= \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\bigr)^2 + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdot\, i_2\cdot} + \overline{Y}_{\cdots}\bigr)^2\,.
\end{aligned}$$

- Hieraus und aus (90) folgt, dass

$$SSE_H - SSE = r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdot\, i_2\cdot} + \overline{Y}_{\cdots}\bigr)^2\,. \tag{93}$$

- Für die in Theorem 3.16 betrachtete Testgröße $T_{\mathbf{H}}$ gilt also, dass

$$\begin{aligned}
T_{\mathbf{H}} &= \frac{(SSE_H - SSE)/(k_1k_2 - 1)}{SSE/\bigl(k_1k_2(r-1)\bigr)} \\
&= \frac{k_1k_2(r-1)r\sum\limits_{i_1=1}^{k_1}\sum\limits_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdot\, i_2\cdot} + \overline{Y}_{\cdots}\bigr)^2}{(k_1k_2-1)\sum\limits_{i_1=1}^{k_1}\sum\limits_{i_2=1}^{k_2}\sum\limits_{j=1}^{r}\bigl(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\bigr)^2} \sim \mathrm{F}_{k_1k_2-1,\,k_1k_2(r-1)}\,.
\end{aligned}$$

3.4.3 Zweifaktorielle Varianzanalyse mit hierarchischer Klassifikation

- Anstelle des in Abschnitt 3.1.3 eingeführten Modells der zweifaktoriellen Varianzanalyse mit Wechselwirkungen wird manchmal das folgende Modell der zweifaktoriellen Varianzanalyse mit *hierarchischer Klassifikation* der Stufenpaare i_1, i_2 der beiden Einflussfaktoren betrachtet.

- Dabei betrachten wir nun die Darstellung
$$\theta_{i_1 i_2} = \mu + \alpha^{(1)}_{i_1} + \alpha^{(2|1)}_{i_2|i_1}\,, \qquad \forall\, i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2 \tag{94}$$
der Erwartungswerte $\theta_{i_1 i_2} = \mathbb{E}\, Y_{i_1 i_2 j}$ der Stichprobenvariablen $Y_{i_1 i_2 j}$.
- Mit anderen Worten: In jede der k_1 Stufen des ersten, d.h. übergeordneten Einflussfaktors sind k_2 Stufen des zweiten (untergeordneten) Einflussfaktors eingebettet.
- Diese Situation kann beispielsweise bei klinischen Studien auftreten, die gleichzeitig in k_1 Ländern (übergeordneter Einflussfaktor) und dabei jeweils in k_2 Krankenhäusern (untergeordneter Einflussfaktor) durchgeführt werden.
- Der Parametervektor $\boldsymbol{\beta}$ hat dann die Dimension $m = 1 + k_1 + k_1 k_2$ mit
$$\boldsymbol{\beta} = \bigl(\mu, \alpha^{(1)}_1, \ldots, \alpha^{(1)}_{k_1}, \alpha^{(2|1)}_{1|1}, \ldots, \alpha^{(2|1)}_{k_2|k_1}\bigr)^\top .$$
- Dabei wird
  - μ erneut als *allgemeines Mittel* der Erwartungswerte $\mathbb{E}\, Y_{i_1 i_2 j}$ der Stichprobenvariablen $Y_{i_1 i_2 j}$ aufgefasst,
  - $\alpha^{(1)}_{i_1}$ wird *Effekt* der i_1-ten Stufe des übergeordneten Einflussfaktors genannt, und
  - $\alpha^{(2|1)}_{i_2|i_1}$ heißt *Effekt* der i_2-ten Stufe des untergeordneten Einflussfaktors bei Vorliegen der i_1-ten Stufe des übergeordneten Einflussfaktors.
- Wir betrachten wiederum lediglich den balancierten Fall, d.h., wir setzen voraus, dass sämtliche $k_1 \cdot k_2$ Teilstichproben $(Y_{i_1 i_2 j},\ j = 1,\ldots,n_{i_1 i_2})$ identische Stichprobenumfänge besitzen.
- Es gilt also $n_{i_1 i_2} = r$ für alle $i_1 = 1,\ldots,k_1$ und $i_2 = 1,\ldots,k_2$ mit $r = n/(k_1 k_2)$, die Designmatrix $\mathbf{X}$ hat die Dimension $n \times m$ mit $n = r k_1 k_2$ und $m = 1 + k_1 + k_1 k_2$, und die Eintragungen von $\mathbf{X}$ bestehen nur aus Nullen und Einsen; $\mathrm{rg}(\mathbf{X}) = k_1 k_2 < m$.

Signifikanz des übergeordneten Einflussfaktors

- Genauso wie in Abschnitt 3.4.2 lässt sich zunächst ein Test zur Untersuchung der Frage konstruieren, ob die Stufen des übergeordneten Einflussfaktors signifikant sind. Hierfür prüfen wir die Hypothese, ob die gemittelten Effekte $\alpha^{(1)*}_{i_1}$ gleich sind, wobei
$$\alpha^{(1)*}_{i_1} = \alpha^{(1)}_{i_1} + \frac{1}{k_2}\sum_{i_2=1}^{k_2} \alpha^{(2|1)}_{i_2|i_1}\,.$$
- Mit anderen Worten: Wir testen die Hypothese
$$H_0 : \alpha^{(1)*}_1 - \alpha^{(1)*}_{i_1} = 0 \quad \forall\, i_1 \in \{1,\ldots,k_1\} \qquad \text{versus} \qquad H_1 : \alpha^{(1)*}_1 - \alpha^{(1)*}_{i_1} \neq 0 \quad \text{für ein } i_1 \in \{1,\ldots,k_1\}\,.$$
- Man kann zeigen, dass die Nullhypothese die Form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ hat, wobei $\mathbf{H}$ eine $(k_1 - 1) \times m$ Matrix mit vollem Zeilenrang $\mathrm{rg}(\mathbf{H}) = k_1 - 1$ ist und sämtliche Komponenten des Vektors $\mathbf{H}\boldsymbol{\beta}$ schätzbare Funktionen von $\boldsymbol{\beta}$ sind.
- Zur Verifizierung der Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ kann somit die in Theorem 3.16 betrachtete Testgröße
$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k_1 - 1)}{SSE/(k_1 k_2 (r-1))} \sim \mathrm{F}_{k_1 - 1,\, k_1 k_2 (r-1)}$$
verwendet werden mit
$$SSE = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1 i_2 j} - \overline{Y}_{i_1 i_2 \cdot}\bigr)^2, \quad SSE_H - SSE = r k_2 \sum_{i_1=1}^{k_1}\bigl(\overline{Y}_{i_1 \cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\bigr)^2, \tag{95}$$
wobei die Formeln in (95) genauso wie (90) bzw. (91) bewiesen werden.

Signifikanz des untergeordneten Einflussfaktors

- Um zu prüfen, ob die Stufen des untergeordneten Einflussfaktors signifikant sind, kann man ähnlich wie bei dem letzten Test in Abschnitt 3.4.2 (auf Signifikanz der Wechselwirkungen) vorgehen. Hierfür testen wir die Hypothese
$$H_0 : \alpha^{(2|1)*}_{1|1} - \alpha^{(2|1)*}_{i_2|i_1} = 0 \quad \forall\,(i_1, i_2) \in \{1, \ldots, k_1\} \times \{1, \ldots, k_2\}\,, \tag{96}$$
wobei
$$\alpha^{(2|1)*}_{i_2|i_1} = \alpha^{(2|1)}_{i_2|i_1} - \overline{\alpha}_{i_1} + \overline{\alpha}\,, \qquad \overline{\alpha}_{i_1} = \frac{1}{k_2}\sum_{i_2=1}^{k_2} \alpha^{(2|1)}_{i_2|i_1}\,, \qquad \overline{\alpha} = \frac{1}{k_1 k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} \alpha^{(2|1)}_{i_2|i_1}\,.$$
- Man kann zeigen, dass sich die in (92) betrachtete Hypothese in der Form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ schreiben lässt, wobei $\mathbf{H}$ eine $k_1(k_2-1) \times m$ Matrix mit vollem Zeilenrang $\mathrm{rg}(\mathbf{H}) = k_1(k_2-1)$ ist und sämtliche Komponenten des Vektors $\mathbf{H}\boldsymbol{\beta}$ schätzbare Funktionen von $\boldsymbol{\beta}$ sind.
- Zur Verifizierung der Hypothese $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ kann somit die in Theorem 3.16 betrachtete Testgröße
$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k_1(k_2-1))}{SSE/(k_1k_2(r-1))}$$
verwendet werden, wobei sich die in (74) bzw. (75) definierten Quadratsummen SSE und SSE_H wie folgt bestimmen lassen.
- Und zwar gilt so wie bisher
$$SSE = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\bigr)^2\,, \tag{97}$$
und die Menge $\{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \Theta_H\}$ in
$$SSE_H = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H) = \min_{\boldsymbol{\beta}\in\Theta_H}\bigl|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\bigr|^2$$
lässt sich wie folgt durch die Menge $\mathbb{R}\times\mathbb{R}^{k_1}_H \subset \mathbb{R}^{1+k_1}$ derjenigen Vektoren $\mathbf{x} = (x, x^{(1)}_1, \ldots, x^{(1)}_{k_1}) \in \mathbb{R}^{1+k_1}$ ersetzen, für die $\sum_{i_1=1}^{k_1} x^{(1)}_{i_1} = 0$, so dass
$$\begin{aligned}
SSE_H &= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}^{k_1}_H}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j} - (x + x^{(1)}_{i_1})\bigr)^2\\
&= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}^{k_1}_H}\Bigl\{\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\bigr)^2 + k_1k_2r\bigl(\overline{Y}_{\cdots} - x\bigr)^2\\
&\qquad + k_2r\sum_{i_1=1}^{k_1}\bigl(\overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdots} - x^{(1)}_{i_1}\bigr)^2 + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\bigr)^2\Bigr\}\\
&= \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\bigl(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\bigr)^2 + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\bigr)^2\,.
\end{aligned}$$
- Hieraus und aus (97) folgt, dass
$$SSE_H - SSE = r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\bigr)^2\,.$$
- Somit gilt
$$T_{\mathbf{H}} = \frac{k_1k_2(r-1)r\sum\limits_{i_1=1}^{k_1}\sum\limits_{i_2=1}^{k_2}\bigl(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\bigr)^2}{k_1(k_2-1)\sum\limits_{i_1=1}^{k_1}\sum\limits_{i_2=1}^{k_2}\sum\limits_{j=1}^{r}\bigl(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\bigr)^2} \sim \mathrm{F}_{k_1(k_2-1),\,k_1k_2(r-1)}\,.$$

4 Verallgemeinerte lineare Modelle

- In den Kapiteln 2 und 3 hatten wir über das lineare Modell $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ stets vorausgesetzt,
  - dass $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{o}$, d.h., $\bigl(\mathbb{E}\,Y_1, \ldots, \mathbb{E}\,Y_n\bigr)^\top = \mathbf{X}\boldsymbol{\beta}$,
  - wobei außerdem $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ gilt, wenn $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I})$.
- Wir verallgemeinern nun dieses Modell und lassen zu, dass die Erwartungswerte $\mathbb{E}\,Y_1, \ldots, \mathbb{E}\,Y_n$ der Stichprobenvariablen $Y_1, \ldots, Y_n$
  - über eine beliebige monotone Funktion $g : G \subset \mathbb{R} \to \mathbb{R}$, die so genannte *Linkfunktion*, durch die Komponenten des Vektors $\mathbf{X}\boldsymbol{\beta}$ ausgedrückt werden können, so dass
  $$\bigl(g(\mathbb{E}\,Y_1), \ldots, g(\mathbb{E}\,Y_n)\bigr)^\top = \mathbf{X}\boldsymbol{\beta}\,, \tag{1}$$
  - wobei der Definitionsbereich G von g noch genauer spezifiziert wird.
- Außerdem müssen jetzt die (unabhängigen) Stichprobevariablen $Y_1, \ldots, Y_n$ nicht notwendig normalverteilt sein, denn wir nehmen nur an, dass die Verteilungen von $Y_1, \ldots, Y_n$ zu einer *Exponentialfamilie* gehören.
- Wir setzen in diesem Kapitel jedoch (so wie in Kapitel 2) stets voraus, dass die Designmatrix $\mathbf{X}$ vollen Spaltenrang hat, d.h. $\mathrm{rg}(\mathbf{X}) = m$.

Genauso wie bei den linearen Modellen, die in den Kapiteln 2 und 3 betrachtet wurden, besteht eine Zielstellung darin, den Parametervektor $\boldsymbol{\beta}$ durch die Beobachtung der Zufallsstichprobe $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ zu schätzen, wobei angenommen wird, dass die Linkfunktion $g : G \subset \mathbb{R} \to \mathbb{R}$ bekannt ist.

4.1 Definition und grundlegende Eigenschaften

4.1.1 Exponentialfamilie

Wir nehmen an, dass die Stichprobenvariablen $Y_1, \ldots, Y_n$ unabhängig (jedoch i. a. *nicht* identisch verteilt) sind,

- wobei ihre Verteilungen zu einer einparametrischen *Exponentialfamilie* gehören, d.h., ihre Dichten bzw. Wahrscheinlichkeitsfunktionen besitzen die folgende Form: Für jedes $i \in \{1, \ldots, n\}$ gilt
  - im absolutstetigen Fall
  $$f(y; \theta_i) = \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta_i + a(y, \tau) - b(\theta_i)\bigr)\Bigr)\,, \qquad \forall\, y \in \mathbb{R}\,, \tag{2}$$
  - im diskreten Fall
  $$\mathbb{P}_{\theta_i}(Y_i = y) = \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta_i + a(y, \tau) - b(\theta_i)\bigr)\Bigr)\,, \qquad \forall\, y \in C\,, \tag{3}$$

  wobei $a : \mathbb{R} \times (0, \infty) \to \mathbb{R}$ und $b : \Theta \to \mathbb{R}$ gewisse Funktionen sind und $C \subset \mathbb{R}$ die kleinste abzählbare Teilmenge von $\mathbb{R}$ ist, für die $\mathbb{P}_{\theta_i}(Y_i \in C) = 1$ gilt.
- Dabei ist $\tau^2 > 0$ ein so genannter *Störparameter*, der nicht vom Index i abhängt, wobei oft angenommen wird, dass τ^2 bekannt ist.
- Dann ist
  $$\Theta = \Bigl\{\theta \in \mathbb{R} : \int_{-\infty}^{\infty} \exp\Bigl(\frac{y\theta + a(y, \tau)}{\tau^2}\Bigr)\,dy < \infty\Bigr\} \tag{4}$$
  bzw.
  $$\Theta = \Bigl\{\theta \in \mathbb{R} : \sum_{y \in C} \exp\Bigl(\frac{y\theta + a(y, \tau)}{\tau^2}\Bigr) < \infty\Bigr\} \tag{5}$$
  der *natürliche Parameterraum*, wobei wir stets annehmen, dass die Integrierbarkeitsbedingung in (4) bzw. (5) für mindestens zwei verschiedene $\theta_1, \theta_2 \in \mathbb{R}$ erfüllt ist.

Beachte Im absolutstetigen Fall kann der Störparameter τ^2 die Rolle eines zusätzlichen Varianzparameters spielen, während τ^2 im diskreten Fall meistens gleich 1 gesetzt wird.

Lemma 4.1 *Der in* (4) *bzw.* (5) *gegebene Parameterraum* $\Theta \subset \mathbb{R}$ *ist ein Intervall in* $\mathbb{R}$.

Beweis

- Wir betrachten hier nur den absolutstetigen Fall, denn im diskreten Fall verläuft der Beweis analog.
- Man kann sich leicht überlegen, dass für beliebige $x_1, x_2 \in \mathbb{R}$ und $\alpha \in (0,1)$

$$\left(e^{x_1}\right)^\alpha \left(e^{x_2}\right)^{1-\alpha} \le \max_{i=1,2,} \left(e^{x_i}\right)^\alpha \left(e^{x_i}\right)^{1-\alpha} = \max_{i=1,2,} e^{x_i} \le e^{x_1} + e^{x_2}.$$

- Hieraus ergibt sich mit der Schreibweise $\theta = \alpha\theta_1 + (1-\alpha)\theta_2$, dass für beliebige $\theta_1, \theta_2 \in \Theta$ und $\alpha \in (0,1)$

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\Bigl(\frac{y\theta + a(y,\tau)}{\tau^2}\Bigr)\, dy &= \int_{-\infty}^{\infty} \Bigl(\exp\Bigl(\frac{y\theta_1 + a(y,\tau)}{\tau^2}\Bigr)\Bigr)^\alpha \Bigl(\exp\Bigl(\frac{y\theta_2 + a(y,\tau)}{\tau^2}\Bigr)\Bigr)^{1-\alpha} dy \\ &\le \int_{-\infty}^{\infty} \Bigl(\exp\Bigl(\frac{y\theta_1 + a(y,\tau)}{\tau^2}\Bigr) + \exp\Bigl(\frac{y\theta_2 + a(y,\tau)}{\tau^2}\Bigr)\Bigr)\, dy < \infty\,.\end{aligned}$$

- Somit gilt auch $\theta \in \Theta$. □

Wegen Lemma 4.1 können (und werden) wir in diesem Kapitel stets annehmen, dass $\Theta \subset \mathbb{R}$ ein *offenes* Intervall ist, so dass die Integrierbarkeitsbedingung in (4) bzw. (5) für jedes $\theta \in \Theta$ erfüllt ist.

Lemma 4.2

- *Die Verteilung der Zufallsvariablen* $Y : \Omega \to \mathbb{R}$ *sei durch* (2) *bzw.* (3) *für ein beliebiges* $\theta \in \Theta$ *gegeben, wobei*

$$\mathbb{E}\,(Y^2) < \infty \qquad \forall\, \theta \in \Theta \tag{6}$$

  gelte und die Funktion $b : \Theta \to \mathbb{R}$ *zweimal stetig differenzierbar sei.*
- *Dann gilt*

$$\mathbb{E}\, Y = b^{(1)}(\theta) \qquad \textit{und} \qquad \operatorname{Var} Y = \tau^2\, b^{(2)}(\theta)\,. \tag{7}$$

Beweis

- Wir betrachten erneut lediglich den absolutstetigen Fall, denn im diskreten Fall verläuft der Beweis analog. Dabei gilt

$$\begin{aligned}\mathbb{E}\, Y &= \int_{-\infty}^{\infty} y \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta + a(y,\tau) - b(\theta)\bigr)\Bigr)\, dy = e^{-b(\theta)/\tau^2} \int_{-\infty}^{\infty} y \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta + a(y,\tau)\bigr)\Bigr)\, dy \\ &= e^{-b(\theta)/\tau^2}\, \tau^2 \int_{-\infty}^{\infty} \frac{d}{d\theta} \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta + a(y,\tau)\bigr)\Bigr)\, dy \\ &= e^{-b(\theta)/\tau^2}\, \tau^2\, \frac{d}{d\theta} \int_{-\infty}^{\infty} \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta + a(y,\tau)\bigr)\Bigr)\, dy \\ &= e^{-b(\theta)/\tau^2}\, \tau^2\, \frac{d}{d\theta}\, e^{b(\theta)/\tau^2} \underbrace{\int_{-\infty}^{\infty} \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta + a(y,\tau) - b(\theta)\bigr)\Bigr)\, dy}_{=1} \\ &= b^{(1)}(\theta)\,.\end{aligned}$$

- Auf ähnliche Weise ergibt sich, dass $\mathbb{E}\,(Y^2) = \tau^2\, b^{(2)}(\theta) + (\mathbb{E}\, Y)^2$. □

4.1.2 Verknüpfung der Parameter; natürliche Linkfunktion

- Von nun an setzen wir voraus, dass die Funktion $b : \Theta \to \mathbb{R}$ zweimal stetig differenzierbar ist mit $b^{(2)}(\theta) > 0$ für jedes $\theta \in \Theta$.
- Außerdem sei $G = \{b^{(1)}(\theta) : \theta \in \Theta\}$, und die Linkfunktion $g : G \to \mathbb{R}$ sei zweimal stetig differenzierbar, so dass $g^{(1)}(x) \neq 0$ für jedes $x \in G$. Die Umkehrfunktion von g bezeichnen wir mit $h = g^{-1}$.
- Wir betrachten das in (1) gegebene *verallgemeinerte lineare Modell* (GLM = generalized linear model), d.h., es gelte
$$\big(g(\mathbb{E}\,Y_1), \ldots, g(\mathbb{E}\,Y_n)\big)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{8}$$
  - Mit der Schreibweise $\mathbf{X} = (x_{ij})$ und $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^\top$ bzw. $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$, wobei $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, ergibt sich dann aus (8) für die Erwartungswerte $\mu_i = \mathbb{E}\,Y_i$ $(i = 1, \ldots, n)$, dass
$$\mu_i = h(\eta_i) = h\big(\mathbf{x}_i^\top \boldsymbol{\beta}\big) \qquad \text{bzw.} \qquad \boldsymbol{\mu} = \big(h(\eta_1), \ldots, h(\eta_n)\big)^\top, \tag{9}$$
    wobei $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$.
  - Wegen (7) und (8) sind die Parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ und $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ wie folgt miteinander verknüpft: Es gilt
$$\big(g\big(b^{(1)}(\theta_1)\big), \ldots, g\big(b^{(1)}(\theta_n)\big)\big)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{10}$$
- Hieraus und aus (9) ergibt sich, dass
$$\big(b^{(1)}(\theta_1), \ldots, b^{(1)}(\theta_n)\big) = \big(h(\mathbf{x}_1^\top \boldsymbol{\beta}), \ldots, h(\mathbf{x}_n^\top \boldsymbol{\beta})\big)\,.$$
  bzw.
$$(\theta_1, \ldots, \theta_n) = \big(\psi\big(h(\mathbf{x}_1^\top \boldsymbol{\beta})\big), \ldots, \psi\big(h(\mathbf{x}_n^\top \boldsymbol{\beta})\big)\big)\,, \tag{11}$$
  wobei $\psi = \big(b^{(1)}\big)^{-1}$ die Umkehrfunktion von $b^{(1)}$ ist.
- Außerdem lässt sich auch die Varianz $\sigma_i^2 = \operatorname{Var} Y_i$ der Stichprobenvariablen Y_i für jedes $i = 1, \ldots, n$ als Funktion $\sigma_i^2(\boldsymbol{\beta})$ von $\boldsymbol{\beta}$ ausdrücken, denn aus Lemma 4.2 und aus (11) ergibt sich, dass
$$\sigma_i^2(\boldsymbol{\beta}) = \tau^2\, b^{(2)}\big(\psi\big(h(\mathbf{x}_i^\top \boldsymbol{\beta})\big)\big) \qquad \forall\, i = 1, \ldots, n\,. \tag{12}$$

Beachte Die Linkfunktion $g : G \to \mathbb{R}$ heißt *natürlich*, wenn $g = \psi$. In diesem Fall gilt $\theta_i = \psi(\mu_i)$ bzw. $\theta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ für jedes $i = 1, \ldots, n$, d.h.,
$$(\theta_1, \ldots, \theta_n)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{13}$$

4.2 Beispiele

4.2.1 Lineares Modell mit normalverteilten Störgrößen

- Für das in Abschnitt 2.2 betrachtete lineare Modell
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{14}$$
  mit normalverteilten Störgrößen $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \sim \mathrm{N}(\mathbf{o}, \sigma^2 \mathbf{I})$ gilt
$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2) \quad \text{mit } \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}\,, \qquad \forall\, i = 1, \ldots, n, \tag{15}$$
  wobei wir voraussetzen, dass σ^2 bekannt ist.

- Die Verteilung von Y_i gehört dann zu der in Abschnitt 4.1.1 betrachteten einparametrischen Exponentialfamilie, denn die Dichte $f(y;\theta_i)$ von Y_i lässt sich in der folgenden Form darstellen, wobei $\theta_i = \mu_i$ für jedes $i = 1, \ldots, n$:
  - Es gilt
  $$f(y;\theta_i) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\Bigl(-\,\frac{1}{2\sigma^2}\,(y-\mu_i)^2\Bigr) \;=\; \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta_i + a(y,\tau) - b(\theta_i)\bigr)\Bigr) \qquad \forall\, y \in \mathbb{R}\,,$$
  - wobei
  $$\tau^2 = \sigma^2\,, \qquad a(y) = -\,\frac{y^2}{2} \qquad \text{und} \qquad b(\theta_i) = \frac{\theta_i^2}{2} + \log\sqrt{2\pi}\sigma^3\,. \tag{16}$$
- Wegen (15) gilt für die Linkfunktion $g : \mathbb{R} \to \mathbb{R}$ offenbar, dass $g(x) = x$ für jedes $x \in \mathbb{R}$.
  - Außerdem ergibt sich aus (16), dass $x = b^{(1)}(x)$ für jedes $x \in \mathbb{R}$.
  - Somit gilt $g(x) = x = \psi(x)$ für jedes $x \in \mathbb{R}$, d.h., durch $g(x) = x$ ist die natürliche Linkfunktion gegeben.

4.2.2 Binäre kategoriale Regression

- In diesem Abschnitt betrachten wir den Fall, dass die Stichprobenvariablen $Y_1, \ldots, Y_n$ Bernoulli–verteilt sind, d.h., sie können nur die Werte 0 bzw. 1 mit positiver Wahrscheinlichkeit annehmen.
  - Dabei verwenden wir die Schreibweise
  $$\pi_i = \mathbb{P}(Y_i = 1) \qquad \Bigl(= \mu_i \;=\; \mathbb{E}\,Y_i\Bigr) \qquad \forall\, i = 1, \ldots, n\,,$$
  wobei vorausgesetzt wird, dass $0 < \pi_i < 1$ für jedes $i = 1, \ldots, n$.
  - In diesem Fall werden also die Wahrscheinlichkeiten $\pi_1, \ldots, \pi_n$ über eine Linkfunktion $g : (0,1) \to \mathbb{R}$ mit dem Parametervektor $\boldsymbol{\beta}$ verknüpft, d.h.,
  $$\bigl(g(\pi_1), \ldots, g(\pi_n)\bigr)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{17}$$
- Für jedes $i = 1, \ldots, n$ gehört die $\mathrm{Bin}(1,\pi_i)$–Verteilung zu der Exponentialfamilie, die in Abschnitt 4.1.1 eingeführt worden ist, wobei $\theta_i = \log(\pi_i/(1-\pi_i))$.
  - Denn für $y = 0, 1$ gilt
  $$\mathbb{P}_{\theta_i}(Y_i = y) = \pi_i^y\,(1-\pi_i)^{1-y} = \exp\Bigl(y\log\Bigl(\frac{\pi_i}{1-\pi_i}\Bigr) + \log(1-\pi_i)\Bigr) = \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta_i + a(y,\tau) - b(\theta_i)\bigr)\Bigr)\,,$$
  - wobei
  $$\tau^2 = 1\,, \qquad a(y) = 0 \qquad \text{und} \qquad b(\theta_i) = \log(1 + e^{\theta_i})\,. \tag{18}$$

Beachte

- Aus (18) ergibt sich, dass $(b^{(1)})^{-1}(x) = \log(x/(1-x))$ für jedes $x \in (0,1)$, d.h., die natürliche Linkfunktion $g : (0,1) \to \mathbb{R}$ ist gegeben durch
  $$g(x) = \log\Bigl(\frac{x}{1-x}\Bigr) \qquad \forall\, x \in (0,1)\,. \tag{19}$$
  - Das in (17) betrachtete GLM mit der in (19) gegebenen natürlichen Linkfunktion wird dann (binäres) *logistisches Regressionsmodell* genannt.

  - In diesem Fall ist die Abhängigkeit der Wahrscheinlichkeiten $\pi_i = \pi_i(\boldsymbol{\beta})$ von den Linearkombinationen $\mathbf{x}_i^\top \boldsymbol{\beta}$ gegeben durch

$$\pi_i = \frac{1}{1+\exp(-\mathbf{x}_i^\top \boldsymbol{\beta})} \qquad \forall\, i = 1,\ldots,n\,. \tag{20}$$

- Eine andere (nicht natürliche) Linkfunktion $g : (0,1) \to \mathbb{R}$, die in diesem Zusammenhang betrachtet wird, ist gegeben durch

$$g = \Phi^{-1}\,, \tag{21}$$

  - wobei $\Phi : \mathbb{R} \to (0,1)$ die Verteilungsfunktion der N(0,1)–Verteilung ist.
  - Dann gilt $\pi_i = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$ für jedes $i = 1,\ldots,n$, und man spricht vom Modell der *Probitanalyse*.

4.2.3 Poisson–verteilte Stichprobenvariablen mit natürlicher Linkfunktion

- Die Stichprobenvariablen $Y_1,\ldots,Y_n$ seien nun Poisson–verteilt, d.h., es gelte $Y_i \sim \mathrm{Poi}(\lambda_i)$ mit $0 < \lambda_i < \infty$ für jedes $i = 1,\ldots,n$.
- Die $\mathrm{Poi}(\lambda_i)$–Verteilung gehört ebenfalls zu der Exponentialfamilie, die in Abschnitt 4.1.1 eingeführt worden ist, wobei $\theta_i = \log \lambda_i$.
  - Denn für jedes $y = 0,1,\ldots$ gilt

$$\mathbb{P}_{\theta_i}(Y_i = y) = \frac{\lambda_i^y e^{-\lambda_i}}{y!} = \exp\bigl(y\log\lambda_i - \log(y!) - \lambda_i\bigr) = \exp\Bigl(\frac{1}{\tau^2}\,\bigl(y\theta_i + a(y,\tau) - b(\theta_i)\bigr)\Bigr)\,,$$

  - wobei

$$\tau^2 = 1\,, \qquad a(y) = -\log(y!) \qquad \text{und} \qquad b(\theta_i) = e^{\theta_i}\,.$$

- Die natürliche Linkfunktion $g : (0,\infty) \to \mathbb{R}$ ist gegeben durch

$$g(x) = \log x \qquad \forall\, x > 0\,. \tag{22}$$

4.3 Maximum–Likelihood–Schätzer für $\boldsymbol{\beta}$

- Weil wir annehmen, dass die Verteilungen der Stichprobenvariablen $Y_1,\ldots,Y_n$ zu einer Exponentialfamilie gehören, kann der Parametervektor $\boldsymbol{\beta}$ mit der Maximum–Likelihood–Methode geschätzt werden.
- Um dies zu zeigen, diskutieren wir zunächst einige Eigenschaften der Loglikelihood–Funktion $\log L(\mathbf{Y},\boldsymbol{\theta})$ der Zufallsstichprobe $\mathbf{Y} = (Y_1,\ldots,Y_n)^\top$ bzw. ihrer partiellen Ableitungen nach den Komponenten $\beta_1,\ldots,\beta_m$ von $\boldsymbol{\beta}$.

4.3.1 Loglikelihood–Funktion und ihre partiellen Ableitungen

- Aus (2) – (3) bzw. aus (11) ergibt sich, dass die Loglikelihood–Funktion $\log L(\mathbf{Y},\boldsymbol{\theta})$ der Zufallsstichprobe $\mathbf{Y} = (Y_1,\ldots,Y_n)^\top$ als eine Funktion $\log L(\mathbf{Y},\boldsymbol{\beta})$ von $\boldsymbol{\beta}$ geschrieben werden kann.
  - Und zwar folgt aus (2) – (3), dass

$$\log L(\mathbf{Y},\boldsymbol{\theta}) = \sum_{i=1}^n \frac{1}{\tau^2}\,\bigl(Y_i\theta_i + a(Y_i,\tau) - b(\theta_i)\bigr)\,. \tag{23}$$

– Hieraus und aus (11) ergibt sich nun, dass

$$\log L(\mathbf{Y},\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\tau^2} \Bigl(Y_i\psi\bigl(h(\mathbf{x}_i^\top\boldsymbol{\beta})\bigr) + a(Y_i,\tau) - b\bigl(\psi\bigl(h(\mathbf{x}_i^\top\boldsymbol{\beta})\bigr)\bigr)\Bigr)\,. \tag{24}$$

- Für verallgemeinerte lineare Modelle mit natürlicher Linkfunktion ergibt sich aus (13) und (23), dass

$$\log L(\mathbf{Y},\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\tau^2} \bigl(Y_i\mathbf{x}_i^\top\boldsymbol{\beta} + a(Y_i,\tau) - b(\mathbf{x}_i^\top\boldsymbol{\beta})\bigr)\,. \tag{25}$$

Zur Bestimmung von Maximum–Likelihood–Schätzern ist die Kenntnis der so genannten Scorefunktionen, d.h. der partiellen Ableitungen der Loglikelihood–Funktion, sowie der Fisher–Informationsmatrix nützlich, die wie folgt definiert ist.

Definition Für beliebige $i,j=1,\ldots,m$ sei

$$U_i(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_i}\log L(\mathbf{Y},\boldsymbol{\beta}) \qquad \text{und} \qquad I_{ij}(\boldsymbol{\beta}) = \mathbb{E}\bigl(U_i(\boldsymbol{\beta})U_j(\boldsymbol{\beta})\bigr)\,.$$

Dann wird der m–dimensionale Zufallsvektor $\mathbf{U}(\boldsymbol{\beta}) = \bigl(U_1(\boldsymbol{\beta}),\ldots,U_m(\boldsymbol{\beta})\bigr)^\top$ bzw. die (deterministische) $m\times m$–Matrix $\mathbf{I}(\boldsymbol{\beta}) = \bigl(I_{ij}(\boldsymbol{\beta})\bigr)$ der *Scorevektor* bzw. die *Fisher–Informationsmatrix* genannt.

Mit der Schreibweise

$$\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta}) \;=\; \frac{dh(s)}{ds}\Big|_{s=\eta_i} \;=\; \Bigl(\frac{dg(t)}{dt}\Bigr)^{-1}\Big|_{t=h(\eta_i)} \tag{26}$$

ergibt sich das folgende Resultat.

Theorem 4.1 *Für beliebige $j,k=1,\ldots,m$ gilt*

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij}\bigl(Y_i-\mu_i(\boldsymbol{\beta})\bigr)\,\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\,\frac{1}{\sigma_i^2(\boldsymbol{\beta})} \tag{27}$$

und

$$I_{jk}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij}x_{ik}\Bigl(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Bigr)^2\,\frac{1}{\sigma_i^2(\boldsymbol{\beta})} \tag{28}$$

bzw. in Matrix–Schreibweise

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{V}^{-1}(\boldsymbol{\beta})\,\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\,(\mathbf{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})) \qquad \textit{und} \qquad \mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{V}^{-1}(\boldsymbol{\beta})\,\Bigl(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\Bigr)^2\,\mathbf{X}\,, \tag{29}$$

wobei

$$\mathbf{V}(\boldsymbol{\beta}) \;=\; \operatorname{diag}\bigl(\sigma_i^2(\boldsymbol{\beta})\bigr) \qquad \textit{und} \qquad \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta}) \;=\; \operatorname{diag}\Bigl(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Bigr)\,.$$

Beweis

- Die in (23) bzw. (24) gegebene Loglikelihood–Funktion lässt sich in der Form

$$\log L(\mathbf{Y},\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{1}{\tau^2}\,\ell^{(i)}(\theta_i)$$

schreiben, wobei $\ell^{(i)}(\theta_i) = Y_i\theta_i + a(Y_i,\tau) - b(\theta_i)$ und $\theta_i = \psi\bigl(h(\mathbf{x}_i^\top\boldsymbol{\beta})\bigr)$.

- Somit gilt für jedes $j = 1, \ldots, m$, dass
$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{1}{\tau^2}\, \frac{\partial \ell^{(i)}}{\partial \beta_j}(\theta_i)\,, \tag{30}$$
wobei sich durch die mehrfache Anwendung der Kettenregel ergibt, dass
$$\frac{\partial \ell^{(i)}}{\partial \beta_j} = \frac{\partial \ell^{(i)}}{\partial \theta_i}\, \frac{\partial \theta_i}{\partial \mu_i}\, \frac{\partial \mu_i}{\partial \eta_i}\, \frac{\partial \eta_i}{\partial \beta_j}\,. \tag{31}$$
  - Andererseits gilt offenbar, dass $\partial \eta_i / \partial \beta_j = x_{ij}$, und aus Lemma 4.2 ergibt sich, dass
$$\frac{\partial \ell^{(i)}}{\partial \theta_i} = Y_i - b^{(1)}(\theta_i) \overset{\text{Lemma 4.2}}{=} Y_i - \mu_i$$
bzw.
$$\Bigl(\frac{\partial \theta_i}{\partial \mu_i}\Bigr)^{-1} = \frac{\partial \mu_i}{\partial \theta_i} \overset{\text{Lemma 4.2}}{=} b^{(2)}(\theta_i) \overset{\text{Lemma 4.2}}{=} \frac{1}{\tau^2}\, \sigma_i^2\,.$$
  - Hieraus und aus (30) – (31) ergibt sich die Gültigkeit von (27).
- Um (28) zu zeigen, genügt es zu beachten, dass für beliebige $i, j = 1, \ldots, n$
$$\mathbb{E}\bigl((Y_i - \mu_i)(Y_j - \mu_j)\bigr) = \begin{cases} \sigma_i^2 & \text{für } i = j, \\ 0 & \text{für } i \neq j. \end{cases}$$
wegen der Unabhängigkeit der Stichprobenvariablen $Y_1, \ldots, Y_n$.
  - Hieraus und aus (27) ergibt sich, dass
$$\begin{aligned} I_{jk}(\boldsymbol{\beta}) &= \mathbb{E}\bigl(U_j(\boldsymbol{\beta}) U_k(\boldsymbol{\beta})\bigr) = \sum_{i=1}^n x_{ij} x_{ik} \Bigl(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Bigr)^2 \frac{1}{\sigma_i^4(\boldsymbol{\beta})}\, \mathbb{E}\bigl((Y_i - \mu_i)^2\bigr) \\ &= \sum_{i=1}^n x_{ij} x_{ik} \Bigl(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Bigr)^2 \frac{1}{\sigma_i^2(\boldsymbol{\beta})}\,. \end{aligned}$$
  - Damit ist (28) bewiesen. □

Korollar 4.1 *Sei* $\bigl(g(\mathbb{E}\, Y_1), \ldots, g(\mathbb{E}\, Y_n)\bigr)^\top = \mathbf{X}\boldsymbol{\beta}$ *ein GLM mit natürlicher Linkfunktion* $g : G \to \mathbb{R}$*. Dann gilt für beliebige* $j, k = 1, \ldots, m$
$$U_j(\boldsymbol{\beta}) = \frac{1}{\tau^2} \sum_{i=1}^n x_{ij} \bigl(Y_i - \mu_i(\boldsymbol{\beta})\bigr) \qquad \textit{bzw.} \qquad \mathbf{U}(\boldsymbol{\beta}) = \frac{1}{\tau^2}\, \mathbf{X}^\top (\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \tag{32}$$
und
$$I_{jk}(\boldsymbol{\beta}) = \frac{1}{\tau^4} \sum_{i=1}^n x_{ij} x_{ik} \sigma_i^2(\boldsymbol{\beta}) \qquad \textit{bzw.} \qquad \mathbf{I}(\boldsymbol{\beta}) = \frac{1}{\tau^4}\, \mathbf{X}^\top \mathbf{V}(\boldsymbol{\beta}) \mathbf{X}\,. \tag{33}$$

Beweis Weil $g : G \to \mathbb{R}$ eine natürliche Linkfunktion ist, gilt $\theta_i = \eta_i$ für jedes $i = 1, \ldots, n$. Hieraus und aus Lemma 4.2 ergibt sich, dass
$$\frac{d\mu_i}{d\eta_i} = b^{(2)}(\theta_i) = \frac{1}{\tau^2}\, \sigma_i^2\,.$$
Die Behauptung ergibt sich somit aus Theorem 4.1. □

4.3.2 Hesse–Matrix

Neben dem (Score-) Vektor $\mathbf{U}(\boldsymbol{\beta})$ der ersten partiellen Ableitungen der Loglikelihood–Funktion $\log L(\mathbf{Y},\boldsymbol{\beta})$ wird auch ihre *Hesse–Matrix*, d.h., die $m\times m$–Matrix

$$\mathbf{W}(\boldsymbol{\beta}) \;=\; \bigl(W_{ij}(\boldsymbol{\beta})\bigr) \;=\; \left(\frac{\partial^2}{\partial\beta_i\partial\beta_j}\,\log L(\mathbf{Y},\boldsymbol{\beta})\right)$$

der zweiten partiellen Ableitungen benötigt.

Theorem 4.2

- *Für jedes GLM gilt*
$$\mathbf{W}(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{R}(\boldsymbol{\beta})\mathrm{diag}\bigl(Y_i-\mu_i(\boldsymbol{\beta})\bigr)\mathbf{X}-\mathbf{I}(\boldsymbol{\beta})\,, \tag{34}$$
wobei $\mathbf{I}(\boldsymbol{\beta})$ *die in* (29) *gegebene Fisher–Informationsmatrix und* $\mathbf{R}(\boldsymbol{\beta})=\mathrm{diag}\bigl(v_i(\boldsymbol{\beta})\bigr)$ *eine* $(n\times n)$*–Diagonalmatrix ist mit*
$$v_i(\boldsymbol{\beta}) \;=\; \frac{1}{\tau^2}\,\frac{d^2u(s)}{ds^2}\Big|_{s=\mathbf{x}_i^\top\boldsymbol{\beta}} \qquad \textit{und} \qquad u=\psi\circ h\,.$$
- *Für GLM mit natürlicher Linkfunktion gilt insbesondere*
$$\mathbf{W}(\boldsymbol{\beta}) = -\mathbf{I}(\boldsymbol{\beta})\,. \tag{35}$$

Beweis

- Aus Formel (27) in Theorem 4.1 ergibt sich, dass für beliebige $j,k=1,\ldots,m$
$$\begin{aligned} W_{jk}(\boldsymbol{\beta}) \;&=\; \frac{\partial}{\partial\beta_k}\,U_j(\boldsymbol{\beta}) \overset{(27)}{=} \frac{\partial}{\partial\beta_k}\,\sum_{i=1}^n x_{ij}\bigl(Y_i-\mu_i(\boldsymbol{\beta})\bigr)\,\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\,\frac{1}{\sigma_i^2(\boldsymbol{\beta})}\\ &=\; \sum_{i=1}^n x_{ij}\left(\bigl(Y_i-\mu_i(\boldsymbol{\beta})\bigr)\,\frac{\partial}{\partial\beta_k}\Bigl(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\,\frac{1}{\sigma_i^2(\boldsymbol{\beta})}\Bigr) - \frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\,\frac{1}{\sigma_i^2(\boldsymbol{\beta})}\,\frac{\partial\mu_i}{\partial\beta_k}\,(\boldsymbol{\beta})\right). \end{aligned}$$
  - Dabei ergibt sich mit der Schreibweise $\eta_i=\mathbf{x}_i^\top\boldsymbol{\beta}$ aus Lemma 4.2, dass
$$\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\,\frac{1}{\sigma_i^2(\boldsymbol{\beta})} \;=\; b^{(2)}\bigl(\psi\circ h(\eta_i)\bigr)\,(\psi\circ h)^{(1)}(\eta_i)\,\frac{1}{\tau^2\,b^{(2)}\bigl(\psi\circ h(\eta_i)\bigr)} \;=\; \frac{1}{\tau^2}\,(\psi\circ h)^{(1)}(\eta_i)$$
  und somit
$$\frac{\partial}{\partial\beta_k}\Bigl(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\,\frac{1}{\sigma_i^2(\boldsymbol{\beta})}\Bigr) \;=\; \frac{1}{\tau^2}\,(\psi\circ h)^{(2)}(\eta_i)x_{ik}\,.$$
  - Außerdem gilt
$$\frac{\partial\mu_i}{\partial\beta_k} \;=\; \frac{d\mu_i}{d\eta_i}\,\frac{\partial\eta_i}{\partial\beta_k}\,.$$
- Insgesamt ergibt sich also, dass
$$W_{jk}(\boldsymbol{\beta}) \;=\; \sum_{i=1}^n x_{ij}x_{ik}(Y_i-\mu_i)v_i - \sum_{i=1}^n x_{ij}x_{ik}\Bigl(\frac{d\mu_i}{d\eta_j}\Bigr)^2\,\frac{1}{\sigma_i^2}\,.$$
- Hieraus und aus der Darstellungsformel (29) für die Fisher–Informationsmatrix $\mathbf{I}(\boldsymbol{\beta})$ ergibt sich (34).
- Weil für GLM mit natürlicher Linkfunktion die Superposition $u=\psi\circ h$ die Identitätsabbildung ist, gilt in diesem Fall $\mathbf{R}(\boldsymbol{\beta})=\mathbf{0}$. Somit ergibt sich (35) aus (34). $\square$

Beachte Für die Beispiele von GLM, die in Abschnitt 4.2 betrachtet worden sind, ergeben sich aus den Theoremen 4.1 und 4.2 bzw. aus Korollar 4.1 die folgenden Formeln für $\mathbf{U}(\boldsymbol{\beta})$ und $\mathbf{W}(\boldsymbol{\beta})$.

1. Für das lineare Modell $\mathbb{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ mit normalverteilten Stichprobenvariablen (und mit der Linkfunktion $g(x) = x$) ist $(d\boldsymbol{\mu}/d\boldsymbol{\eta})(\boldsymbol{\beta})$ die Einheitsmatrix. Somit gilt
$$\mathbf{U}(\boldsymbol{\beta}) \;=\; \frac{1}{\sigma^2}\,\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\,, \qquad \mathbf{W}(\boldsymbol{\beta}) \;=\; -\,\frac{1}{\sigma^2}\,\mathbf{X}^\top\mathbf{X}\,, \tag{36}$$
vgl. auch Abschnitt 2.2.
2. Für das logistische Regressionsmodell (mit der natürlichen Linkfunktion) gilt
$$\mathbf{U}(\boldsymbol{\beta}) \;= \mathbf{X}^\top(\mathbf{Y}-\boldsymbol{\pi})\,, \qquad \mathbf{W}(\boldsymbol{\beta}) \;=\; -\mathbf{X}^\top\mathrm{diag}\big(\pi_i(1-\pi_i)\big)\mathbf{X}\,, \tag{37}$$
wobei $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_n)^\top$ und die Wahrscheinlichkeiten π_i so wie in (20) durch $\boldsymbol{\beta}$ ausgedrückt werden können.
3. Für Poisson–verteilte Stichprobenvariablen mit natürlicher Linkfunktion gilt
$$\mathbf{U}(\boldsymbol{\beta}) \;= \mathbf{X}^\top(\mathbf{Y}-\boldsymbol{\lambda})\,, \qquad \mathbf{W}(\boldsymbol{\beta}) \;=\; -\mathbf{X}^\top\mathrm{diag}\big(\lambda_i\big)\mathbf{X}\,, \tag{38}$$
wobei $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_n)^\top$ und $\lambda_i = e^{\mathbf{x}_i^\top\boldsymbol{\beta}}$.

4.3.3 Maximum–Likelihood–Gleichung und numerische Lösungsansätze

- Zur Bestimmung eines Maximum–Likelihood–Schätzers für $\boldsymbol{\beta}$ wird die *Maximum–Likelihood–Gleichung*
$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{o} \tag{39}$$
betrachtet, die im allgemeinen nichtlinear ist und deshalb oft nur mit iterativen Methoden gelöst werden kann.
- Wegen Theorem 4.1 ist die Gleichung (39) äquivalent mit
$$\mathbf{X}^\top\mathbf{V}^{-1}(\boldsymbol{\beta})\,\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\,(\mathbf{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{o}\,. \tag{40}$$

Beachte

- Aus Korollar 4.1 ergibt sich, dass sich (40) im Fall einer natürlichen Linkfunktion vereinfacht zu:
$$\mathbf{X}^\top\;(\mathbf{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{o}\,. \tag{41}$$
- Weil wir außerdem voraussetzen, dass $0 < \sigma_i^2(\boldsymbol{\beta}) < \infty$ für jedes $i = 1,\ldots,n$ und dass die Designmatrix $\mathbf{X}$ vollen Spaltenrang hat, ist die Matrix $\mathbf{W}(\boldsymbol{\beta}) \;=\; -\,\tau^{-4}\mathbf{X}^\top\mathbf{V}(\boldsymbol{\beta})\mathbf{X}$ der zweiten partiellen Ableitungen negativ definit.
- Somit gilt: Wenn (41) eine Lösung hat, dann ist diese Lösung ein (eindeutig bestimmter) Maximum–Likelihood–Schätzer $\widehat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$.

Wir diskutieren nun die Grundideen von zwei numerischen Iterationsmethoden zur Lösung der Maximum–Likelihood–Gleichung (39). Dabei betrachten wir eine Folge von Zufallsvektoren $\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, \ldots : \Omega \to \mathbb{R}^m$, die unter gewissen Bedingungen gegen einen Zufallsvektor $\widehat{\boldsymbol{\beta}}$ konvergieren, so dass $\widehat{\boldsymbol{\beta}}$ Lösung von (39) ist.

1. *Newton–Verfahren*

   - Sei $\widehat{\boldsymbol{\beta}}_0 : \Omega \to \mathbb{R}^m$ ein geeignet gewählter Startvektor, und die Iterationen $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_k$ seien bereits berechnet worden.
   - Zur Berechnung der $(k+1)$–ten Iteration $\widehat{\boldsymbol{\beta}}_{k+1}$ aus $\widehat{\boldsymbol{\beta}}_k$ wird die linke Seite $\mathbf{U}(\boldsymbol{\beta})$ der Maximum–Likelihood–Gleichung (39) ersetzt durch
     - die ersten beiden Glieder $\mathbf{U}(\widehat{\boldsymbol{\beta}}_k) + \mathbf{W}(\widehat{\boldsymbol{\beta}}_k)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_k)$ der Taylor–Reihenentwicklung von $\mathbf{U}(\boldsymbol{\beta})$ an der Stelle $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_k$.
     - Die $(k+1)$–te Iteration $\widehat{\boldsymbol{\beta}}_{k+1}$ ist also Lösung der Gleichung

$$\mathbf{U}(\widehat{\boldsymbol{\beta}}_k) + \mathbf{W}(\widehat{\boldsymbol{\beta}}_k)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_k) = \mathbf{o}\,. \tag{42}$$

   - Wenn die Matrix $\mathbf{W}(\widehat{\boldsymbol{\beta}}_k)$ invertierbar ist, dann ergibt sich aus (42), dass

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k - \mathbf{W}^{-1}(\widehat{\boldsymbol{\beta}}_k)\mathbf{U}(\widehat{\boldsymbol{\beta}}_k)\,, \tag{43}$$

   - Damit die so konstruierte Folge $\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, \ldots$ gegen $\widehat{\boldsymbol{\beta}}$ konvergiert, muss $\widehat{\boldsymbol{\beta}}$ Lösung von (39) sein und der Startvektor $\widehat{\boldsymbol{\beta}}_0$ muss genügend nahe bei $\widehat{\boldsymbol{\beta}}$ liegen.

2. *Fisher–Scoring*

   - Wir betrachten nun eine Variante des Newton–Verfahrens, die so genannte *Scoring–Methode von Fisher*, bei der die Hesse–Matrix $\mathbf{W}(\boldsymbol{\beta})$ in (42) durch die Erwartungswertmatrix $\mathbb{E}\,\mathbf{W}(\boldsymbol{\beta})$ ersetzt wird.
     - Dies hat den Vorteil, dass die $(m \times m)$–Matrix $\mathbb{E}\,\mathbf{W}(\boldsymbol{\beta})$ invertierbar ist.
     - Aus den Theoremen 4.1 und 4.2 ergibt sich nämlich, dass

$$\begin{aligned}
\mathbb{E}\,\mathbf{W}(\boldsymbol{\beta}) &\overset{(34)}{=} \mathbb{E}\Bigl(\mathbf{X}^\top \mathbf{R}(\boldsymbol{\beta})\mathrm{diag}\bigl(Y_i - \mu_i(\boldsymbol{\beta})\bigr)\mathbf{X} - \mathbf{I}(\boldsymbol{\beta})\Bigr)\\
&= -\mathbf{I}(\boldsymbol{\beta})\\
&\overset{(29)}{=} -\mathbf{X}^\top \mathbf{V}^{-1}(\boldsymbol{\beta})\,\Bigl(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\Bigr)^2 \mathbf{X}\,,
\end{aligned}$$

       wobei sich die zweite Gleichheit aus der Identität $\mathbb{E}\,Y_i = \mu_i(\boldsymbol{\beta})$ ergibt.
     - Dabei ist der letzte Ausdruck eine invertierbare $(m \times m)$–Matrix, weil wir voraussetzen, dass die Designmatrix $\mathbf{X}$ vollen Spaltenrang hat und dass $(d\mu_i/d\eta_i)(\boldsymbol{\beta}) \neq 0$ für jedes $i = 1, \ldots, n$.
   - Anstelle von (43) wird somit die folgende Iterationsgleichung betrachtet:

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k + \bigl(\mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\mathbf{X}\bigr)^{-1}\Biggl(\mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\Bigl(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\Bigr)(\widehat{\boldsymbol{\beta}}_k)\bigl(\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_k)\bigr)\Biggr)\,, \tag{44}$$

   wobei

$$\mathbf{Z}(\boldsymbol{\beta}) = \mathbf{V}^{-1}(\boldsymbol{\beta})\,\Bigl(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\Bigr)^2 \qquad \text{und} \qquad \Bigl(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\Bigr)(\boldsymbol{\beta}) = \Bigl(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}\Bigr)^{-1}(\boldsymbol{\beta})\,.$$

   - Bei natürlicher Linkfunktion ergibt sich aus Lemma 4.2, dass

$$\frac{d\mu_i}{d\eta_i} = b^{(2)}(\theta_i) = \frac{1}{\tau^2}\,\sigma_i^2 \qquad \text{bzw.} \qquad \mathbf{Z}(\boldsymbol{\beta}) = \frac{1}{\tau^4}\,\mathbf{V}(\boldsymbol{\beta})\,.$$

   - In diesem Fall hat dann die Iterationsgleichung (44) die Form:

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k + \tau^2\bigl(\mathbf{X}^\top \mathbf{V}(\widehat{\boldsymbol{\beta}}_k)\mathbf{X}\bigr)^{-1}\Bigl(\mathbf{X}^\top\bigl(\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_k)\bigr)\Bigr)\,.$$

Beachte

- Wenn in (44) die Zufallsstichprobe $\mathbf{Y}$ durch die so genannte *Pseudo-Zufallsstichprobe*

$$\mathbf{Y}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta} + \Bigl(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\Bigr)(\boldsymbol{\beta})\bigl(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\bigr)$$

  ersetzt wird, dann lässt sich die Iterationsgleichung (44) in der folgenden Form schreiben:

$$\bigl(\mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\mathbf{X}\bigr)\widehat{\boldsymbol{\beta}}_{k+1} = \mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\mathbf{Y}(\widehat{\boldsymbol{\beta}}_k)\,.$$

- Diese Gleichung kann als *gewichtete Normalengleichung* für $\widehat{\boldsymbol{\beta}}_{k+1}$ bezüglich der Pseudo-Zufallsstichprobe $\mathbf{Y}(\widehat{\boldsymbol{\beta}}_k)$ aufgefasst werden, wobei die Gewichte, d.h., die Eintragungen der Diagonalmatrix $\mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)$ ebenfalls von der k-ten Iteration $\widehat{\boldsymbol{\beta}}_k$ abhängen.

4.3.4 Asymptotische Normalverteiltheit von ML–Schätzern; asymptotische Tests

- Der Begriff der Verteilungskonvergenz von Zufallsvektoren wird wie folgt definiert.
  - Sei $m \in \mathbb{N}$ eine beliebige natürliche Zahl, und seien $\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \ldots : \Omega \to \mathbb{R}^m$ beliebige Zufallsvektoren. Man sagt, dass $\{\mathbf{Z}_n\}$ *in Verteilung* gegen $\mathbf{Z}$ konvergiert, wenn

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{Z}_n \le \mathbf{x}) = \mathbb{P}(\mathbf{Z} \le \mathbf{x}) \tag{45}$$

    für jedes $\mathbf{x} \in \mathbb{R}^m$ mit $\mathbb{P}(\mathbf{Z} = \mathbf{x}) = 0$. Schreibweise: $\mathbf{Z}_n \stackrel{\mathrm{d}}{\longrightarrow} \mathbf{Z}$.
- Wir diskutieren nun asymptotische (Verteilungs–) Eigenschaften von Maximum–Likelihood–Schätzern $\widehat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ bzw. asmyptotische Tests, wenn der Stichprobenumfang n gegen unendlich geht.
  - Dabei betrachten wir lediglich den Fall der natürlichen Linkfunktion $g : G \to \mathbb{R}$
  - und indizieren die Zufallsstichprobe $\mathbf{Y}$, die Loglikelihood–Funktion $\log L(\mathbf{Y}, \boldsymbol{\beta})$, den Scorevektor $\mathbf{U}(\boldsymbol{\beta})$, die Fisher–Informationsmatrix $\mathbf{I}(\boldsymbol{\beta})$ bzw. den ML–Schätzer $\widehat{\boldsymbol{\beta}}$ jeweils mit n.

1. *Asymptotische Verteilungseigenschaften*

   Unter gewissen Bedingungen (vgl. Abschnitt VII.2.6 in Pruscha (2000)) kann man zeigen: Für jedes $\boldsymbol{\beta} \in \mathbb{R}^m$ mit $\mathbf{x}_i^\top \boldsymbol{\beta} \in \Theta$ für $i = 1, 2, \ldots$ gibt es

   - einen konsistenten ML–Schätzer $\widehat{\boldsymbol{\beta}}_n$ für $\boldsymbol{\beta}$, d.h., für jedes $\varepsilon > 0$ gilt

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}}\bigl(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| \le \varepsilon, \mathbf{U}_n(\widehat{\boldsymbol{\beta}}_n) = \mathbf{o}\bigr) \;=\; 1\,, \tag{46}$$

   - eine Folge $\{\boldsymbol{\Gamma}_n\}$ von invertierbaren $(m \times m)$–Matrizen, die von $\boldsymbol{\beta}$ abhängen können und für die $\lim_{n\to\infty} \boldsymbol{\Gamma}_n = \mathbf{0}$ gilt,
   - sowie eine symmetrische und positiv definite $(m \times m)$–Matrix $\mathbf{K}(\boldsymbol{\beta})$, so dass

$$\lim_{n\to\infty} \boldsymbol{\Gamma}_n^\top \mathbf{I}_n(\boldsymbol{\beta})\boldsymbol{\Gamma}_n = \mathbf{K}^{-1}(\boldsymbol{\beta}) \tag{47}$$

   und

$$\boldsymbol{\Gamma}_n^{-1}\bigl(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\bigr) \stackrel{\mathrm{d}}{\longrightarrow} \mathrm{N}\bigl(\mathbf{o}, \mathbf{K}(\boldsymbol{\beta})\bigr) \qquad \text{bzw.} \qquad 2\bigl(\log L_n(\mathbf{Y}_n, \widehat{\boldsymbol{\beta}}_n) - \log L_n(\mathbf{Y}_n, \boldsymbol{\beta})\bigr) \stackrel{\mathrm{d}}{\longrightarrow} \chi^2_m\,. \tag{48}$$

2. *Asymptotische Tests*

   - Bei großem n kann zur Konstruktion eines asymptotischen Tests für das Hypothesenpaar
   $$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \qquad \text{vs.} \qquad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$$
   die Testgröße
   $$T_n = 2\bigl(\log L_n(\mathbf{Y}_n, \widehat{\boldsymbol{\beta}}_n) - \log L_n(\mathbf{Y}_n, \boldsymbol{\beta}_0)\bigr)$$
   betrachtet werden. Wegen (48) wird dabei H_0 abgelehnt, wenn $T_n > \chi^2_{m,1-\alpha}$.
   - Von besonderem Interesse ist die Nullhypothese $H_0 : \boldsymbol{\beta} = \mathbf{o}$. Wenn diese abgelehnt wird, dann können speziellere Hypothesen getestet werden, zum Beispiel für jedes $i = 1, \ldots, m$ die Hypothese $H_0 : \beta_i = 0$.

Beachte

- Wenn $\mathbf{I}_n(\boldsymbol{\beta})$ positiv definit für jedes hinreichend große n ist und wenn
$$\lim_{n\to\infty} \mathbf{I}_n^{-1}(\boldsymbol{\beta}) = \mathbf{0}\,, \tag{49}$$
dann kann $\boldsymbol{\Gamma}_n = \mathbf{I}_n^{-1/2}$ in (47) und (48) gesetzt werden, so dass $\mathbf{K}(\boldsymbol{\beta})$ die Einheitsmatrix ist.
- In (37) hatten wir gezeigt, dass im logistischen Regressionsmodell $\mathbf{I}_n(\boldsymbol{\beta}) = \mathbf{X}^\top \operatorname{diag}\bigl(\pi_i(1-\pi_i)\bigr)\mathbf{X}$ gilt.
  - Weil wir voraussetzen, dass $0 < \pi_i(\boldsymbol{\beta}) < 1$ für jedes $i = 1, 2, \ldots$ und dass die Designmatrix $\mathbf{X}$ vollen Spaltenrang hat, ist die Matrix $\mathbf{I}_n(\boldsymbol{\beta})$ in diesem Fall positiv definit.
  - Wenn außerdem $\inf_{i\geq 1} \pi_i(1-\pi_i) > 0$ und wenn die Eintragungen x_{ij} der Designmatrix $\mathbf{X}$ so gewählt werden, dass $\lim_{n\to\infty} \bigl(\mathbf{X}^\top\mathbf{X}\bigr)^{-1} = \mathbf{0}$, dann gilt auch (49).
- Sei nun $\mathbf{K}(\boldsymbol{\beta})$ die Einheitsmatrix. Wegen (47) und (48) wird dann $H_0 : \beta_i = 0$ abgelehnt, wenn
$$\frac{\bigl|(\widehat{\boldsymbol{\beta}}_n)_i\bigr|}{\sqrt{\bigl(\mathbf{I}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\bigr)_{ii}}} > z_{1-\alpha/2}\,, \tag{50}$$
wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$–Quantil der $\mathrm{N}(0,1)$–Verteilung ist.

4.4 Gewichteter KQ–Schätzer bei kategorialer Regression

Anstelle des in Abschnitt 4.3 diskutierten Maximum–Likelihood–Ansatzes zur Schätzung des Parametervektors $\boldsymbol{\beta}$ betrachten wir nun für das kategoriale Regressionsmodell noch einen *gewichteten KQ–Schätzer* für $\boldsymbol{\beta}$.

4.4.1 Schätzung des Erwartungswertvektors

Zur Erinnerung (vgl. Abschnitt 4.2.2): Im binären kategorialen Regressionsmodell sind die Stichprobenvariablen $Y_1, \ldots, Y_n$ Bernoulli–verteilt, d.h., sie können nur die Werte 0 bzw. 1 mit positiver Wahrscheinlichkeit annehmen.

- Dabei verwenden wir so wie bisher die Schreibweise
$$\pi_i = \mathbb{P}(Y_i = 1) \qquad \Bigl(= \mu_i \;=\; \mathbb{E}\, Y_i\Bigr) \qquad \forall\, i = 1, \ldots, n\,,$$
wobei vorausgesetzt wird, dass $0 < \pi_i < 1$ für jedes $i = 1, \ldots, n$.
- In diesem Fall werden also die Wahrscheinlichkeiten $\pi_1, \ldots, \pi_n$ über eine Linkfunktion $g : (0,1) \to \mathbb{R}$ mit dem Parametervektor $\boldsymbol{\beta}$ verknüpft, d.h.,
$$\bigl(g(\pi_1), \ldots, g(\pi_n)\bigr)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{51}$$

- Um die Vektoren $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ bzw. $\mathbf{g}(\boldsymbol{\pi}) = \big(g(\pi_1), \ldots, g(\pi_n)\big)^\top$ schätzen zu können, nehmen wir an, dass wir für jedes $i = 1, \ldots, n$ jeweils $n_i > 0$ unabhängige und identisch verteilte „Kopien" $Y_{i1}, \ldots, Y_{in_i}$ von Y_i beobachten können. Der Gesamtstichprobenumfang ist dann gleich $\sum_{i=1}^n n_i$.
- Für jedes $i = 1, \ldots, n$ ist
$$\widehat{\pi}_i \;=\; \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \tag{52}$$
ein natürlicher Schätzer für π_i.
- Hieraus ergeben sich die Schätzer $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_n)^\top$ bzw. $\mathbf{g}(\widehat{\boldsymbol{\pi}}) = \big(g(\widehat{\pi}_1), \ldots, g(\widehat{\pi}_n)\big)^\top$ für $\boldsymbol{\pi}$ bzw. $g(\boldsymbol{\pi})$.

Man kann sich leicht überlegen, dass der Schätzer $\widehat{\boldsymbol{\pi}}$ erwartungstreu für $\boldsymbol{\pi}$ ist und dass seine Kovarianzmatrix $\mathbf{K}(\widehat{\boldsymbol{\pi}}) = \big(\mathrm{Cov}\,(\widehat{\pi}_i, \widehat{\pi}_j)\big)$ die folgende Form besitzt.

Lemma 4.3 *Es gilt*
$$\mathbb{E}\,\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}\,, \qquad \textit{und} \qquad \mathrm{Var}\,\widehat{\pi}_i = \pi_i(1-\pi_i)/n_i \tag{53}$$
und
$$\mathbf{K}(\widehat{\boldsymbol{\pi}}) = \mathrm{diag}\big(\mathrm{Var}\,\widehat{\pi}_i\big)\,. \tag{54}$$

Beweis Die Behauptung ergibt sich unmittelbar aus der Tatsache, dass die Zufallsvariablen $n_1\widehat{\pi}_1, \ldots, n_n\widehat{\pi}_n$ unabhängig und binomialverteilt sind mit $n_i\widehat{\pi}_i \sim \mathrm{B}(n_i, \pi_i)$ für jedes $i = 1, \ldots, n$. □

Außerdem ergibt sich aus dem folgenden *zentralen Grenzwertsatz*, dass der Schätzer $\mathbf{g}(\widehat{\boldsymbol{\pi}}) = \big(g(\widehat{\pi}_1), \ldots, g(\widehat{\pi}_n)\big)^\top$ asymptotisch normalverteilt ist.

Theorem 4.3 *Wenn $n_i \to \infty$ für jedes $i = 1, \ldots, n$, so dass*
$$\frac{\sum_{j=1}^n n_j}{n_i} \to \lambda_i \in [1, \infty) \qquad \forall\, i = 1, \ldots, n\,, \tag{55}$$
dann gilt
$$\Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{g}(\boldsymbol{\pi})\big) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{o}, \mathbf{K})\,, \tag{56}$$
wobei
$$\mathbf{K} = \mathrm{diag}(\alpha_i) \qquad \textit{und} \qquad \alpha_i = \lambda_i (g^{(1)}(\pi_i))^2 \pi_i (1-\pi_i)\,. \tag{57}$$

Beweis

- Weil wir voraussetzen, dass die Linkfunktion $g : (0,1) \to \mathbb{R}$ zweimal stetig differenzierbar ist, ergibt sich durch Taylor-Reihenentwicklung, dass für jedes $i = 1, \ldots, n$
$$g(\widehat{\pi}_i) - g(\pi_i) \;=\; g^{(1)}(\pi_i)\big(\widehat{\pi}_i - \pi_i\big) + g^{(2)}(Z_i)\big(\widehat{\pi}_i - \pi_i\big)^2 \;=\; g^{(1)}(\pi_i)\big(\widehat{\pi}_i - \pi_i\big) + R_i\,,$$
wobei $R_i = g^{(2)}(Z_i)\big(\widehat{\pi}_i - \pi_i\big)^2$ und $Z_i : \Omega \to \mathbb{R}$ eine Zufallsvariable ist, deren Werte zwischen $\widehat{\pi}_i$ und π_i liegen.
- Aus dem zentralen Grenzwertsatz für Summen von unabhängigen und identisch verteilten Zufallsvariablen (vgl. Theorem WR-5.16) ergibt sich, dass
$$n_i^{1/2}\big(\widehat{\pi}_i - \pi_i\big) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \pi_i(1-\pi_i)) \qquad \forall\, i = 1, \ldots, n\,. \tag{58}$$

- Weil $\widehat{\pi}_i - \pi_i \to 0$ und somit auch $Z_i - \pi_i \to 0$ bzw. $g^{(2)}(Z_i) \to g^{(2)}(\pi_i)$ mit Wahrscheinlichkeit 1, gilt außerdem, dass $n_i^{1/2} R_i \xrightarrow{\mathrm{P}} 0$ bzw.

$$\Bigl(\sum\nolimits_{j=1}^n n_j\Bigr)^{1/2} R_i = \left(\frac{\sum_{j=1}^n n_j}{n_i}\right)^{1/2} n_i^{1/2} R_i \xrightarrow{\mathrm{P}} 0\,.$$

- Insgesamt ergibt sich also mit Hilfe des Satzes von Slutsky (vgl. die Theoreme WR-5.9 und WR-5.11), dass

$$\begin{aligned}\Bigl(\sum\nolimits_{j=1}^n n_j\Bigr)^{1/2}\bigl(g(\widehat{\pi}_i) - g(\pi_i)\bigr) &= \left(\frac{\sum_{j=1}^n n_j}{n_i}\right)^{1/2} g^{(1)}(\pi_i) n_i^{1/2}(\widehat{\pi}_i - \pi_i) + \Bigl(\sum\nolimits_{j=1}^n n_j\Bigr)^{1/2} R_i \\ &\xrightarrow{\mathrm{d}} \mathrm{N}\bigl(0, \lambda_i (g^{(1)}(\pi_i))^2 \pi_i(1-\pi_i)\bigr)\,.\end{aligned}$$

- Weil die Zufallsvariablen $g(\widehat{\pi}_1) - g(\pi_1), \ldots, g(\widehat{\pi}_i) - g(\pi_i)$ unabhängig sind, ist damit die Behauptung bewiesen. □

4.4.2 Asymptotische Normalverteiltheit des KQ–Schätzers

Durch die Gestalt der asymptotischen Kovarianzmatrix $\mathbf{K}$ in Theorem 4.3 wird der folgende Ansatz zur Schätzung des Parametervektors $\boldsymbol{\beta}$ motiviert.

- Ähnlich wie in Abschnitt 2.1 betrachten wir hierfür die Methode der kleinsten Quadrate zur Bestimmung eines Schätzers $\widehat{\boldsymbol{\beta}}$ für die unbekannten Regressionskoeffizienten $\beta_1, \ldots, \beta_m$.
- Und zwar soll ein Zufallsvektor $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)^\top$ bestimmt werden, so dass der *gewichtete quadratische Fehler*

$$e(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{\bigl(g(\widehat{\pi}_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\bigr)^2}{\widehat{\sigma}_{ii}^2} \tag{59}$$

für $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ minimal wird, wobei $\widehat{\sigma}_{ii}^2 = \bigl(\sum_{j=1}^n n_j/n_i\bigr)(g^{(1)}(\widehat{\pi}_i))^2 \widehat{\pi}_i(1-\widehat{\pi}_i)$ und vorausgesetzt wird, dass die Gewichte $\widehat{\sigma}_{ii}^2$ positiv sind.

Beachte

- Die gewichtete Summe $e(\boldsymbol{\beta})$ der quadrierten Residuen $\bigl(g(\widehat{\pi}_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\bigr)^2$ in (59) kann wie folgt dargestellt werden: Mit der Schreibweise $\widehat{\mathbf{K}} = \mathrm{diag}\bigl(\widehat{\sigma}_{ii}^2\bigr)$ gilt

$$e(\boldsymbol{\beta}) = \bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}\bigr)^\top \widehat{\mathbf{K}}^{-1}\bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}\bigr)\,. \tag{60}$$

- Genauso wie im Beweis von Theorem 2.1 lässt sich zeigen, dass der gewichtete quadratische Fehler $e(\boldsymbol{\beta})$ genau dann minimal ist, wenn $\boldsymbol{\beta}$ Lösung der folgenden Normalengleichung ist:

$$\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{g}(\widehat{\boldsymbol{\pi}})\,. \tag{61}$$

- Weil die Matrix $\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X}$ invertierbar ist, hat (61) die eindeutig bestimmte Lösung

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{g}(\widehat{\boldsymbol{\pi}})\,. \tag{62}$$

Wir zeigen nun, dass der gewichtete KQ–Schätzer $\widehat{\boldsymbol{\beta}}$ in (62) asymptotisch normalverteilt ist, wenn die (Teil–) Stichprobenumfänge n_i für jedes $i = 1, \ldots, n$ unbegrenzt wachsen.

Hierfür benötigen wir die folgenden vektoriellen Versionen des Satzes von Slutsky (vgl. die Theoreme WR–5.9 und WR–5.11) sowie des „Continuous Mapping Theorems" (vgl. Theorem WR–5.12).

Lemma 4.4

- *Sei $m \in \mathbb{N}$, seien $\mathbf{Y}, \mathbf{Y}_n, \mathbf{Z}_n : \Omega \to \mathbb{R}^m$ beliebige Zufallsvektoren über einunddemselben Wahrscheinlichkeitsraum, und sei $c \in \mathbb{R}^m$.*
- *Wenn $\mathbf{Y}_n \xrightarrow{\mathrm{d}} \mathbf{Y}$ und $\mathbf{Z}_n \xrightarrow{\mathrm{d}} c$, dann gilt $\mathbf{Y}_n + \mathbf{Z}_n \xrightarrow{\mathrm{d}} \mathbf{Y} + c$ und $\mathbf{Y}_n^\top \mathbf{Z}_n \xrightarrow{\mathrm{d}} c^\top \mathbf{Y}$.*

Lemma 4.5

- *Sei $m \in \mathbb{N}$, seien $\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \ldots : \Omega \to \mathbb{R}^m$ beliebige Zufallsvektoren, und sei $\varphi : \mathbb{R}^m \to \mathbb{R}$ eine stetige Funktion.*
- *Dann gilt $\varphi(\mathbf{Z}_n) \xrightarrow{\mathrm{d}} \varphi(\mathbf{Z})$, wenn $\mathbf{Z}_n \xrightarrow{\mathrm{d}} \mathbf{Z}$.*

Die *Beweise* der Lemmas 4.4 und 4.5 verlaufen ähnlich wie die Beweise der Theoreme WR–5.9, WR–5.11 bzw. WR–5.12. Sie werden deshalb hier weggelassen.

Theorem 4.4 *Wenn $n_i \to \infty$ für jedes $i = 1, \ldots, n$, so dass*

$$\frac{\sum_{j=1}^n n_j}{n_i} \to \lambda_i \in [1, \infty) \qquad \forall\, i = 1, \ldots, n\,, \tag{63}$$

dann gilt

$$\Bigl(\sum_{j=1}^n n_j\Bigr)^{1/2} \bigl(\widehat{\boldsymbol\beta} - \boldsymbol\beta\bigr) \xrightarrow{\mathrm{d}} \mathrm{N}\bigl(\mathbf{o}, \bigl(\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X}\bigr)^{-1}\bigr)\,, \tag{64}$$

wobei $\mathbf{K} = \mathrm{diag}(\alpha_i)$ die in Theorem 4.3 betrachtete Diagonalmatrix ist.

Beweis

- Aus der Definitionsgleichung von $\widehat{\boldsymbol\beta}$ in (62) ergibt sich, dass
$$\begin{aligned}\widehat{\boldsymbol\beta} - \boldsymbol\beta &= (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{g}(\widehat{\boldsymbol\pi}) - \boldsymbol\beta \;=\; (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \bigl(\mathbf{g}(\widehat{\boldsymbol\pi}) - \mathbf{X}\boldsymbol\beta\bigr)\\ &= (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \bigl(\mathbf{g}(\widehat{\boldsymbol\pi}) - \mathbf{g}(\boldsymbol\pi)\bigr)\,,\end{aligned}$$
  wobei in der letzten Gleichheit verwendet wurde, dass $\mathbf{g}(\boldsymbol\pi) = \mathbf{X}\boldsymbol\beta$; vgl. (51).
- In Theorem 4.3 hatten wir gezeigt, dass
$$\Bigl(\sum_{j=1}^n n_j\Bigr)^{1/2} \bigl(\mathbf{g}(\widehat{\boldsymbol\pi}) - \mathbf{g}(\boldsymbol\pi)\bigr) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{o}, \mathbf{K})\,,$$
  wobei die asymptotische Kovarianzmatrix $\mathbf{K}$ in (57) gegeben ist.
- Außerdem gilt $\widehat{\mathbf{K}} \xrightarrow{\mathrm{P}} \mathbf{K}$, wenn $n_i \to \infty$ für jedes $i = 1, \ldots, n$.
- Insgesamt ergibt sich hieraus mit Hilfe
  - des Satzes von Slutsky (vgl. Lemma 4.4),
  - des „Continuous Mapping Theorems" (vgl. Lemma 4.5) sowie
  - des Theorems 1.3 über die Lineartransformation normalverteilter Zufallsvektoren,

  dass
$$\begin{aligned}\Bigl(\sum_{j=1}^n n_j\Bigr)^{1/2} \bigl(\widehat{\boldsymbol\beta} - \boldsymbol\beta\bigr) &= \Bigl(\sum_{j=1}^n n_j\Bigr)^{1/2} (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \bigl(\mathbf{g}(\widehat{\boldsymbol\pi}) - \mathbf{g}(\boldsymbol\pi)\bigr)\\ &= \Bigl(\sum_{j=1}^n n_j\Bigr)^{1/2} \underbrace{(\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \bigl((\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1}\bigr)^{-1}}_{\xrightarrow{\mathrm{P}} \mathbf{I}}\\ &\qquad\qquad \times (\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1} \bigl(\mathbf{g}(\widehat{\boldsymbol\pi}) - \mathbf{g}(\boldsymbol\pi)\bigr)\\ &\xrightarrow{\mathrm{d}} \mathrm{N}\bigl(\mathbf{o}, \bigl((\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1}\bigr)\, \mathbf{K}\, \bigl((\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1}\bigr)^\top\bigr)\\ &\qquad = \mathrm{N}\bigl(\mathbf{o}, \bigl(\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X}\bigr)^{-1}\bigr)\,. \end{aligned}$$

$\square$

Beachte

- Wenn $n_1 + \ldots + n_n$ eine große Zahl ist, dann kann zur Konstruktion eines asymptotischen Tests für das Hypothesenpaar $H_0 : \beta_i = 0$ vs. $H_1 : \beta \neq 0$ die Testgröße
$$T = \Bigl(\sum\nolimits_{j=1}^{n} n_j\Bigr)^{1/2} \widehat{\beta}_i \Big/ \sqrt{\widetilde{k}_{ii}}$$
betrachtet werden, wobei $\widetilde{k}_{ii}$ das i-te Diagonalelement der Matrix $\widetilde{\mathbf{K}} = \bigl(\mathbf{X}^\top \widehat{\mathbf{K}}^{-1}\mathbf{X}\bigr)^{-1}$ ist.
- Wegen Theorem 4.4 wird dabei H_0 abgelehnt, wenn $|T| > z_{1-\alpha/2}$.

4.4.3 Bewertung der Anpassungsgüte

Ein wichtiges Problem ist die Wahl einer geeigneten Designmatrix $\mathbf{X}$, um das Modell $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$ möglichst gut an die vorliegenden Daten anzupassen. Das folgende Resultat kann zur Beantwortung dieser Frage dienen.

Theorem 4.5 *Unter den Bedingungen von Theorem* 4.3 *gilt*
$$\sum_{j=1}^{n} n_j \,\bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr)^\top \widehat{\mathbf{K}}^{-1} \bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr) \xrightarrow{\mathrm{d}} \chi^2_{n-m}\,. \tag{65}$$

Beweis

- In Theorem 4.3 hatten wir gezeigt, dass der Zufallsvektor $\bigl(\sum_{j=1}^{n} n_j\bigr)^{1/2}\bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}\bigr)$ näherungsweise $\mathrm{N}(\mathbf{o}, \mathbf{K})$ verteilt ist.
- Weil $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$ und $\widehat{\mathbf{K}} \xrightarrow{\mathrm{P}} \mathbf{K}$, ergibt sich die Behauptung mit Hilfe
  - des Satzes von Slutsky (vgl. Lemma 4.4),
  - des „Continuous Mapping Theorems” (vgl. Lemma 4.5) sowie
  - des Theorems 1.9 über quadratische Formen von normalverteilten Zufallsvektoren. □

Beachte

- Wegen Theorem 4.5 kann die Größe $\sum_{j=1}^{n} n_j \,\bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr)^\top \widehat{\mathbf{K}}^{-1} \bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr)$ als Maßzahl für die Güte der Anpassung des Models $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$ an die vorliegenden Daten aufgefasst werden.
- Dabei wird die Anpassungsgüte als hinreichend gut eingeschätzt, wenn
$$\sum_{j=1}^{n} n_j \,\bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr)^\top \widehat{\mathbf{K}}^{-1} \bigl(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\bigr) \;<\; \chi^2_{n-m,1-\alpha}\,.$$
- Andererseits sollten die Dimensionen von $\mathbf{X}$ möglichst klein sein, d.h. insbesondere, dass für jedes $i = 1, \ldots, m$ die Nullhypothese des Tests $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ klar abgelehnt werden sollte.

5 Tests von Verteilungsannahmen

- In diesem Kapitel bezeichnen wir die Stichprobenvariablen mit $X_1, \ldots, X_n$, wobei wir von jetzt an stets voraussetzen, dass $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ eine Folge von unabhängigen *und* identisch verteilten Zufallsvariablen ist.
- Die Annahmen, die wir bisher über die Verteilung P der Stichprobenvariablen $X_1, \ldots, X_n$ gemacht haben, waren entweder *rein qualitativ* (diskrete bzw. absolutstetige Verteilung) oder *parametrisch*, wobei in diesem Fall vorausgesetzt wurde,
  - dass P zu einer parametrischen Familie $\{P_\theta,\, \theta \in \Theta\}$ von Verteilungen mit $\Theta \subset \mathbb{R}^m$ gehört
  - und dass lediglich der Parametervektor $\theta = (\theta_1, \ldots, \theta_m)^\top$ bzw. ein Teil seiner Komponenten unbekannt ist.
- Im Folgenden diskutieren wir nun so genannte *Anpassungstests*, die in der englischsprachigen Literatur *Goodness–of–Fit–Tests* genannt werden.
  - Dabei betrachten wir zunächst einen Test, um die Hypothese $H_0 : P = P_0$ zu verifizieren, dass die Verteilung P der Stichprobenvariablen gleich einer vorgegebenen (hypothetischen) Verteilung P_0 ist.
  - Anschließend konstruieren wir Tests, um zu prüfen, ob P zu einer vorgegebenen (parametrischen) *Klasse* von Verteilungen $\{P_\theta,\, \theta \in \Theta\}$ gehört.

5.1 Kolmogorow–Smirnow–Test

5.1.1 Empirische Verteilungsfunktion; KS–Teststatistik

- In der Literatur werden verschiedene Tests vorgeschlagen, um die Hypothese $H_0 : P = P_0$ zu verifizieren, dass die Verteilung P der unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ gleich einer vorgegebenen Verteilung P_0 ist.
- Ein solches Verfahren ist der *Kolmogorow–Smirnow–Test*, der auf der Untersuchung der in Abschnitt I–1.5 eingeführten *empirischen Verteilungsfunktion* $\widehat{F}_n : \mathbb{R} \times \mathbb{R}^n \to [0,1]$ beruht, wobei

$$\widehat{F}_n(t; x_1, \ldots, x_n) = \frac{\#\{i : 1 \le i \le n,\, x_i \le t\}}{n}\,. \tag{1}$$

- Dabei wird die Stichprobenfunktion $T_n : \mathbb{R}^n \to [0, \infty)$ betrachtet mit

$$T_n(x_1, \ldots, x_n) = \sqrt{n}\, \sup_{t \in \mathbb{R}} |\, \widehat{F}_n(t; x_1, \ldots, x_n) - F_0(t)|\,. \tag{2}$$

- In Abschnitt I–1.5.3 hatten wir gezeigt, dass die Verteilung der *KS–Teststatistik* $T_n(X_1, \ldots, X_n)$ nicht von P_0 abhängt, wenn vorausgesetzt wird, dass die zu P_0 gehörende Verteilungsfunktion $F_0 : \mathbb{R} \to [0,1]$ stetig ist, vgl. Theorem I–1.19.
- Sei $s_{n,1-\alpha}$ das $(1-\alpha)$–Quantil der Verteilung von $T_n(X_1, \ldots, X_n)$ unter einer beliebigen stetigen Verteilungsfunktion F_0. Der Kolmogorow–Smirnow–Test verwirft die Nullhypothese $H_0 : P = P_0$, wenn

$$T_n(x_1, \ldots, x_n) > s_{n,1-\alpha}\,. \tag{3}$$

Beachte

- Die Quantile $s_{n,1-\alpha}$ lassen sich durch Monte–Carlo–Simulation bestimmen, wobei die Verteilungsfunktion F_0 der Standard–Gleichverteilung in $[0,1]$ zugrunde gelegt werden kann, vgl. Korollar I–1.3.

- Wenn nicht vorausgesetzt wird, dass F_0 stetig ist, dann liefert die in (3) betrachtete Entscheidungsregel einen Test, dessen Niveau kleiner als α sein kann.
- Wenn jedoch das Quantil $s'_{n,1-\alpha}$ von $T_n(X_1,\ldots,X_n)$ unter F_0 (beispielsweise durch MC–Simulation) bestimmt werden kann, dann ist durch $T_n(x_1,\ldots,x_n) > s'_{n,1-\alpha}$ auch bei unstetigem F_0 ein Test gegeben, der das Niveau α ausschöpft.

5.1.2 Asymptotische Verteilung

Wir untersuchen nun die asymptotische Verteilung der in (2) eingeführten KS–Teststatistik $T_n(X_1,\ldots,X_n)$, wenn $n\to\infty$. Hierfür stellen wir zunächst einige Hilfsmittel bereit.

Insbesondere benötigen wir den folgenden *Stetigkeitssatz* für charakteristische Funktionen von Zufallsvektoren, der eine mehrdimensionale Verallgemeinerung von Theorem WR–5.20 ist und den wir hier ohne Beweis angeben.

Lemma 5.1 *Sei $m\in\mathbb{N}$, und seien $\mathbf{Z},\mathbf{Z}_1,\mathbf{Z}_2,\ldots:\Omega\to\mathbb{R}^m$ beliebige Zufallsvektoren mit den charakteristischen Funktionen $\varphi_{\mathbf{Z}_n}$ bzw $\varphi_{\mathbf{Z}}$. Es gilt $\mathbf{Z}_n \xrightarrow{\mathrm{d}} \mathbf{Z}$ genau dann, wenn*

$$\lim_{n\to\infty}\varphi_{\mathbf{Z}_n}(\mathbf{t})=\varphi_{\mathbf{Z}}(\mathbf{t})\qquad\forall\,\mathbf{t}\in\mathbb{R}^m\,. \tag{4}$$

Außerdem benötigen wir einen *multivariaten zentralen Grenzwertsatz* für Summen von unabhängigen und identisch verteilten Zufallsvektoren,

- dessen Beweis mit Hilfe von Lemma 5.1 auf den entsprechenden zentralen Grenzwertsatz für reellwertige Zufallsvariablen (vgl. Theorem WR–5.16) zurückgeführt werden kann.
- Dieser Ansatz wird in der englischsprachigen Literatur *Cramèr–Wold–Device* genannt.

Lemma 5.2

- *Sei $m\in\mathbb{N}$, und sei $\mathbf{Z}_1,\mathbf{Z}_2,\ldots:\Omega\to\mathbb{R}^m$ eine Folge von unabhängigen und identisch verteilten Zufallsvektoren mit Erwartungswertvektor $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_m)^\top$ und Kovarianzmatrix $\mathbf{K}$.*
- *Dann gilt*
$$\lim_{n\to\infty}\mathbb{P}\Bigl(\frac{(\mathbf{Z}_1+\ldots+\mathbf{Z}_n)-n\boldsymbol{\mu}}{\sqrt{n}}\le\mathbf{x}\Bigr)=\Phi_{\mathbf{K}}(\mathbf{x})\qquad\forall\,\mathbf{x}\in\mathbb{R}^m\,, \tag{5}$$
wobei $\Phi_{\mathbf{K}}:\mathbb{R}^m\to[0,1]$ die Verteilungsfunktion der $\mathrm{N}(\mathbf{o},\mathbf{K})$–Verteilung bezeichnet.

Beweis

- Sei $\mathbf{Z}_n=(Z_{n1},\ldots,Z_{nm})^\top$. Wegen Lemma 5.1 ist die Verteilungskonvergenz (5) damit äquivalent, dass für jedes $\mathbf{t}\in\mathbb{R}^m$
$$\lim_{n\to\infty}\varphi_n(\mathbf{t})=\varphi(\mathbf{t})\,, \tag{6}$$
  - wobei $\varphi_n(\mathbf{t})$ die charakteristische Funktion von $(\mathbf{Z}_1+\ldots+\mathbf{Z}_n-n\boldsymbol{\mu})/\sqrt{n}$ ist mit
$$\varphi_n(\mathbf{t})=\mathbb{E}\,\exp\Bigl(\mathrm{i}\sum_{j=1}^m t_j\;\frac{(Z_{1j}+\ldots+Z_{nj})-n\mu_j}{\sqrt{n}}\Bigr)$$
  - und $\varphi(\mathbf{t})$ die charakteristische Funktion der $\mathrm{N}(\mathbf{o},\mathbf{K})$–Verteilung ist mit
$$\varphi(\mathbf{t})=\exp\Bigl(-\,\frac{1}{2}\,\mathbf{t}^\top\mathbf{K}\mathbf{t}\Bigr)\,. \tag{7}$$

- Außerdem kann man sich leicht überlegen, dass

$$\varphi_n(\mathbf{t}) = \mathbb{E}\exp\Bigl(\mathrm{i}\sum_{k=1}^{n}\frac{\sum\limits_{j=1}^{m} t_j(Z_{kj}-\mu_j)}{\sqrt{n}}\Bigr) \qquad \forall\, \mathbf{t}=(t_1,\ldots,t_m)^\top \in \mathbb{R}^m \tag{8}$$

  und

$$\mathbb{E}\Bigl(\sum_{j=1}^{m} t_j(Z_{kj}-\mu_j)\Bigr) = 0\,, \qquad \mathrm{Var}\Bigl(\sum_{j=1}^{m} t_j(Z_{kj}-\mu_j)\Bigr) = \mathbf{t}^\top\mathbf{K}\mathbf{t} \qquad \forall\, k \in \mathbb{N}\,. \tag{9}$$

- Wenn $\mathbf{t}^\top\mathbf{K}\mathbf{t} = 0$, dann ergibt sich aus (9),
  - dass $\sum_{j=1}^{m} t_j(Z_{kj}-\mu_j) = 0$ mit Wahrscheinlichkeit 1 für beliebige $k = 1,\ldots,n$ und $n \ge 1$.
  - Hieraus und aus (7) – (8) folgt, dass $\varphi_n(\mathbf{t}) = 1 = \varphi(\mathbf{t})$ für jedes $n \ge 1$, d.h., (6) gilt.
- Sei nun $\mathbf{t}^\top\mathbf{K}\mathbf{t} > 0$.
  - Aus (8) ergibt sich, dass die Zahl $\varphi_n(\mathbf{t})$ der Wert der charakteristischen Funktion der reellwertigen Zufallsvariable $\sum_{k=1}^{n}\sum_{j=1}^{m} t_j(Z_{kj}-\mu_j)/\sqrt{n}$ an der Stelle 1 ist.
  - Außerdem ergibt sich aus (7), dass $\varphi(\mathbf{t})$ der Wert der charakteristischen Funktion der (eindimensionalen) Normalverteilung $\mathrm{N}(0,\mathbf{t}^\top\mathbf{K}\mathbf{t})$ an der Stelle 1 ist.
  - Andererseits folgt aus Theorem WR–5.16, d.h., aus dem (1–dimensionalen) zentralen Grenzwertsatz für Summen von unabhängigen und identisch verteilten (reellwertigen) Zufallsvariablen, dass für $n \to \infty$

$$\sum_{k=1}^{n}\frac{\sum\limits_{j=1}^{m} t_j(Z_{kj}-\mu_j)}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,\mathbf{t}^\top\mathbf{K}\mathbf{t})\,. \tag{10}$$

  - Hieraus, aus (7) – (8) und aus Theorem WR-5.20, d.h., aus dem Stetigkeitssatz für charakteristische Funktionen von reellwertigen Zufallsvariablen ergibt sich nun die Gültigkeit von (6). □

Der folgende Grenzwertsatz, der bereits in Abschnitt I–1.5.3 erwähnt wurde, liefert eine Näherungsformel für die Verteilungsfunktion von $T_n(X_1,\ldots,X_n)$ bei großem Stichprobenumfang n.

Theorem 5.1 *Die Verteilungsfunktion $F_0 : \mathbb{R} \to [0,1]$ sei stetig. Unter $H_0 : P = P_0$ gilt dann*

$$\lim_{n\to\infty} \mathbb{P}\bigl(T_n(X_1,\ldots,X_n) \le x\bigr) = K(x) \qquad \forall\, x \in \mathbb{R}\,,$$

wobei $K : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der sogenannten Kolmogorow–Verteilung ist mit

$$K(x) = \begin{cases} 1 - 2\sum\limits_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2x^2)\,, & \textit{wenn } x > 0, \\ 0\,, & \textit{wenn } x \le 0. \end{cases} \tag{11}$$

Beweis

- Wir skizzieren hier lediglich die Beweisidee, denn der komplette Beweis von Theorem 5.1 (vgl. z.B. A. van der Vaart und J. Wellner (1996)) geht über den Rahmen dieser Vorlesung hinaus,
  - weil er relativ tiefliegende Hilfsmittel aus der Theorie der stochastischen Prozesse erfordert, die in den Kursvorlesungen nicht behandelt werden.
  - Dabei wird insbesondere der Begriff der *Verteilungskonvergenz in Funktionenräumen* sowie ein so genannter *funktionaler zentraler Grenzwertsatz* benötigt,
  - der als eine (unendlich dimensionale) Verallgemeinerung der klassischen zentralen Grenzwertsätze für Summen von rellwertigen Zufallsvariablen (vgl. Abschnitt WR–5.3) bzw. von endlich–dimensionalen Zufallsvektoren (vgl. Lemma 5.2) aufgefasst werden kann.

- Weil die Verteilung von $T_n(X_1,\ldots,X_n)$ nicht von F_0 abhängt (vgl. Theorem I–1.19), können wir o.B.d.A. annehmen, dass F_0 die Verteilungsfunktion der Gleichverteilung in $[0,1]$ ist, d.h., $F_0(t)=t$ für jedes $t\in[0,1]$.
  - Um die asymptotische Verteilung von $T_n(X_1,\ldots,X_n)$ für $n\to\infty$ zu untersuchen, verwenden wir dabei die abkürzende Schreibweise
$$B_n(t)=\sqrt{n}\,\bigl(\widehat F_n(t;X_1,\ldots,X_n)-F_0(t)\bigr)\qquad\forall\, t\in[0,1]\,, \tag{12}$$
  - wobei die Familie von Zufallsvariablen $\{B_n(t),\,t\in[0,1]\}$ ein stochastischer Prozess ist, der in der Literatur *empirischer Prozess* genannt wird.
- Für beliebige $t_1,\ldots,t_m\in[0,1]$ gilt dann $\sqrt{n}\bigl(B_n(t_1),\ldots,B_n(t_m)\bigr)=\sum_{i=1}^n\bigl(Y_i(t_1)-t_1,\ldots,Y_i(t_m)-t_m\bigr)$, wobei
$$Y_i(t_j)=\begin{cases}1\,, & \text{wenn } X_i\le t_j,\\ 0\,, & \text{wenn } X_i>t_j.\end{cases}$$
  - Der Zufallsvektor $\sqrt{n}\bigl(B_n(t_1),\ldots,B_n(t_m)\bigr)$ lässt sich somit für jedes $n\ge 1$ als Summe von n unabhängigen und identisch verteilten Zufallsvektoren mit Erwartungswertvektor $\mathbf{o}$ darstellen, dessen Kovarianzmatrix $\mathbf{K}=(\sigma_{ij}^2)$ durch $\sigma_{ij}^2=\min\{t_i,t_j\}-t_it_j$ gegeben ist.
  - Aus Lemma 5.2 ergibt sich nun, dass für $n\to\infty$
$$\bigl(B_n(t_1),\ldots,B_n(t_m)\bigr)\stackrel{\mathrm{d}}{\longrightarrow}\bigl(B(t_1),\ldots,B(t_m)\bigr)\,, \tag{13}$$
  wobei $\bigl(B(t_1),\ldots,B(t_m)\bigr)$ ein normalverteilter Zufallsvektor ist mit $\bigl(B(t_1),\ldots,B(t_m)\bigr)\sim\mathrm{N}(\mathbf{o},\mathbf{K})$.
  - Hieraus und aus dem Continuous–Mapping–Theorem für Zufallsvektoren (vgl. Lemma 4.5) folgt, dass
$$\max_{i=1,\ldots,m}\sqrt{n}\bigl|\widehat F_n(t_i;X_1,\ldots,X_n)-F_0(t_i)\bigr|\stackrel{\mathrm{d}}{\longrightarrow}\max_{i=1,\ldots,m}\bigl|B(t_i)\bigr|\,, \tag{14}$$
- Man kann sich leicht überlegen, dass die Verteilung $\mathrm{N}(\mathbf{o},\mathbf{K})$ des Zufallsvektors $\bigl(B(t_1),\ldots,B(t_m)\bigr)$ als
  - endlich–dimensionale Verteilung des so genannten *Brownschen Brückenprozesses* $\{B(t),\,t\in[0,1]\}$ mit $B(t)=X(t)-tX(1)$ aufgefasst werden kann, wobei $\{X(t),\,t\in[0,1]\}$ ein (Standard–) Wiener–Prozess ist,
  - d.h., $\{X(t),\,t\in[0,1]\}$ ist ein stochastischer Prozess mit stetigen Trajektorien und unabhängigen Zuwächsen, so dass $X(0)=0$ und $X(t_2)-X(t_1)\sim\mathrm{N}(0,t_2-t_1)$ für beliebige $t_1,t_2\in[0,1]$ mit $t_1<t_2$, vgl. Abschnitt 2.4 des Skriptes zur Vorlesung „Wahrscheinlichkeitstheorie".
- Mit Hilfe der Theorie der Verteilungskonvergenz in Funktionenräumen sowie eines entsprechenden funktionalen zentralen Grenzwertsatzes kann man nun zeigen, dass nicht nur die „endlich–dimensionalen" Konvergenzen (13) und (14) gelten, sondern dass darüber hinaus auch
$$\bigl(B_n(t),\,t\in[0,1]\bigr)\stackrel{\mathrm{d}}{\longrightarrow}\bigl(B(t),\,t\in[0,1]\bigr) \tag{15}$$
  bzw.
$$\max_{t\in[0,1]}\sqrt{n}\bigl|\widehat F_n(t;X_1,\ldots,X_n)-F_0(t)\bigr|\stackrel{\mathrm{d}}{\longrightarrow}\max_{t\in[0,1]}\bigl|B(t)\bigr|\,. \tag{16}$$
- Außerdem kann man zeigen, dass die Verteilungsfunktion des Maximums $\max_{t\in[0,1]}\bigl|B(t)\bigr|$ der Brownschen Brücke $\{B(t),\,t\in[0,1]\}$ durch (11) gegeben ist. $\square$

Beachte

- Wegen Theorem 5.1 wird bei hinreichend großem Stichprobenumfang (als „Faustregel" gilt $n>40$, vgl. die Bemerkung am Ende von Abschnitt I–1.5.3) die Hypothese $H_0: P=P_0$ abgelehnt, wenn
$$T_n(x_1,\ldots,x_n)>\xi_{1-\alpha}\,,$$
- wobei $\xi_{1-\alpha}$ das $(1-\alpha)$–Quantil der in (11) gegebenen Kolmogorow–Verteilung bezeichnet, d.h., $\xi_{1-\alpha}$ ist Lösung der Gleichung $K(\xi_{1-\alpha})=1-\alpha$.

5.1.3 Güteeigenschaften; punktweise und gleichmäßige Konsistenz

In diesem Abschnitt betrachten wir einige Güteeigenschaften des Kolmogorow–Smirnow–Tests.

Um die (punktweise) Konsistenz des KS–Tests zu zeigen, benötigen wir den *Satz von Gliwenko–Cantelli* (vgl. Theorem I–1.18), d.h., dass

$$\mathbb{P}_{F_0}\Bigl(\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\bigl|\widehat{F}_n(t;X_1,\ldots,X_n)-F_0(t)\bigr|=0\Bigr)=1\,. \tag{17}$$

Theorem 5.2 *Die Verteilungsfunktion $F_0:\mathbb{R}\to[0,1]$ sei stetig. Dann ist der Kolmogorow–Smirnow–Test punktweise konsistent für jede Verteilungsfunktion F der Stichprobenvariablen mit $F\neq F_0$, d.h., es gilt*

$$\lim_{n\to\infty}\mathbb{P}_F\bigl(T_n(X_1,\ldots,X_n)>s_{n,1-\alpha}\bigr)=1\,. \tag{18}$$

Beweis

- Aus (17) ergibt sich, dass für jedes $F\neq F_0$
$$\mathbb{P}_F\Bigl(\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\bigl|\widehat{F}_n(t;X_1,\ldots,X_n)-F_0(t)\bigr|>0\Bigr)=1\,.$$
- Hieraus folgt, dass $T_n(X_1,\ldots,X_n)\to\infty$ mit Wahrscheinlichkeit 1 unter $F\neq F_0$.
- Weil $s_{n,1-\alpha}\to\xi_{1-\alpha}<\infty$ für $n\to\infty$, wobei $\xi_{1-\alpha}$ das $(1-\alpha)$–Quantil der in (11) gegebenen Kolmogorow–Verteilung ist, gilt auch, dass $T_n(X_1,\ldots,X_n)-(s_{n,1-\alpha}-\xi_{1-\alpha})\stackrel{\text{f.s.}}{\longrightarrow}\infty$ und somit
$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}_F\bigl(T_n(X_1,\ldots,X_n)>s_{n,1-\alpha}\bigr) &= \lim_{n\to\infty}\mathbb{P}_F\bigl(T_n(X_1,\ldots,X_n)-(s_{n,1-\alpha}-\xi_{1-\alpha})>\xi_{1-\alpha}\bigr)\\ &= \lim_{n\to\infty}\mathbb{P}_F\bigl(T_n(X_1,\ldots,X_n)>\xi_{1-\alpha}\bigr)\;=\;1\,.\end{aligned}$$
□

Beachte

- In Verschärfung von Theorem 5.2 kann man zeigen, dass der KS–Test auch *gleichmäßig konsistent* ist, wenn der *Kolmogorow–Abstand*
$$d_K(\Delta_n;F_0)=\inf_{F\in\Delta_n}\sup_{t\in\mathbb{R}}|F(t)-F_0(t)| \tag{19}$$
zwischen der Familie Δ_n der alternativen Verteilungsfunktionen und der (hypothetischen) Verteilungsfunktion F_0 mit wachsendem Stichprobenumfang n nicht zu schnell gegen 0 konvergiert.
- Hierfür benötigen wir die folgende Verschärfung des Satzes von Gliwenko–Cantelli, die in der Literatur die *Ungleichung von Dworetsky–Kiefer–Wolfowitz* genannt wird und die wir hier ohne Beweis angeben.

Lemma 5.3 *Für beliebige $c>0$ und $n\geq 1$ gilt*

$$\mathbb{P}_F\Bigl(\sup_{t\in\mathbb{R}}\bigl|\widehat{F}_n(t;X_1,\ldots,X_n)-F(t)\bigr|>c\Bigr)\leq C\exp\bigl(-2nc^2\bigr)\,, \tag{20}$$

wobei $C\leq 2$ eine universelle Konstante ist, die nicht von F abhängt.

Beachte

- Aus Lemma 5.3 folgt, dass es für jedes $\varepsilon>0$ ein $c'>0$ gibt, das nicht von F abhängt und dass
$$\inf_{n\geq 1}\mathbb{P}_F\Bigl(\sqrt{n}\sup_{t\in\mathbb{R}}\bigl|\widehat{F}_n(t;X_1,\ldots,X_n)-F(t)\bigr|\leq c'\Bigr)\geq 1-\varepsilon\,. \tag{21}$$

- Um dies zu sehen, genügt es für $\varepsilon \in (0,1)$, den Schwellenwert c in (20) so zu wählen, dass $c = c'/\sqrt{n}$, wobei
$$c' = \sqrt{-\frac{1}{2}\log\Bigl(\frac{\varepsilon}{C}\Bigr)}\,.$$
- Weil c' nicht von F abhängt, gilt außerdem, dass es für jedes $\varepsilon > 0$ ein $c' > 0$ gibt, so dass
$$\inf_{n\ge 1} \mathbb{P}_{F_n}\Bigl(\sqrt{n}\sup_{t\in\mathbb{R}} \bigl|\widehat{F}_n(t; X_1,\ldots,X_n) - F_n(t)\bigr| \le c'\Bigr) \ge 1-\varepsilon\,, \tag{22}$$
wobei $\{F_n\}$ eine beliebige Folge von Verteilungsfunktionen ist.

Mit diesen Hilfsmitteln können wir die bereits oben erwähnte gleichmäßige Konsistenz–Eigenschaft des KS–Tests für den Fall zeigen, dass der Kolmogorow–Abstand $d_K(\Delta_n; F_0)$ zwischen der Familie Δ_n der alternativen Verteilungsfunktionen und der (hypothetischen) Verteilungsfunktion F_0 mit wachsendem Stichprobenumfang n nicht zu schnell gegen 0 konvergiert.

Theorem 5.3 *Wenn es eine Folge $\{\delta_n\}$ positiver Zahlen mit $\delta_n \to \infty$ gibt, so dass*
$$\sqrt{n}\, d_K(\Delta_n; F_0) \ge \delta_n \qquad \forall\, n \ge 1\,, \tag{23}$$
dann gilt
$$\lim_{n\to\infty} \inf_{F\in\Delta_n} \mathbb{P}_F\bigl(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\bigr) = 1\,. \tag{24}$$

Beweis

- Sei $\{\delta_n\}$ eine Folge positiver Zahlen mit $\delta_n \to \infty$, für die (23) gilt, und sei $\{F_n\}$ eine beliebige Folge von Verteilungsfunktionen, so dass für jedes $n \ge 1$
$$F_n \in \Delta_n \qquad \text{und somit} \qquad \sqrt{n}\, d_K(F_n; F_0) \ge \delta_n\,, \tag{25}$$
wobei $d_K(F_n; F_0) = \sup_{t\in\mathbb{R}} |F_n(t) - F_0(t)|$.
- Es genügt zu zeigen, dass
$$\lim_{n\to\infty} \mathbb{P}_{F_n}\bigl(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\bigr) = 1\,. \tag{26}$$
  - Aus der Dreiecksungleichung ergibt sich, dass
  $$d_K(F_n, F_0) \le d_K(F_n, \widehat{F}_n) + d_K(\widehat{F}_n, F_0)\,.$$
  - Hieraus und aus (25) folgt, dass
  $$T_n(X_1,\ldots,X_n) \ge \delta_n - \sqrt{n}\, d_K(F_n, \widehat{F}_n)\,.$$
  - Folglich gilt
  $$\mathbb{P}_{F_n}\bigl(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\bigr) \ge \mathbb{P}_{F_n}\bigl(\sqrt{n}\, d_K(F_n, \widehat{F}_n) < \delta_n - s_{n,1-\alpha}\bigr)\,. \tag{27}$$
- Weil $s_{n,1-\alpha} \to \xi_{1-\alpha} < \infty$ und somit $\delta_n - s_{n,1-\alpha} \to \infty$ für $n \to \infty$, ergibt sich die Gültigkeit von (26) aus (22) und (27). $\square$

Beachte

- Die Bedingung (23) ist insbesondere dann erfüllt, wenn $d_K(\Delta_n; F_0) \ge \delta$ für jedes $n \ge 1$ und $\delta > 0$ eine Konstante ist, die nicht von n abhängt.

- Wenn $\sqrt{n}\, d_K(F_n;F_0) \ge \delta_n$ und $\delta_n > s_{n,1-\alpha}$, dann ergibt sich aus (27), wobei die rechte Seite dieser Ungleichung mit Lemma 5.3 (für $F = F_n$) weiter nach unten abgeschätzt wird, dass für jedes $n \ge 1$ die folgende (nicht asymptotische) untere Schranke für die *Macht* des KS–Tests gilt:

$$\mathbb{P}_{F_n}\bigl(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\bigr) \ge 1 - 2\exp\Bigl(-2(\delta_n - s_{n,1-\alpha})^2\Bigr). \tag{28}$$

- Bei vorgegebenem (endlichem) Stichprobenumfang $n < \infty$ kann die „Ablehnungswahrscheinlichkeit" $\mathbb{P}_{F_0}\bigl(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\bigr)$ jedoch kleiner als α sein.

Andererseits kann die (asymptotische) Macht des KS–Tests beliebig klein werden, d.h., beliebig nahe bei α sein, wenn der Kolmogorow–Abstand $d_K(\Delta_n;F_0)$ zwischen der Familie Δ_n der alternativen Verteilungsfunktionen und der (hypothetischen) Verteilungsfunktion F_0 mit wachsendem Stichprobenumfang n hinreichend schnell gegen 0 konvergiert.

Theorem 5.4

- *Sei $\{F_n\}$ eine beliebige Folge von stetigen Verteilungsfunktionen, so dass*

$$\lim_{n\to\infty} \sqrt{n}\, d_K(F_n;F_0) \to 0\,. \tag{29}$$

- *Dann gilt*

$$\limsup_{n\to\infty} \mathbb{P}_{F_n}\bigl(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\bigr) \le \alpha\,. \tag{30}$$

Beweis

- Aus der Dreiecksungleichung ergibt sich, dass

$$\mathbb{P}_{F_n}\bigl(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\bigr) \le \mathbb{P}_{F_n}\bigl(\sqrt{n}\, d_K(\widehat{F}_n;F_n) + \sqrt{n}\, d_K(F_n;F_0) > s_{n,1-\alpha}\bigr)\,.$$

- Hieraus und aus (29) ergibt sich die Gültigkeit von (30), weil die Verteilung von $\sqrt{n}\, d_K(\widehat{F}_n;F_n)$ unter F_n nicht von n abhängt. $\square$

5.2 χ^2–Anpassungstest

Wir diskutieren nun einen asymptotischen Anpassungstest, wobei eine Testgröße betrachtet wird, die bei großem Stichprobenumfang näherungsweise χ^2–verteilt ist. Dabei wird jedoch im allgemeinen nicht die in Abschnitt 5.1 analysierte Hypothese

$$H_0 : P = P_0 \qquad \Bigl(\text{versus} \quad H_1 : P \ne P_0\Bigr) \tag{31}$$

betrachtet, denn wir „vergröbern" das Modell der Zufallsstichprobe $(X_1,\ldots,X_n)$ durch Klassenbildung.

5.2.1 Klassenbildung; Pearson–Statistik

- Für eine (hinreichend große) natürliche Zahl r zerlegen wir den Wertebereich der Zufallsvariablen $X_1,\ldots,X_n$ in r Klassen $(a_1,b_1],\ldots,(a_r,b_r]$ mit

$$-\infty \le a_1 < b_1 = a_2 < b_2 = \ldots = a_r < b_r \le \infty\,.$$

- Anstelle der Stichprobenvariablen $X_1,\ldots,X_n$ betrachten wir die „Klassenstärken" $Z_1,\ldots,Z_r$, wobei

$$Z_j = \#\{i:\ 1 \le i \le n,\ a_j < X_i \le b_j\} \qquad \forall\, j = 1,\ldots,r\,. \tag{32}$$

Wir zeigen zunächst, dass der Zufallsvektor $(Z_1,\ldots,Z_r)$ multinomialverteilt ist mit den Parametern $n\ge 1$ und

$$\mathbf{p}=(p_1,\ldots,p_{r-1})^\top\in[0,1]^{r-1}\,,\qquad \text{wobei } p_j=\mathbb{P}(a_j<X_1\le b_j)\qquad \forall\, j=1,\ldots,r-1\,.$$

Lemma 5.4 *Für beliebige natürliche Zahlen $k_1,\ldots,k_r\ge 0$ mit $k_1+\ldots+k_r=n$ gilt*

$$\mathbb{P}(Z_1=k_1,\ldots,Z_r=k_r)=\frac{n!}{k_1!\cdot\ldots\cdot k_r!}\,p_1^{k_1}\ldots p_r^{k_r}\,, \tag{33}$$

wobei $p_r=1-(p_1+\ldots+p_{r-1})$.

Beweis

- Weil die Zufallsvariablen $X_1,\ldots,X_n$ unabhängig und identisch verteilt sind, gilt

$$\mathbb{P}\bigl(X_1\in(a_{i_1},b_{i_1}],\ldots,X_n\in(a_{i_n},b_{i_n}]\bigr)=\prod_{j=1}^n\mathbb{P}(a_{i_j}<X_1\le b_{i_j})=p_1^{k_1}\ldots p_r^{k_r} \tag{34}$$

  für jede Folge von Intervallen $(a_{i_1},b_{i_1}],\ldots,(a_{i_n},b_{i_n}]$, die k_1-mal das Intervall $(a_1,b_1],\ldots,$ k_r-mal das Intervall $(a_r,b_r]$ enthält.
- Die Behauptung (33) ergibt sich nun durch Summation der in (34) betrachteten Wahrscheinlichkeiten über sämtliche Permutationen von Folgen $(a_{i_1},b_{i_1}],\ldots,(a_{i_n},b_{i_n}]$ dieser Art. $\square$

Beachte

- Die Multinomialverteilung mit den Parametern $n\ge 1$ und $\mathbf{p}=(p_1,\ldots,p_{r-1})^\top\in[0,1]^{r-1}$ bezeichnen wir mit $\mathrm{M}_{r-1}(n,\mathbf{p})$. Man kann sich leicht überlegen, dass für $r=2$ die Multinomialverteilung $\mathrm{M}_1(n,p_1)$ mit der Binomialverteilung $\mathrm{Bin}(n,p_1)$ übereinstimmt.
- Anstelle das Testproblem (31) zu untersuchen, prüfen wir die Hypothese $H_0:\mathbf{p}=\mathbf{p}_0$ (gegen die Alternative $H_1:\mathbf{p}\ne\mathbf{p}_0$) für einen vorgegebenen (hypothetischen) Parametervektor

$$\mathbf{p}_0=(p_{0,1},\ldots,p_{0,r-1})^\top\in(0,1)^{r-1}\qquad\text{mit}\quad \sum_{i=1}^{r-1}p_{0,i}<1.$$

  - Wir zerlegen also die Familie Δ der insgesamt in Betracht gezogenen Verteilungen der Stichprobenvariablen $X_1,\ldots,X_n$ in die Teilmengen

$$\Delta_0=\{P:\ \mathbb{P}_P(a_j<X_1\le b_j)=p_{0,j}\ \forall j=1,\ldots,r-1\}\qquad\text{bzw.}\qquad \Delta_1=\Delta\setminus\Delta_0\,. \tag{35}$$

  - Dabei betrachten wir die Stichprobenfunktion $T_n:\mathbb{R}^n\to[0,\infty)$ mit

$$T_n(x_1,\ldots,x_n)=\sum_{j=1}^r\frac{\bigl(Z_j(x_1,\ldots,x_n)-np_{0,j}\bigr)^2}{np_{0,j}}\,, \tag{36}$$

  - wobei $Z_j(x_1,\ldots,x_n)$ die Anzahl derjenigen Stichprobenwerte $x_1,\ldots,x_n$ bezeichnet, die im Intervall $(a_j,b_j]$ liegen.
- Unter $H_0:\mathbf{p}=\mathbf{p}_0$ gilt $\mathbb{E}\,Z_j(X_1,\ldots,X_n)=np_{0,j}$ für jedes $j\in\{1,\ldots,r\}$.
  - Es ist deshalb sinnvoll, die Hypothese $H_0:\mathbf{p}=\mathbf{p}_0$ abzulehnen, wenn $T_n(x_1,\ldots,x_n)$ signifikant größer als 0 ist.
  - Um dies entscheiden zu können, benötigen wir Kenntnisse über die Verteilung der in (36) eingeführten Testgröße $T_n(X_1,\ldots,X_n)$, die *Pearson–Statistik* genannt wird.

5.2.2 Asymptotische Verteilung

Wir zeigen, dass $T_n(X_1,\ldots,X_n)$ in Verteilung gegen die χ^2–Verteilung mit $r-1$ Freiheitsgraden strebt, wenn $n\to\infty$. Dies ist die Grundlage des χ^2-*Anpassungstests*, der von Karl Pearson (1857–1936) eingeführt worden ist.

Theorem 5.5 *Für jedes $P\in\Delta_0$ gilt*

$$\lim_{n\to\infty}\mathbb{P}_P\bigl(T_n(X_1,\ldots,X_n)>\chi^2_{r-1,1-\alpha}\bigr)=\alpha\,,\qquad\forall\,\alpha\in(0,1)\,,\tag{37}$$

wobei $\chi^2_{r-1,1-\alpha}$ das $(1-\alpha)$–Quantil der χ^2–Verteilung mit $r-1$ Freiheitsgraden bezeichnet.

Beweis

- In Lemma 5.4 hatten wir gezeigt, dass der in (32) gegebene Zufallsvektor $\mathbf{Z}_n=(Z_{n1},\ldots,Z_{nr})^\top$, wobei $Z_{nj}=Z_j(X_1,\ldots,X_n)$, multinomialverteilt ist unter $P\in\Delta_0$ mit den Parametern $n\in\mathbb{N}$ und

$$\mathbf{p}_0=(p_{01},\ldots,p_{0,r-1})^\top\in[0,1]^{r-1}\,,\qquad\text{wobei }\;p_{0j}=\mathbb{P}_P(a_j<X_1\le b_j)\quad\forall\,j=1,\ldots,r-1\,.$$

  - Hieraus folgt insbesondere, dass für beliebige $i,j\in\{1,\ldots,r\}$

$$\mathbb{E}\,Z_{ni}=np_{0i}\,,\qquad\operatorname{Cov}(Z_{ni},Z_{nj})=\begin{cases}-np_{0i}p_{0j}\,, & \text{wenn } i\ne j,\\ np_{0i}(1-p_{0i})\,, & \text{wenn } i=j.\end{cases}\tag{38}$$

  - Außerdem ergibt sich aus (32), dass $Z_j=\sum_{i=1}^n \mathbf{1}_{\{a_j<X_i\le b_j\}}$, wobei $\mathbf{1}_{\{a_j<X_i\le b_j\}}$ der Indikator des Ereignisses $\{a_j<X_i\le b_j\}$ ist, d.h., $\mathbf{Z}_n$ lässt sich als Summe von n unabhängigen und identisch verteilten Zufallsvektoren darstellen.

- Mit der Schreibweise

$$\mathbf{Z}'_n=\Bigl(\frac{Z_{n1}}{\sqrt{n}}-\sqrt{n}p_{01},\;\ldots,\;\frac{Z_{n,r-1}}{\sqrt{n}}-\sqrt{n}p_{0,r-1}\Bigr)^\top\tag{39}$$

  ergibt sich somit aus Lemma 5.2, dass für $n\to\infty$

$$\mathbf{Z}'_n\xrightarrow{\mathrm{d}}\mathbf{Z}'\sim\mathrm{N}(\mathbf{o},\mathbf{K})\,,\tag{40}$$

  - wobei der $(r-1)$–dimensionale Zufallsvektor $\mathbf{Z}'$ eine (reguläre) multivariate Normalverteilung hat,
  - deren Kovarianzmatrix $\mathbf{K}=(\sigma^2_{ij})$ gegeben ist durch

$$\sigma^2_{ij}=\begin{cases}-p_{0i}p_{0j}\,, & \text{wenn } i\ne j,\\ p_{0i}(1-p_{0i})\,, & \text{wenn } i=j.\end{cases}\tag{41}$$

- Man kann sich leicht überlegen, dass $\mathbf{K}$ invertierbar ist, wobei die Eintragungen a_{ij} der inversen Matrix $\mathbf{A}=\mathbf{K}^{-1}$ gegeben sind durch

$$a_{ij}=\begin{cases}\dfrac{1}{p_{0r}}\,, & \text{wenn } i\ne j,\\[2mm] \dfrac{1}{p_{0i}}+\dfrac{1}{p_{0r}}\,, & \text{wenn } i=j.\end{cases}\tag{42}$$

  - Aus (40) und aus den Eigenschaften von Lineartransformationen normalverteilter Zufallsvektoren (vgl. Theorem 1.3) ergibt sich nun mit Hilfe von Lemma 4.5, dass $\mathbf{A}^{1/2}\mathbf{Z}'_n\xrightarrow{\mathrm{d}}\mathrm{N}(\mathbf{o},\mathbf{I}_{r-1})$, wobei $\mathbf{I}_{r-1}$ die $(r-1)\times(r-1)$–dimensionale Einheitsmatrix ist.
  - Die erneute Anwendung von Lemma 4.5 ergibt somit, dass

$$\bigl(\mathbf{A}^{1/2}\mathbf{Z}'_n\bigr)^\top\bigl(\mathbf{A}^{1/2}\mathbf{Z}'_n\bigr)\xrightarrow{\mathrm{d}}\chi^2_{r-1}\,.$$

- Es genügt nun zu beachten, dass

$$\left(\mathbf{A}^{1/2}\mathbf{Z}'_n\right)^\top\left(\mathbf{A}^{1/2}\mathbf{Z}'_n\right) = T_n(X_1,\ldots,X_n)\,.$$

  - Es gilt nämlich

$$\begin{aligned}\left(\mathbf{A}^{1/2}\mathbf{Z}'_n\right)^\top\left(\mathbf{A}^{1/2}\mathbf{Z}'_n\right) &= \left(\mathbf{Z}'_n\right)^\top\mathbf{A}\mathbf{Z}'_n\\ &= n\sum_{j=1}^{r-1}\frac{1}{p_{0j}}\left(\frac{Z_{nj}}{n}-p_{0j}\right)^2 + \frac{n}{p_{0r}}\sum_{i=1}^{r-1}\sum_{j=1}^{r-1}\left(\frac{Z_{ni}}{n}-p_{0i}\right)\left(\frac{Z_{nj}}{n}-p_{0j}\right),\end{aligned}$$

  - wobei sich der zweite Summand des letzten Ausdruckes schreiben lässt in der Form

$$\frac{n}{p_{0r}}\sum_{i=1}^{r-1}\sum_{j=1}^{r-1}\left(\frac{Z_{ni}}{n}-p_{0i}\right)\left(\frac{Z_{nj}}{n}-p_{0j}\right) = \frac{n}{p_{0r}}\left(\sum_{j=1}^{r-1}\left(\frac{Z_{nj}}{n}-p_{0j}\right)\right)^2 = \frac{n}{p_{0r}}\left(\frac{Z_{nr}}{n}-p_{0r}\right)^2,$$

  - denn offenbar gilt $\sum_{j=1}^{r-1} Z_{nj} = n - Z_{nr}$ und $\sum_{j=1}^{r-1} p_{0j} = 1 - p_{0r}$. □

Beachte

- Bei der praktischen Durchführung des χ^2-*Anpassungstests* zur Prüfung der Hypothese $H_0 : \mathbf{p} = \mathbf{p}_0$ ist zunächst der Wert der in (36) definierten Testgröße $T_n(x_1,\ldots,x_n)$ zu berechnen.
  - Bei hinreichend großem Stichprobenumfang n wird die Hypothese $H_0 : \mathbf{p} = \mathbf{p}_0$ abgelehnt, wenn

$$T_n(x_1,\ldots,x_n) > \chi^2_{r-1,1-\alpha}\,,$$

  - wobei $\chi^2_{r-1,1-\alpha}$ das $(1-\alpha)$–Quantil der χ^2–Verteilung mit $(r-1)$-Freiheitsgraden bezeichnet.
- Eine „Faustregel" dafür, dass n hinreichend groß ist, ist die Gültigkeit der Ungleichung $np_{0,j} \ge a$ für jedes $j \in \{1,\ldots,r\}$ und für eine Konstante $a > 0$.
  - Über die erforderliche Größe von $a > 0$ gibt es unterschiedliche Auffassungen in der Literatur, die von $a = 2$ bis $a = 5$ reichen. Manche Autoren fordern sogar, dass $a = 10$.
  - Andere Autoren meinen, dass bei einer großen Zahl von Klassen (etwa $r \ge 10$) auch schon für $a = 1$ die Approximation hinreichend gut ist.

5.2.3 Güteeigenschaften; lokale Alternativen

Es ist nicht schwierig, die folgende (punktweise) Konsistenz des χ^2—Anpassungstests zu zeigen.

Theorem 5.6 *Der χ^2—Anpassungstest ist punktweise konsistent gegen jeden Vektor $\mathbf{p} = (p_1,\ldots,p_{r-1})^\top \in [0,1]^{r-1}$ mit $\mathbf{p} \ne \mathbf{p}_0$, d.h., es gilt*

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}}\left(T_n(X_1,\ldots,X_n) > \chi^2_{r-1,1-\alpha}\right) = 1\,. \tag{43}$$

Beweis

- Aus $\mathbf{p} \ne \mathbf{p}_0$ folgt, dass für ein $j \in \{1,\ldots,r-1\}$

$$p_j \ne p_{0,j}\,. \tag{44}$$

- Außerdem ergibt sich aus dem starken Gesetz der großen Zahlen (vgl. Theorem WR–5.15), dass $Z_{nj}/n \xrightarrow{\text{f.s.}} p_j$ für $n \to \infty$ unter $\mathbb{P}_{\mathbf{p}}$.

- Hieraus und aus (44) folgt, dass unter $\mathbb{P}_{\mathbf{p}}$

$$T_n(X_1,\ldots,X_n) \ge n\Big(\frac{Z_{nj}}{n} - p_{0,j}\Big)^2 \xrightarrow{\text{f.s.}} \infty\,.$$

- Damit ist (43) bewiesen. □

Beachte

- Anstelle eines fest vorgegebenen Vektors $\mathbf{p} \neq \mathbf{p}_0$ sind auch lokale Alternativen $\mathbf{p}_n = (p_{n1},\ldots,p_{n,r-1})^\top$ der Form
$$p_{nj} = p_{0j} + \frac{h_j}{\sqrt{n}} \qquad \forall\, j = 1,\ldots,r-1 \tag{45}$$
denkbar, die vom Stichprobenumfang n abhängen können, wobei
$$\sum_{j=1}^{r} h_j = 0\,. \tag{46}$$

- Dann kann man zeigen, dass für $n \to \infty$ die asymptotische Macht des χ^2—Anpassungstests gegen solche Alternativen kleiner als 1 sein kann.

Um diese Behauptung zu beweisen, benötigen wir als Hilfsmittel die folgende Abschätzung, die in der Literatur die *Ungleichung von Berry–Esséen* genannt wird.

Lemma 5.5 *Sei $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen mit $\mathbb{E}\left(|Y_1|^3\right) < \infty$. Wenn $\mathbb{E}\,Y_1 = 0$ und $\operatorname{Var} Y_1 = 1$, dann gilt für jedes $n \ge 1$*

$$\sup_{x\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{Y_1+\ldots+Y_n}{\sqrt{n}} \le x\Big) - \Phi(x)\Big| \le C\,\frac{\mathbb{E}\left(|Y_1|^3\right)}{\sqrt{n}}\,, \tag{47}$$

wobei $\Phi : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der $N(0,1)$–Verteilung bezeichnet und $C < \infty$ eine universelle Konstante ist, die nicht von der Verteilung der Zufallsvariablen $Y_1, Y_2, \ldots$ abhängt.

Theorem 5.7 *Sei $\{\mathbf{p}_n\}$ eine Folge von Vektoren, die durch (45) und (46) gegeben sind.*

- *Dann gilt für jedes $x \ge 0$*
$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\big(T_n(X_1,\ldots,X_n) \le x\big) = F_{r-1,\lambda}(x)\,, \tag{48}$$
wobei $F_{r-1,\lambda} : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der nichtzentralen χ^2–Verteilung mit $r-1$ Freiheitsgraden ist, deren Nichtzentralitätsparameter λ gegeben ist durch
$$\lambda = \sum_{j=1}^{r} \frac{h_j^2}{p_{0j}}\,. \tag{49}$$

- *Wenn $h_j \neq 0$ für ein $j = 1,\ldots,r$, dann konvergiert die Macht des χ^2—Anpassungstests bei Betrachtung der lokalen Alternativen $\{\mathbf{p}_n\}$ gegen einen Grenzwert, der größer als α und kleiner als 1 ist, d.h.,*
$$\alpha < \lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\big(T_n(X_1,\ldots,X_n) > \chi^2_{r-1,1-\alpha}\big) < 1\,. \tag{50}$$

Beweis

- Der Beweis der ersten Teilaussage verläuft analog zum Beweis von Theorem 5.5, denn für den in (39) eingeführten Zufallsvektor $\mathbf{Z}'_n$ gilt wegen (45) und (46), dass

$$\mathbf{Z}'_n = \sqrt{n}\Bigl(\frac{Z_{n1}}{n} - p_{n1}, \ldots, \frac{Z_{n,r-1}}{n} - p_{n,r-1}\Bigr)^\top + \mathbf{h}, \qquad \text{wobei} \quad \mathbf{h} = (h_1, \ldots, h_{r-1})^\top. \tag{51}$$

- Wegen (51) kann man so wie beim Beweis des multivariaten zentralen Grenzwertsatzes in Lemma 5.2 zeigen, dass

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\bigl(\mathbf{Z}'_n \le \mathbf{x}\bigr) = F_{\mathrm{N}(\mathbf{h},\mathbf{K})}(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbb{R}^{r-1}. \tag{52}$$

- Dabei genügt es lediglich zu beachten, dass man mit der Ungleichung von Berry–Esséen in Lemma 5.5 zeigen kann, dass (in Analogie zu Formel (10) im Beweis von Lemma 5.2)
  - für beliebige $\mathbf{t} = (t_1, \ldots, t_{r-1})^\top \in \mathbb{R}^{r-1}$ und $x \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\Bigl(n^{-1/2}\sum_{j=1}^{r-1} t_j\bigl(Z_{nj} - np_{nj}\bigr) \le x\Bigr) = F_{\mathrm{N}(0,\mathbf{t}^\top\mathbf{K}\mathbf{t})}(x),$$

  - wobei $\mathbf{K}$ die in (41) eingeführte Kovarianzmatrix ist.
- Genauso wie im Beweis von Theorem 5.5 erhalten wir nun aus (52), dass

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\bigl(\mathbf{A}^{1/2}\mathbf{Z}'_n \le \mathbf{x}\bigr) = F_{\mathrm{N}(\mathbf{A}^{1/2}\mathbf{h},\mathbf{I}_{r-1})}(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbb{R}^{r-1}.$$

  - Hieraus und aus der Definition der nichtzentralen χ^2–Verteilung in Abschnitt 1.3.2 ergibt sich, dass

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\bigl((\mathbf{A}^{1/2}\mathbf{Z}'_n)^\top(\mathbf{A}^{1/2}\mathbf{Z}'_n) \le x\bigr) = F_{r-1,\lambda}(x) \qquad \forall\, x \in \mathbb{R},$$

  - wobei $\mathbf{A}$ die inverse Matrix $\mathbf{A} = \mathbf{K}^{-1}$ in (42) ist und der Nichtzentralitätsparameter λ gegeben ist durch

$$\lambda = (\mathbf{A}^{1/2}\mathbf{h})^\top(\mathbf{A}^{1/2}\mathbf{h}) = \mathbf{h}^\top\mathbf{A}\mathbf{h} = \sum_{j=1}^{r} \frac{h_j^2}{p_{0j}}.$$

- Damit ist (48) bewiesen, und wegen $\alpha < 1 - F_{r-1,\lambda}(\chi^2_{r-1,1-\alpha}) < 1$ ergibt sich hieraus auch die Gültigkeit von (50). □

5.3 χ^2–Anpassungstest von Pearson–Fisher

- Die Nullhypothese $H_0 : \mathbf{p} = \mathbf{p}_0$, die in Abschnitt 5.2 betrachtet wurde, ist in Wirklichkeit eine zusammengesetzte Hypothese, denn sie ist äquivalent mit der Hypothese

$$H_0 : \; P \in \Delta_0,$$

  wobei Δ_0 die in (35) eingeführte Teilmenge von Verteilungen der Stichprobenvariablen ist.

- Wenn geprüft werden soll, ob die Verteilung P der unabhängigen und identisch verteilten Stichprobenvariablen $X_1, \ldots, X_n$ zu einer vorgegebenen (parametrischen) *Klasse von Verteilungen* $\{P_{\boldsymbol\theta}, \boldsymbol\theta \in \Theta\}$ gehört mit $\Theta \subset \mathbb{R}^m$, dann kann ähnlich wie bei dem in Abschnitt 5.2 diskutieren χ^2-Anpassungstest vorgegangen werden.

- Die Stichprobenfunktion $T_n : \mathbb{R}^n \to [0,\infty)$, die bei der Definition der Pearson–Statistik $T_n(X_1, \ldots, X_n)$ in (36) betrachtet wurde, wird dabei durch eine modifizierte Stichprobenfunktion $\widehat{T}_n : \mathbb{R}^n \to [0,\infty)$ ersetzt.

5.3.1 Pearson–Fisher–Teststatistik

- So wie in Abschnitt 5.2.1 „vergröbern" wir das Modell, d.h.,
  - wir zerlegen den Wertebereich der Stichprobenvariablen $X_1,\ldots,X_n$ in r Klassen $(a_1,b_1],\ldots,(a_r,b_r]$ mit $-\infty\le a_1<b_1=a_2<b_2=\ldots=a_r<b_r\le\infty$, wobei r eine (hinreichend große) natürliche Zahl ist.
  - Anstelle der Stichprobenvariablen $X_1,\ldots,X_n$ betrachten wir erneut die „Klassenstärken" $Z_1,\ldots,Z_r$, die bereits in (32) eingeführt wurden, wobei

$$Z_j=\#\{i:\,1\le i\le n,\,a_j<X_i\le b_j\}\qquad\forall\,j=1,\ldots,r\,.$$

- Gemäß Lemma 5.4 gilt dann $(Z_1,\ldots,Z_r)\sim\mathrm{M}_{r-1}(n,\mathbf{p})$, wobei wir jetzt annehmen,
  - dass der Parameter $\mathbf{p}=(p_1,\ldots,p_{r-1})^\top\in[0,1]^{r-1}$ der Multinomialverteilung $M_{r-1}(n,\mathbf{p})$
  - eine (bekannte) Funktion $\boldsymbol{\theta}\mapsto\mathbf{p}(\boldsymbol{\theta})$ des (unbekannten) Parametervektors $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)^\top\in\Theta\subset\mathbb{R}^m$ mit $m<r-1$ ist.
- Getestet werden soll die Hypothese $H_0:\mathbf{p}\in\{\mathbf{p}(\boldsymbol{\theta}),\,\boldsymbol{\theta}\in\Theta\}$.
  - Um bei der Verifizierung dieser Hypothese ähnlich wie in Abschnitt 5.2 vorgehen zu können, muss zunächst ein Schätzer $\widehat{\boldsymbol{\theta}}=(\widehat{\theta}_1,\ldots,\widehat{\theta}_m)^\top$ für $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)^\top$ bestimmt werden.
  - Damit ist auch gleichzeitig ein Schätzer $(\widehat{p}_1,\ldots,\widehat{p}_r)=(p_1(\widehat{\boldsymbol{\theta}}),\ldots,p_r(\widehat{\boldsymbol{\theta}}))$ für die Wahrscheinlichkeiten $(p_1,\ldots,p_r)=(p_1(\boldsymbol{\theta}),\ldots,p_r(\boldsymbol{\theta}))$ gegeben, wobei

$$p_j(\boldsymbol{\theta})=\mathbb{P}_{\boldsymbol{\theta}}(a_j<X_1\le b_j)\qquad\forall\,j=1,\ldots,r\,.$$

Definition Die Zufallsvariable $\widehat{T}_n(X_1,\ldots,X_n)$, die durch die Stichprobenfunktion $\widehat{T}_n:\mathbb{R}^n\to[0,\infty)$ mit

$$\widehat{T}_n(x_1,\ldots,x_n)=\sum_{j=1}^r\frac{\bigl(Z_j(x_1,\ldots,x_n)-n\widehat{p}_j(x_1,\ldots,x_n)\bigr)^2}{n\widehat{p}_j(x_1,\ldots,x_n)}\tag{53}$$

gegeben ist, heißt *Pearson–Fisher–Statistik* .

Beachte

- Wenn die Abbildung $\boldsymbol{\theta}\mapsto\mathbf{p}(\boldsymbol{\theta})$ stetig und $\widehat{\boldsymbol{\theta}}$ ein (schwach) konsistenter Schätzer für $\boldsymbol{\theta}$ ist,
  - dann ergibt sich aus dem Gesetz der großen Zahlen (vgl. Theorem WR–5.15), dass für beliebige $j\in\{1,\ldots,r\}$ und $\boldsymbol{\theta}\in\Theta$

$$\lim_{n\to\infty}\mathbb{E}_{\boldsymbol{\theta}}\Bigl|\,\frac{1}{n}\,Z_j(X_1,\ldots,X_n)-\widehat{p}_j(X_1,\ldots,X_n)\Bigr|=0\,.$$

  - Es ist deshalb sinnvoll, die Nullhypothese $H_0:\mathbf{p}\in\{\mathbf{p}(\boldsymbol{\theta}),\,\boldsymbol{\theta}\in\Theta\}$ abzulehnen, wenn $\widehat{T}_n(x_1,\ldots,x_n)$ signifikant größer als 0 ist.
- Um dies entscheiden zu können,
  - diskutieren wir zunächst Bedingungen an die Abbildung $\boldsymbol{\theta}\mapsto\mathbf{p}(\boldsymbol{\theta})$, die die Konstruktion einer Folge von konsistenten (Maximum–Likelihood–) Schätzern $\widehat{\boldsymbol{\theta}}_n$ für $\boldsymbol{\theta}$ ermöglichen, die asymptotisch normalverteilt sind,
  - und bestimmen danach die (asymptotische Grenz–) Verteilung der in (53) eingeführten Testgröße $\widehat{T}_n(X_1,\ldots,X_n)$ für $n\to\infty$.

5.3.2 Multivariater zentraler Grenzwertsatz für ML–Schätzer

Ähnlich wie in Abschnitt I–2.4.2, wo der Fall $m = 1$ betrachtet wurde, lässt sich ein *multivariater zentraler Grenzwertsatz* für konsistente Folgen von Maximum–Likelihood–Schätzern des Parametervektors $\boldsymbol{\theta}$ herleiten.

Dabei werden die folgenden *Regularitätsbedingungen* benötigt.

- Die Familie $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ bestehe entweder nur aus diskreten Verteilungen oder nur aus absolutstetigen Verteilungen, wobei $\Theta \subset \mathbb{R}^m$ eine offene Menge sei.
- Es gelte
$$P_{\boldsymbol{\theta}} \neq P_{\boldsymbol{\theta}'} \qquad \text{genau dann, wenn} \qquad \boldsymbol{\theta} \neq \boldsymbol{\theta}'\,.$$
- Die Menge $B = \{x \in \mathbb{R} : L(x; \boldsymbol{\theta}) > 0\}$ hänge nicht von $\boldsymbol{\theta} \in \Theta$ ab, wobei die Likelihood–Funktion $L(x; \boldsymbol{\theta})$ gegeben ist durch
$$L(x;\boldsymbol{\theta}) = \begin{cases} p(x;\boldsymbol{\theta}) & \text{im diskreten Fall,} \\ f(x;\boldsymbol{\theta}) & \text{im absolutstetigen Fall} \end{cases}$$
und $p(x; \boldsymbol{\theta})$ bzw. $f(x; \boldsymbol{\theta})$ die Wahrscheinlichkeitsfunktion bzw. Dichte von $P_{\boldsymbol{\theta}}$ ist.
- Außerdem sei die Abbildung $\boldsymbol{\theta} \to L(x; \boldsymbol{\theta})$ für jedes $x \in B$ dreimal stetig differenzierbar, und für jedes $x \in B$ gelte
$$\frac{\partial^k}{\partial\theta_{i_1}\ldots\partial\theta_{i_k}} \int\limits_B L(x;\boldsymbol{\theta})\,dx = \int\limits_B \frac{\partial^k}{\partial\theta_{i_1}\ldots\partial\theta_{i_k}} L(x;\boldsymbol{\theta})\,dx \qquad \forall\, k \in \{1,2,3\},\ i_1,\ldots,i_k \in \{1,\ldots,m\},\ \boldsymbol{\theta} \in \Theta, \tag{54}$$
wobei die Integrale im diskreten Fall durch die entsprechenden Summen zu ersetzen sind.
- Für jedes $\boldsymbol{\theta}_0 \in \Theta$ gebe es eine Konstante $c_{\boldsymbol{\theta}_0} > 0$ und eine messbare Funktion $g_{\boldsymbol{\theta}_0} : B \to [0,\infty)$, so dass für jedes Tripel $(i_1, i_2, i_3) \in \{1, \ldots, m\}^3$
$$\Bigl|\frac{\partial^3}{\partial\theta_{i_1}\partial\theta_{i_2}\partial\theta_{i_3}} \log L(x;\boldsymbol{\theta})\Bigr| \le g_{\boldsymbol{\theta}_0}(x) \qquad \forall\, x \in B,\quad \forall\, \boldsymbol{\theta} \in \Theta \text{ mit } |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < c_{\boldsymbol{\theta}_0} \tag{55}$$
und
$$\mathbb{E}_{\boldsymbol{\theta}_0} g_{\boldsymbol{\theta}_0}(X_1) < \infty\,. \tag{56}$$

Beachte

- *Zur Erinnerung* :
  - Im Allgemeinen wird der Maximum–Likelihood–Schätzer $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$ für $\boldsymbol{\theta}$ als Lösung des folgenden Optimierungsproblems definiert (vgl. Abschnitt I–2.2.2).
  - Dabei ist $\widehat{\boldsymbol{\theta}} : \mathbb{R}^n \to \Theta \subset \mathbb{R}^m$ eine Stichprobenfunktion mit
$$L(x_1,\ldots,x_n;\boldsymbol{\theta}) \le L(x_1,\ldots,x_n;\widehat{\boldsymbol{\theta}}(x_1,\ldots,x_n)) \qquad \forall\,(x_1,\ldots,x_n) \in \mathbb{R}^n,\ \boldsymbol{\theta} \in \Theta \tag{57}$$
  und
$$L(x_1,\ldots,x_n;\boldsymbol{\theta}) = \begin{cases} p(x_1;\boldsymbol{\theta})\ldots p(x_n;\boldsymbol{\theta}) & \text{im diskreten Fall,} \\ f(x_1;\boldsymbol{\theta})\ldots f(x_n;\boldsymbol{\theta}) & \text{im absolutstetigen Fall.} \end{cases}$$
- Unter den obengenannten Regularitätsbedingungen kann man zeigen, dass $\widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n)$ für beliebige $x_1, \ldots, x_n \in \mathbb{R}$ dem folgenden Gleichungssystem genügt:
$$\frac{\partial}{\partial\theta_i} L\bigl(x_1,\ldots,x_n;\widehat{\boldsymbol{\theta}}(x_1,\ldots,x_n)\bigr) = 0 \qquad \forall\, i = 1,\ldots,m\,. \tag{58}$$

- Um den multivariaten zentralen Grenzwertsatz formulieren zu können, benötigen wir den Begriff der *Fischer-Informationsmatrix*, der bereits in Abschnitt 4.3.1 eingeführt wurde.
  - Für jedes $\boldsymbol{\theta} \in \Theta$ wird dabei die $m \times m$ Matrix $\mathbf{I}(\boldsymbol{\theta}) = (I_{ij}(\boldsymbol{\theta}))$ betrachtet mit

$$I_{ij}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\Bigl(\frac{\partial}{\partial\theta_i}\log L(X_1;\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log L(X_1;\boldsymbol{\theta})\Bigr). \tag{59}$$

  - wobei vorausgesetzt wird, dass der Erwartungswert in (59) für beliebige $i, j \in \{1, \ldots, m\}$ existiert (und eine endliche reelle Zahl ist).

In Verallgemeinerung von Theorem I–2.11, wo der 1–dimensionale Fall betrachtet wurde, lässt sich für schwach konsistente Folgen von Maximum–Likelihood–Schätzern $\{\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n),\ n \ge 1\}$ des Parametervektors $\boldsymbol{\theta}$, die dem Gleichungssystem (58) genügen, der folgende multivariate zentrale Grenzwertsatz herleiten.

Theorem 5.8

- *Die Fisher–Informationsmatrix $\mathbf{I}(\boldsymbol{\theta})$ sei für jedes $\boldsymbol{\theta} \in \Theta$ positiv definit (und damit invertierbar), und sei $\{\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n),\ n \ge 1\}$ eine schwach konsistente Folge von Maximum–Likelihood–Schätzern für $\boldsymbol{\theta}$.*
- *Dann gilt für $n \to \infty$*

$$\sqrt{n}\bigl(\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n) - \boldsymbol{\theta}\bigr) \xrightarrow{\mathrm{d}} \mathrm{N}\bigl(\mathbf{o}, \mathbf{I}^{-1}(\boldsymbol{\theta})\bigr). \tag{60}$$

Der *Beweis* von Theorem 5.8 verläuft ähnlich wie der Beweis von Theorem I–2.11. Er wird deshalb hier weggelassen, vgl. beispielsweise E.L. Lehmann und G. Casella (1998) *The Theory of Point Estimation*, Springer–Verlag, New York.

5.3.3 Fisher–Informationsmatrix und zentraler Grenzwertsatz im vergröberten Modell

- Wir kehren nun zu dem „vergröberten" Modell zurück, das bereits in Abschnitt 5.3.1 betrachtet wurde.
  - Dabei setzen wir voraus, dass die Likelihood–Funktion $L : \mathbb{R} \times \Theta \to (0, 1)$ mit

$$L(x; \boldsymbol{\theta}) = p_j(\boldsymbol{\theta}), \qquad \text{wenn } x \in (a_j, b_j], \tag{61}$$

  - wobei die Wahrscheinlichkeiten $p_j(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(a_j < X_1 \le b_j)$ positiv und kleiner als 1 seien.
- Über die in (61) gegebene Likelihood–Funktion setzen wir außerdem voraus, dass die in Abschnitt 5.3.2 formulierten Regularitätsbedingungen erfüllt sind.

Lemma 5.6 *Für die Fisher–Informationsmatrix $\mathbf{I}(\boldsymbol{\theta})$ gilt dann*

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})^\top \mathbf{C}(\boldsymbol{\theta}), \tag{62}$$

wobei

$$\mathbf{C}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial p_1(\boldsymbol{\theta})/\partial\theta_1}{\sqrt{p_1(\boldsymbol{\theta})}} & \frac{\partial p_1(\boldsymbol{\theta})/\partial\theta_2}{\sqrt{p_1(\boldsymbol{\theta})}} & \cdots & \frac{\partial p_1(\boldsymbol{\theta})/\partial\theta_m}{\sqrt{p_1(\boldsymbol{\theta})}} \\ \frac{\partial p_2(\boldsymbol{\theta})/\partial\theta_1}{\sqrt{p_2(\boldsymbol{\theta})}} & \frac{\partial p_2(\boldsymbol{\theta})/\partial\theta_2}{\sqrt{p_2(\boldsymbol{\theta})}} & \cdots & \frac{\partial p_2(\boldsymbol{\theta})/\partial\theta_m}{\sqrt{p_2(\boldsymbol{\theta})}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial p_r(\boldsymbol{\theta})/\partial\theta_1}{\sqrt{p_r(\boldsymbol{\theta})}} & \frac{\partial p_r(\boldsymbol{\theta})/\partial\theta_2}{\sqrt{p_r(\boldsymbol{\theta})}} & \cdots & \frac{\partial p_r(\boldsymbol{\theta})/\partial\theta_m}{\sqrt{p_r(\boldsymbol{\theta})}} \end{pmatrix}. \tag{63}$$

Beweis

- Wegen (61) gilt für jedes $x \in \mathbb{R}$
$$\log L(x;\boldsymbol{\theta}) = \sum_{j=1}^{r} \mathbb{I}_{\{a_j < x \le b_j\}} \log p_j(\boldsymbol{\theta})\,.$$
- Hieraus ergibt sich für die Eintragungen $I_{ij}(\boldsymbol{\theta})$ von $\mathbf{I}(\boldsymbol{\theta})$, dass
$$\begin{aligned} I_{ij}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}\Bigl(\frac{\partial}{\partial\theta_i}\log L(X_1;\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log L(X_1;\boldsymbol{\theta})\Bigr) = \sum_{k=1}^{r}\Bigl(\frac{\partial}{\partial\theta_i}\log p_k(\boldsymbol{\theta})\Bigr)\Bigl(\frac{\partial}{\partial\theta_j}\log p_k(\boldsymbol{\theta})\Bigr)p_k(\boldsymbol{\theta}) \\ &= \sum_{k=1}^{r}\Bigl(\frac{\partial}{\partial\theta_i}p_k(\boldsymbol{\theta})\Bigr)\bigl(p_k(\boldsymbol{\theta})\bigr)^{-1}\Bigl(\frac{\partial}{\partial\theta_j}p_k(\boldsymbol{\theta})\Bigr) = \Bigl(\mathbf{C}(\boldsymbol{\theta})^\top\mathbf{C}(\boldsymbol{\theta})\Bigr)_{ij}\,. \end{aligned}$$ □

Aus Theorem 5.8 ergibt sich somit das folgende Resultat.

Korollar 5.1 *Wenn die in* (63) *gegebene Matrix* $\mathbf{I}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})^\top\mathbf{C}(\boldsymbol{\theta})$ *für jedes* $\boldsymbol{\theta} \in \Theta$ *positiv definit ist, dann gilt*
$$\sqrt{n}\bigl(\widehat{\boldsymbol{\theta}}(X_1,\ldots,X_n) - \boldsymbol{\theta}\bigr) \xrightarrow{\mathrm{d}} \mathrm{N}\Bigl(\mathbf{o}, \bigl(\mathbf{C}(\boldsymbol{\theta})^\top\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\Bigr) \tag{64}$$
für jede schwach konsistente Folge $\{\widehat{\boldsymbol{\theta}}(X_1,\ldots,X_n),\ n \ge 1\}$ *von Maximum–Likelihood–Schätzern für* $\boldsymbol{\theta}$*, die durch die Beobachtung des „vergröberten" Modells gewonnen werden.*

Beachte

- Aus (61) ergibt sich für die Likelihood–Funktion $L(x_1,\ldots,x_n;\boldsymbol{\theta})$, dass
$$L(x_1,\ldots,x_n;\boldsymbol{\theta}) = \prod_{j=1}^{r} p_j(\boldsymbol{\theta})^{Z_j(x_1,\ldots,x_n)}\,,$$
bzw. für die Loglikelihood–Funktion $\log L(x_1,\ldots,x_n;\boldsymbol{\theta})$, dass
$$\log L(x_1,\ldots,x_n;\boldsymbol{\theta}) = \sum_{j=1}^{r} Z_j(x_1,\ldots,x_n)\log p_j(\boldsymbol{\theta})\,. \tag{65}$$
- Jede Maximum–Likelihood–Schätzung $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}\bigl(Z_1(x_1,\ldots,x_n),\ldots,Z_r(x_1,\ldots,x_n)\bigr)$ für $\boldsymbol{\theta}$, die aus den vergröberten Daten $Z_1(x_1,\ldots,x_n),\ldots,Z_r(x_1,\ldots,x_n)$ gewonnen wird, genügt wegen der obengenannten Regularitätsbedingungen dem Gleichungssystem
$$\frac{\partial \log L(x_1,\ldots,x_n;\boldsymbol{\theta})}{\partial\theta_i} = 0\,, \qquad \forall\, i = 1,\ldots,m\,. \tag{66}$$
- Dabei ergibt sich aus (65), dass für beliebige $i = 1,\ldots,m$ und $\boldsymbol{\theta} \in \Theta$
$$\frac{\partial \log L(x_1,\ldots,x_n;\boldsymbol{\theta})}{\partial\theta_i} = \sum_{j=1}^{r}\frac{Z_j(x_1,\ldots,x_n)}{p_j(\boldsymbol{\theta})}\,\frac{\partial p_j(\boldsymbol{\theta})}{\partial\theta_i}$$
bzw.
$$\frac{\partial \log L(x_1,\ldots,x_n;\boldsymbol{\theta})}{\partial\theta_i} = \sum_{j=1}^{r}\frac{Z_j(x_1,\ldots,x_n) - np_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})}\,\frac{\partial p_j(\boldsymbol{\theta})}{\partial\theta_i}\,, \tag{67}$$
wobei sich die letzte Gleichheit aus der Tatsache ergibt, dass
$$\sum_{j=1}^{r}\frac{\partial p_j(\boldsymbol{\theta})}{\partial\theta_i} = 0\,, \qquad \forall\, i = 1,\ldots,m\,. \tag{68}$$

5.3.4 Asymptotische Verteilung der Pearson–Fisher–Statistik

Das folgende Theorem ist die Grundlage des χ^2-*Anpassungstests von Pearson-Fisher*. Dabei setzen wir voraus, dass

- die in (61) betrachtete Likelihood–Funktion des vergröberten Modells den Regularitätsbedingungen von Abschnitt 5.3.2 genügt,
- die in (62) gegebene Fisher–Informationsmatrix $I(\boldsymbol{\theta})$ positiv definit ist und dass
- $\{\widehat{\boldsymbol{\theta}}_n\} = \{\widehat{\boldsymbol{\theta}}(X_1,\ldots,X_n),\, n \ge 1\}$ eine schwach konsistente Folge von ML–Schätzern für $\boldsymbol{\theta}$ ist, die durch die Beobachtung des vergröberten Modells gewonnen werden.

Theorem 5.9

- *Sei $\widehat{T}_n(X_1,\ldots,X_n)$ die in* (53) *eingeführte Pearson–Fisher–Teststatistik, d.h.,*
$$\widehat{T}_n(X_1,\ldots,X_n) = \sum_{j=1}^{r} \frac{\bigl(Z_j(X_1,\ldots,X_n) - n\widehat{p}_j(X_1,\ldots,X_n)\bigr)^2}{n\widehat{p}_j(X_1,\ldots,X_n)}\,, \tag{69}$$
wobei $\widehat{p}_j(X_1,\ldots,X_n) = p_j\bigl(\widehat{\boldsymbol{\theta}}(X_1,\ldots,X_n)\bigr)$.
- *Dann gilt*
$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}}\bigl(\widehat{T}_n(X_1,\ldots,X_n) > \chi^2_{r-1-m,1-\alpha}\bigr) = \alpha\,, \qquad \forall\, \alpha \in (0,1) \tag{70}$$
für jedes $\boldsymbol{\theta} \in \Theta$, wobei $\chi^2_{r-1-m,1-\alpha}$ das $(1-\alpha)$–Quantil der χ^2–Verteilung mit $r-1-m$ Freiheitsgraden bezeichnet.

Ein mathematisch strikter Beweis von Theorem 5.9 kann durch Reinterpretation des χ^2–Anpassungstests von Pearson–Fisher als *Likelihood–Quotiententest* geführt werden, vgl. beispielsweise Abschnitt 4.7 in H. Pruscha (2000) *Vorlesungen über mathematische Statistik*, Teubner–Verlag, Stuttgart.

Weil diese Beweistechnik jedoch relativ komplex ist, geben wir hier lediglich eine Herleitung von Theorem 5.9 an, die teilweise *heuristisch* ist.

- Und zwar sei $\mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}),\ldots,p_r(\boldsymbol{\theta}))^\top$ und $\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) = \bigl(\widetilde{Z}_{n1}(\boldsymbol{\theta}),\ldots,\widetilde{Z}_{nr}(\boldsymbol{\theta})\bigr)^\top$ mit
$$\widetilde{Z}_{nj}(\boldsymbol{\theta}) = \frac{Z_j(X_1,\ldots,X_n) - np_j(\boldsymbol{\theta})}{\sqrt{np_j(\boldsymbol{\theta})}}\,, \qquad j = 1,\ldots,r\,. \tag{71}$$
- Weil $\mathbb{E}\,\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) = \mathbf{o}$ und weil sich $\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta})$ als Summe von n unabhängigen und identisch verteilten Zufallsvektoren darstellen lässt, ergibt sich aus dem multivariaten zentralen Grenzwertsatz (genauso wie im Beweis von Theorem 5.5), dass
$$\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) \stackrel{\mathrm{d}}{\longrightarrow} \widetilde{\mathbf{Z}}(\boldsymbol{\theta}) \sim \mathrm{N}(\mathbf{o}, \mathbf{B}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta}))\,, \tag{72}$$
wobei
$$\mathbf{B}(\boldsymbol{\theta}) = \begin{pmatrix} 1/\sqrt{p_1(\boldsymbol{\theta})} & 0 & \ldots & 0 \\ 0 & 1/\sqrt{p_2(\boldsymbol{\theta})} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1/\sqrt{p_r(\boldsymbol{\theta})} \end{pmatrix}, \qquad \mathbf{K}(\boldsymbol{\theta}) = \begin{cases} -p_i(\boldsymbol{\theta})p_j(\boldsymbol{\theta})\,, & \text{wenn } i \ne j, \\[2ex] p_i(\boldsymbol{\theta})(1-p_j(\boldsymbol{\theta}))\,, & \text{wenn } i = j. \end{cases}$$

– Für die Kovarianzmatrix $\mathbf{B}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})$ in (72) gilt somit

$$\mathbf{B}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta}) = \mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta}), \qquad \text{wobei } \mathbf{q}(\boldsymbol{\theta}) = \bigl(\sqrt{p_1(\boldsymbol{\theta})}, \ldots, \sqrt{p_r(\boldsymbol{\theta})}\bigr)^\top . \tag{73}$$

– Für die in (63) eingeführte Matrix $\mathbf{C}(\boldsymbol{\theta})$ gilt wegen (68), dass $\mathbf{q}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta}) = \mathbf{o}$ und somit

$$\begin{aligned}
&\Bigl(\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Bigr)\Bigl(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Bigr)\Bigl(\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Bigr)^\top \\
&= \Bigl(\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Bigr)\Bigl(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Bigr)\mathbf{C}(\boldsymbol{\theta})\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1} \\
&= \bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1} .
\end{aligned}$$

– Aus (72) und (73) ergibt sich nun, dass

$$\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\widetilde{\mathbf{Z}}(\boldsymbol{\theta}) \sim \mathrm{N}\bigl(\mathbf{o}, \bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\bigr) . \tag{74}$$

- Außerdem ergibt sich aus Korollar 5.1 durch Taylor–Reihenentwicklung, dass

$$\sqrt{n}\,\mathbf{B}(\boldsymbol{\theta})\bigl(\mathbf{p}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{p}(\boldsymbol{\theta})\bigr) = \sqrt{n}\,\mathbf{C}(\boldsymbol{\theta})\bigl(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\bigr) + o\bigl(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\bigr) \xrightarrow{\mathrm{d}} \mathrm{N}\bigl(\mathbf{o}, \mathbf{C}(\boldsymbol{\theta})\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\bigr) .$$

– Hieraus und aus (74) folgt, dass

$$\sqrt{n}\,\mathbf{B}(\boldsymbol{\theta})\bigl(\mathbf{p}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{p}(\boldsymbol{\theta})\bigr) \xrightarrow{\mathrm{d}} \mathbf{C}(\boldsymbol{\theta})\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\widetilde{\mathbf{Z}}(\boldsymbol{\theta}) . \tag{75}$$

– Andererseits ergibt sich aus (69) und (71), dass

$$\begin{aligned}
\widehat{T}_n(X_1, \ldots, X_n) &= \sum_{j=1}^r \bigl(\widetilde{Z}_{nj}(\widehat{\boldsymbol{\theta}}_n)\bigr)^2 \\
&= \sum_{j=1}^r \left(\widetilde{Z}_{nj}(\boldsymbol{\theta}) + \widetilde{Z}_{nj}(\boldsymbol{\theta})\Bigl(\sqrt{\frac{p_j(\boldsymbol{\theta})}{p_j(\widehat{\boldsymbol{\theta}}_n)}} - 1\Bigr) - \frac{\sqrt{n}}{\sqrt{p_j(\widehat{\boldsymbol{\theta}}_n)}}\bigl(p_j(\widehat{\boldsymbol{\theta}}_n) - p_j(\boldsymbol{\theta})\bigr)\right)^2 \\
&= \sum_{j=1}^r \left(\widetilde{Z}_{nj}(\boldsymbol{\theta}) - \frac{\sqrt{n}}{\sqrt{p_j(\widehat{\boldsymbol{\theta}}_n)}}\bigl(p_j(\widehat{\boldsymbol{\theta}}_n) - p_j(\boldsymbol{\theta})\bigr) + o(1)\right)^2 ,
\end{aligned}$$

– wobei sich die letzte Gleichheit aus der Null-Konvergenz

$$\widetilde{Z}_{nj}(\boldsymbol{\theta})\Bigl(\sqrt{\frac{p_j(\boldsymbol{\theta})}{p_j(\widehat{\boldsymbol{\theta}}_n)}} - 1\Bigr) \xrightarrow{\mathrm{P}} 0 , \qquad \forall\, j = 1, \ldots, r$$

ergibt, die aus $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\theta}$ und dem Continuous–Mapping–Theorem für Zufallsvektoren (vgl. Lemma 4.5) folgt.

- Mit anderen Worten: Mit der Schreibweise $\varphi(z_1, \ldots, z_r) = \sum_{j=1}^r z_j^2$ gilt

$$\widehat{T}_n(X_1, \ldots, X_n) = \varphi\Bigl(\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) - \sqrt{n}\,\mathbf{B}(\boldsymbol{\theta})\bigl(\mathbf{p}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{p}(\boldsymbol{\theta})\bigr) + o(1)\Bigr) . \tag{76}$$

– Zusammen mit (72) und (75) suggeriert die asymptotische Näherungsformel (76) die *Vermutung*, dass für $n \to \infty$

$$\widehat{T}_n(X_1, \ldots, X_n) \xrightarrow{\mathrm{d}} \varphi\Bigl(\Bigl(\mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\bigl(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\bigr)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Bigr)\widetilde{\mathbf{Z}}(\boldsymbol{\theta})\Bigr) . \tag{77}$$

– Die Verteilungskonvergenz (77) ergibt sich jedoch *nicht direkt* aus (72), (75) und (76), sondern sie erfordert einen *separaten Beweis*, der hier weggelassen wird.

- Wir zeigen nun noch, dass

$$\varphi\Big(\Big(\mathbf{I}_r - \mathbf{C}(\boldsymbol\theta)\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta)\Big)\widetilde{\mathbf{Z}}(\boldsymbol\theta)\Big) \sim \chi^2_{r-1-m}\,. \tag{78}$$

- In (72) und (73) hatten wir gezeigt, dass

$$\widetilde{\mathbf{Z}}(\boldsymbol\theta) \sim \mathrm{N}\big(\mathbf{o}, \mathbf{I}_r - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\big)\,, \qquad \text{wobei } \mathbf{q}(\boldsymbol\theta) = \big(\sqrt{p_1(\boldsymbol\theta)}, \ldots, \sqrt{p_r(\boldsymbol\theta)}\big)^\top.$$

  - Außerdem ergibt sich aus $\mathbf{q}^\top(\boldsymbol\theta)\mathbf{q}(\boldsymbol\theta) = 1$, dass

$$\begin{aligned}\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\Big)^2 &= \mathbf{I}_r - 2\mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta) + \mathbf{q}(\boldsymbol\theta)\underbrace{\mathbf{q}^\top(\boldsymbol\theta)\mathbf{q}(\boldsymbol\theta)}_{=1}\mathbf{q}^\top(\boldsymbol\theta)\\ &= \mathbf{I}_r - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\,,\end{aligned}$$

  - d.h., die Kovarianzmatrix

$$\mathbf{I}_r - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)) = \begin{pmatrix} 1-p_1(\boldsymbol\theta) & -\sqrt{p_1(\boldsymbol\theta)p_2(\boldsymbol\theta)} & -\sqrt{p_1(\boldsymbol\theta)p_3(\boldsymbol\theta)} & \ldots & -\sqrt{p_1(\boldsymbol\theta)p_r(\boldsymbol\theta)} \\ -\sqrt{p_1(\boldsymbol\theta)p_2(\boldsymbol\theta)} & 1-p_2(\boldsymbol\theta) & -\sqrt{p_2(\boldsymbol\theta)p_3(\boldsymbol\theta)} & \ldots & -\sqrt{p_2(\boldsymbol\theta)p_r(\boldsymbol\theta)} \\ \vdots & \vdots & \vdots & & \vdots \\ -\sqrt{p_1(\boldsymbol\theta)p_r(\boldsymbol\theta)} & -\sqrt{p_2(\boldsymbol\theta)p_r(\boldsymbol\theta)} & -\sqrt{p_3(\boldsymbol\theta)p_r(\boldsymbol\theta)} & \ldots & 1-p_r(\boldsymbol\theta) \end{pmatrix}$$

    des Zufallsvektors $\widetilde{\mathbf{Z}}(\boldsymbol\theta)$ ist symmetrisch und idempotent.
  - Hieraus und aus der ersten Teilaussage von Theorem 1.4 ergibt sich nun die Darstellungsformel

$$\widetilde{\mathbf{Z}}(\boldsymbol\theta) \stackrel{\mathrm{d}}{=} \Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\Big)\boldsymbol\varepsilon\,, \tag{79}$$

    wobei $\boldsymbol\varepsilon \sim \mathrm{N}(\mathbf{o}, \mathbf{I}_r)$.

- Außerdem ist auch die Matrix $\mathbf{I}_r - \mathbf{C}(\boldsymbol\theta)\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta)$ symmetrisch und idempotent, und aus $\mathbf{q}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta) = \mathbf{o}$ ergibt sich, dass

$$\Big(\mathbf{I}_r - \mathbf{C}(\boldsymbol\theta)\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta)\Big)\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\Big) = \mathbf{I}_r - \mathbf{C}(\boldsymbol\theta)\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta) - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\,.$$

  - Hieraus folgt, dass die Matrix $\mathbf{R} = \Big(\mathbf{I}_r - \mathbf{C}(\boldsymbol\theta)\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta)\Big)\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\Big)$ ebenfalls symmetrisch und idempotent ist.
  - Aus (79) und aus Theorem 1.9 ergibt sich nun, dass

$$\varphi\Big(\big(\mathbf{I}_r - \mathbf{C}(\boldsymbol\theta)\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta)\big)\widetilde{\mathbf{Z}}(\boldsymbol\theta)\Big) \stackrel{\mathrm{d}}{=} \varphi\big(\mathbf{R}\boldsymbol\varepsilon\big) = \boldsymbol\varepsilon^\top\mathbf{R}\boldsymbol\varepsilon \sim \chi^2_{\mathrm{rg}(\mathbf{R})}\,.$$

  - Mit Hilfe von Lemma 1.3 ergibt sich für den Rang $\mathrm{rg}(\mathbf{R})$ der symmetrischen und idempotenten Matrix $\mathbf{R}$, dass

$$\begin{aligned}\mathrm{rg}(\mathbf{R}) &= \mathrm{sp}(\mathbf{R})\\ &= \mathrm{sp}(\mathbf{I}_r) - \mathrm{sp}\Big(\mathbf{C}(\boldsymbol\theta)\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta)\Big) - \mathrm{sp}\Big(\mathbf{q}(\boldsymbol\theta)\mathbf{q}^\top(\boldsymbol\theta)\Big)\\ &= \mathrm{sp}(\mathbf{I}_r) - \mathrm{sp}\Big(\big(\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\big)^{-1}\mathbf{C}^\top(\boldsymbol\theta)\mathbf{C}(\boldsymbol\theta)\Big) - \mathrm{sp}\Big(\mathbf{q}^\top(\boldsymbol\theta)\mathbf{q}(\boldsymbol\theta)\Big)\\ &= r - m - 1\,.\end{aligned}$$

- Damit ist die Gültigkeit von (78) bewiesen.

Beachte Bei der praktischen Durchführung des χ^2-*Anpassungstests von Pearson–Fisher* kann ähnlich wie in Abschnitt 5.2.2 vorgegangen werden, um die Hypothese $H_0 : P \in \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ zu prüfen.

- Zunächst wird eine ML–Schätzung $\widehat{\boldsymbol{\theta}}(x_1,\ldots,x_n) = (\widehat{\theta}_1(x_1,\ldots,x_n),\ldots,\widehat{\theta}_m(x_1,\ldots,x_n))^\top$ für $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)^\top$ durch Lösung des Gleichungssystems (66) bestimmt.
- Dann wird der Wert der in (53) definierten Testgröße $T_n(x_1,\ldots,x_n)$ berechnet.
- Bei hinreichend großem Stichprobenumfang n wird $H_0 : P \in \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ abgelehnt, wenn

$$T_n(x_1,\ldots,x_n) > \chi^2_{r-1-m,1-\alpha}\,,$$

- wobei $\chi^2_{r-1-m,1-\alpha}$ das $(1-\alpha)$–Quantil der χ^2–Verteilung mit $(r-1-m)$-Freiheitsgraden bezeichnet.

5.4 Beispiele

5.4.1 χ^2–Anpassungstest auf Poisson–Verteilung

- Durch die Beobachtung der (unabhängigen und identisch verteilten) Stichprobenvariablen $X_1,\ldots,X_n$ soll geprüft werden, ob die Verteilung P von X_i zur Familie der Poisson–Verteilungen gehört.
  - Sei also $\Theta = (0,\infty)$ mit $\boldsymbol{\theta} = \lambda$, und sei $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\} = \{\,\mathrm{Poi}(\lambda),\, \lambda > 0\}$ die Familie der Poisson–Verteilungen.
  - Wir betrachten die folgenden r Klassen $\{0\},\{1\},\ldots,\{r-2\}$ und $\{r-1,r,r+1,\ldots\}$, d.h.

$$(a_1,b_1] = (-\infty,0], \quad (a_2,b_2] = (0,1], \quad \ldots \quad (a_{r-1},b_{r-1} = (r-3,r-2], \quad (a_r,b_r] = (r-2,\infty)\,.$$

  - Die Wahrscheinlichkeiten $p_j(\lambda) = \mathbb{P}_\lambda(a_j < X_1 \le b_j)$ sind dann gegeben durch

$$p_j(\lambda) = \frac{\lambda^{j-1}}{(j-1)!}\, e^{-\lambda} \quad \forall\, j = 1,\ldots,r-1 \qquad \text{und} \qquad p_r(\lambda) = \sum_{i=r}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}\, e^{-\lambda} \tag{80}$$

- Gemäß (66) genügt jede Maximum–Likelihood–Schätzung $\widehat{\lambda}$ für λ, die aus den gruppierten Daten gewonnen wird, der Gleichung

$$\sum_{j=1}^{r} Z_j(x_1,\ldots,x_n)\, \frac{\dfrac{d}{d\lambda}\, p_j(\lambda)}{p_j(\lambda)} = 0\,. \tag{81}$$

  - Dabei ergibt sich aus (80), dass

$$\frac{\dfrac{d}{d\lambda}\, p_j(\lambda)}{p_j(\lambda)} = \frac{j-1}{\lambda} - 1 \quad \forall\, j = 1,\ldots,r-1 \qquad \text{und} \qquad \frac{\dfrac{d}{d\lambda}\, p_r(\lambda)}{p_r(\lambda)} = \frac{\displaystyle\sum_{i=r}^{\infty} \frac{\Bigl(\dfrac{i-1}{\lambda} - 1\Bigr)\lambda^{i-1}}{(i-1)!}}{\displaystyle\sum_{i=r}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}}\,.$$

  - Hieraus und aus (81) folgt, dass die ML–Schätzung $\widehat{\lambda}$ der folgenden Gleichung genügt:

$$\sum_{j=1}^{r-1} Z_j(x_1,\ldots,x_n) \Bigl(\frac{j-1}{\lambda} - 1\Bigr) + Z_r(x_1,\ldots,x_n)\, \frac{\displaystyle\sum_{i=r}^{\infty} \frac{\Bigl(\dfrac{i-1}{\lambda} - 1\Bigr)\lambda^{i-1}}{(i-1)!}}{\displaystyle\sum_{i=r}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}} = 0\,. \tag{82}$$

- Für jedes n gibt es ein $r_0 = r_0(n) \in \mathbb{N}$, so dass $Z_r(x_1,\ldots,x_n) = 0$ für jedes $r > r_0$. Hieraus und aus (82) folgt, dass für $r \to \infty$

$$\widehat{\lambda}_n = \widehat{\lambda}(x_1,\ldots,x_n) \to \overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i\,.$$

  - Bei einer hinreichend großen Anzahl r von Klassen $\{0\}$, $\{1\}$, ..., $\{r-2\}$, $\{r-1, r, r+1, \ldots\}$ bildet also das Stichprobenmittel $\overline{x}_n$, das eine ML–Schätzung für λ im ungruppierten Poisson-Modell ist, eine gute Näherung für die ML–Schätzung $\widehat{\lambda}_n$ für λ im gruppierten Poisson–Modell.
  - Die Nullhypothese $H_0 : P \in \{\,\mathrm{Poi}(\lambda),\ \lambda > 0\}$ wird somit abgelehnt, wenn

$$\widehat{T}_n(x_1,\ldots,x_n) = \sum_{j=1}^r \frac{\bigl(Z_j(x_1,\ldots,x_n) - n\widehat{p}_j(x_1,\ldots,x_n)\bigr)^2}{n\widehat{p}_j(x_1,\ldots,x_n)} > \chi^2_{r-2,1-\alpha}\,,$$

  wobei $\widehat{p}_j(x_1,\ldots,x_n) = p_j(\overline{x}_n)$ mit der in (80) gegebenen Funktion $p_j : (0,\infty) \to [0,1]$ und der Schätzung $\overline{x}_n$ für λ.

5.4.2 χ^2–Anpassungstest auf Normalverteilung

- Sei nun $\Theta = \mathbb{R} \times (0,\infty)$ mit $\boldsymbol{\theta} = (\mu,\sigma^2)^\top$, und sei $\{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\} = \{\,\mathrm{N}(\mu,\sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$ die Familie der (eindimensionalen) Normalverteilungen.
  - Die Wahrscheinlichkeiten $p_j(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(a_j < X_1 \le b_j)$ sind dann gegeben durch

$$p_j(\boldsymbol{\theta}) = \int\limits_{a_j}^{b_j} f(x;\boldsymbol{\theta})\,dx\,, \qquad \text{wobei } f(x;\boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}}\,\exp\Bigl(-\frac{(x-\mu)^2}{2\sigma^2}\Bigr). \tag{83}$$

  - Gemäß (66) genügt jede Maximum–Likelihood–Schätzung

$$\widehat{\boldsymbol{\theta}}(x_1,\ldots,x_n) = (\widehat{\mu}(x_1,\ldots,x_n), \widehat{\sigma}^2(x_1,\ldots,x_n))^\top$$

  für $\boldsymbol{\theta} = (\mu,\sigma^2)^\top$, die aus den gruppierten Daten gewonnen wird, dem Gleichungssystem

$$\sum_{j=1}^r Z_j(x_1,\ldots,x_n)\,\frac{\int_{a_j}^{b_j} \frac{\partial}{\partial\theta_i} f(x;\boldsymbol{\theta})\,dx}{\int_{a_j}^{b_j} f(x;\boldsymbol{\theta})\,dx} = 0 \qquad \text{für } i = 1,2. \tag{84}$$

  - Dabei ergibt sich aus (83), dass

$$\frac{\partial}{\partial\mu} f(x;\boldsymbol{\theta}) = \frac{x-\mu}{\sigma^2}\,f(x;\boldsymbol{\theta}) \qquad \text{bzw.} \qquad \frac{\partial}{\partial\sigma^2} f(x;\boldsymbol{\theta}) = -\,\frac{1}{2}\,f(x;\boldsymbol{\theta})\Bigl((\sigma^2)^{-3/2} + (\sigma^2)^{-5/2}(x-\mu)^2\Bigr)\,.$$

- Hieraus und aus (84) folgt, dass die ML–Schätzung $\widehat{\boldsymbol{\theta}}(x_1,\ldots,x_n)$ dem folgenden Gleichungssystem genügt:

$$\sum_{j=1}^r Z_j(x_1,\ldots,x_n)\,\frac{\int_{a_j}^{b_j} (x-\mu) f(x;\boldsymbol{\theta})\,dx}{\int_{a_j}^{b_j} f(x;\boldsymbol{\theta})\,dx} = 0\,, \qquad \sum_{j=1}^r Z_j(x_1,\ldots,x_n)\,\frac{\int_{a_j}^{b_j} (x-\mu)^2 f(x;\boldsymbol{\theta})\,dx}{\int_{a_j}^{b_j} f(x;\boldsymbol{\theta})\,dx} - n\sigma^2 = 0\,,$$

  wobei die erste Gleichung dieses Gleichungssystems äquivalent ist mit

$$\sum_{j=1}^r Z_j(x_1,\ldots,x_n)\,\frac{\int_{a_j}^{b_j} x f(x;\boldsymbol{\theta})\,dx}{\int_{a_j}^{b_j} f(x;\boldsymbol{\theta})\,dx} - \mu\sum_{j=1}^r Z_j(x_1,\ldots,x_n) = 0\,.$$

- Die ML–Schätzung $\widehat{\boldsymbol{\theta}}(x_1,\ldots,x_n) = (\widehat{\mu}(x_1,\ldots,x_n), \widehat{\sigma}^2(x_1,\ldots,x_n))^\top$ genügt deshalb dem Gleichungssystem

$$\mu = \frac{1}{n}\sum_{j=1}^r Z_j(x_1,\ldots,x_n)\frac{\int\limits_{a_j}^{b_j} x f(x;\mu,\sigma^2)\,dx}{\int\limits_{a_j}^{b_j} f(x;\mu,\sigma^2)\,dx}\,, \qquad \sigma^2 = \frac{1}{n}\sum_{j=1}^r Z_j(x_1,\ldots,x_n)\frac{\int\limits_{a_j}^{b_j} (x-\mu)^2 f(x;\mu,\sigma^2)\,dx}{\int\limits_{a_j}^{b_j} f(x;\mu,\sigma^2)\,dx}\,,$$

  - das sich bei einer hinreichend großen Anzahl r von Klassen $(a_1,b_1],\ldots,(a_r,b_r]$ wie folgt näherungsweise lösen lässt:

$$\widehat{\mu} \approx \frac{1}{n}\sum_{j=1}^r c_j Z_j(x_1,\ldots,x_n)\,, \qquad \widehat{\sigma}^2 \approx \frac{1}{n}\sum_{j=1}^r (c_j-\widehat{\mu})^2 Z_j(x_1,\ldots,x_n)\,, \tag{85}$$

  - wobei $c_1 = b_1$, $c_r = b_{r-1}$ der rechte bzw. linke Endpunkt der ersten bzw. r-ten Klasse und $c_j = (b_{j-1}+b_j)/2$ die Klassenmittelpunkte für $j = 2,\ldots,r-1$ sind.

- Die Nullhypothese $H_0 : P \in \{\,\mathrm{N}(\mu,\sigma^2),\ \mu\in\mathbb{R}, \sigma^2>0\}$ wird abgelehnt, wenn

$$\widehat{T}_n(x_1,\ldots,x_n) = \sum_{j=1}^r \frac{\bigl(Z_j(x_1,\ldots,x_n) - n\widehat{p}_j(x_1,\ldots,x_n)\bigr)^2}{n\widehat{p}_j(x_1,\ldots,x_n)} > \chi^2_{r-3,1-\alpha}\,,$$

  wobei $\widehat{p}_j(x_1,\ldots,x_n) = p_j(\widehat{\mu},\widehat{\sigma}^2)$ mit der in (83)) gegebenen Funktion $p_j : \mathbb{R}\times(0,\infty)\to[0,1]$ und der in (85) gegebenen Schätzung $(\widehat{\mu},\widehat{\sigma}^2)$ für (μ,σ^2).

Beachte

- Die Näherungslösung (85) des Gleichungssystems (84) sollte nur dann verwendet werden, wenn die Anzahl r der Klassen hinreichend groß ist.
  - Dies setzt einen hinreichend großen Stichprobenumfang n voraus.
  - Mit anderen Worten: Wenn der Stichprobenumfang n klein ist, dann ist der χ^2-Anpassungstest auch aus diesem Grund nicht geeignet, um die Hypothese der Normalverteiltheit zu verifizieren.
- Alternative Tests auf Normalverteilung sind die folgenden *Anpassungstests vom Shapiro–Wilk–Typ*, die auch bei kleinem Stichprobenumfang n zu akzeptablen Ergebnissen führen.

5.4.3 Anpassungstests vom Shapiro–Wilk–Typ

- In diesem Abschnitt diskutieren wir zwei Anpassungstests vom Shapiro–Wilk–Typ, mit denen ebenfalls die Hypothese $H_0 : P \in \{\,\mathrm{N}(\mu,\sigma^2),\ \mu\in\mathbb{R}, \sigma^2>0\}$ verifiziert werden kann.
- Hierfür werden die *Ordnungsstatistiken* $X_{(1)},\ldots,X_{(n)}$ der (unabhängigen und identisch verteilten) Stichprobenvariablen $X_1,\ldots,X_n$ betrachtet, die bereits in Abschnitt I–1.4 eingeführt worden sind.
  - *Zur Erinnerung*: Die Ordnungsstatistiken werden mit Hilfe der Stichprobenfunktion $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ definiert, wobei

$$(x_1,\ldots,x_n) \to (x_{(1)},\ldots,x_{(n)}) = \varphi(x_1,\ldots,x_n) \qquad \text{mit } x_{(i)} = \min\bigl\{x_j : \#\{k : x_k \le x_j\} \ge i\bigr\} \tag{86}$$

    für jedes $i \in \{1,\ldots,n\}$.
  - Dabei ist die in (86) gegebene Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ ist eine *Permutation* der Komponenten des Vektors $(x_1,\ldots,x_n)$, so dass

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}\,.$$

- Für jedes $\omega \in \Omega$ sei dann

$$(X_{(1)}(\omega), \ldots, X_{(n)}(\omega)) = \varphi\big(X_1(\omega), \ldots, X_n(\omega)\big)$$

die in (86) gegebene (messbare) Permutation von $(X_1(\omega), \ldots, X_n(\omega))$, so dass

$$X_{(1)}(\omega) \le \ldots \le X_{(n)}(\omega)\,. \tag{87}$$

- Die auf diese Weise definierten Zufallsvariablen $X_{(1)}, \ldots, X_{(n)} : \Omega \to \mathbb{R}$ heißen die *Ordnungsstatistiken* von $(X_1, \ldots, X_n)$.

- Wenn $X_i \sim \mathrm{N}(\mu, \sigma^2)$ für gewisse $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, dann kann man sich leicht überlegen, dass für die Erwartungswerte $b_i = \mathbb{E}\, X_{(i)}$ der Ordnungsstatisiken $X_{(i)}$ die folgende Darstellungsformel gilt:

$$b_i = \mu + \sigma a_i \qquad \forall\, i = 1, \ldots, n\,, \tag{88}$$

  - wobei a_i der Erwartungswert der i–ten Ordnungsstatistik bei $\mathrm{N}(0,1)$–verteilten Stichprobenvariablen ist.
  - Der Nutzen der Darstellungsformel (88) besteht darin, dass die Erwartungswerte $a_1, \ldots, a_n$ in Form von Tabellen vorliegen bzw. durch Monte–Carlo–Simulation bestimmt werden können.

- Weil sich die Vektoren $(b_1, \ldots, b_n)$ und $(X_{(1)}, \ldots, X_{(n)})$ unter H_0 nur relativ wenig voneinander unterscheiden sollten, wird zur Verifizierung der Nullhypothese $H_0 : P \in \{\mathrm{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$ der folgende *empirische Korrelationskoeffizient* betrachtet:

$$\widetilde{T}(X_1, \ldots, X_n) = \frac{\sum\limits_{i=1}^n \big(b_i - \overline{b}\big)\big(X_{(i)} - \overline{X}\big)}{\sqrt{\sum\limits_{i=1}^n (b_i - \overline{b})^2}\sqrt{\sum\limits_{i=1}^n \big(X_{(i)} - \overline{X}\big)^2}}\,, \tag{89}$$

wobei $\overline{b} = \sum_{i=1}^n b_i/n$ und $\overline{X} = \sum_{i=1}^n X_i/n$.

1. *Shapiro–Francia–Test*

   - Weil Korrelationskoeffizienten invariant gegenüber Lineartransformationen sind, können wir in (89) die b_i's durch die a_i's ersetzen, wobei dann $\overline{a} = \sum_{i=1}^n a_i/n = 0$.
   - Außerdem gilt

$$\sum_{i=1}^n \big(X_{(i)} - \overline{X}\big)^2 = \sum_{i=1}^n \big(X_i - \overline{X}\big)^2 \qquad \text{und} \qquad \sum_{i=1}^n a_i \overline{X} = 0\,,$$

   d.h., die Definitionsgleichung (89) von $\widetilde{T}(X_1, \ldots, X_n)$ ist äquivalent mit

$$\widetilde{T}(X_1, \ldots, X_n) = \frac{\sum\limits_{i=1}^n a_i X_{(i)}}{\sqrt{\sum\limits_{i=1}^n a_i^2}\sqrt{\sum\limits_{i=1}^n \big(X_i - \overline{X}\big)^2}}\,. \tag{90}$$

   - Weil stets $|\widetilde{T}(X_1, \ldots, X_n)| \le 1$ gilt, wird H_0 abgelehnt, wenn $\widetilde{T}(X_1, \ldots, X_n) < q_{n,1-\alpha}$, wobei $q_{n,1-\alpha}$ das $(1-\alpha)$–Quantil der Verteilung von $\widetilde{T}(X_1, \ldots, X_n)$ bezeichnet.
   - Das ist der sogenannte *Shapiro–Francia–Test* auf Normalverteilung, wobei die Quantile $q_{n,1-\alpha}$ der Verteilung von $\widetilde{T}(X_1, \ldots, X_n)$ entweder aus Tabellen entnommen oder per Monte–Carlo–Simulation bestimmt werden können.

2. *Shapiro–Wilk–Test*

   - Anstelle der Konstanten $a_1, \ldots, a_n$ kann in (90) die Lineartransformation

   $$(a_1', \ldots, a_n')^\top = \mathbf{K}^{-1}(a_1, \ldots, a_n)^\top$$

   betrachtet werden, wobei die Kovarianzmatrix $\mathbf{K} = (k_{ij})$ gegeben ist durch

   $$k_{ij} = \mathbb{E}\left((Y_{(i)} - a_i)(Y_{(j)} - a_j)\right) \qquad \text{mit } Y_i \sim \mathrm{N}(0,1).$$

   - Der auf diese Weise konstruierte Test wird *Shapiro–Wilk–Test* genannt.

6 Nichtparametrische Lokalisationstests

6.1 Zwei einfache Beispiele von Einstichproben–Problemen

6.1.1 Binomialtest

- Der in Abschnitt 5.2 betrachtete χ^2–Anpassungstest kann durch den folgenden *Binomialtest* ersetzt werden, wenn $r = 2$, d.h., wenn nur zwei Klassen betrachtet werden (beispielsweise bei binären Alternativdaten).
  - Wir zerlegen also den Wertebereich der (unabhängigen und identisch verteilten) Stichprobenvariablen $X_1, \ldots, X_n$ in zwei Teilmengen $(a_1, b_1]$ und $(a_2, b_2]$, so dass
  $$(a_1, b_1] \cap (a_2, b_2] = \emptyset \qquad \text{und} \qquad \mathbb{P}\big(X_1 \in (a_1, b_1] \cup (a_2, b_2]\big) = 1\,,$$
  und betrachten die „Klassenstärke"
  $$T(X_1, \ldots, X_n) = \#\{i : 1 \le i \le n,\, a_1 < X_i \le b_1\}\,.$$
  - Man kann sich leicht überlegen, dass $T = T(X_1, \ldots, X_n)$ binomialverteilt ist, d.h.,
  $$T \sim \text{Bin}(n, p)\,, \qquad \text{wobei } p = \mathbb{P}(a_1 < X_1 \le b_1). \tag{1}$$
- Wir betrachten zunächst das Testproblem $H_0 : p = p_0$ versus $H_1 : p \ne p_0$, wobei $p_0 \in (0, 1)$ eine beliebige positive Zahl ist.
  - Wegen (1) wird H_0 abgelehnt, wenn $T \le t_{\alpha_1}$ oder $T \ge t_{1-\alpha_2}$,
  - wobei die „kritischen Werte" t_{α_1} und $t_{1-\alpha_2}$ für beliebige $\alpha_1, \alpha_2 \in (0, 1)$ mit $\alpha_1 + \alpha_2 = \alpha$ gegeben sind durch
  $$\begin{aligned} t_{\alpha_1} &= \max\{t \in \mathbb{R} : \mathbb{P}_{p_0}(T \le t) \le \alpha_1\} \\ &= \max\Big\{k \in \{0, 1, \ldots, n\} : \sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i} \le \alpha_1\Big\} \end{aligned}$$
  bzw.
  $$\begin{aligned} t_{1-\alpha_2} &= \min\{t \in \mathbb{R} : \mathbb{P}_{p_0}(T \ge t) \le \alpha_2\} \\ &= \min\Big\{k \in \{0, 1, \ldots, n\} : \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \le \alpha_2\Big\}. \end{aligned}$$
  - Für $p_0 = 0.5$ wird dabei normalerweise $\alpha_1 = \alpha_2 = \alpha/2$ gewählt. Wenn p_0 nahe bei 0 bzw. 1 liegt, dann ist es zweckmäßig α_1 kleiner bzw. größer als α_2 zu wählen.
  - Die Quantile t_{α_1} bzw. $t_{1-\alpha_2}$ der Binomialverteilung $\text{Bin}(n, p_0)$ können entweder aus Tabellen entnommen oder per Monte–Carlo–Simulation bestimmt werden.
- Das (einseitige) Testproblem $H_0 : p \le p_0$ versus $H_1 : p > p_0$ kann ähnlich behandelt werden. Dabei wird H_0 abgelehnt, wenn $T \ge t_{1-\alpha}$.
- Völlig analog ergibt sich eine Enscheidungsregel für das (einseitige) Testproblem $H_0 : p \ge p_0$ versus $H_1 : p < p_0$, wobei H_0 abgelehnt wird, wenn $T \le t_\alpha$.

Beachte

- Der oben beschriebene Binomialtest wird auch *Vorzeichentest* genannt, weil die Bildung von 2 Klassen als Binarisierung der ursprünglich vorliegenden Daten aufgefasst werden kann.

- Bei den beiden einseitigen Testproblemen erfolgt die Bestimmung der kritischen Werte $t_{1-\alpha}$ bzw. t_α für $p = p_0$, obwohl die Nullhypothese $H_0 : p \le p_0$ bzw. $H_0 : p \ge p_0$ lautet.
  - Die Tatsache, dass dennoch die Werte $t_{1-\alpha}$ bzw. t_α betrachtet werden, steht nicht damit im Widerspruch, dass für jedes einzelne $p < p_0$ bzw. $p > p_0$ der kritische Wert kleiner als $t_{1-\alpha}$ bzw. größer als t_α wäre und dass dann H_0 öfter abgelehnt werden müsste.
  - Die Erklärung für die Wahl der kritischen Werte $t_{1-\alpha}$ bzw. t_α ist, dass nicht ein einzelnes p mit $p < p_0$ bzw. $p > p_0$ betrachtet wird, sondern dass p beliebig nahe bei p_0 liegen kann und dass insbesondere auch $p = p_0$ zugelassen wird.
- Wenn der Stichprobenumfang n groß ist und wenn p_0 nahe bei 0 oder 1 liegt,
  - dann ist die direkte Berechnung der Quantile $t_{1-\alpha}$ bzw. t_α der Binomialverteilung $\mathrm{Bin}(n, p_0)$ schwierig.
  - Aus dem Gesetz der seltenen Ereignisse (vgl. Abschnitt WR–3.2.2) ergibt sich, dass $t_{1-\alpha}$ bzw. t_α in diesem Fall durch Quantile der Poisson–Verteilung $\mathrm{Poi}(\lambda)$ approximiert werden können, wobei $\lambda = np_0$ bzw. $\lambda = n(1 - p_0)$.
- Außerdem kann $t_{1-\alpha}$ bzw. t_α für jedes beliebige $p_0 \in (0, 1)$ durch geeignet transformierte Quantile der $\mathrm{N}(0, 1)$–Verteilung approximiert werden können, wenn der Stichprobenumfang n „hinreichend groß" ist.
  - Aus dem zentralen Grenzwertsatz von DeMoivre–Laplace (vgl. Theorem WR–3.6) ergibt sich dann, dass die transformierte Testgröße
$$T' = \frac{T - np_0}{\sqrt{np_0(1 - p_0)}}$$
näherungsweise $\mathrm{N}(0, 1)$–verteilt ist, d.h., dass
$$\mathbb{P}(T \le t) = \mathbb{P}(T' \le t') \approx \Phi(t')\,, \qquad \text{wobei} \quad t' = \frac{t - np_0}{\sqrt{np_0(1 - p_0)}}$$
und $\Phi : \mathbb{R} \to [0, 1]$ die Verteilungsfunktion der $\mathrm{N}(0, 1)$–Verteilung ist.
  - Es gilt also, dass $t_\alpha \approx np_0 + z_\alpha \sqrt{np_0(1 - p_0)}$, wobei z_α das α–Quantil der $\mathrm{N}(0, 1)$–Verteilung ist.
- Als ein mögliches Kriterium für „hinreichend groß" werden dabei in der Literatur beispielsweise die Bedingungen $n \ge 20$ und $10 \le np_0 \le n - 10$ angegeben.
  - Bei der Untersuchung des (zweiseitigen) Testproblems $H_0 : p = p_0$ versus $H_1 : p \ne p_0$ wird dann H_0 abgelehnt, wenn
$$T \le np_0 + z_{\alpha_1} \sqrt{np_0(1 - p_0)} \qquad \text{oder} \qquad T \ge np_0 + z_{1-\alpha_2} \sqrt{np_0(1 - p_0)}\,.$$
  - Ähnliche Näherungsformeln ergeben sich für die kritischen Werte der obenerwähnten einseitigen Tests.

Beispiel

- Die Verteilungsfunktion $F : \mathbb{R} \to [0, 1]$ der Stichprobenvariablen $X_1, \ldots, X_n$ sei stetig, und es sei γ_p das p–Quantil von F, d.h., es gelte $F(\gamma_p) = p$ für $p \in (0, 1)$.
- Um die Hypothese $H_0 : \gamma_p = \gamma_p^0$ zu testen, kann man die „vergröberte" Zufallsstichprobe $(Y_1, \ldots, Y_n)$ betrachten mit
$$Y_i = \begin{cases} 1\,, & \text{wenn } X_i \le \gamma_p, \\ 0\,, & \text{wenn } X_i > \gamma_p. \end{cases}$$
- Dann gilt $Y_i \sim \mathrm{Bin}(1, p)$ für jedes $i = 1, \ldots, n$, und die Hypothese $H_0 : \gamma_p = \gamma_p^0$ bezüglich $(X_1, \ldots, X_n)$ ist äquivalent mit der Hypothese $H_0 : p = p_0$ bezüglich $(Y_1, \ldots, Y_n)$.
- Mit dem Binomialtest kann also insbesondere die Hypothese $H_0 : \gamma_{0.5} = 0$ getestet werden.

6.1.2 Iterationstest auf Zufälligkeit

- In diesem Abschnitt wird *nicht* vorausgesetzt, dass die Stichprobenvariablen $X_1, \ldots, X_n$ unabhängig sind.
  - Wir nehmen nämlich an, dass $X_1, \ldots, X_n$ nur die Werte 0 oder 1 annehmen können, wobei n_1–mal der Wert 0 und n_2–mal der Wert 1 auftreten möge; $n_2 = n - n_1$.
  - Insgesamt gibt es dann $\binom{n}{n_1}$ mögliche Realisierungen der Zufallsstichprobe $(X_1, \ldots, X_n)$.
  - Dabei soll die Nullhypothese H_0 geprüft werden, ob jede dieser $\binom{n}{n_1}$ Realisierungen die gleiche Wahrscheinlichkeit hat.
  - Mit anderen Worten: Es soll geprüft werden, ob die Lokalisation, d.h. die Reihenfolge „rein zufällig" ist, in der die n_1 Einsen bzw. die n_2 Nullen angeordnet sind.
- Als Testgröße $T : \Omega \to \{0, 1, \ldots\}$ betrachten wir die Anzahl $T(\omega)$ von *Iterationen* in der (konkreten) Stichprobe $\omega = (x_1, \ldots, x_n)$, d.h. die Anzahl von (Teil–) Folgen aufeinanderfolgender gleicher Zeichen in $\omega = (x_1, \ldots, x_n)$.

Beispiel

- Sei $n = 20$ mit $n_1 = 12$ und $n_2 = 8$. Für
$$\omega = (1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,0,0,1) \tag{2}$$
gilt dann $T(\omega) = 7$.
- Wir untersuchen nun die Frage, ob die in (2) gegebenen Daten mit der Hypothese H_0 vereinbart sind, dass die Reihenfolge „rein zufällig" ist, oder, ob H_0 verworfen wird.
- Hierfür bestimmen wir die Verteilung von T, indem wir einen geeignet gewählten Laplaceschen Wahrscheinlichkeitsraum betrachten, vgl. Abschnitt WR–2.4.1.

Theorem 6.1 *Unter H_0 gilt für jedes $i = 1, 2, \ldots, \min\{n_1, n_2\}$*

$$\mathbb{P}(T = k) = \begin{cases} \dfrac{2\binom{n_1 - 1}{i - 1}\binom{n_2 - 1}{i - 1}}{\binom{n}{n_1}}\,, & \text{wenn } k = 2i, \\[3ex] \dfrac{\binom{n_1 - 1}{i}\binom{n_2 - 1}{i - 1} + \binom{n_1 - 1}{i - 1}\binom{n_2 - 1}{i}}{\binom{n}{n_1}}\,, & \text{wenn } k = 2i + 1. \end{cases} \tag{3}$$

Außerdem gilt

$$\mathbb{E}\, T = 1 + \frac{2 n_1 n_2}{n} \qquad \text{und} \qquad \operatorname{Var} T = \frac{2 n_1 n_2 (2 n_1 n_2 - n)}{n^2 (n - 1)}\,. \tag{4}$$

Beweis

- Wir zeigen die Gültigkeit von (3) nur für den Fall $k = 2i$, denn der Beweis für den Fall $k = 2i + 1$ verläuft analog.
  - Sei also $k = 2i$. Dann gibt es je i Iterationen, die aus Einsen bzw. aus Nullen bestehen.
  - Für die Zerlegung der n_1 Nullen in i Teilmengen gibt es $\binom{n_1 - 1}{i - 1}$ Möglichkeiten.
  - Für jede dieser Zerlegungen gibt es $\binom{n_2 - 1}{i - 1}$ Möglichkeiten, die n_2 Einsen in i Teilmengen zu zerlegen.

  - Wenn nun noch beachtet wird, dass die Stichprobe $\omega = (x_1, \ldots, x_n)$ entweder mit $x_1 = 0$ oder mit $x_1 = 1$ beginnen kann, dann ergeben sich insgesamt $2\binom{n_1-1}{i-1}\binom{n_2-1}{i-1}$ Zerlegungsmöglichkeiten.
  - Damit ist (3) für den Fall $k = 2i$ bewiesen.

- Bei der Bestimmung des Erwartungswertes $\mathbb{E}\, T$ nutzen wir die folgende Überlegung.
  - Für jedes $j = 2, \ldots, n$ betrachten wir die Indikatorvariable $Y_j : \Omega \to \{0, 1\}$ mit

$$Y_j = \begin{cases} 1\,, & \text{wenn an der } j\text{–ten Stelle eine Iteration beginnt,} \\ 0\,, & \text{sonst.} \end{cases}$$

  - Dann gilt $\{\omega \in \Omega :\ Y_j(\omega) = 1\} = \{\omega \in \Omega :\ X_{j-1}(\omega) \neq X_j(\omega)\}$, d.h., es gibt $2\binom{n-2}{n_i-1}$ Möglichkeiten, dass an der j–ten Stelle eine Iteration beginnt.
  - Somit gilt

$$\mathbb{E}\, Y_j \;=\; \mathbb{P}(Y_j = 1) \;=\; 2\,\frac{\binom{n-2}{n_i-1}}{\binom{n}{n_1}} \;=\; 2\,\frac{(n-2)!\,(n-n_1)!\,n_1!}{(n-n_1-1)!(n_1-1)!n!} \;=\; 2\,\frac{n_1(n-n_1)}{n(n-1)}\,.$$

  - Hieraus und aus der Identität

$$T = 1 + \sum_{j=2}^{n} Y_j \tag{5}$$

ergibt sich, dass

$$\mathbb{E}\, T \;=\; 1 + \sum_{j=2}^{n} \mathbb{E}\, Y_j \;=\; 1 + 2\,\frac{n_1(n-n_1)}{n}\,.$$

- Die Varianzformel in (4) lässt sich auf ähnliche Weise beweisen, denn aus (5) ergibt sich, dass

$$\begin{aligned}
\operatorname{Var} T \;&=\; \mathbb{E}\Bigl(\sum_{j=2}^{n} \mathbb{E}\, Y_j\Bigr)^2 - \Bigl(\sum_{j=2}^{n} \mathbb{E}\, Y_j\Bigr)^2 \\
&=\; \sum_{j=2}^{n} \mathbb{E}\, Y_j^2 \;+ \sum_{2\le j_1, j_2 \le n,\, j_1 \ne j_2} \mathbb{E}\,\bigl(Y_{j_1} Y_{j_2}\bigr) \;-\; \Bigl(\sum_{j=2}^{n} \mathbb{E}\, Y_j\Bigr)^2 \\
&=\; \sum_{j=2}^{n} \mathbb{E}\, Y_j \;+ \sum_{2\le j_1, j_2 \le n,\, j_1 \ne j_2} \mathbb{E}\,\bigl(Y_{j_1} Y_{j_2}\bigr) \;-\; \Bigl(\sum_{j=2}^{n} \mathbb{E}\, Y_j\Bigr)^2 ,
\end{aligned}$$

so dass lediglich noch die Momente $\mathbb{E}\,\bigl(Y_{j_1} Y_{j_2}\bigr)$ zu bestimmen sind. □

Beachte

- Eine mögliche Alternative zu der Nullhypothese H_0 der „rein zufälligen" Lokalisation der Nullen und Einsen ist die Tendenz zur Klumpen– bzw. Clusterbildung.
- Als Ablehnungsbereich von H_0 wird dann das linke Ende der Verteilung von T gewählt.
- Mit anderen Worten: H_0 wird abgelehnt, wenn $T \le r_\alpha(n_1; n_2)$, wobei

$$r_\alpha(n_1; n_2) = \max\Bigl\{r \in \{1, 2, \ldots\} :\ \mathbb{P}(T \le r) \le \alpha\Bigr\}$$

das α–Quantil der Verteilung der Testgröße T ist.
- Die Quantile $r_\alpha(n_1; n_2)$ können mit Hilfe der in Theorem 6.1 gegebenen Formeln für die Wahrscheinlichkeiten $\mathbb{P}(T = k)$ berechnet werden. Sie können aus Tafeln entnommen werden, die in der Literatur gegeben sind.

Beispiel (Fortsetzung) Für $\alpha = 0.1$ und $n_1 = 12$, $n_2 = 8$ ergibt sich, dass $r_{0.1}(12;8) = 7$. Andererseits gilt für die in (2) betrachtete Stichprobe

$$T(\omega) = 7 \ \Bigl(\le r_{0.1}(12;8)\Bigr),$$

d.h., H_0 wird abgelehnt.

Wenn die (Teil–) Stichprobenumfänge n_1 und n_2 groß sind, dann ist die Bestimmung der Quantile $r_\alpha(n_1;n_2)$ von $T = T_{n_1,n_2}$ mit erheblichem Rechenaufwand verbunden. Einen Ausweg bietet dann der folgenden zentrale Grenzwertsatz, den wir hier ohne Beweis angeben.

Theorem 6.2 *Wenn $n_1, n_2 \to \infty$, so dass $n_1/(n_1+n_2) \to p$ bzw. $n_2/(n_1+n_2) \to 1-p$ für ein $p \in (0,1)$, dann gilt*

$$\lim_{n_1,n_2\to\infty} \frac{1}{n_1+n_2}\,\mathbb{E}\,T_{n_1,n_2} = 2p(1-p) \qquad \textit{und} \qquad \lim_{n_1,n_2\to\infty} \frac{1}{n_1+n_2}\,\mathrm{Var}\,T_{n_1,n_2} = 4p^2(1-p)^2 \tag{6}$$

sowie

$$\lim_{n_1,n_2\to\infty} \mathbb{P}\Bigl(\frac{T_{n_1,n_2} - 2(n_1+n_2)p(1-p)}{2\sqrt{n_1+n_2}\,p(1-p)} \le x\Bigr) = \Phi(x) \qquad \forall\, x \in \mathbb{R}, \tag{7}$$

wobei $\Phi : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der $\mathrm{N}(0,1)$–Verteilung ist.

Beachte Wegen Theorem 6.2 wird H_0 für große n_1, n_2 abgelehnt, wenn

$$\frac{T_{n_1,n_2} - 2n_1n_2/(n_1+n_2)}{2n_1n_2/(n_1+n_2)^{3/2}} \le z_\alpha\,, \tag{8}$$

wobei z_α das α–Quantil der $\mathrm{N}(0,1)$–Verteilung ist.

6.2 Vorzeichenrangtest von Wilcoxon

6.2.1 Modellbeschreibung; Mediantest

- Wir kehren nun wieder zu dem Fall zurück, dass die Stichprobenvariablen $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind, mit der Verteilungsfunktion $F : \mathbb{R} \to [0,1]$.
  - Am Ende von Abschnitt 6.1.1 hatten wir im Zusammenhang mit dem Binomial– bzw. Vorzeichentest einen *Mediantest* diskutiert, um die Hypothese
    $$H_0 : \gamma_{0.5} = 0 \tag{9}$$
    zu verifizieren, wobei $\gamma_{0.5}$ ein Median von F ist, d.h., $F(\gamma_{0.5}) = 0.5$.
  - In diesem Abschnitt betrachten wir einen weiteren (effizienteren) Ansatz, um die in (9) gegebene Hypothese zu testen.
- Dabei nehmen wir an, dass die Verteilungsfunktion F der Stichprobenvariablen $X_1, \ldots, X_n$ zu der folgenden (nichtparametrischen) Klasse von Verteilungsfunktionen gehört.
  - Sei $G : \mathbb{R} \to [0,1]$ eine beliebige stetige Verteilungsfunktion, die die folgende Symmetrieeigenschaft bezüglich des Nullpunktes besitzt: Für jedes $x \in \mathbb{R}$ gelte $G(-x) = 1 - G(x)$.
  - Hieraus folgt insbesondere, dass $G(0) = 1/2$, d.h., der Nullpunkt ist ein Median von G.

- Die Familie Δ von Verteilungsfunktionen der Stichprobenvariablen $X_1,\ldots,X_n$, die beim (zweiseitigen) Wilcoxon–Test in Betracht gezogenen wird, sei gegeben durch $\Delta = \bigl\{F_\delta:\ F_\delta(x) = G(x-\delta)\ \forall\, x,\delta\in\mathbb{R}\bigr\}$.
- Weil G stetig ist, gilt dann für jedes $x\in\mathbb{R}$ und für beliebige $i,j=1,\ldots,n$ mit $i\neq j$

$$\mathbb{P}(X_i = x) = \mathbb{P}(X_i = X_j) = 0\,. \tag{10}$$

- Wir diskutieren das (zweiseitige) Testproblem $H_0:\ \delta=\delta_0$ vs. $H_1:\ \delta\neq\delta_0$ für ein $\delta_0\in\mathbb{R}$.
  - Dabei können wir (o.B.d.A.) $\delta_0=0$ setzen; ansonsten können die transformierten Stichprobenvariablen $X_1',\ldots,X_n'$ mit $X_i' = X_i-\delta_0$ betrachtet werden.
  - Auf ähnliche Weise kann auch das (einseitige) Testproblem $H_0:\ \delta=0$ vs. $H_1:\ \delta>0$ behandelt werden.
- Zur Verifizierung der Nullhypothese $H_0:\ \delta=0$ betrachten wir die *Ränge* $R_1,\ldots,R_n$ der Zufallsvariablen $|X_1|,\ldots,|X_n|$ mit

$$R_i = \sum_{j=1}^n \mathbf{1}_{\{|X_j|\le|X_i|\}}\qquad \forall\, i=1,\ldots,n\,,$$

wobei die Indikatorvariable $\mathbf{1}_{\{|X_j|\le|X_i|\}}:\Omega\to\{0,1\}$ gegeben ist durch

$$\mathbf{1}_{\{|X_j|\le|X_i|\}}(\omega) = \begin{cases} 1\,, & \text{wenn } |X_j(\omega)|\le|X_i(\omega)|,\\ 0\,, & \text{sonst.}\end{cases}$$

- Dabei betrachten wir die Teststatistiken

$$T_n^+ = \sum_{i=1}^n R_i\mathbf{1}_{\{X_i>0\}}\qquad\text{bzw.}\qquad T_n^- = \sum_{i=1}^n R_i\mathbf{1}_{\{X_i<0\}}\,. \tag{11}$$

Beachte

- Wegen (10) gilt mit Wahrscheinlichkeit 1

$$T_n^- \;=\; \sum_{i=1}^n R_i - T_n^+ \;=\; \binom{n+1}{2} - T_n^+\,. \tag{12}$$

- Man kann zeigen, dass $T_n^- \stackrel{\mathrm{d}}{=} T_n^+$ unter $H_0:\ \delta=0$ gilt; vgl. (18).
- Unter $H_0:\ \delta=0$ sollten daher die Teststatistiken T_n^+ und T_n^- etwa gleich große Werte annehmen. Wegen (12) bedeutet dies, dass dann $T_n^+\approx\binom{n+1}{2}/2$.
- Sehr kleine oder sehr große Werte von T_n^+ sprechen daher für die Alternativhypothese $H_1:\ \delta\neq\delta_0$, d.h., $H_0:\ \delta=0$ wird abgelehnt, wenn

$$T_n^+\le t_{\alpha/2}\qquad\text{oder}\qquad T_n^+\ge t_{1-\alpha/2}\,, \tag{13}$$

wobei die „kritischen Werte" $t_{\alpha/2}$ und $t_{1-\alpha/2}$ das $(\alpha/2)$–Quantil bzw. das $(1-\alpha/2)$–Quantil der Verteilung von T_n^+ sind.

6.2.2 Verteilung der Teststatistik T_n^+ für kleine Stichprobenumfänge

- Wenn der Stichprobenumfang n nicht zu groß ist, dann lassen sich die Quantile $t_{\alpha/2}$ und $t_{1-\alpha/2}$ in (13) durch kombinatorische Überlegungen bestimmen.
  - Wegen (10) ist der Zufallsvektor $\mathbf{R}=(R_1,\ldots,R_n)$ der Ränge $R_1,\ldots,R_n$ von $|X_1|,\ldots,|X_n|$ eine (zufällige) Permutation der Zahlen $1,\ldots,n$.

- Dabei lässt sich die in (11) gegebene Testgröße T_n^+ wie folgt darstellen:

$$T_n^+ = \sum_{i=1}^n i\, Z_i\,, \qquad \text{wobei } Z_i = \mathbf{1}_{\{X_{R_i^{-1}}>0\}} \tag{14}$$

- und $\mathbf{R}^{-1} = (R_1^{-1},\ldots,R_n^{-1})$ die zu $\mathbf{R}$ inverse Permutation bezeichnet, d.h., wenn $R_i = j$, dann gilt $R_j^{-1} = i$.

- Außerdem ist der folgende Hilfssatz nützlich, um die Verteilung von T_n^+ zu bestimmen.

Lemma 6.1 *Unter $H_0 : \delta = 0$ gilt:*

- *Die Zufallsvektoren $\bigl(\mathbf{1}_{\{X_1>0\}},\ldots,\mathbf{1}_{\{X_n>0\}}\bigr)$ und $\mathbf{R} = (R_1,\ldots,R_n)$ sind unabhängig.*
- *Die Komponenten $Z_1,\ldots,Z_n$ von $(Z_1,\ldots,Z_n)$ sind unabhängig und identisch verteilt mit $Z_i \sim \mathrm{Bin}(1,1/2)$.*

Beweis

- Wir zeigen zunächst, dass die Zufallsvariablen $\mathbf{1}_{\{X_i>0\}}$ und $|X_i|$ für jedes $i = 1,\ldots,n$ unabhängig sind.
  - Für jedes $x \ge 0$ gilt

$$\mathbb{P}\bigl(\mathbf{1}_{\{X_i>0\}} = 1,\, |X_i| \le x\bigr) \;=\; \mathbb{P}\bigl(0 < X_i \le x\bigr) \;=\; G(x) - \frac{1}{2}\,.$$

  und

$$\mathbb{P}\bigl(\mathbf{1}_{\{X_i>0\}} = 1\bigr)\,\mathbb{P}\bigl(|X_i| \le x\bigr) \;=\; \frac{1}{2}\,\bigl(G(x) - G(-x)\bigr) \;=\; \frac{1}{2}\,\bigl(G(x) - (1 - G(x))\bigr) \;=\; G(x) - \frac{1}{2}\,.$$

  - Außerdem gilt offenbar für jedes $x < 0$

$$\mathbb{P}\bigl(\mathbf{1}_{\{X_i>0\}} = 1,\, |X_i| \le x\bigr) \;=\; 0 \;=\; \mathbb{P}\bigl(\mathbf{1}_{\{X_i>0\}} = 1\bigr)\,\mathbb{P}\bigl(|X_i| \le x\bigr)\,.$$

  - Auf die gleiche Weise lässt sich zeigen, dass für jedes $x \in \mathbb{R}$

$$\mathbb{P}\bigl(\mathbf{1}_{\{X_i>0\}} = 0,\, |X_i| \le x\bigr) \;=\; \mathbb{P}\bigl(\mathbf{1}_{\{X_i>0\}} = 0\bigr)\,\mathbb{P}\bigl(|X_i| \le x\bigr)\,.$$

- Weil aus der Unabhängigkeit der Stichprobenvariablen $X_1,\ldots,X_n$ folgt, dass die Zufallsvektoren $\bigl(\mathbf{1}_{\{X_1>0\}},\, |X_1|\bigr),\ldots,\bigl(\mathbf{1}_{\{X_n>0\}},\, |X_n|\bigr)$ unabhängig sind,
  - ergibt sich nun insgesamt, dass $\bigl(\mathbf{1}_{\{X_1>0\}},\ldots,\mathbf{1}_{\{X_n>0\}}\bigr)$ und $\bigl(|X_1|,\ldots,|X_n|\bigr)$ unabhängige Zufallsvektoren sind.
  - Weil $\mathbf{R} = (R_1,\ldots,R_n)$ eine Borel–messbare Funktion von $\bigl(|X_1|,\ldots,|X_n|\bigr)$ ist, sind damit auch die Zufallsvektoren $\bigl(\mathbf{1}_{\{X_1>0\}},\ldots,\mathbf{1}_{\{X_n>0\}}\bigr)$ und $\mathbf{R}$ unabhängig.
- Hieraus folgt, dass für beliebige $i \in \{1,\ldots,n\}$ und $z \in \{0,1\}$

$$\begin{aligned}
\mathbb{P}(Z_i = z) &= \mathbb{P}\bigl(\mathbf{1}_{\{X_{R_i^{-1}}>0\}} = z\bigr)\\
&= \sum_{\mathbf{r}} \mathbb{P}\bigl(\mathbf{1}_{\{X_{R_i^{-1}}>0\}} = z \mid \mathbf{R} = \mathbf{r}\bigr)\,\mathbb{P}(\mathbf{R} = \mathbf{r})\\
&= \sum_{\mathbf{r}} \mathbb{P}\bigl(\mathbf{1}_{\{X_{r_i^{-1}}>0\}} = z \mid \mathbf{R} = \mathbf{r}\bigr)\,\mathbb{P}(\mathbf{R} = \mathbf{r})\\
&= \sum_{\mathbf{r}} \mathbb{P}\bigl(\mathbf{1}_{\{X_{r_i^{-1}}>0\}} = z\bigr)\,\mathbb{P}(\mathbf{R} = \mathbf{r}) \;=\; \frac{1}{2}\,\sum_{\mathbf{r}} \mathbb{P}(\mathbf{R} = \mathbf{r}) \;==\; \frac{1}{2}\,,
\end{aligned}$$

  wobei sich die Summation über alle Permutationen $\mathbf{r} = (r_1,\ldots,r_n)$ der Zahlen $1,\ldots,n$ erstreckt.

- Damit ergibt sich, dass für beliebige $\mathbf{z} = (z_1, \ldots, z_n) \in \{0,1\}^n$

$$\begin{aligned}
\mathbb{P}\bigl(Z_1 = z_1, \ldots, Z_n = z_n\bigr) &= \mathbb{P}\bigl(\mathbf{1}_{\{X_{R_1^{-1}}>0\}} = z_1, \ldots, \mathbf{1}_{\{X_{R_n^{-1}}>0\}} = z_n\bigr)\\
&= \sum_{\mathbf{r}} \mathbb{P}\bigl(\mathbf{1}_{\{X_{R_1^{-1}}>0\}} = z_1, \ldots, \mathbf{1}_{\{X_{R_n^{-1}}>0\}} = z_n \mid \mathbf{R} = \mathbf{r}\bigr)\, \mathbb{P}\bigl(\mathbf{R} = \mathbf{r}\bigr)\\
&= \sum_{\mathbf{r}} \mathbb{P}\bigl(\mathbf{1}_{\{X_{r_1^{-1}}>0\}} = z_1, \ldots, \mathbf{1}_{\{X_{r_n^{-1}}>0\}} = z_n \mid \mathbf{R} = \mathbf{r}\bigr)\, \mathbb{P}\bigl(\mathbf{R} = \mathbf{r}\bigr)\\
&= \sum_{\mathbf{r}} \mathbb{P}\bigl(\mathbf{1}_{\{X_1>0\}} = z_{r_1}, \ldots, \mathbf{1}_{\{X_n>0\}} = z_{r_n} \mid \mathbf{R} = \mathbf{r}\bigr)\, \mathbb{P}\bigl(\mathbf{R} = \mathbf{r}\bigr)\\
&= \sum_{\mathbf{r}} \mathbb{P}\bigl(\mathbf{1}_{\{X_1>0\}} = z_{r_1}, \ldots, \mathbf{1}_{\{X_n>0\}} = z_{r_n}\bigr)\, \mathbb{P}\bigl(\mathbf{R} = \mathbf{r}\bigr)\\
&= \frac{1}{2^n} = \mathbb{P}(Z_1 = z_1) \ldots \mathbb{P}(Z_n = z_n)\,. \qquad \square
\end{aligned}$$

Theorem 6.3 *Unter $H_0: \delta = 0$ ist die Verteilung von T_n^+ gegeben durch*

$$\mathbb{P}(T_n^+ = k) = \frac{a_k}{2^n} \qquad \forall\, k = 0, 1, \ldots, n\,, \tag{15}$$

wobei

$$a_k = \#\Bigl\{\mathbf{z} = (z_1, \ldots, z_n) \in \{0,1\}^n : \sum_{i=1}^n i z_i = k\Bigr\}\,. \tag{16}$$

Außerdem gilt dann

$$\mathbb{E}\, T_n^+ = \frac{n(n+1)}{4} \qquad \textit{und} \qquad \operatorname{Var} T_n^+ = \frac{n(n+1)(2n+1)}{24}\,. \tag{17}$$

Beweis

- Aus der Darstellungsformel (14) für T_n^+ und aus Lemma 6.1 ergibt sich, dass für jedes $k = 0, 1, \ldots, n$

$$\mathbb{P}(T_n^+ = k) = \mathbb{P}\Bigl(\sum_{i=1}^n i\, Z_i = k\Bigr) = \sum_{\mathbf{z}=(z_1,\ldots,z_n)\in\{0,1\}^n:\, \sum_{i=1}^n i z_i = k} \mathbb{P}\bigl(Z_1 = z_1, \ldots, Z_n = z_n\bigr) = \frac{a_k}{2^n}\,.$$

- Außerdem ergibt sich auf diese Weise, dass

$$\mathbb{E}\, T_n^+ = \mathbb{E}\Bigl(\sum_{i=1}^n i\, Z_i\Bigr) = \sum_{i=1}^n i\, \mathbb{E}\, Z_i = \frac{1}{2}\binom{n+1}{2} = \frac{n(n+1)}{4}$$

und

$$\operatorname{Var} T_n^+ = \operatorname{Var}\Bigl(\sum_{i=1}^n i\, Z_i\Bigr) = \sum_{i=1}^n i^2 \operatorname{Var} Z_i = \frac{1}{4}\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{24}\,. \qquad \square$$

Beachte

- Aus (12) und (14) ergibt sich darüber hinaus mit Hilfe von Lemma 6.1, dass

$$T_n^- = \binom{n+1}{2} - T_n^+ = \binom{n+1}{2} - \sum_{i=1}^n i\, Z_i = \sum_{i=1}^n i\,(1 - Z_i) \stackrel{\mathrm{d}}{=} \sum_{i=1}^n i\, Z_i = T_n^+\,,$$

d.h., unter $H_0: \delta = 0$ gilt

$$T_n^- \stackrel{\mathrm{d}}{=} T_n^+\,. \tag{18}$$

- Somit ergibt sich aus (12), dass für jedes $k = 0, 1, \ldots, n$

$$\mathbb{P}(T_n^+ = k) \;=\; \mathbb{P}\Bigl(T_n^- = \frac{n(n+1)}{2} - k\Bigr) \;=\; \mathbb{P}\Bigl(T_n^+ = \frac{n(n+1)}{2} - k\Bigr),$$

  d.h., die Verteilung von T_n^+ ist symmetrisch bezüglich des Erwartungswertes $\mathbb{E}\, T_n^+ = n(n+1)/4$.
- Dies bedeutet, dass auch die Quantile $t_{\alpha,n} = \max\{t \in \mathbb{R} : \mathbb{P}(T_n^+ \le t) \le \alpha\}$ diese Symmetrieeigenschaft besitzen, d.h., für jedes $\alpha \in (0,1)$ gilt

$$t_{\alpha,n} \;=\; \frac{n(n+1)}{2} \;-\, t_{1-\alpha,n}\,.$$

- Die Quantile $t_{\alpha,n}$ können entweder aus Tabellen entnommen oder per Monte–Carlo–Simulation bestimmt werden.

6.2.3 Asymptotische Verteilung

- Wenn der Stichprobenumfang n groß ist, dann ist die direkte Bestimmung der Quantile $t_{\alpha/2}$ und $t_{1-\alpha/2}$ mit Hilfe von Theorem 6.3 schwierig.
  - Ein anderer Zugang zur (näherungsweisen) Bestimmung der Verteilung der Teststatistik T_n^+ beruht auf der Darstellungsformel (14).
  - Dabei wird die Tatsache genutzt, dass $T_n^+ = \sum_{i=1}^n i\, Z_i$ wegen Lemma 6.1 eine Summe von unabhängigen Zufallsvariablen ist.
  - Und zwar kann mit Hilfe eines zentralen Grenzwertsatzes für Summen von unabhängigen (jedoch nicht notwendig identisch verteilten) Zufallsvariablen gezeigt werden, dass T_n^+ asymptotisch normalverteilt ist.
- Hierfür betrachten wir das folgende stochastische Modell: Für jedes $n \ge 1$ sei $X_{n1}, \ldots, X_{nn} : \Omega \to \mathbb{R}$ eine Folge von unabhängigen Zufallsvariablen,
  - wobei wir (o.B.d.A.) voraussetzen, dass für jedes $k \in \{1, \ldots, n\}$

$$\mathbb{E}\, X_{nk} = 0\,, \qquad 0 < \sigma_{nk}^2 = \operatorname{Var} X_{nk} < \infty \qquad \text{und} \qquad \sum_{k=1}^n \sigma_{nk}^2 = 1\,. \tag{19}$$

  - Wenn die Zufallsvariablen $X_{n1}, \ldots, X_{nn}$ die in (19) formulierten Bedingungen nicht erfüllen, dann gehen wir zu den transformierten Zufallsvariablen $X'_{n1}, \ldots, X'_{nn}$ über mit

$$X'_{nk} = \frac{X_{nk} - \mathbb{E}\, X_{nk}}{\sqrt{n \operatorname{Var} X_{nk}}}\,. \tag{20}$$

  - Die Verteilungsfunktion von X_{nk} bezeichnen wir mit F_{nk}, wobei nicht ausgeschlossen wird, dass F_{nk} für jedes $k \in \{1, \ldots, n\}$ auch von der Anzahl n der insgesamt betrachteten Zufallsvariablen $X_{n1}, \ldots, X_{nn}$ abhängen kann.

Der folgende *zentrale Grenzwertsatz von Lindeberg* (vgl.Theorem WR–5.22) bildet die Grundlage, um zu zeigen, dass T_n^+ asymptotisch normalverteilt ist.

Lemma 6.2

- *Für jedes $n \in \mathbb{N}$ sei $X_{n1}, \ldots, X_{nn} : \Omega \to \mathbb{R}$ eine Folge von unabhängigen Zufallsvariablen, die den Bedingungen* (19) *genügen.*

- *Wenn außerdem für jedes $\varepsilon > 0$*
$$\lim_{n\to\infty} \sum_{k=1}^{n} \int_{\mathbb{R}\setminus(-\varepsilon,\varepsilon)} x^2\, dF_{nk}(x) = 0\,, \tag{21}$$
dann gilt für jedes $x \in \mathbb{R}$
$$\lim_{n\to\infty} P\big(X_{n1} + \ldots + X_{nn} \le x\big) = \Phi(x)\,, \tag{22}$$
wobei $\Phi : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der $\mathrm{N}(0,1)$–Verteilung ist.

Theorem 6.4 *Unter $H_0 : \delta = 0$ gilt*
$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{T_n^+ - \mathbb{E}\,T_n^+}{\sqrt{\operatorname{Var} T_n^+}} \le x\Big) \;=\; \Phi(x) \qquad \forall\, x \in \mathbb{R}\,. \tag{23}$$

Beweis

- Wegen (14) genügt es zu zeigen, dass die Zufallsvariablen $X_{n1}, \ldots, X_{nn}$ mit
$$X_{nk} = \frac{kZ_k - k\,\mathbb{E}\,Z_k}{\sqrt{\operatorname{Var} T_n^+}} \tag{24}$$
den Bedingungen von Lemma 6.2 genügen.
  - Dabei ergibt sich das Erfülltsein von (19) unmittelbar aus der Definitionsgleichung (24).
  - Es muss also lediglich noch gezeigt werden, dass die Lindeberg–Bedingung (22) erfüllt ist.
- Mit Hilfe von Lemma 6.1 ergibt sich für die Verteilungsfunktion $F_{nk} : \mathbb{R} \to [0,1]$ der in (24) eingeführten Zufallsvariablen X_{nk}, dass
$$F_{nk}(x) = \begin{cases} 0\,, & \text{wenn } x < \dfrac{-k}{2\sqrt{\operatorname{Var} T_n^+}}\,, \\[2mm] \dfrac{1}{2}\,, & \text{wenn } \dfrac{-k}{2\sqrt{\operatorname{Var} T_n^+}} \le x < \dfrac{k}{2\sqrt{\operatorname{Var} T_n^+}}\,, \\[2mm] 1\,, & \text{wenn } \dfrac{k}{2\sqrt{\operatorname{Var} T_n^+}} \le x\,. \end{cases}$$
  - Hieraus folgt, dass $\int_{\mathbb{R}\setminus(-\varepsilon,\varepsilon)} x^2\, dF_{nk}(x) = 0$ für jedes $k \in \{1,\ldots,n\}$, wenn n so gewählt wird, dass
$$\frac{n^2}{4\operatorname{Var} T_n^+} \;=\; \frac{6n^2}{n(n+1)(2n+1)} \;<\; \varepsilon^2\,,$$
wobei sich die letzte Gleichheit aus der Formel für $\operatorname{Var} T_n^+$ in Theorem 6.3 ergibt.
  - Damit ist die Gültigkeit der Lindeberg–Bedingung (22) gezeigt. □

Beachte

- Wegen Theorem 6.4 wird bei dem (zweiseitigen) Testproblem $H_0 : \delta = 0$ vs. $H_1 : \delta \ne 0$ der folgenden kritische Bereich betrachtet.
- Bei hinreichend großem n wird $H_0 : \delta = 0$ abgelehnt, wenn
$$\Big|\, \frac{T_n^+ - \mathbb{E}\,T_n^+}{\sqrt{\operatorname{Var} T_n^+}} \,\Big| \ge z_{1-\alpha/2}\,, \tag{25}$$
wobei $\mathbb{E}\,T_n^+$ bzw. $\operatorname{Var} T_n^+$ in Theorem 6.3 gegeben sind und $z_{1-\alpha/2}$ das $(1-\alpha/2)$–Quantil der $\mathrm{N}(0,1)$–Verteilung ist.
- Als ein mögliches Kriterium für „hinreichend groß" wird dabei in der Literatur die Bedingung $n \ge 20$ angegeben.

6.3 Zweistichproben–Probleme

- In diesem Abschnitt diskutieren wir nichtparametrische Tests für den Fall, dass zwei unabhängige Zufallsstichproben $(X_1, \ldots, X_{n_1})$ und $(Y_1, \ldots, Y_{n_2})$ beobachtet werden.
- Mit anderen Worten: Wir nehmen an, dass die Zufallsvariablen $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ vollständig unabhängig sind mit den (unbekannten) Verteilungsfunktionen F bzw. G, d.h.,
$$F(x) = \mathbb{P}(X_i \le x) \qquad \text{und} \qquad G(y) = \mathbb{P}(Y_j \le y) \qquad \forall\, x, y \in \mathbb{R},\ i = 1, \ldots, n,\ j = 1, \ldots, m\,.$$
- Ein (zweiseitiges) Testproblem ist dann beispielsweise gegeben durch
$$H_0 : F(x) = G(x) \quad \forall\, x \in \mathbb{R} \qquad \text{vs.} \qquad H_1 : F(x) \ne G(x) \quad \exists\, x \in \mathbb{R}\,. \tag{26}$$
- Als einseitige Alternativen zu $H_0 : F(x) = G(x)\ \forall\, x \in \mathbb{R}$ können zum Beispiel die folgenden Hypothesen betrachtet werden:
$$H_1 : F(x) \ge G(x) \quad \forall\, x \in \mathbb{R} \quad \text{und} \quad F(x) > G(x) \quad \exists\, x \in \mathbb{R} \tag{27}$$
bzw.
$$H_1 : F(x) \le G(x) \quad \forall\, x \in \mathbb{R} \quad \text{und} \quad F(x) < G \quad \exists\, x \in \mathbb{R} \tag{28}$$

6.3.1 Iterationstest von Wald–Wolfowitz

- Zur Untersuchung des in (26) gegebenen Testproblems kann der Iterationstest auf Zufälligkeit angewendet werden, der in Abschnitt 6.1.2 diskutiert worden ist.
  - Hierfür vereinigen wir die Stichprobenvariablen $X_1, \ldots, X_{n_1}$ und $Y_1, \ldots, Y_{n_2}$ zu einer Zufallsstichprobe
  $$(X'_1, \ldots, X'_n) = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})\,, \qquad \text{wobei } n = n_1 + n_2,$$
  und betrachten die geordnete Stichprobe $(X'_{(1)}, \ldots, X'_{(n)})$.
  - Dabei setzen wir voraus, dass die Verteilungsfunktionen F und G stetig sind, d.h., die Abbildung
  $$(X'_1, \ldots, X'_n) \mapsto (X'_{(1)}, \ldots, X'_{(n)})$$
  ist mit Wahrscheinlichkeit 1 eindeutig festgelegt.
- Unter $H_0 : F(x) = G(x)\ \forall\, x \in \mathbb{R}$ ist zu erwarten, dass die X_i's und Y_j's in $(X'_{(1)}, \ldots, X'_{(n)})$ „gut gemischt” sind,
  - weil dann die Stichprobenvariablen $X'_{(1)}, \ldots, X'_{(n)}$ unabhängig und identisch verteilt sind.
  - Wenn als Alternative die Tendenz zur „Klumpen– bzw. Clusterbildung” betrachtet wird, dann wird H_0 abgelehnt, wenn die Anzahl T der Iterationen in der (binären) Stichprobe $(Z_1, \ldots, Z_n)$ „zu klein” ist, wobei $Z_i = 0$, wenn $X'_{(i)} = X_j$ für ein $j \in \{1, \ldots, n\}$, und $Z_i = 1$, wenn $X'_{(i)} = Y_j$ für ein $j \in \{1, \ldots, n\}$.

Beispiel

- Im Rahmen einer medizinischen Studie über Schulanfänger wurde die Körpergröße von $n_1 = 8$ Mädchen und $n_2 = 10$ Jungen untersucht.
- Dabei ergaben sich die folgenden Messergebnisse:

x_i	117	121	122	124	125	126	128	132		
y_j	110	113	114	115	116	118	119	120	123	127

- Wenn wir diese Messwerte der Größe nach ordnen und dabei den Körpergrößen der Jungen jeweils eine 0 bzw. den Körpergrößen der Mädchen jeweils eine 1 zuordnen, dann ergibt sich die Folge

$$\omega = (0,0,0,0,0,1,0,0,0,1,1,0,1,1,1,0,1,1) \qquad \text{mit } T(\omega) = 8. \tag{29}$$

- Andererseits ergibt sich aus Theorem 6.1, dass für das α–Quantil $r_\alpha(n_1; n_2)$ der Verteilung von T beispielsweise $r_{0.05}(8;10) = 6$ für $\alpha = 0.05$ gilt.
- In diesem Fall wird somit H_0 nicht abgelehnt, weil $T(\omega) = 8 > 6 = r_{0.05}(8;10)$.

Beachte

- Der in diesem Abschnitt betrachtete Iterationstest kann einseitige Alternativen vom Typ (27) bzw. (28) nicht erkennen.
- Dies wird durch das in (29) gegebene Beispiel klar: Denn die Anzahl der Iterationen $T(\omega) = 8$ ändert sich nicht, wenn wir (umgekehrt zu der bisherigen Vorgehensweise) den Körpergrößen der Jungen jeweils eine 1 bzw. den Körpergrößen der Mädchen jeweils eine 0 zuordnen.
- Auch bei zweiseitigen Alternativen sollte der Iterationstest von Wald–Wolfowitz, der ein so genannter „Omnibustest" ist, nur dann verwendet werden, wenn die Form der Alternative nicht näher spezifiziert wird.
- Beim Vorliegen spezieller Alternativen, die beispielsweise nur Lage– oder Variabilitätskenngrößen betreffen, sind andere Testverfahren effizienter, vgl. Abschnitt 6.3.2.

6.3.2 Rangsummentest von Wilcoxon für Lagealternativen

- Wir diskutieren nun einen weiteren nichtparametrischen Test für den Fall, dass zwei unabhängige Zufallsstichproben $(X_1, \ldots, X_{n_1})$ und $(Y_1, \ldots, Y_{n_2})$ beobachtet werden.
- Dabei werden jedoch jetzt speziellere Alternativen als in (26) – (28) betrachtet.
  - Wir nehmen an, dass die Zufallsvariablen $X_1, \ldots, X_{n_1}$ und $Y_1, \ldots, Y_{n_2}$ vollständig unabhängig sind mit den (unbekannten) stetigen Verteilungsfunktionen F bzw. G.
  - Ähnlich wie in Abschnitt 6.2 wird vorausgesetzt, dass es ein $\delta \in \mathbb{R}$ gibt, so dass

$$F(x) = G(x+\delta) \qquad \forall\, x \in \mathbb{R}\,.$$

  - Ein (zweiseitiges) Testproblem, das dem oben erwähnten, allgemeineren Testproblem (26) entspricht, ist dann gegeben durch

$$H_0 : \delta = 0 \qquad \text{vs.} \qquad H_1 : \delta \neq 0\,. \tag{30}$$

  - Als einseitige Alternativen zu $H_0 : \delta = 0$ können die folgenden Hypothesen betrachtet werden:

$$H_1 : \delta > 0 \qquad \text{bzw.} \qquad H_1 : \delta < 0\,. \tag{31}$$

- Genauso wie in Abschnitt 6.3.1 vereinigen wir die Stichprobenvariablen $X_1, \ldots, X_{n_1}$ und $Y_1, \ldots, Y_{n_2}$ zu einer kombinierten Zufallsstichprobe $(X'_1, \ldots, X'_n) = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$, wobei $n = n_1 + n_2$.
  - Außerdem betrachten wir den (Zufalls–) Vektor der Ränge $\mathbf{R}' = (R'_1, \ldots, R'_n)$ der Stichprobenvariablen $X'_1, \ldots, X'_n$ in der kombinierten Stichprobe, wobei

$$R'_i = \sum_{j=1}^{n} \mathbf{1}_{\{X'_j \le X'_i\}} \qquad \forall\, i = 1, \ldots, n\,.$$

- So wie in Abschnitt 6.3.1 ist unter $H_0 : \delta = 0$ zu erwarten, dass die X_i's und Y_j's in der kombinierten Stichprobe $(X'_{(1)}, \ldots, X'_{(n)})$ „gut gemischt" sind, weil dann die Stichprobenvariablen $X'_{(1)}, \ldots, X'_{(n)}$ unabhängig und identisch verteilt sind.
- Daher wird H_0 bei dem zweiseitigen Testproblem in (30) abgelehnt, wenn die *Rangsumme*

$$T_{n_1,n_2} = \sum_{i=1}^{n_1} R'_i \tag{32}$$

„zu klein" oder „zu groß" ist.

- Um den Test praktisch durchführen zu können, muss die Verteilung der in (32) eingeführten Teststatistik T_{n_1,n_2} bestimmt werden. Hierfür ist der folgende Hilfssatz nützlich.

Lemma 6.3

- *Sei $X : \Omega \to \{\ldots, -1, 0, 1, \ldots\}$ eine diskrete Zufallsvariable, so dass $\mathbb{E}\,|X| < \infty$ und dass für ein $\mu \in \mathbb{R}$ die folgende Symmetrieeigenschaft erfüllt ist:*

$$\mathbb{P}(X = \mu - k) = \mathbb{P}(X = \mu + k) \qquad \forall\, k \in \{\ldots, -1, 0, 1, \ldots\}\,. \tag{33}$$

- *Dann gilt $\mathbb{E}\,X = \mu$.*

Beweis

- Wir können o.B.d.A. annehmen, dass $\mu = 0$, weil ansonsten die transformierte Zufallsvariable $X' = X - \mu$ betraxchtet werden kann.
- Dann ergibt sich aus (33) mit $\mu = 0$, dass

$$\mathbb{E}\,X = \sum_{k=-\infty}^{\infty} k\,\mathbb{P}(X = k) \;=\; -\sum_{k=1}^{\infty} k\,\mathbb{P}(X = -k) \;+\; \sum_{k=1}^{\infty} k\,\mathbb{P}(X = k) \overset{(33)}{=} 0\,.$$

□

Theorem 6.5

- *Unter $H_0 : \delta = 0$ ist die Verteilung von T_{n_1,n_2} gegeben durch*

$$\mathbb{P}(T_{n_1,n_2} = k) \;=\; \frac{a_{k,n_1,n_2}}{\binom{n_1+n_2}{n_1}} \qquad \forall\, k = \frac{n_1(n_1+1)}{2}, \ldots, n_1 n_2 + \frac{n_1(n_1+1)}{2}\,, \tag{34}$$

wobei

$$a_{k,n_1,n_2} = \#\Big\{\mathbf{z} = (z_1, \ldots, z_{n_1+n_2}) \in \{0,1\}^{n_1+n_2} : \#\{i : z_i = 1\} = n_1,\ \sum_{i=1}^{n_1+n_2} i z_i = k\Big\}\,. \tag{35}$$

- *Außerdem gilt dann*

$$\mathbb{P}(T_{n_1,n_2} = k) \;=\; \mathbb{P}(T_{n_1,n_2} = 2\mu - k) \qquad \forall\, k \in \{\ldots, -1, 0, 1, \ldots\} \tag{36}$$

und somit

$$\mathbb{E}\,T_{n_1,n_2} = \mu\,, \tag{37}$$

wobei $\mu = n_1(n_1 + n_2 + 1)/2$.

Beweis

- Unter $H_0 : \delta = 0$ sind die Stichprobenvariablen $X'_1, \ldots, X'_{n_1+n_2}$ unabhängig und identisch verteilt.
  - Somit hat jede der $\binom{n_1+n_2}{n_1}$ Aufteilungen der n_1 Variablen $X_1, \ldots, X_{n_1}$ auf die $n_1 + n_2$ insgesamt vorhandenen Rangplätze die gleiche Wahrscheinlichkeit.
  - Außerdem gilt für den Minimal– bzw. Maximalwert $t_{\min}$ bzw. $t_{\max}$ von T_{n_1,n_2}, dass
$$t_{\min} = \sum_{i=1}^{n_1} i \;=\; \frac{n_1(n_1+1)}{2} \qquad \text{und} \qquad t_{\max} = \sum_{i=n_2+1}^{n_2+n_1} i \;=\; n_1 n_2 \;+\; \frac{n_1(n_1+1)}{2}\,.$$
  - Hieraus ergibt sich die Gültigkeit von (34) – (35).
- Um (36) zu beweisen, nutzen wir die folgende Symmetrieeigenschaft.
  - Jedem $\mathbf{z} = (z_1, \ldots, z_{n_1+n_2}) \in \{0,1\}^{n_1+n_2}$ mit
$$\#\{i : z_i = 1\} = n_1 \qquad \text{und} \qquad \sum_{i=1}^{n_1+n_2} i z_i = k$$
    entspricht ein $\widetilde{\mathbf{z}} = (\widetilde{z}_1, \ldots, \widetilde{z}_{n_1+n_2}) \in \{0,1\}^{n_1+n_2}$ mit
$$\#\{i : \widetilde{z}_{n_1+n_2+1-i} = 1\} = n_1 \qquad \text{und} \qquad \sum_{i=1}^{n_1+n_2} (n_1+n_2+1-i)\,\widetilde{z}_i = n_1(n_1+n_2+1) - k\,.$$
  - Weil die Stichprobenvariablen $X'_{(1)}, \ldots, X'_{(n)}$ unabhängig und identisch verteilt sind, ergibt sich somit, dass für jedes $k \in \{\ldots, -1, 0, 1, \ldots\}$
$$\mathbb{P}(T_{n_1,n_2} = k) = \mathbb{P}(T_{n_1,n_2} = n_1(n_1+n_2+1) - k) = \mathbb{P}(T_{n_1,n_2} = 2\mu - k)\,, \tag{38}$$
    wobei $2\mu = n_1(n_1+n_2+1)$.
- Um (37) zu zeigen, genügt es in (38) die Substitution $k = \mu - i$ einzusetzen.
  - Dann ergibt sich aus (38), dass
$$\mathbb{P}(T_{n_1,n_2} = \mu - i) = \mathbb{P}(T_{n_1,n_2} = \mu + i) \qquad \forall\, i \in \{\ldots, -1, 0, 1, \ldots\}\,.$$
  - Hieraus und aus Lemma 6.3 folgt die Gültigkeit von (37). □

Beachte

- Aus (38) ergibt sich die folgende Symmetrieeigenschaft für die Quantile t_{α,n_1,n_2} von T_{n_1,n_2}.
  - Für jedes $\alpha \in (0,1)$ gilt
$$t_{\alpha,n_1,n_2} \;=\; n_1(n_1+n_2+1) - t_{1-\alpha,n_1,n_2}\,.$$
  - Die Quantile t_{α,n_1,n_2} können entweder aus Tabellen entnommen oder per Monte–Carlo–Simulation bestimmt werden.
- Die Nullhypothese $H_0 : \delta = 0$ wird zugunsten von $H_1 : \delta \neq 0$ abgelehnt, wenn
$$T_{n_1,n_2} \le t_{\alpha/2,n_1,n_2} \qquad \text{oder} \qquad T_{n_1,n_2} \ge n_1(n_1+n_2+1) - t_{\alpha/2,n_1,n_2}\,.$$
- Analog wird die Nullhypothese $H_0 : \delta = 0$ zugunsten von $H_1 : \delta > 0$ bzw. $H_1 : \delta < 0$ abgelehnt, wenn
$$T_{n_1,n_2} \ge n_1(n_1+n_2+1) - t_{\alpha,n_1,n_2} \qquad \text{bzw.} \qquad T_{n_1,n_2} \le t_{\alpha,n_1,n_2}\,.$$

Wenn die Stichprobenumfänge n_1 und n_2 groß sind, dann ist die direkte Bestimmung der Quantile t_{α,n_1,n_2} mit Hilfe von Theorem 6.5 schwierig. Die (näherungsweise) Bestimmung der Verteilung der Teststatistik T_{n_1,n_2} ist dann jedoch mit Hilfe des folgenden zentralen Grenzwertsatzes, den wir hier ohne Beweis angeben.

Theorem 6.6 *Wenn $n_1, n_2 \to \infty$, so dass $n_1/(n_1+n_2) \to p$ bzw. $n_2/(n_1+n_2) \to 1-p$ für ein $p \in (0,1)$, dann gilt*

$$\lim_{n_1,n_2\to\infty} \mathbb{P}\Big(\frac{T_{n_1,n_2} - \mathbb{E}\,T_{n_1,n_2}}{\sqrt{\operatorname{Var} T_{n_1,n_2}}} \le x\Big) = \Phi(x) \qquad \forall\, x \in \mathbb{R}\,, \tag{39}$$

wobei

$$\mathbb{E}\,T_{n_1,n_2} = \frac{n_1(n_1+n_2+1)}{2}\,, \qquad \operatorname{Var} T_{n_1,n_2} = \frac{n_1 n_2(n_1+n_2+1)}{12}$$

und $\Phi : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der $\mathrm{N}(0,1)$–Verteilung ist.

Beachte

- Wegen Theorem 6.6 wird für große n_1, n_2 die Nullhypothese $H_0 : \delta = 0$ zugunsten von $H_1 : \delta \neq 0$ abgelehnt, wenn
$$\Big|\frac{T_{n_1,n_2} - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2(n_1+n_2+1)/12}}\Big| \ge z_{1-\alpha/2}\,,$$
wobei z_α das α–Quantil der $\mathrm{N}(0,1)$–Verteilung ist.
- Analog wird $H_0 : \delta = 0$ zugunsten von $H_1 : \delta > 0$ bzw. $H_1 : \delta < 0$ abgelehnt, wenn
$$\frac{T_{n_1,n_2} - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2(n_1+n_2+1)/12}} \ge z_{1-\alpha} \qquad \text{bzw.} \qquad \frac{T_{n_1,n_2} - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2(n_1+n_2+1)/12}} \le -z_{1-\alpha}\,.$$