Lineare Modell

Prof. Dr. Hajo Holzmann
Fb 12, Universität Marburg

18. April 2011

Inhaltsverzeichnis

1	Multivariate Normalverteilung und quadratische Formen	1
1.1	Zufallsvektoren und zufällige quadratische Formen	1
1.2	Die multivariate Normalverteilung	2
1.3	Bedingte Verteilungen	4
1.4	Aus der Normalverteilung abgeleitete Verteilungen	6
1.5	Verteilung quadratischer Formen	7
2	Lineare Regression	11
2.1	Lineare Regression und Modellbildung	11
2.1.1	Regression und das lineare Modell	11
2.1.2	Modellierung des Einflusses der Kovariablen	12
2.2	Kleinste Quadrate Schätzung	15
2.2.1	Methode der kleinsten Quadrate	15
2.2.2	Verteilungstheorie bei normalverteilten Fehlern	19
2.2.3	Asymptotische Verteilungstheorie	22
2.3	Andere Schätzverfahren	26
2.3.1	Mittlerer quadratischer Fehler und Ridge-Regression	26
2.3.2	Median - und Quantilregression	28
2.4	Modelldiagnostik	29
2.4.1	Das Bestimmtheitsmaß R^2	29
2.4.2	Residuen-Analyse	30
2.4.3	Weitere diagnostische Plots	31
2.4.4	Transformation	31
2.4.5	Ausreißer, Leverage- und Influenz-Punkte	33

2.5	Variablen-Selektion und Shrinkage Schätzung	37
2.5.1	Selektionsmethoden	37
2.5.2	Test-basierte Methoden	38
2.5.3	Informationskriterien	38
2.5.4	Shrinkage Schätzer	45
2.5.5	Inferenz nach Modellwahl	48
2.6	Verallgemeinerte kleinste Quadrate und Varianzschätzen	52
2.6.1	Verallgemeinerte kleinste Quadrate	52
2.6.2	Varianzschätzen	53

1 Multivariate Normalverteilung und quadratische Formen

1.1 Zufallsvektoren und zufällige quadratische Formen

Sei $\mathbf{X} = (X_1, \dots, X_n)^T \in \mathbb{R}^d$ ein d -variater Zufallsvektor, wobei X_i Zufallsvariable seien. Der Erwartungswertvektor von $\mathbf{X}$ ist definiert durch $E\mathbf{X} = (EX_1, \dots, EX_n)^T$, falls die Erwartungswerte EX_i existieren. Die Kovarianzmatrix von $\mathbf{X}$ ist gegeben durch

$$\text{Cov } \mathbf{X} = (\text{Cov}(X_i, X_j))_{i,j=1,\dots,n},$$

falls die X_i endliche Varianzen haben. Für einen Vektor $\mathbf{a} \in \mathbb{R}^d$ gilt

$$\text{Var}(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \text{Cov } \mathbf{X} \mathbf{a}.$$

Da die Varianz auf der linken Seite stets nicht-negativ ist, folgt, dass die Kovarianzmatrix stets positiv semidefinit ist. Weiter ist $\text{Cov } \mathbf{X}$ genau dann degeneriert, falls die X_i (als Abbildungen auf dem zugrundeliegenden W-Raum) linear abhängig sind (fast sicher).

Satz 1.1 (lineare Transformationen)

Sei $\mathbf{X} \in \mathbb{R}^n$ ein Zufallsvektor mit endlichem Erwartungswertvektor $E\mathbf{X}$ und endlicher Kovarianzmatrix $\text{Cov } \mathbf{X}$. Für $A \in \mathbb{R}^{m \times n}$ gilt dann

$$E(A\mathbf{X}) = A E\mathbf{X}, \quad \text{Cov}(A\mathbf{X}) = A \text{Cov } \mathbf{X} A^T.$$

Der Beweis ist ein einfaches Nachrechnen. Allgemeiner definieren wir für Zufallsvektoren $X \in \mathbb{R}^d$ und $Y \in \mathbb{R}^q$ die Kovarianzmatrix

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = (\text{Cov}(X_i, Y_j))_{i=1,\dots,d, j=1,\dots,q} \in \mathbb{R}^{d \times q}.$$

Falls $\mathbf{X}$ und $\mathbf{Y}$ unabhängig sind, so gilt (analog zu zwei Zufallsvariablen) $\text{Cov}(\mathbf{X}, \mathbf{Y}) = 0$.

Für einen Zufallsvektor $\mathbf{X} \in \mathbb{R}^d$ und eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ betrachten wir die zufällige quadratische Form $\mathbf{X}^T A \mathbf{X}$, eine Zufallsvariable.

Satz 1.2 (Erwartungswert quadratischer Formen)

Sei $\mathbf{X} \in \mathbb{R}^n$ ein Zufallsvektor mit $E\mathbf{X} = \boldsymbol{\mu} \in \mathbb{R}^n$, $\text{Cov } \mathbf{X} = \Sigma \in \mathbb{R}^{n \times n}$, und sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Betrachte die quadratische Form $Q = \mathbf{X}^T A \mathbf{X}$. Dann gilt

$$EQ = \text{Spur}(A \cdot \Sigma) + \boldsymbol{\mu}^T A \boldsymbol{\mu}$$

Beweis

$$Q = \sum_{i,j=1}^n a_{ij} X_i X_j, \quad EX_i X_j = \Sigma_{ij} + \mu_i \mu_j$$

Somit

$$EQ = \sum_{i,j=1}^n a_{ij} (\Sigma_{ij} + \mu_i \mu_j) = \boldsymbol{\mu}^T A \boldsymbol{\mu} + \sum_{i,j=1}^n a_{ij} \Sigma_{ij} = \boldsymbol{\mu}^T A \boldsymbol{\mu} + \text{Spur}(A \Sigma)$$

[Beachte: Sowohl A als auch Σ sind symmetrisch.] ■

1.2 Die multivariate Normalverteilung

Definition

Sei $\boldsymbol{\mu} \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d \times d}$ symmetrisch und positiv definit. Die **multivariate Normalverteilung** $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ ist gegeben durch die Dichte

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

wobei $|\Sigma| = \det \Sigma$.

Für $\boldsymbol{\mu} = 0$ und $\Sigma = I_d$ spricht man von der d -variaten Standardnormalverteilung. Wir prüfen kurz, dass $f(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ tatsächlich eine Dichte auf $\mathbb{R}^d$ definiert, also zu 1 integriert. Für $f(\mathbf{x}; 0, I_d)$ ist dies klar, da es sich um eine d -fache Produktdichte von univariaten Dichten der Standardnormalverteilung handelt. Den allgemeinen Fall führen wir hierauf durch Variablen Transformation zurück. Dazu betrachte die Spektralzerlegung von Σ . Es existieren eine orthogonale Matrix Q und $\lambda_1 > 0, \dots, \lambda_d > 0$ (die Eigenwerte von Σ , diese sind > 0 , da Σ positiv definit ist), so dass

$$\Sigma = Q \operatorname{diag}(\lambda_1, \dots, \lambda_d) Q^T.$$

Man definiert dann die Wurzel aus Σ durch

$$\Sigma^{\frac{1}{2}} = Q \operatorname{diag}(\lambda_1^{\frac{1}{2}}, \dots, \lambda_d^{\frac{1}{2}}) Q^T,$$

diese ist eindeutig bestimmt und es gilt $\Sigma^{\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}} = \Sigma$. Ausserdem gelten

$$\begin{aligned} \Sigma^{-1} &= Q \operatorname{diag}(\lambda_1^{-1}, \dots, \lambda_d^{-1}) Q^T, \\ \Sigma^{-\frac{1}{2}} &= (\Sigma^{\frac{1}{2}})^{-1} = Q \operatorname{diag}(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_d^{-\frac{1}{2}}) Q^T. \end{aligned}$$

Zur Berechnung von $\int f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) d\mathbf{x}$ führt man eine Variablentransformation $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ durch und erhält die d -variate Standardnormalverteilung.

Satz 1.3 (Lineare Transformation)

Sei $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $p \leq d$, $A \in \mathbb{R}^{p \times d}$ mit vollem Rang und $\mathbf{b} \in \mathbb{R}^p$. Dann ist $\mathbf{Y} = A\mathbf{X}$ verteilt nach $\mathbf{Y} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

Beweis

Wir behandeln nur den Fall $\boldsymbol{\mu} = 0$, $\mathbf{b} = 0$, die nötigen Translationen können leicht behandelt werden.

Wir betrachten zunächst den Fall $\Sigma = I_d$. Ist etwa $A^T = (\mathbf{v}_1, \dots, \mathbf{v}_p)$, $\mathbf{v}_i \in \mathbb{R}^d$, so setze $V = \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_p\} \subset \mathbb{R}^d$. Dann ist V ein p -dimensionaler Untervektorraum von $\mathbb{R}^d$, und wir können eine orthonormale Basis $\mathbf{e}_1, \dots, \mathbf{e}_{d-p}$ des orthogonalen Komplements $V^\perp$ wählen. Setze $\tilde{A}$, derart, dass $\tilde{A}^T = (\mathbf{v}_1, \dots, \mathbf{v}_p, \mathbf{e}_1, \dots, \mathbf{e}_{d-p}) \in \mathbb{R}^{d \times d}$. Dann hat $\tilde{A}$ vollen Rang. Wir betrachten nun $\mathbf{W} = \tilde{A}\mathbf{X}$. Dann ist $\mathbf{W} = (\mathbf{Y}^T, \mathbf{Z}^T)^T$ für geeignetes $\mathbf{Z} \in \mathbb{R}^d$. Zum Beweis können wir also die Verteilung von $\mathbf{W}$ und dann die Randverteilung von $\mathbf{Y}$ berechnen. Eine Variablentransformation liefert für die Dichte von $\mathbf{W}$

$$f_{\mathbf{W}}(\mathbf{w}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\tilde{A}|} \exp\left(-\frac{1}{2}\mathbf{w}^T (\tilde{A}\tilde{A}^T)^{-1}\mathbf{w}\right).$$

Nach Konstruktion von $\tilde{A}$ erhalten wir die Blockgestalt (I_r ist die r -dimensionale Einheitsmatrix)

$$\tilde{A}\tilde{A}^T = \begin{pmatrix} AA^T & \mathbf{0} \\ \mathbf{0} & I_{d-p} \end{pmatrix} =: \text{diag}(AA^T, I_{d-p}).$$

Ausserdem ist

$$|\tilde{A}| = |\tilde{A}\tilde{A}^T|^{1/2} = |AA^T|^{1/2}.$$

Setzen wir $\mathbf{w} = (\mathbf{y}^T, \mathbf{z}^T)^T$, so erhalten wir

$$f_{\mathbf{w}}(\mathbf{y}, \mathbf{z}) = \frac{1}{(2\pi)^{p/2} |AA^T|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{y}^T (AA^T)^{-1} \mathbf{y}\right) \frac{1}{(2\pi)^{(d-p)/2}} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z}\right).$$

Wir können nun $\mathbf{z}$ herausintegrieren (Standardnormalverteilung) und erhalten für $\mathbf{Y}$ die Dichte in gewünschter Form.

Den Fall für allgemeines Σ führt man nun leicht auf obigen Fall zurück. Dafür setze $\tilde{\mathbf{X}} = \Sigma^{-1/2} \mathbf{X}$, dann ist $\tilde{\mathbf{X}} \sim N(0, I_d)$, und $\mathbf{Y} = A\Sigma^{1/2} \tilde{\mathbf{X}} \sim N(0, A\Sigma^{1/2} \Sigma^{1/2} A^T)$ nach dem bereits behandelten Fall, wie behauptet. ■

Momente Ist $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, so ist $E\mathbf{X} = \boldsymbol{\mu}$, $\text{Cov } \mathbf{X} = \Sigma$. Zum Beweis benutzen wir Satz 1.3. Zunächst bemerken wir, dass die Formeln für die d -variate Standardnormalverteilung richtig sind (direkt aus der univariaten Normalverteilung). Ist $\mathbf{Y} \sim N(0, I_d)$, so ist $\mathbf{X} = \Sigma^{1/2} \mathbf{Y} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \Sigma)$. Mit Satz 1.1 folgt die Behauptung.

Marginalverteilungen und Unabhängigkeit Sei $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, wobei $\mathbf{X} = (\mathbf{Y}^T, \mathbf{Z}^T)^T$, $\mathbf{Y} \in \mathbb{R}^r$, $\mathbf{Z} \in \mathbb{R}^{d-r}$, für $1 \leq r < d$. Sei

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

mit $\boldsymbol{\mu}_1 \in \mathbb{R}^r$, $\boldsymbol{\mu}_2 \in \mathbb{R}^{d-r}$, $\Sigma_{11} \in \mathbb{R}^{r \times r}$, $\Sigma_{22} \in \mathbb{R}^{(d-r) \times (d-r)}$, $\Sigma_{12} \in \mathbb{R}^{r \times (d-r)}$.

Satz 1.4

Es sind $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$, $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_{22})$, und $\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \Sigma_{12}$. Weiter gilt

$$\mathbf{Y} \text{ und } \mathbf{Z} \text{ unabhängig} \Leftrightarrow \Sigma_{12} = 0.$$

Beweis

Für die Verteilung von $\mathbf{Y}$ verwende Satz 1.3 mit $A = (I_r, 0) \in \mathbb{R}^{r \times d}$, $\mathbf{Y} = A\mathbf{X}$, für $\mathbf{Z}$ nutze $\mathbf{Z} = B\mathbf{X}$, $B = (0, I_{d-r}) \in \mathbb{R}^{(d-r) \times d}$. Für den zweiten Teil beachte zunächst, dass unabhängige Zufallsvektoren Kovarianzmatrix 0 haben. Falls $\Sigma_{12} = 0$, so ist $\Sigma = \text{diag}(\Sigma_1, \Sigma_2)$ eine Blockmatrix mit inverser $\Sigma^{-1} = \text{diag}(\Sigma_1^{-1}, \Sigma_2^{-1})$, und die Dichte $f(\mathbf{y}, \mathbf{z})$ von $(\mathbf{Y}^T, \mathbf{Z}^T)^T$ berechnet sich leicht zur Produktdichte der beiden multivariaten Normalverteilungen, was die Unabhängigkeit zeigt. ■

Satz 1.5

Sei $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $p, q \leq d$, $B \in \mathbb{R}^{p \times d}$ mit vollem Rang, $C \in \mathbb{R}^{q \times d}$ mit vollem Rang. Dann gilt:

$$B\mathbf{X} \text{ und } C\mathbf{X} \text{ sind unabhängig} \Leftrightarrow B\Sigma C^T = 0$$

Beweis

„ $\Rightarrow$ “. Klar, da

$$\text{Cov}(B\mathbf{X}, C\mathbf{X}) = E(B(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T C^T) = B\Sigma C^T = 0$$

bei Unabhängigkeit.

„ $\Leftarrow$ “. Betrachte zunächst den Fall $\Sigma = I_d$. Schreibe

$$B = (\mathbf{b}_1, \dots, \mathbf{b}_p)^T, \quad C = (\mathbf{c}_1, \dots, \mathbf{c}_q)^T, \quad \mathbf{b}_i, \mathbf{c}_j \in \mathbb{R}^d$$

Dann: $B\Sigma C^T = BC^T = 0 \Leftrightarrow \mathbf{b}_i^T \mathbf{c}_j = 0$. Somit hat

$$A = (\mathbf{b}_1, \dots, \mathbf{b}_p, \mathbf{c}_1, \dots, \mathbf{c}_q)^T$$

vollen Rang und

$$A\mathbf{X} = \begin{pmatrix} B\mathbf{X} \\ C\mathbf{X} \end{pmatrix} \sim \mathcal{N}(A\boldsymbol{\mu}, AA^T) = \mathcal{N}\left(A\boldsymbol{\mu}, \begin{pmatrix} BB^T & 0 \\ 0 & CC^T \end{pmatrix}\right)$$

da $\Sigma = I_d$ und $BC^T = 0$. Also sind $B\mathbf{X}$ und $C\mathbf{X}$ unabhängig.

Zum allgemeinen Fall: Betrachte $\mathbf{Y} = \Sigma^{-\frac{1}{2}}\mathbf{X}$ mit $\mathbf{Y} \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}\boldsymbol{\mu}, \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}) = \mathcal{N}(\Sigma^{-\frac{1}{2}}\boldsymbol{\mu}, I_d)$. Weiter

$$B\mathbf{X} = \underbrace{B\Sigma^{\frac{1}{2}}}_{=: \tilde{B}} \mathbf{Y}, \quad C\mathbf{X} = \underbrace{C\Sigma^{\frac{1}{2}}}_{=: \tilde{C}} \mathbf{Y}$$

und $\tilde{B}\tilde{C}^T = B\Sigma^{\frac{1}{2}}(\Sigma^{\frac{1}{2}})^T C^T = B\Sigma C^T = 0 \Leftrightarrow \tilde{B}\tilde{C}^T = 0$ und $\tilde{B}\mathbf{Y} = B\mathbf{X}$ und $\tilde{C}\mathbf{Y} = C\mathbf{X}$ nach obigem Fall unabhängig. $\blacksquare$

Relevante R Befehle. `eigen` (berechnet Eigenwerte und Vektoren einer Matrix), `qr.solve` (berechnet Inverse einer Matrix). Die Library `MASS` enthält die Funktion `mvrnorm` (Erzeugung von multivariaten normalverteilten Zufallszahlen). Die Dichte muss per Hand berechnet werden.

1.3 Bedingte Verteilungen

Wir wiederholen zunächst die *bedingten Dichte* bei Zufallsvektoren. Es sei $\mathbf{X} = (\mathbf{Y}^T, \mathbf{Z}^T)^T \in \mathbb{R}^d$ mit Dichte $f(\mathbf{y}, \mathbf{z})$, $\mathbf{Y} \in \mathbb{R}^r$, $\mathbf{Z} \in \mathbb{R}^{d-r}$, dann hat $\mathbf{Y}|\mathbf{Z} = \mathbf{z}$ die Dichte

$$f(\mathbf{y}|\mathbf{z}) = \frac{f(\mathbf{y}, \mathbf{z})}{f(\mathbf{z})}. \quad (1)$$

Im Folgenden sei wie oben $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, wobei $\mathbf{X} = (\mathbf{Y}^T, \mathbf{Z}^T)^T$, $\mathbf{Y} \in \mathbb{R}^r$, $\mathbf{Z} \in \mathbb{R}^{d-r}$, für $1 \leq r < d$. Sei

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

mit $\boldsymbol{\mu}_1 \in \mathbb{R}^r$, $\boldsymbol{\mu}_2 \in \mathbb{R}^{d-r}$, $\Sigma_{11} \in \mathbb{R}^{r \times r}$, $\Sigma_{22} \in \mathbb{R}^{(d-r) \times (d-r)}$, $\Sigma_{12} \in \mathbb{R}^{r \times (d-r)}$.

Satz 1.6

Es ist $\mathbf{Y}|\mathbf{Z} = \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}}, \Sigma_{1|2})$, wobei

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}} &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_2^{-1}(\mathbf{z} - \boldsymbol{\mu}_2), \\ \Sigma_{1|2} &= \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}.\end{aligned}$$

Die Matrix $\Sigma_{1|2}$ heisst das *Schur Komplement*, diese ist positiv definit, falls Σ positiv definit ist.

Beweis

Nach (1) ist

$$f(\mathbf{y}|\mathbf{z}) = \frac{\exp(-\frac{1}{2}((\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) - (\mathbf{z} - \boldsymbol{\mu}_2)^T \Sigma_2^{-1}(\mathbf{z} - \boldsymbol{\mu}_2)))}{\left(\frac{|\Sigma|}{|\Sigma_2|}\right)^{\frac{1}{2}} \cdot (2\pi)^{\bar{2}}}.$$

Wir berechne Σ^{-1} in Blockform durch Spaltenumformungen¹.

$$\begin{array}{ccc} \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix} & & \begin{pmatrix} I_1 & 0 \\ 0 & I_2 \end{pmatrix} \\ & \Sigma_2^{-1} \cdot 2 & \\ & \downarrow & \\ \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_2^{-1}\Sigma_{21} & I_2 \end{pmatrix} & & \begin{pmatrix} I_1 & 0 \\ 0 & \Sigma_2^{-1} \end{pmatrix} \\ & 1 - \Sigma_{12} \cdot 2 & \\ & \downarrow & \\ \begin{pmatrix} \Sigma_{1|2} & 0 \\ \Sigma_2^{-1}\Sigma_{21} & I_2 \end{pmatrix} & & \begin{pmatrix} I_1 & -\Sigma_{12}\Sigma_2^{-1} \\ 0 & \Sigma_2^{-1} \end{pmatrix} \\ & \Sigma_{1|2}^{-1} \cdot 1 & \\ & \downarrow & \\ \begin{pmatrix} I_1 & 0 \\ \Sigma_2^{-1}\Sigma_{21} & I_2 \end{pmatrix} & & \begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_2^{-1} \\ 0 & \Sigma_2^{-1} \end{pmatrix} \\ & 2 - \Sigma_2^{-1}\Sigma_{21} \cdot 1 & \\ & \downarrow & \\ \begin{pmatrix} I_1 & 0 \\ 0 & I_2 \end{pmatrix} & & \begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_2^{-1} \\ -\Sigma_2^{-1}\Sigma_{21}\Sigma_{1|2}^{-1} & \Sigma_2^{-1} - \Sigma_2^{-1}\Sigma_{21}\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_2^{-1} \end{pmatrix} \end{array}$$

Somit ist

$$\begin{aligned}& (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) - (\mathbf{z} - \boldsymbol{\mu}_2)^T \Sigma_2^{-1}(\mathbf{z} - \boldsymbol{\mu}_2) \\ &= \left(\mathbf{y} - \boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_2^{-1}(\mathbf{z} - \boldsymbol{\mu}_2) \right)^T \Sigma_{1|2}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_2^{-1}(\mathbf{z} - \boldsymbol{\mu}_2) \right).\end{aligned}$$

Wegen $|\Sigma^{-1}| = |\Sigma_2^{-1}| \cdot |\Sigma_{1|2}^{-1}|$ folgt der Satz. ■

¹Gaußscher Algorithmus ($(\Sigma|I)$ wird umgeformt zu $(I|\Sigma^{-1})$)

Regression to the mean

Für

$$\boldsymbol{\mu} = (\mu_1, \mu_2)^T, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad -1 < \rho < 1,$$

sei $(X, Y) \sim N(\boldsymbol{\mu}, \Sigma)$ bivariat normalverteilt. Dann gilt nach Satz 1.6 für die bedingte Verteilung

$$Y|X = x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), (1 - \rho^2)\sigma_2^2\right).$$

Insbesondere ist für

$$\sigma_1^2 = \sigma_2^2 = \sigma^2, \quad \mu_1 = \mu_2 = \mu, \quad (2)$$

$E(Y|X = x) = \mu + \rho(x - \mu)$ und somit für $0 < \rho < 1$,

$$\mu < E(Y|X = x) < x \quad \text{falls } x > \mu, \quad x < E(Y|X = x) < \mu \quad \text{falls } x < \mu.$$

Selbst wenn also X und Y positiv korreliert sind, nähert sich im bedingten Mittel Y wieder μ an. Analog gilt für die Änderung $Y - X|X$ unter (2),

$$Y - X|X = x \sim N((1 - \rho)(\mu - x), (1 - \rho^2)\sigma^2).$$

Also hat der Regressionskoeffizient von der Änderung $Y - X$ (abhängige Variable) regressiert auf X (Kovariable) negatives Vorzeichen auch bei positivem ρ .

1.4 Aus der Normalverteilung abgeleitete Verteilungen

a. Ist $\mathbf{X} \sim N(\boldsymbol{\mu}, I_d)$, so hat $\mathbf{X}^T \mathbf{X} = \sum_{i=1}^d X_i^2$ die **nichtzentrale χ^2 -Verteilung** mit d Freiheitsgraden und Nichtzentralitätsparameter $\frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\mu}$. Schreibweise: $\chi^2(d; \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\mu})$. In der Tat hängt die Dichte von $X^T X$,

$$f(u) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{k!} \frac{u^{\frac{1}{2}d+k-1} e^{-\frac{1}{2}u}}{2^{\frac{1}{2}d+k} \Gamma(\frac{1}{2}d+k)}, \quad \lambda = \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\mu},$$

nur von λ und nicht von ganz $\boldsymbol{\mu}$ ab. Für $\lambda = 0$ (bzw. $\boldsymbol{\mu} = 0$) ergibt sich die zentrale χ^2 -Verteilung mit d Freiheitsgraden, Bezeichnung $\chi^2(d)$, diese hat die Dichte (Beweis!)

$$f_{\chi^2, d}(u) = \frac{u^{d/2-1} e^{-u/2}}{2^{d/2} \Gamma(d/2)}. \quad (3)$$

Man kann zeigen, dass (3) auch für nicht ganzes d eine Dichte definiert, daher kann man die Freiheitsgrade in $(0, \infty)$ variieren lassen.

Weitere Notation:

$\chi_\alpha^2(n)$: Das α -Quantil der zentralen χ^2 Verteilung mit n Freiheitsgraden ($0 < \alpha < 1$).

$\chi^2(n)(x)$: Wert der Verteilungsfunktion der zentralen χ^2 Verteilung mit n Freiheitsgraden bei x ($x > 0$).

Relevante R Befehle. `dchisq` (Dichte), `pchisq` (Verteilungsfunktion), `qchisq` (Quantile) und

`rchisq` (Zufallszahlen).

b. Ist $U_1 \sim \chi^2(d_1; \lambda)$, $U_2 \sim \chi^2(d_2)$, U_1, U_2 unabhängig, so hat

$$V = \frac{U_1/d_1}{U_2/d_2} \sim F(d_1, d_2; \lambda)$$

die **nichtzentrale F-Verteilung** mit Freiheitsgraden d_1 und d_2 und Nichtzentralitätsparameter λ . Für $\lambda = 0$ erhält man die zentrale F-Verteilung, diese hat die Dichte (Beweis)

$$f_{F,d_1,d_2}(u) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma(d_1/2)\Gamma(d_2/2)} d_1^{d_1/2} d_2^{d_2/2} \frac{u^{d_1/2}}{(d_2 + d_1 u)^{(d_1+d_2)/2}}.$$

Weitere Notation:

$F_\alpha(n, m; \lambda)$: Das α -Quantil der F Verteilung mit n und m Freiheitsgraden ($0 < \alpha < 1$) und Nichtzentralitätsparameter λ .

$F(n, m; \lambda)(x)$: Wert der Verteilungsfunktion der F Verteilung mit n Freiheitsgraden bei x ($x > 0$) und Nichtzentralitätsparameter λ .

Relevante R Befehle. `df` (Dichte), `pf` (Verteilungsfunktion), `qf` (Quantile) und `rf` (Zufallszahlen).

c. Ist $X \sim \mathcal{N}(\mu, 1)$, $U \sim \chi^2(d)$, so hat

$$V = \frac{X}{\sqrt{U/d}}$$

die **t-Verteilung** mit d Freiheitsgraden und Nichtzentralitätsparameter μ , Bezeichnung $t(n; \mu)$. Für $\mu = 0$ erhält man die zentrale t -Verteilung, diese hat die Dichte (Beweis)

$$f_{t,d}(u) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{\pi d}} (1 + u^2/d)^{(d+1)/2}.$$

Weitere Notation:

$t_\alpha(n; \mu)$: Das α -Quantil der t Verteilung mit n und m Freiheitsgraden ($0 < \alpha < 1$) und Nichtzentralitätsparameter μ .

$t(n; \mu)(x)$: Wert der Verteilungsfunktion der t Verteilung mit n Freiheitsgraden und Nichtzentralitätsparameter μ bei x ($x > 0$).

Relevante R Befehle. `dt` (Dichte), `pt` (Verteilungsfunktion), `qt` (Quantile) und `rt` (Zufallszahlen).

Ist bei einer dieser Verteilung der Nichtzentralitätsparameter = 0, so lässt man diesen in der Notation einfach weg.

1.5 Verteilung quadratischer Formen

Satz 1.7

Sei $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $A \in \mathbb{R}^{d \times d}$ positiv semidefinit². Ist $A\Sigma$ idempotent, d.h. $(A\Sigma)^2 = A\Sigma$, so gilt

$$\mathbf{X}^T A \mathbf{X} \sim \chi^2(r(A), \frac{1}{2} \boldsymbol{\mu}^T A \boldsymbol{\mu})$$

²setzt Symmetrie voraus!

($r(A)$ ist der Rang von A)

Bemerkung

Es gilt auch die Rückrichtung.

Beweis

a. Zunächst sei wieder $\Sigma = I_d$. Wegen $A = A^2$ hat die Spektralzerlegung von A die Form

$$A = Q^T \underbrace{\text{diag}(1, \dots, 1, 0, \dots, 0)}_{r(A) \text{ mal}} Q$$

mit orthogonaler Matrix Q . Somit

$$\mathbf{X}^T \mathbf{A} \mathbf{X} = \mathbf{X}^T Q^T \underbrace{\text{diag}(1, \dots, 1, 0, \dots, 0)}_{=: D} \underbrace{Q \mathbf{X}}_{=: \mathbf{Y}} = \mathbf{Y}^T D \mathbf{Y} = Y_1^2 + \dots + Y_{r(A)}^2$$

wobei $\mathbf{Y} \sim \mathcal{N}(\underbrace{Q\boldsymbol{\mu}}_{=: \mathbf{v}}, I_d)$. Somit gilt:

$$\mathbf{X}^T \mathbf{A} \mathbf{X} \sim \chi^2(r(A), \underbrace{\frac{1}{2}(v_1^2 + \dots + v_{r(A)}^2)}_{=: \mathbf{v}^T D \mathbf{v} = \boldsymbol{\mu}^T A \boldsymbol{\mu}}) = \chi^2(r(A), \frac{1}{2} \boldsymbol{\mu}^T A \boldsymbol{\mu})$$

b. Allgemeiner Fall: Ist $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, so gilt $\mathbf{Y} = \Sigma^{-\frac{1}{2}} \mathbf{X} \sim \mathcal{N}(\Sigma^{-\frac{1}{2}} \boldsymbol{\mu}, I_d)$ und $\mathbf{X}^T \mathbf{A} \mathbf{X} = \mathbf{Y}^T \Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} \mathbf{Y}$.

Es ist $\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}$ idempotent, denn

$$\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} = \Sigma^{-\frac{1}{2}} \Sigma A \Sigma A \Sigma^{\frac{1}{2}} = \Sigma^{-\frac{1}{2}} \Sigma A \Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}.$$

Nach (a) gilt somit

$$\begin{aligned} \mathbf{X}^T \mathbf{A} \mathbf{X} &\sim \chi^2(r(\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}), \frac{1}{2} (\Sigma^{-\frac{1}{2}} \boldsymbol{\mu})^T \Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} (\Sigma^{-\frac{1}{2}} \boldsymbol{\mu})) \\ &= \chi^2(r(A), \frac{1}{2} \boldsymbol{\mu}^T A \boldsymbol{\mu}). \end{aligned}$$

da $\Sigma^{\frac{1}{2}}$ vollen Rang hat. ■

Beispiel 1.8

Es seien $X_1, \dots, X_n$ unabhängig und $N(\mu, \sigma^2)$ verteilt. Als Schätzer für Erwartungswert und Varianz betrachtet man

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Offenbar ist $\bar{X}_n \sim N(\mu, \sigma^2/n)$. Wir zeigen

$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1). \quad (4)$$

Dazu setze $\mathbf{1}_n = (1, \dots, 1)^T \in \mathbb{R}^n$ und $P_n = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ (die Zentrierungsmatrix). Für $\mathbf{X} = (X_1, \dots, X_n)^T$ ist in der Tat $P_n \mathbf{X} = (X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^T$. Weiter $P_n^2 = P_n$ (Beweis), also nach Satz 1.7

$$\frac{n-1}{\sigma^2} S_n^2 = \frac{1}{\sigma^2} \mathbf{X}^T P_n \mathbf{X} \sim \chi^2(r(P_n), \mu^2 \mathbf{1}_n^T P_n \mathbf{1}_n / 2).$$

Da $P_n^2 = P_n$ und $P_n^T = P_n$, ist $r(P_n) = \text{Spur } P_n = n - 1$. Ausserdem ist $P_n \mathbf{1}_n = 0$. Dies zeigt (4).

Satz 1.9 (Craig und Sakamoto)

Sei $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

a. Ist $A \in \mathbb{R}^{d \times d}$ positiv semidefinit, $B \in \mathbb{R}^{p \times d}$, so gilt

$$B \Sigma A = 0 \Rightarrow \mathbf{X}^T A \mathbf{X} \text{ und } B \mathbf{X} \text{ sind unabhängig}$$

b. Ist auch $B \in \mathbb{R}^{d \times d}$ positiv semidefinit, so gilt

$$B \Sigma A = 0 \Rightarrow \mathbf{X}^T A \mathbf{X} \text{ und } \mathbf{X}^T B \mathbf{X} \text{ sind unabhängig}$$

Bemerkung Es gilt jeweils auch die Rückrichtung (dies ist der schwerere, aber weniger relevante Teil).

Beweis

a. Spektralzerlegung von A

$$A = Q \text{diag}(\lambda_1, \dots, \lambda_{r(A)}, 0, \dots, 0) Q^T$$

mit Q orthogonal, $\lambda_i > 0$.

Sei $Q = (\mathbf{q}_1, \dots, \mathbf{q}_d)$, $\tilde{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_{r(A)}) \in \mathbb{R}^{d \times r(A)}$. Dann

$$A = \tilde{Q} D D \tilde{Q}^T, \quad D = \text{diag}(\lambda_1^{\frac{1}{2}}, \dots, \lambda_{r(A)}^{\frac{1}{2}})$$

Setze $L := \tilde{Q} D$, dann ist $A = L L^T$. Weiter gilt

$$L^T L = D \underbrace{\tilde{Q}^T \tilde{Q}}_{=I_{r(A)}} D = D^2$$

ist invertierbar, also

$$B \Sigma A = B \Sigma L L^T = 0 \Rightarrow B \Sigma L (L^T L) (L^T L)^{-1} = B \Sigma L = 0.$$

Nach Satz 1.5 sind somit die Vektoren $B \mathbf{X}$ und $L^T \mathbf{X}$ unabhängig und somit auch $B \mathbf{X}$ und $\mathbf{X}^T L L^T \mathbf{X} = \mathbf{X}^T A \mathbf{X}$ (ist Funktion von $L^T \mathbf{X}$).

b. Analog. (Zerlege A und B.) ■

Fortsetzung von Beispiel 4. Da $\bar{X}_n = \mathbf{1}_n^T \mathbf{X}/n$ und $\mathbf{1}_n^T P_n = 0$, sind $\bar{X}_n$ und S_n^2 bei normalverteilten X_i unabhängig.

Literatur

Discroll, M. F. (1999). An improved result relating quadratic forms and chi square distributions. *Amer. Statist.* **53**, 273-275.

Discroll, M. F. and Grundberg, W. R. (1986). A history of the development of Craig's theorem. *Amer. Statist.* **40**, 65-69.

Discroll, M. F. and Krasnicka, B. (1995). An accessible proof of Craig's theorem in the general case. *Amer. Statist.* **49**, 59-62.

2 Lineare Regression

2.1 Lineare Regression und Modellbildung

2.1.1 Regression und das lineare Modell

In der Regressionsanalyse geht es darum, den Einfluss einer Reihe von erklärenden Variablen $x_1, \dots, x_r$, sogenannte *Kovariablen*, auf eine abhängige Variable Y , die *Zielvariable*, zu modellieren bzw. zu schätzen. Dieser Zusammenhang drückt sich in Form einer Funktion $y = f(x_1, \dots, x_r)$ aus. Nun wird aber nicht angenommen, dass diese Beziehung exakt gilt. Vielmehr ist sie durch zufällige Störgrößen ϵ überlagert, d.h. es gilt

$$Y = f(x_1, \dots, x_r) + \epsilon.$$

In der linearen Regressionsanalyse nimmt man an, dass der Einfluss der Kovariablen, zumindest nach geeigneter Transformation dieser Variablen, in einer linearen Form

$$Y = b_0 + b_1 x_1 + \dots + b_r x_r + \epsilon.$$

Dabei ist ϵ eine Zufallsvariable (bzw. deren Realisierung) mit Erwartungswert $E\epsilon = 0$ und endlicher Varianz $\text{Var } \epsilon = \sigma^2$, und somit ist auch die Zilegröße Y eine Zufallsvariable (bzw. deren Realisierung). Ziel ist dann die Schätzung der Parameter $b_0, \dots, b_r$. Diese fasst man in einem Vektor zusammen. Wir schreiben $\boldsymbol{\beta} = (b_0, \dots, b_r) \in \mathbb{R}^p$, also $p = r + 1$, und für die Komponenten von $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ gilt dann $\beta_i = b_{i-1}$. Weiter schreibt man $\mathbf{x} = (1, x_1, \dots, x_r)^T \in \mathbb{R}^p$, so dass $Y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$.

Ziel ist nun in einem ersten Schritt die Schätzung der unbekannt Parameter des Modells, insbesondere von $\boldsymbol{\beta}$. Dazu nimmt man an, es werden Daten $(Y_i, x_{i,1}, \dots, x_{i,r})$, $i = 1, \dots, n$, beobachtet, so dass

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \mathbf{x}_i = (1, x_{i,1}, \dots, x_{i,r})^T.$$

Für die Fehler ϵ_i nimmt man dabei an, dass diese unabhängig oder zumindest unkorreliert sind, also dass $\text{Cov}(\epsilon_i, \epsilon_j) = 0$, $i \neq j$. Falls darüber hinaus die Varianzen $\sigma_i^2 = \text{Var } \epsilon_i$ alle gleich sind, also $\sigma_1^2 = \dots = \sigma_n^2$, so spricht man von einer *homoskedastischen* Fehlerstruktur, ansonsten von einer *heteroskedastischen* Fehlerstruktur.

Die Analyse eines homoskedastischen linearen Regressionsmodells findet nun im Rahmen der Theorie linearer Modelle statt. Dazu schreiben wir das Modell in Vektor- und Matrixform wie folgt.

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \quad X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n.$$

Es gilt dann $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Für die Kovariablen nimmt man noch an, dass sie dergestalt sind, dass die sogenannte *Designmatrix* X vollen Rang p hat. Der Achsenabschnitt wird meistens, aber nicht immer in das lineare Regressionsmodell mit aufgenommen. Das lineare Regressionsmodell fällt unter die folgende allgemeinere Definition.

Definition 2.1

Das Modell

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (5)$$

heißt *lineares Modell*, falls $\boldsymbol{\beta} \in \mathbb{R}^p$ ein (konstanter, unbekannter) Parametervektor, $X \in \mathbb{R}^{n \times p}$ eine bekannte Matrix (Designmatrix), $\mathbf{Y}$ ein beobachteter Zufallsvektor (Zielgrößen) und $\boldsymbol{\epsilon}$ ein nichtbeobachteter Zufallsvektor (Störgrößen) mit $E\boldsymbol{\epsilon} = \mathbf{0}$ und $\text{Cov } \boldsymbol{\epsilon} = \sigma^2 I_n$ sind. Sind darüber hinaus die Fehler normalverteilt, also $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, dann spricht man von einem linearen Modell mit normalverteilten Fehlern.

Wir werden in den folgenden Abschnitten alle Resultate im allgemeinen linearen Modell (5) formulieren, aber wir denken dabei stets an ein homoskedastisches lineares Regressionsmodell. Lineare Modelle treten darüber hinaus insbesondere in dem verwandten Kontext der *Varianzanalyse* auf.

2.1.2 Modellierung des Einflusses der Kovariablen**a. Einfluss einer metrischen Kovariable**

Angenommen, neben der Zielvariable Y wird noch eine metrische Kovariable x beobachtet. Das einfachste Modell ist nun die direkt lineare Regression von Y auf x , die *einfache lineare Regression*

$$Y = b_0 + b_1 x + \varepsilon,$$

also $\mathbf{x}^T = (1, x)^T$ und $\boldsymbol{\beta}^T = (b_0, b_1)^T$. Manchmal liegt ein linearer Zusammenhang auch erst vor, nachdem x geeignet transformiert wurde, etwa zu $\tilde{x} = \log(x)$ (falls $x > 0$) und dann $Y = b_0 + b_1 \tilde{x} + \varepsilon$, also $\mathbf{x}^T = (1, \tilde{x})^T$.

Man kann aus einer metrischen Kovariable auch mehrere metrische Kovariablen machen durch Anwendung linear unabhängiger Funktionen $f_1, \dots, f_r$, also

$$Y = b_0 + b_1 f_1(x) + \dots + b_r f_r(x) + \varepsilon,$$

also $\mathbf{x} = (1, f_1(x), \dots, f_r(x))^T$. Beispiele sind die polynomiale Regression, bei der $f_i(x) = x^i$ gewählt wird, also

$$Y = b_0 + b_1 x + \dots + b_r x^r + \varepsilon,$$

und $\mathbf{x} = (1, x, \dots, x^r)^T$, oder auch für $x \in [0, 1]$ die trigonometrische Regression, bei der $f_{2j-1}(x) = \sin(2j\pi x)$ und $f_{2j}(x) = \cos(2j\pi x)$, $j = 1, \dots, q$ gewählt wird, also

$$Y = b_0 + \sum_{j=1}^q (b_{2j-1} \sin(2j\pi x) + b_{2j} \cos(2j\pi x)),$$

und $\mathbf{x} = (1, \sin(2\pi x), \cos(2\pi x), \sin(2q\pi x), \cos(2q\pi x))^T$, $\boldsymbol{\beta} = (b_0, b_1, b_2, \dots, b_{2q})^T$, also $p = 2q + 1$.

b. Einfluss einer kategoriellen Kovariable

Bei *kategoriellen Kovariablen* unterscheidet man *nominale* Kovariablen, bei denen die Kategorien nicht geordnet sind (etwa Autotypen), und *ordinale* Kovariable, bei denen die Kategorien

in einer natürlichen Reihenfolge vorliegen (etwa Schulnoten).

Wir betrachten zunächst die Modellierung einer nominale Kovariablen mit den Kategorien $i = 1, \dots, I$. Um nicht zu viele Parameter ins Modell aufzunehmen, damit also die Designmatrix X noch vollen Rang hat, wählt man eine Referenzkategorie, z.B. $i = 1$, und für künstliche Kovariablen ein, deren Koeffizient den Unterschied zwischen der betrachteten Kategorie $i = 2, \dots, I$ und der Referenzkategorie beschreibt. Hier sind insbesondere zwei Kodierungen üblich.

Kodierung durch Dummy Variablen Wird die Kategorie x beobachtet und ist 1 die Referenzkategorie, so setze $\mathbf{x} = (1, 1_{x=2}, \dots, 1_{x=I})^T \in \mathbb{R}^I$, d.h. falls eine der Kategorien $i = 2, \dots, I$ vorliegt, kommt eine 1 hinzu, ansonsten gibt es nur den Achsenabschnitt. Im Koeffizientenvektor $\boldsymbol{\beta} = (\beta_1, \dots, \beta_I)^T$ beschreibt β_i den Unterschied des Einflusses von Kategorie $i \geq 2$ gegenüber der Referenzkategorie, und $\beta_1 + \beta_i$ den Gesamteinfluss von Kategorie $i \geq 2$.

Effektkodierung Wird die Kategorie x beobachtet und ist 1 die Referenzkategorie, so setze $\mathbf{x} = (1, 1_{x=2} - 1_{x=1}, \dots, 1_{x=I} - 1_{x=1})^T \in \mathbb{R}^I$.

Die Software R verwendet standardmäßig die Dummy Kodierung.

Handelt es sich bei x um eine ordinale Kovariable, so kann man versuchen, den geordneten Kategorien konkrete Zahlen (etwa den Schulnoten die Zahlen 1 – 6) zuzuordnen, und diese dann wie eine metrische Kovariable zu benutzen. Dies hat den Vorteil, dass in dem Modell wesentlich weniger Parameter (nur ein Parameter β für Kovariable x statt $I - 1$ Parameter) verwendet werden müssen. Dabei müssen die zugeordneten Zahlen (insbesondere das Verhältnis von deren Abständen) aber sorgfältig gewählt werden. Falls dies nicht adäquat möglich ist, sollte die Kovariable lieber wie eine nominale Kovariable und mit der Dummy Kodierung behandelt werden.

relevante R Befehle kategorielle Kovariablen müssen bei der Funktion `lm`, die lineare Regression mit kleinsten Quadraten anpasst, als Faktor vorliegen. Dazu kann man den Typ mit `str` erfahren, und gegebenenfall mit `as.factor` zu einem Faktor umwandeln.

c. Interaktionen

Interaktionen zwischen einer kategoriellen und einer stetigen Kovariable

Ist x eine kategorielle (nominale) Kovariable mit den Kategorien $i = 1, \dots, I$ und t eine stetige Kovariable, die direkt (linear) in die Zielgröße eingeht, so können die Kategorienausprägungen von x auch den Koeffizienten von t beeinflussen. Dies nennt man Interaktionen, man setzt dann bei Referenzkategorie 1 und Dummykodierung von x

$$\mathbf{x} = (1, 1_{x=2}, \dots, 1_{x=I}, t, 1_{x=2}t, \dots, 1_{x=I}t)^T.$$

Im Koeffizientenvektor $\boldsymbol{\beta} = (\beta_1, \dots, \beta_I, \beta_{I+1}, \dots, \beta_{2I})^T$ beschreibt dann $\beta_{I+1} + \beta_{2I}$ die Steigung von t bei Vorliegen von Kategorie $i \geq 2$, und β_{I+1} die Steigung bei Vorliegen der Referenzkategorie 1. Man muss dabei natürlich nicht alle Interaktionen in das Modell aufnehmen.

Falls sowohl stetige als auch kategorielle Kovariable auftreten, spricht man manchmal statt von der Regressionsanalyse auch von der *Kovarianzanalyse*.

Interaktionen zwischen zwei kategoriellen Kovariablen

Ist x eine kategorielle (nominale) Kovariable mit den Kategorien $i = 1, \dots, I$ und t eine kategorielle Kovariable mit Kategorien $j = 1, \dots, J$, so kann man Interaktionen für gemeinsames

Vorliegen von $x = i$ und $t = j$ modellieren. Sind $i = 1$ und $j = 1$ die Referenzkategorien, so bildet man in Dummy Kodierung

$$\mathbf{x} = (1, 1_{x=2}, \dots, 1_{x=I}, 1_{t=2}, \dots, 1_{t=J}, 1_{x=2}1_{t=2}, \dots, 1_{x=2}1_{t=J}, \dots, 1_{x=I}1_{t=J})^T \in \mathbb{R}^{IJ}.$$

Die Terme $1_{x=i}1_{t=j}$, $i = 2, \dots, I$, $j = 2, \dots, J$, entsprechen dann den Interaktionen, diese sind wieder als Abweichungen gegenüber den Haupteffekten $1_{x=i}$ und $1_{t=j}$ zu interpretieren.

Interaktionen zwischen zwei metrischen Kovariablen

Interaktionen zwischen zwei metrischen Kovariablen x und t müssen durch Aufnahme bestimmter gemeinsamer nichtlinearer Funktionen, etwa xt oder $e^x e^t$, modelliert werden. Man benutzt häufig gemeinsame Polynome niedrigen Grades.

In welcher Form metrische Kovariablen aufgenommen werden, und welche Interaktionen mit kategoriellen oder anderen metrischen Kovariablen aufgenommen werden, muss innerhalb der *Modellwahl* und der *Modelldiagnostik* bestimmt werden. Wie nehmen zunächst an, dass ein lineares Modell der Form (5) in seiner korrekten Form gegeben ist.

2.2 Kleinste Quadrate Schätzung

2.2.1 Methode der kleinsten Quadrate

Der bekannteste Schätzer von β im linearen Modell (5) ergibt sich aus der Methode der kleinsten Quadrate. Wann wählt dabei β derart, dass $\|\mathbf{Y} - X\beta\|^2 = \sum_{i=1}^n (Y_i - x_i^T \beta)^2$ minimal wird, also

$$\hat{\beta} = \hat{\beta}_{LS} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - X\beta\|^2$$

Dabei stehet LS für least squares = kleinste Quadrate. Wir schreiben für die Komponenten von $\hat{\beta}_{LS}$ explizit $\hat{\beta}_{LS} = (\hat{\beta}_{1,LS}, \dots, \hat{\beta}_{p,LS})$. Wir wollen $\hat{\beta}_{LS}$ in expliziter Form auf zwei Arten herleiten.

Normalgleichungen. Ableiten von $\|\mathbf{Y} - X\beta\|^2$ und gleich $\mathbf{0}$ setzen liefert

$$X^T(\mathbf{Y} - X\beta) = \mathbf{0}.$$

Dies nennt man auch die Normalgleichungen, diese sind eine notwendige Bedingung für ein lokales Extremum. Da X vollen Rang p hat, ist $X^T X \in \mathbb{R}^{p \times p}$ invertierbar und man erhält

$$\hat{\beta}_{LS} = \hat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}. \quad (6)$$

Dass $\hat{\beta}_{LS}$ das einzige lokale und somit globale Minimum von $\|\mathbf{Y} - X\beta\|^2$ ist, sieht man leicht daran, dass die Hessische Matrix (Matrix der zweiten Ableitungen) gleich $X^T X$ und somit positiv definit ist.

Geometrische Herleitung: Ein $\hat{\beta}$ minimiert die Funktion $\|\mathbf{Y} - X\beta\|^2$ genau dann, wenn $X\hat{\beta}$ die orthogonale Projektion von $\mathbf{Y}$ auf den von den Spaltenvektoren von $X = [\mathbf{v}_1, \dots, \mathbf{v}_p]$, $\mathbf{v}_i \in \mathbb{R}^n$, erzeugten Unterraum $V = \operatorname{span}\{v_1, \dots, v_p\}$ im $\mathbb{R}^n$ ist. In der Tat: Für jedes andere β gilt nach Pythagoras:

$$\|\mathbf{Y} - X\beta\|^2 = \underbrace{\|\mathbf{Y} - X\hat{\beta}\|}_{\perp \mathbf{v}_1, \dots, \mathbf{v}_p}^2 + \|X(\hat{\beta} - \beta)\|^2 = \|\mathbf{Y} - X\hat{\beta}\|^2 + \|X(\hat{\beta} - \beta)\|^2 \geq \|\mathbf{Y} - X\hat{\beta}\|^2$$

Da X vollen Rang hat, sind $\mathbf{v}_1, \dots, \mathbf{v}_p$ linear unabhängig und somit ist der Koeffizientenvektor $\hat{\beta}_{LS}$ eindeutig bestimmt.

Um den Schätzer $\hat{\beta}_{LS}$ in der expliziten Form (6) zu erhalten, betrachten wir die Matrix $P_X = X(X^T X)^{-1} X^T \in \mathbb{R}^{n \times n}$ (die sogenannte *hat matrix*). Es ist

$$P_X : \mathbb{R}^n \rightarrow V \quad \mathbf{z} \mapsto P_X \mathbf{z}$$

die orthogonale Projektion auf V . Dazu zeigt man durch direkte Rechnung:

- P_X ist orthogonale Projektion: $P_X^2 = P_X$, $P_X^T = P_X$
- P_X lässt V invariant
- $P_X \mathbb{R}^n \subset V$

Somit muss gelten:

$$X\hat{\beta} = P_X \mathbf{Y} = X(X^T X)^{-1} X^T \mathbf{Y}.$$

Da X vollen Rang hat, ergibt sich wieder die Form (6).

Bemerkung

Erwartungswert und Varianz von $\mathbf{Y}$ im linearen Modell (5) hängen von den unbekanntem Parametern $(\boldsymbol{\beta}, \sigma^2)$, ab, höhere Momente sogar von der unbekanntem Verteilung der Störungen $\boldsymbol{\epsilon}$. Daher müsste man diese bei Bildung von Erwartungswert und Varianz eigentlich mitschreiben, also etwa $E_{\boldsymbol{\beta}, \sigma^2}(\cdot)$ und $\text{Cov}_{\boldsymbol{\beta}, \sigma^2}(\cdot)$. Wir werden diese Parameter aber in der Notation im allgemeinen unterdrücken, und einfach E und Cov schreiben.

Satz 2.2

Der kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}_{LS}$ im linearen Modell (5) ist unverfälscht, also $E\hat{\boldsymbol{\beta}}_{LS} = \boldsymbol{\beta}$, und es ist

$$\text{Cov} \hat{\boldsymbol{\beta}}_{LS} = \sigma^2 (X^T X)^{-1}.$$

Beweis

Mit Satz 1.2 folgt

$$\begin{aligned} E\hat{\boldsymbol{\beta}}_{LS} &= E(X^T X)^{-1} X^T \mathbf{Y} \\ &= E(X^T X)^{-1} X^T (X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= E(X^T X)^{-1} X^T X\boldsymbol{\beta} + E(X^T X)^{-1} X^T \boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + (X^T X)^{-1} X^T \underbrace{E\boldsymbol{\epsilon}}_{=0}, \\ &= \boldsymbol{\beta} \\ \text{Cov} \hat{\boldsymbol{\beta}}_{LS} &= \text{Cov}((X^T X)^{-1} X^T (\underbrace{X\boldsymbol{\beta}}_{\text{konst.}} + \boldsymbol{\epsilon})) \\ &= \text{Cov}((X^T X)^{-1} X^T \boldsymbol{\epsilon}) \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

Man nennt $\sigma((X^T X)^{-1})^{1/2}$ und für einen Schätzer $\hat{\sigma}^2$ von σ^2 (s.u.) auch $\hat{\sigma}((X^T X)^{-1})^{1/2}$ den Standardfehler von $\hat{\boldsymbol{\beta}}_{LS}$.

Im Folgenden zeigen wir, dass $\hat{\boldsymbol{\beta}}_{LS}$ der eindeutig bestimmte, lineare unverfälschte Schätzer mit der kleinsten Varianz ist.

Satz 2.3 (Gauß-Markov-Aitken)

- a. Sei $S(\mathbf{Y}) = A\mathbf{Y}$, $A \in \mathbb{R}^{p \times n}$, ein linearer, unverfälschter Schätzer für $\boldsymbol{\beta}$ im linearen Modell (5) (d.h. $ES(\mathbf{Y}) = \boldsymbol{\beta} \forall \boldsymbol{\beta} \in \mathbb{R}^p$). Dann gilt

$$\text{Cov}(S(\mathbf{Y})) \geq \text{Cov}(\hat{\boldsymbol{\beta}}_{LS})$$

im Sinne, dass die Differenz $(\text{Cov}(S(\mathbf{Y})) - \text{Cov}(\hat{\boldsymbol{\beta}}_{LS}))$ positiv semidefinit ist.

- b. Ist $A \neq (X^T X)^{-1} X^T$, so existiert $\mathbf{z} = \mathbf{z}(A) \in \mathbb{R}^p$, so dass

$$\mathbf{z}^T (\text{Cov}(S(\mathbf{Y})) - \text{Cov}(\hat{\boldsymbol{\beta}}_{LS})) \mathbf{z} > 0$$

Beweis

- a. Aus der Unverfälschtheit folgt

$$ES(\mathbf{Y}) = A X \boldsymbol{\beta} \stackrel{!}{=} \boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \in \mathbb{R}^p,$$

also $AX = I_p$. Damit und mit Satz 2.2 ist

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}_{LS}) &= \sigma^2(X^T X)^{-1} = \sigma^2 AX(X^T X)^{-1} X^T A^T = \sigma^2 AP_X A^T, \\ \text{Cov}(S(\mathbf{Y})) &= A\sigma^2 I_n A^T = \sigma^2 AA^T.\end{aligned}$$

Damit erhält man

$$\text{Cov}(S(Y)) - \text{Cov}(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2 A(I_n - P_X)A^T$$

Die Matrix $(I_n - P_X)$ ist idempotent und symmetrisch: $(I_n - P_X)^2 = (I_n - P_X) = (I_n - P_X)^T$. Somit:

$$\sigma^2 \mathbf{z}^T A(I_n - P_X)A^T \mathbf{z} = \sigma^2 \|(I_n - P_X)A^T \mathbf{z}\|^2 \geq 0.$$

b. Angenommen, $\|(I_n - P_X)A^T \mathbf{z}\|^2 = 0 \forall \mathbf{z} \in \mathbb{R}^p$, und somit $(I_n - P_X)A^T = 0$. Sei $A^T = (\mathbf{a}_1, \dots, \mathbf{a}_p)$, $\mathbf{a}_i \in \mathbb{R}^n$, dann erhält man ³: $\mathbf{a}_i \in \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_p)$, also $A^T = XM$ für eine Matrix $M \in \mathbb{R}^{p \times p}$. Wegen $AX = I_p$ folgt $M^T X^T X = I_p$, also $M^T = (X^T X)^{-1} \Rightarrow A = (X^T X)^{-1} X^T$. ■

Wegen Satz 2.3 heißt der Schätzer $\hat{\boldsymbol{\beta}}_{LS}$ auch der *beste lineare unverfälschte Schätzer* (best linear unbiased estimator, BLUE).

Bemerkung 2.4

1. Satz 2.3, b., impliziert, dass es für einen unverfälschten Schätzer $S(\mathbf{Y}) \neq \hat{\boldsymbol{\beta}}_{LS}$ ein $\mathbf{z} \in \mathbb{R}^p$ gibt, so dass gilt $\text{Var}(\mathbf{z}^T S(\mathbf{Y})) > \text{Var}(\mathbf{z}^T \hat{\boldsymbol{\beta}}_{LS})$. (Schätzer für $\mathbf{z}^T \boldsymbol{\beta}$)
2. Satz 2.3, a., folgt $(\text{Cov}(S(Y)))_{jj} \geq (\text{Cov}(\hat{\boldsymbol{\beta}}_{LS}))_{jj}$ für $j = 1, \dots, p$, und daher $\text{Spur}(\text{Cov}(S(Y))) \geq \text{Spur}(\text{Cov}(\hat{\boldsymbol{\beta}}_{LS}))$

Der kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}_{LS}$ als Maximum-Likelihood-Schätzer. Angenommen, im linearen Modell (5) sind die Fehler normalverteilt, also $\boldsymbol{\epsilon} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$ und somit $Y \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$. Dann ist die Likelihood-Funktion gegeben durch

$$L_n(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2\right),$$

und die log-Likelihood-Funktion durch

$$\mathcal{L}_n(\boldsymbol{\beta}, \sigma^2) = \log L_n(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\|\mathbf{Y} - X\boldsymbol{\beta}\|^2}{2\sigma^2}$$

Damit $\mathcal{L}_n(\boldsymbol{\beta}, \sigma^2)$ maximal wird, muss offenbar $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ minimal werden. Daher ist $\hat{\boldsymbol{\beta}}_{LS}$ gleich dem Maximum Likelihood Schätzer $\hat{\boldsymbol{\beta}}_{ML}$ für $\boldsymbol{\beta}$ im linearen Modell mit normalverteilten Fehlern.

Schätzen der Fehlervarianz σ^2 .

Zunächste einige Bezeichnungen:

$$\begin{aligned}\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} &= (\hat{Y}_1, \dots, \hat{Y}_n)^T && \text{angepassten Werte (fitted values)} \\ \hat{\boldsymbol{\epsilon}} &= \mathbf{Y} - \hat{\mathbf{Y}} && \text{Residuen}\end{aligned}$$

³ $\mathbf{a}_i \in \text{Kern}(I_n - P_X) = \text{Bild}(P_X)$; Eigenschaft von Projektionen (hier: P_X)

Idee: Nutze Variation der $\hat{\boldsymbol{\epsilon}} = (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T$ zum Schätzen von σ^2 . Schätze σ^2 durch

$$\hat{\sigma}^2 = \hat{\sigma}_{LS}^2 = \frac{1}{n-p} \sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n-p} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \frac{1}{n-p} \mathbf{Y}^T (I - P_X) \mathbf{Y},$$

wobei wir $\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = P_X \mathbf{Y}$ und $(I - P_X) = (I - P_X)^2 = (I - P_X)^T$ genutzt haben.

Satz 2.5

Im linearen Modell (5) ist $\hat{\sigma}_{LS}^2$ erwartungstreu für σ^2 .

Beweis

Da $E\mathbf{Y} = X\boldsymbol{\mu}$, $\text{Cov } \mathbf{Y} = \sigma^2 I_n$, folgt mit Satz 1.2

$$E(\mathbf{Y}^T (I - P_X) \mathbf{Y}) = \boldsymbol{\mu}^T \underbrace{X^T (I - P_X) X}_{=0} \boldsymbol{\mu} + \text{Spur}(\sigma^2 (I - P_X)).$$

Es gilt:

$$\begin{aligned} \text{Spur}(I_n - P_X) &= n - \text{Spur}(X(X^T X)^{-1} \cdot X^T) \\ &= n - \text{Spur}(X^T \cdot X(X^T X)^{-1}) \\ &= n - \text{Spur } I_p \\ &= n - p. \end{aligned} \tag{7}$$

Somit ist

$$E\hat{\sigma}_{LS}^2 = \frac{1}{n-p} E(\mathbf{Y}^T (I - P_X) \mathbf{Y}) = \sigma^2. \quad \blacksquare$$

Übung Der ML-Schätzer $\hat{\sigma}_{ML}^2$ für σ^2 im linearen Modell mit normalverteilten Fehlern ist gegeben durch $\frac{1}{n} \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{LS}\|^2$.

Relevante R Befehle `lm` führt eine kleinste Quadrate Schätzung im linearen Modell durch. Auf ein dadurch erzeugtes Objekt kann man `summary` (Schätzwerte, Schätzer für σ^2 , Standardfehler und mehr), `fitted.values` (angepasste Werte), `residuals` (Residuen) anwenden.

2.2.2 Verteilungstheorie bei normalverteilten Fehlern

In diesem Abschnitt gehen wir auf die Verteilung der Schätzer sowie relevanter Teststatistiken im linearen Modell (5) mit normalverteilten, homoskedastischen Fehlern, also $\epsilon \sim N(0, \sigma^2 I_n)$, ein. Wir beginnen mit dem grundlegenden

Satz 2.6

Im linearen Modell $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, mit normalverteilten, homoskedastischen Fehlern $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$ gilt

$$\hat{\boldsymbol{\beta}}_{LS} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1}), \quad \frac{n-p}{\sigma^2} \hat{\sigma}^2 \sim \chi^2(n-p),$$

und $\hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$ sind unabhängig.

Beweis

Da $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$, folgt aus $\hat{\boldsymbol{\beta}}_{LS} = (X^T X)^{-1} X^T \mathbf{Y}$ mit Satz 1.3 direkt die Behauptung für $\hat{\boldsymbol{\beta}}_{LS}$. Weiter ist

$$\hat{\sigma}^2 = \frac{1}{\sigma^2} \mathbf{Y}^T (I_n - P_X) \mathbf{Y}.$$

Da $(I_n - P_X)^2 = I_n - P_X$ (die orthogonale Projektion auf das orthogonale Komplement des Spaltenraumes von X), ist Satz 1.7 anwendbar (die Division durch σ^2 standardisiert die $\mathbf{Y}$). Da $(I_n - P_X)X = \mathbf{0}$, ist der Nichtzentralitätsparameter gleich 0. Weiter ist nach (7) $\text{Spur}(I_n - P_X) = n - p$.

Es bleibt die Unabhängigkeit von $\hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$ zu zeigen. Diese folgt direkt aus Satz 1.9, da $(I_n - P_X)X = \mathbf{0}$. ■

Konfidenzintervall für $\mathbf{c}^T \boldsymbol{\beta}$. Für $\mathbf{c} \in \mathbb{R}^p$ gilt nach Satz 2.6 und Satz 1.3

$$\mathbf{c}^T \hat{\boldsymbol{\beta}}_{LS} \sim \mathcal{N}(\mathbf{c}^T \boldsymbol{\beta}, \sigma^2 \mathbf{c}^T (X^T X)^{-1} \mathbf{c})$$

und

$$\frac{\mathbf{c}^T \hat{\boldsymbol{\beta}}_{LS} - \mathbf{c}^T \boldsymbol{\beta}}{\sigma (\mathbf{c}^T (X^T X)^{-1} \mathbf{c})^{\frac{1}{2}}} \sim \mathcal{N}(0, 1).$$

Wegen der Unabhängigkeit von $\hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$ (und somit auch von $\mathbf{c}^T \hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$) folgt direkt aus der Definition der t Verteilung

$$\frac{\mathbf{c}^T \hat{\boldsymbol{\beta}}_{LS} - \mathbf{c}^T \boldsymbol{\beta}}{\hat{\sigma} (\mathbf{c}^T (X^T X)^{-1} \mathbf{c})^{\frac{1}{2}}} \sim t(n-p). \quad (8)$$

Konfidenzintervall für $\mathbf{c}^T \boldsymbol{\beta}$ zum Niveau $\alpha > 0$:

$$\left[\mathbf{c}^T \hat{\boldsymbol{\beta}}_{LS} - \hat{\sigma} (\mathbf{c}^T (X^T X)^{-1} \mathbf{c})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p); \mathbf{c}^T \hat{\boldsymbol{\beta}}_{LS} + \hat{\sigma} (\mathbf{c}^T (X^T X)^{-1} \mathbf{c})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p) \right].$$

Insbesondere für Komponente β_i ($\mathbf{c} = \mathbf{e}_i$, i -te Einheitsvektor):

$$\left[\hat{\beta}_{LS,i} - \hat{\sigma} ((X^T X)^{-1})_{ii}^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p); \hat{\beta}_{LS,i} + \hat{\sigma} ((X^T X)^{-1})_{ii}^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p) \right].$$

Aufgabe Konfidenzintervall für σ^2 .

relevante R Befehle `confint` liefert für ein Objekt aus `lm` die Konfidenzintervalle der einzelnen Komponenten des KQ Schätzers.

Hypothesen Testen mit dem t Test Für $\mathbf{c} \in \mathbb{R}^p$ und $\delta \in \mathbb{R}$ betrachten wir die Hypothese

$$H_{\mathbf{c},\delta} : \mathbf{c}^T \boldsymbol{\beta} = \delta.$$

Unter Gültigkeit der Hypothese $H_{\mathbf{c},\delta}$ gilt nach (8):

$$T_{\mathbf{c},\delta} = \frac{\mathbf{c}^T \hat{\boldsymbol{\beta}}_{LS} - \delta}{\hat{\sigma}(\mathbf{c}^T (X^T X)^{-1} \mathbf{c})^{\frac{1}{2}}} \sim t(n-p).$$

Man verwirft die zweiseitige Hypothese $H_{\mathbf{c},\delta}$ zum Niveau $\alpha > 0$, falls $T_{\mathbf{c},\delta} < t_{\alpha/2}(n-p)$ oder falls $T_{\mathbf{c},\delta} > t_{1-\alpha/2}(n-p)$. Der zweiseitige p-Wert ergibt sich als

$$P = \begin{cases} 2t(n-p)(T_{\mathbf{c},\delta}), & T_{\mathbf{c},\delta} < 0, \\ 2(1-t(n-p)(T_{\mathbf{c},\delta})), & T_{\mathbf{c},\delta} > 0. \end{cases} \quad (9)$$

Analog für einseitige Hypothesen. Wichtigster Spezialfall: $\mathbf{c} = \mathbf{e}_i$, $\delta = 0$. Dann $H_i : \beta_i = 0$, und

$$T_i = \frac{(\hat{\boldsymbol{\beta}}_{LS})_i}{\hat{\sigma}((X^T X)^{-1})_{ii}^{1/2}}.$$

relevante R Befehle `summary` liefert für ein Objekt aus `lm` die zweiseitigen P-Werte für die Hypothese H_i .

Vorhersageintervalle Ein *Konfidenzintervall* bezieht sich auf den Erwartungswert $\mathbf{c}^T \boldsymbol{\beta}$ von $\mathbf{c}^T \hat{\boldsymbol{\beta}}$, wobei $\hat{\boldsymbol{\beta}}$ aus dem linearen Modell (5) berechnet wird.

Bei einem *Vorhersageintervall* (Prediction Interval) hingegen ist eine zusätzliche Kovariablenausprägung $\mathbf{x}_{n+1}$ erforderlich, bei der die abhängige Variable Y_{n+1} vorhergesagt werden soll. Das Vorhersageintervall bezieht sich also nicht auf einen Parameter wie das Konfidenzintervall, sondern auf die Zufallsvariable Y_{n+1} .

Sei $\hat{\boldsymbol{\beta}}_{LS}$ der KQ-Schätzer im linearen Modell (5). Als Vorhersage für Y_{n+1} bei $\mathbf{x}_{n+1}$ betrachten man

$$Y^{\text{Pred}} = \mathbf{x}_{n+1}^T \hat{\boldsymbol{\beta}}_{LS}.$$

Nach dem linearen Modell würde die Beobachtung Y_{n+1} entstehen durch

$$Y_{n+1} = \mathbf{x}_{n+1}^T \boldsymbol{\beta} + \varepsilon_{n+1},$$

wobei ε_{n+1} und $\boldsymbol{\epsilon}$ unabhängig sind. Somit

$$Y^{\text{Pred}} - Y_{n+1} = \mathbf{x}_{n+1}^T (\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}) + \varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2 + \sigma^2 \mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1}),$$

und nach Satz 2.6

$$\frac{Y^{\text{Pred}} - Y_{n+1}}{\hat{\sigma}(1 + \mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1})^{\frac{1}{2}}} \sim t_{n-p}.$$

Als zweiseitigen Vorhersagebereich erhält man

$$\left[Y^{\text{Pred}} - \hat{\sigma}(1 + \mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p), Y^{\text{Pred}} + \hat{\sigma}(1 + \mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p) \right].$$

Vergleich. Das Vorhersageintervall für Y_{n+1} ist breiter als das Konfidenzintervall für $\mathbf{x}_{n+1}^T \boldsymbol{\beta}$, da der zusätzliche Fehler ε_{n+1} in Y_{n+1} mit berücksichtigt werden muss.

relevante R Befehle `predict.lm` anwenden auf Objekt aus `lm` und zusätzliche Kovariable.

Konfidenzbereich für $A\boldsymbol{\beta}$ Die Matrix $A \in \mathbb{R}^{q \times p}$, $1 \leq q \leq p$, habe vollen Rang. Nach Satz 2.6 und Satz 1.3 ist

$$A\hat{\boldsymbol{\beta}}_{LS} \sim \mathcal{N}(A\boldsymbol{\beta}, \sigma^2 S),$$

wobei $S = A(X^T X)^{-1} A^T \in \mathbb{R}^{q \times q}$ vollen Rang hat. Nach Satz 1.7 ist dann

$$(A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}))^T \frac{S^{-1}}{\sigma^2} A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}) \sim \chi^2(q).$$

Da nach Satz 2.6 $\hat{\sigma}^2$ und $\hat{\boldsymbol{\beta}}_{LS}$ (und damit auch eine Funktion von $\hat{\boldsymbol{\beta}}_{LS}$) unabhängig sind, folgt aus der Definition der F Verteilung

$$\frac{(A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}))^T S^{-1} A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})}{q\hat{\sigma}^2} \sim F(q; n - p). \quad (10)$$

Somit ergibt sich als $(1-\alpha)$ -Konfidenzbereich für $A\boldsymbol{\beta}$:

$$\left\{ A\boldsymbol{\beta} : \frac{(A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}))^T S^{-1} A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})}{q\hat{\sigma}^2} \leq F_{1-\alpha}(q; n - p) \right\}.$$

Ein wichtiger Spezialfall entsteht, falls A eine Teilmatrix von I_p ist.

relevante R Befehle Die library `ellipse` enthält den Befehl `ellipse`, welcher zweidimensionale Konfidenzellipsoide berechnet. Plotten einfach mit `plot`.

Testen allgemeiner linearer Hypothesen mit dem F-Test Man möchte manchmal allgemeinere lineare Hypothesen, die nicht von der Form $\mathbf{c}^T \boldsymbol{\beta} = \delta$ sind, testen.

Beispiele linearer Hypothesen.

- a. H: $\boldsymbol{\beta} = \mathbf{0}$ (alle $\beta_i = 0$)
- b. H: $\beta_{i_1} = \dots = \beta_{i_q} = 0$, $1 \leq i_1 < \dots < i_q \leq p$
- c. H: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\beta}_0 \neq \mathbf{0}$.

Allgemeine lineare Hypothese: Für $A \in \mathbb{R}^{q \times p}$, $q \leq p$ mit vollem Rang, $\mathbf{m} \in \mathbb{R}^q$ betrachte

$$H_{A,\mathbf{m}} : A\boldsymbol{\beta} = \mathbf{m}.$$

Es gilt

$$A\hat{\boldsymbol{\beta}} - \mathbf{m} \sim \mathcal{N}(A\boldsymbol{\beta} - \mathbf{m}, \sigma^2 S),$$

wobei wiederum $S = A(X^T X)^{-1} A^T \in \mathbb{R}^{q \times q}$ vollen Rang hat. Dann ist

$$(A\hat{\beta} - \mathbf{m})^T \frac{S^{-1}}{\sigma^2} (A\hat{\beta} - \mathbf{m}) \sim \chi^2(q, \lambda)$$

wobei der Nichtzentralitätsparameter $\lambda = \frac{1}{2}(A\beta - \mathbf{m})^T \frac{S^{-1}}{\sigma^2} (A\beta - \mathbf{m})$, und somit

$$FS = \frac{(A\hat{\beta} - m)^T S^{-1} (A\hat{\beta} - \mathbf{m})}{q\hat{\sigma}} \sim F(q, n - p, \lambda)$$

Unter der Hypothese $H_{A, \mathbf{m}}$ gilt $\lambda = 0$. Daraus bestimmt man den zweiseitigen p-Wert für die lineare Hypothese als $P = 1 - F(q, n - p)(FS)$.

Aufgabe Bestimme den kleinsten Quadrate Schätzer unter der linearen Nebenbedingung $A\beta = m$.

relevante R Befehle `anova` führt den F Test durch, dabei muss das Modell unter $H_{A, \mathbf{m}}$ mit geschätzt worden sein und als Argument übergeben werden. Falls kein zweites Modell übergeben wird, führt `anova` die F-Tests dafür durch, ob bei kategorielle Kovariablen alle Koeffizienten der Dummy Variablen = 0 sind, und gegebenenfalls auch für die Interaktionen.

2.2.3 Asymptotische Verteilungstheorie

In diesem Abschnitt zeigen wir, dass der kleinste Quadrate Schätzer im linearen Regressionsmodell unter geeigneten Bedingungen bei nicht-normalverteilten Fehlern zumindest *asymptotisch* normalverteilt ist. Dazu benötigen wir zunächst ein geeignetes Beobachtungsmodell, in dem die Anzahl der Beobachtungen n wächst ($n \rightarrow \infty$), die Zahl der Parameter p sowie der Parametervektor β und σ^2 fest bleiben. Wir nehmen also an, dass für jedes n Beobachtungen $(Y_{i,n}, x_{i,n})$, $i = 1, \dots, n$, aus dem Modell

$$Y_{i,n} = \mathbf{x}_{i,n}^T \beta + \epsilon_{i,n} \quad i = 1, \dots, n, \quad (11)$$

vorliegen, wobei $\epsilon_{i,n}$ identisch verteilt sind für alle i, n mit $E\epsilon_{i,n} = 0$, $E\epsilon_{i,n}^2 = \sigma^2$, und $\epsilon_{1,n}, \dots, \epsilon_{n,n}$ unabhängig sind für alle n . Setze

$$X_n = (\mathbf{x}_{1,n}^T, \dots, \mathbf{x}_{n,n}^T)^T \in \mathbb{R}^{n \times p}, \quad \mathbf{Y}_n = (Y_{1,n}, \dots, Y_{n,n}) \in \mathbb{R}^n, \quad \epsilon = (\epsilon_{1,n}, \dots, \epsilon_{n,n})^T \in \mathbb{R}^n,$$

wobei wir im Vektor ϵ die Abhängigkeit von n unterdrückt haben.

Annahme. Alle Komponenten der $\mathbf{x}_{j,n}$ (für alle j, n) liegen in einer kompakter Menge K , und es gelte

$$\frac{1}{n} X_n^T X_n \rightarrow Q \in \mathbb{R}^{p \times p} \quad (12)$$

für eine positiv definite Matrix $Q > 0$.

Satz 2.7

In einer Folge von linearen Modellen (11) unter obigen Annahmen gilt für die Folge von Schätzern $\hat{\beta}_{LS,n}$ und σ_n^2 unter Annahme (12)

$$\sqrt{n}(\hat{\beta}_{LS,n} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1}) \quad (13)$$