Materialien zu

Lineare Modelle mit Anwendungen

Dr. Christian Kredler

SS 2003

Inhaltsverzeichnis

Kapitel 1

Einführung

Lineare Modelle mit Anwendungen schließt direkt an die Einführungsvorlesung Stochastik 1 (zum vorausgesetzten Stoff vgl. [Kredler (1999)]) an und behandelt die wohl häufigste Anwendung der mathematischen Statistik in der Praxis.

Die Vorlesung ist für Diplom-Mathematiker, Techno-, Finanz- und Wirtschaftsmathematiker sowie Studierende des Lehramts Mathematik an Gymnasien konzipiert. Sie kann ab dem vierten Semester gehört werden und bietet sich ebenso für Physiker und Ingenieure an.

Der Praxisbezug wird durch eine intensive Arbeit am Datenmaterial hergestellt. Als Software wurde R bzw. S-Plus gewählt. Das Ausbildungskonzept an der TU München beruht auf einer umfassenden Darstellung im Web (siehe www-m4.ma.tum.de/nbu/linmod/). Diese beinhaltet u.a.

- eine Zusammenstellung der notwendigen Vorkenntnisse,
- eine Einführung in R bzw. S-Plus,
- eine ausführliche Diskussion der Vorlesungsbeispiele mit Grafiken und den relevanten S-Plus-Befehlen,
- zum Selbststudium geeignete Übungsaufgaben mit Datenbeispielen,
- eine Sammlung von Beispiel-Datensätzen.

Diese Vorlesungsausarbeitung erhebt keinen Anspruch auf Vollständigkeit. Sie orientiert sich in Aufbau und Notation am Buch von [Myers (1990)]. Viele Beispiele sind einem Skript von Prof. Claudia Czado entnommen, die eine entsprechende Vorlesung im SS 2000 an der TU München hielt.

1.1 Das Modell

Regressionsanalyse (mit Hilfe Linearer Modelle) ist ein statistisches Verfahren, das den Zusammenhang bestimmt zwischen

- einer *abhängigen Variablen* Y (Zielvariable, response)
- und *unabhängigen Variablen* $X_1, \ldots, X_k$ (Kovariablen, Regressoren, erklärende Variablen), wobei eine lineare Beziehung der Form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + E$$

mit einer Fehlervariablen E unterstellt wird.

Beispiele mit mehreren Kovariablen werden in Abschnitt 3.1 diskutiert. Der Spezialfall **Einfache Lineare Regression** ($k = 1$) wird in Kapitel 2 behandelt. Schon am Beispiel

Y: Körpergewicht
X: Körpergröße

sieht man, dass Response Y und Kovariable (hier X) nicht vertauschbar sind. In der Regel wird diejenige Variable als Response gewählt, für die Prognosen zu erstellen sind oder die eine natürliche Variablität aufweist. So liegt die Körpergröße ab einem bestimmten Alter fest, was man vom Körpergewicht meist nicht sagen kann.

Allgemein betrachtet man in der sog. **Multiplen Linearen Regression**

$$(M_k) \qquad Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

x_{ik} gegebene, nicht stochastische Daten

$$E(\epsilon_i) = 0,\ Cov(\epsilon_i, \epsilon_j) = \sigma^2 \delta_{ij},\ \delta_{ij} = \begin{cases} 1, & \text{falls } i = j \\ 0, & \text{sonst} \end{cases}, \tag{1.1}$$

d.h. Fehler sind unkorreliert mit homogener Varianz.

Damit sind die Y_i Zufallsvariablen (ZV) mit homogener Varianz, unkorreliert, aber mit verschiedenen Erwartungswerten.

Das Modell (M_k) ist linear in den unbekannten Parametern $\beta_0, \beta_1, \ldots, \beta_k$.
Die Y_i und die x_{ij} können nichtlineare Transformationen anderer Variablen sein.

Spezialfälle:

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	Einfache Lineare Regression
$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$	linear in $\beta_0, \beta_1, \beta_2$
$Y_i = \beta_0 e^{\beta_1 x_i} \cdot \epsilon_i$	$ln(Y_i)$ linear in $\gamma_0 = ln(\beta_0)$ und $\gamma_1 = ln(\beta_1)$
$Y_i = \beta_0 + e^{\beta_1 x_i} + \epsilon_i$	$ln(Y_i)$ nichtlinear in β_0, β_1

Ziele

1. Finde ein "gutes Modell" für Y (*Modell-Spezifikation*) $\rightarrow$ "kleiner Fehler" ϵ_i
2. Finde "gute" Regressoren (*Variablenselektion*)
3. Schätze die unbekannten Parameter (*Parameterschätzung*)
4. Sage Y für feste x's voraus (*Vorhersage, Prediction*)

Einige Probleme

1. Wichtige Regressoren wurden in den Daten nicht miterhoben.
2. Vorhersagende Y-Werte gehören zu x-Werten, die außerhalb der bisher beobachteten x-Werte liegen.
3. Der wahre Zusammenhang zwischen Y und den x's entspricht nicht dem Ansatz von (M_k), ist z.B. nicht additiv.

1.2 Das Modell mit bedingter Erwartung

In vielen Fällen betrachtet man n unabhängige Wiederholungen des Zufallsvektors $(Y, X_1, \ldots, X_k)$. Dabei seien die Realisierungen $x_{i1}, \ldots, x_{ik}$ der ZV $X_{i1}, \ldots, X_{ik}$, $i = 1, \ldots n, j = 1, \ldots k$, zu einem festgesetzten Status bzw. Zeitpunkt beobachtbar, die Zielvariablen Y_i aber erst später. Nach Eintritt der $X_{ij} = x_{ij}$ sind also nur noch die Y_i stochastisch. In diesem Fall lautet der Modellansatz

$$(M_k') \qquad E(Y_i|X_{i1} = x_{i1}, \ldots, X_{ik} = x_{ik}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

1.3 Notation in der Literatur

Als häufigste Schreibweise findet man (siehe auch Kapitel 3)

$$(M_k'') \qquad y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

zu gegebenen Kovariablen-Daten x_{ik}. Diese werden als nicht stochastisch angesehen. Die y_i sind stochastisch, bezeichnen also ZV. In den meisten Fällen wird nicht zwischen den ZV selbst und deren Realisierungen (Daten) unterschieden.

Kapitel 2

Einfache Lineare Regression

2.1 Modellannahmen

In der Einfachen Linearen Regression liegen Daten der Form $(y_i, x_i)_{i=1,\dots,n}$ vor. Für diese wählt man den Ansatz

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \,, \tag{2.1}$$

(M) $\qquad E(\epsilon_i) = 0$ und $Cov(\epsilon_i, \epsilon_j) = \delta_{ij}\,\sigma^2$, $i = 1, \dots, n$.

β_0 und β_1 sind unbekannte Parameter.
Die Daten y_i werden als Realisierungen der Zufallsvariablen (ZV) Y_i angesehen. Die x_i sind nicht stochastisch, z.B. Messstellen.

2.1.1 Reparametrisierung

Für theoretische Überlegungen erweist sich stets das sog. "zentrierte Modell" als überlegen. Dieses erhält man aus Gleichung 2.1 mit $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ und

$$y_i = \beta_0^* + \beta_1\,(x_i - \bar{x}) + \epsilon_i \,, \tag{2.2}$$

wobei $\beta_0 = \beta_0^* - \beta_1\,\bar{x}$.

2.1.2 Normalverteilung

Die Annahme

$$\epsilon_i \text{ iid } N(0, \sigma^2) \tag{2.3}$$

impliziert (1.1) bzw. (2.1), d.h. $E(\epsilon_i) = 0$ und $Cov(\epsilon_i, \epsilon_j) = \delta_{ij}\sigma^2$.

b_0 (bzw. b_0^*) und b_1 seien statistische Schätzer für die unbekannten Parameter β_0 (bzw. β_0^*) und β_1. Diese werden so gewählt, dass die Regressionsgerade $y = b_0 + b_1 x$ die Punktewolke $(y_i, x_i)_{i=1,\dots,n}$ möglichst "gut" anpasst.

2.1.3 Definitionen und Identitäten

Def. 2.1 (Response und Residuen)
$\hat{Y}_i := b_0 + b_1 x_i = b_0^* + b_1(x_i - \bar{x})$ *heißen* **Y-Schätzungen (fitted response)**
$e_i := Y_i - \hat{Y}_i$ *heißen* **Residuen**.

$\hat{Y}_i$ und e_i sind ZV mit denselben Erwartungswerten wie Y_i und ϵ_i. Sie sind aber nicht mehr unkorreliert.

Def. 2.2 (Sums of Squares)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \;\; \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(x_i - \bar{x})Y_i$$

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i$$

$$SS_e = RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i)^2 = SS_{Res}$$

RSS = residual sum of squares

$Q(\beta_o, \beta_1) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$ *heißt Summe der quadratischen Abweichungen für die Parameter* β_0 *und* β_1.

2.2 Methode der Kleinsten Quadrate, KQ-Schätzer

Nach der Methode der kleinsten Quadrate (**KQ, Least Squares, LS**) wählt man die Schätzungen b_0 und b_1 so, dass

$$Q(b_0, b_1) \leq Q(\beta_0, \beta_1) \quad \forall \beta_0, \beta_1 \,.$$

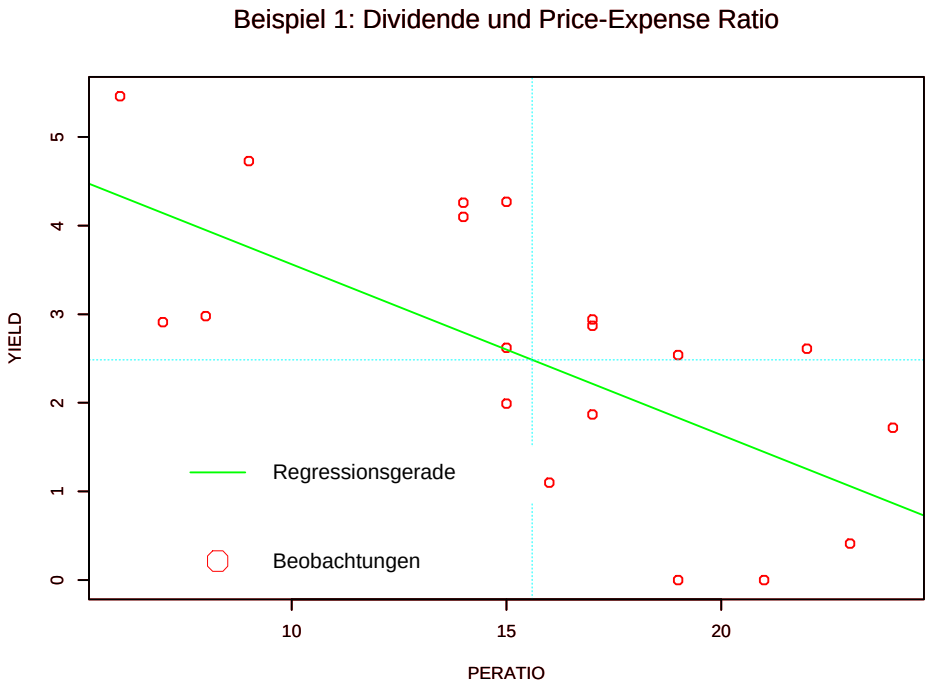

Abbildung 2.1: y = YIELD, x = PERATIO

$YIELD$ = Jährliche Dividendenrate in %
bezogen auf Kurswert der Aktie (share price) am 16.03.90

$$PERATIO = \frac{Kurswert\ (16.03.90)}{EPS}$$

$$EPS = \frac{Nettogewinn - Dividenden}{Anzahl\ der\ Aktien}$$ bzgl. 1989 earnings

Satz 2.3 (KQ-Schätzer)

1. $b_1 = \dfrac{SS_{xy}}{SS_{xx}} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

   $b_0^* = \bar{Y}$

   $b_0 = \bar{Y} - b_1\bar{x}$

2. *Ein unverzerrter Schätzer für* σ^2 *ist*

   $s^2 = \hat{\sigma}^2 = \dfrac{SS_{Res}}{n-2} = \dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$

   n: Anzahl der Beobachtungen
   2: Anzahl der unbekannten Parameter

Satz 2.4 (Eigenschaften der KQ-Schätzer)

1. $E(b_1) = \beta_1$
2. $E(b_0) = \beta_0,\ E(b_0^*) = \beta_0^*$
3. b_1 *und* b_0^* *sind unkorreliert.*
4. (b_0, b_1) *und* $\boldsymbol{e} = (Y_1 - \hat{Y}_1, \ldots, Y_n - \hat{Y}_n)^T$ *sind unkorreliert.*
5. $Var(b_1) = \dfrac{\sigma^2}{SS_{xx}}$
6. $Var(b_0^*) = \dfrac{\sigma^2}{n}$
7. $Var(b_0) = \sigma^2(\dfrac{1}{n} + \dfrac{\bar{x}^2}{SS_{xx}})$
8. *Unter Normalverteilungsannahme gilt:*
   8a) (b_0, b_1) *sind normalverteilt.*
   8b) (b_0, b_1) *und* s^2 *sind unabhängig.*

Seien nun $\boldsymbol{Y} = (Y_1, \ldots Y_n)^T$, $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, $\boldsymbol{e} = (e_1, \ldots e_n)^T$, $\mathbf{1} = (1, \ldots, 1)^T$ und $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$.

Mit $\hat{Y}_i = b_0 + b_1 x_i,\ e_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n$ und
$\hat{\boldsymbol{Y}} \in span(\mathbf{1}, \boldsymbol{x}),\ \boldsymbol{e} \in span(\mathbf{1}, \boldsymbol{x})^{\perp}$
folgen

$$\hat{\boldsymbol{Y}} \perp \boldsymbol{e} \text{ und } \hat{\boldsymbol{Y}} - \bar{Y} \cdot \mathbf{1} \perp \boldsymbol{e}. \tag{2.4}$$

Satz 2.5 (Streuungszerlegung)

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{SS_{Total}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{SS_{Reg}} + \underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{SS_{Res}}$$

totale Variabilität	*durch Regression erklärte Variabilität*	*nicht erklärte Variabilität*

Das **Bestimmtheitsmaß** R^2 (coefficient of determination) misst die Güte des gewählten Modells.

Def. 2.6 (Multipler Korrelationskoeffizient R)

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = 1 - \frac{SS_{Res}}{SS_{Total}}$$

$R = sign(b_1) \cdot \sqrt{R^2}$ *heißt* **multipler Korrelationskoeffizient**.

Es gilt:

$$0 \leq R^2 \leq 1 \tag{2.5}$$

$R^2 = 1 \Leftrightarrow Y_i = b_0 + b_1 x_i, \quad i = 1, \ldots, n$
$R^2 = 0 \Leftrightarrow SS_{Res} = SS_{Total} \Leftrightarrow$ Regression bringt nichts

2.3 Einfache lineare Regression durch den Ursprung

Der Vollständigkeit halber werden kurz die Gleichungen für das Modell ohne "intercept" β_0 eingeschoben.

$$Y_i = \beta x_i + \epsilon_i \tag{2.6}$$

$$b = \hat{\beta} = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2 \tag{2.7}$$

$$\tilde{R}^2 = \sum_{i=1}^{n} \hat{Y}_i^2 / \sum_{i=1}^{n} Y_i^2, \ \hat{Y}_i = b x_i \tag{2.8}$$

Ab jetzt gelte wieder das Modell (2.1).

2.4 Tests und Konfidenzintervalle

2.4.1 Testen von Hypothesen

Satz 2.7 (Verteilung von SS_{Res}, SS_{Total}, SS_{Reg})
Vor.: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i$ *iid* $N(0, \sigma^2),\ i = 1, \ldots, n$
Beh.:

1. $SS_{Res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \sim \sigma^2 \mathcal{X}^2_{n-2}$
2. SS_{Res} *und* SS_{Reg} *sind unabhängig.*
3. *Falls zusätzlich gilt:* $H_0 : \beta_1 = 0$
   $SS_{Reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \sim \sigma^2 \mathcal{X}^2_1$ *(unter* H_0*)*
   $SS_{Total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \sigma^2 \mathcal{X}^2_{n-1}$ *(unter* H_0*)*
   $F = \dfrac{SS_{Reg}}{SS_{Res}/(n-2)} \sim F_{1,n-2}$

Testvorschrift:

$H_0 : \beta_1 = 0,$	$H_1 : \beta_1 \neq 0$
Nullhypothese	Alternative

H_0 ist beim Niveau α (z.B. $\alpha = 0.05, 0.01$) abzulehnen, falls $F > F_{1-\alpha;1,n-2}$

2.4.2 Analysis of Variance, ANOVA-Tabelle

df = degrees of freedom, Freiheitsgrade

Source	df	SS	MS	F
Regression	1	SS_{Reg}	SS_{Reg}	MS_{Reg}/s^2
Residual	$n-2$	SS_{Res}	s^2	
Total	$n-1$	SS_{Total}		

MS = mean square

ANOVA-Tabelle zu Beispiel 1

Source	df	SS	MS	F	P-Wert
Regression	1	18.60	18.60	11.36	0.0034
Residual	18	29.48	1.64		
Total	19	48.08			

H_0 ist zu verwerfen beim Niveau $\alpha = 0.05$ und $\alpha = 0.01$
H_0 ist anzunehmen beim Niveau $\alpha = 0.001$

Def. 2.8 (P-Wert; p-value)
Der P-Wert ist das kleinste α-Niveau, für das die Hypothese noch verworfen werden kann.

2.4.3 t-Test

Generell gilt für eine mit m Freiheitsgraden studentverteilte Zufallsvariable

$$T_m^2 = F_{1,m}. \tag{2.9}$$

Für den Signifikanztest zur Hypothese $H_0 : \beta_1 = 0$ gegen die Alternative $H_1 : \beta_1 \neq 0$ ist

$$t = \frac{b_1 - \beta_1}{se(b_1)}. \tag{2.10}$$

studentverteilt mit n-2 Freiheitsgraden, wobei $se(b_1) = \frac{s}{\sqrt{SS_{xx}}}$ den Standardfehler (standard error) von b_1 bezeichnet.
Der Standardfehler eines Schätzers (hier b_1) ist die Schätzung für seine (unbekannte) Standardabweichung (hier $\sigma/\sqrt{SS_{xx}}$). Dabei muss die entsprechende Schätzung der Varianz (hier von b_1) erwartungstreu sein und sollte selbst möglichst kleine Varianz haben.

2.4.4 Konfidenzintervalle (KI) für Parameter

Mit $a = t_{1-\frac{\alpha}{2};n-2} \cdot se(b_1)$ gilt

$$1 - a = P(b_1 - a \leq \beta_1 \leq b_1 + a) \tag{2.11}$$

allgemein: $a =$ Quantil $\cdot$ Standard-Error Punktschätzer

Analog hierzu erhält man ein KI für β_0 mit
$se(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$

KI für erwarteten Response (mean response) $E(Y \mid x_0)$

$\hat{Y}(x_0) = b_0 + b_1 x_0, \quad se(E(Y \mid x_0)) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$
$c = t_{1-\frac{\alpha}{2};n-2} \cdot se(E(Y \mid x_0))$
$1 - \alpha = P(\hat{Y}(x_0) - c \leq \beta_0 + \beta_1 x_0 \leq \hat{Y}(x_0) + c)$

Prognose-Intervall (prediction interval) für $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$

$\hat{Y}_0 = b_0 + b_1 x_0$
$se(\hat{Y}_0 - Y_0) = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$
$d = t_{1-\frac{\alpha}{2};n-2} \cdot se(\hat{Y}_0 - Y_0)$
$1 - \alpha = P(\hat{Y}_0 - d \leq Y_0 \leq \hat{Y}_0 + d)$

2.5 Variablen-Transformationen

Oft deuten Scatterplots auf Varianzinhomogenität hin; vgl. Mietspiegel-Daten mit $x = Wfl,\ Z = NM$.
In diesem Beispiel führt z.B. die Transformation $Y := \tilde{Z} := lnZ$ zu einem zufriedenstellenderen Ergebnis (Vor. $Z_i > 0$)

Man hätte auch eine Transformation aus der Klasse
$\tilde{Z} := Z^\gamma$, $0 < \gamma < 1$, z.B. $\tilde{Z} = \sqrt{Z}$ wählen können. (Vor. $Z_i \geq 0$)

Am praktischen Beispiel sucht man geeignete Transformationen durch Probieren und Vergleich der Scatterplots transformierter Daten.
Dabei verfolgt man das Ziel: Die Residuen sollen keine erkennbaren Muster aufweisen.

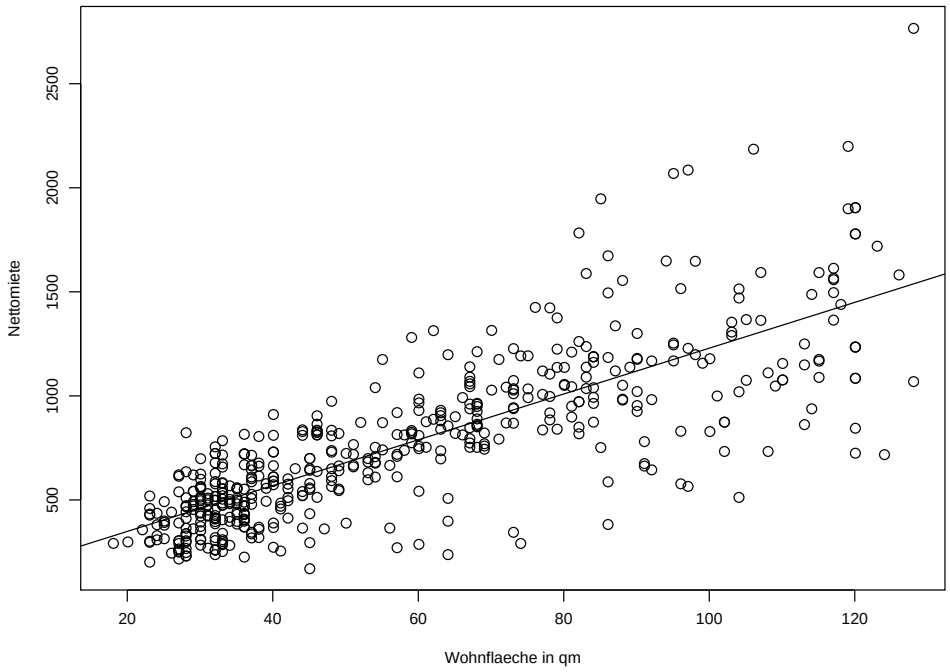

Abbildung 2.2: y = Nettomiete, x = Wohnflaeche

*

2.6 Ein erster Blick auf die Residuen

Seien
$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
$E(\epsilon_i) = 0,\ Cov(\epsilon_i, \epsilon_j) = \delta_{ij}\sigma^2,\ i, j = 1, \ldots, n$ oder ϵ_i iid $N(0, \sigma^2)$

b_0, b_1 KQ-Schätzer für β_0, β_1

$\hat{Y}_i = b_0 + b_1 x_i$, $e_i = Y_i - \hat{Y}_i$ (Residuen).

Achtung:

e_i	$\neq$	ϵ_i
berechenbar		nicht beobachtbar

Die e_i sind korrelierte Zufallsvariablen, verwendet zur

1. Verifizierung der Modellwahl (siehe z.B. "Korngrößen")
2. Überprüfung Varianzhomogenität:
   $Var(\epsilon_i) = \sigma^2$, konstant; verletzt bei "Mietspiegel"
3. Identifizierung von Ausreißern
4. Überprüfung der Normalverteilungs-Annahme (siehe "Korngrößen")

2.6.1 Modellwahl

a) Die Residuen sollen (aufgetragen gegen die x_i oder später gegen die $\hat{Y}_i$) *kein Muster* aufweisen (vgl. Abb. 2.3).

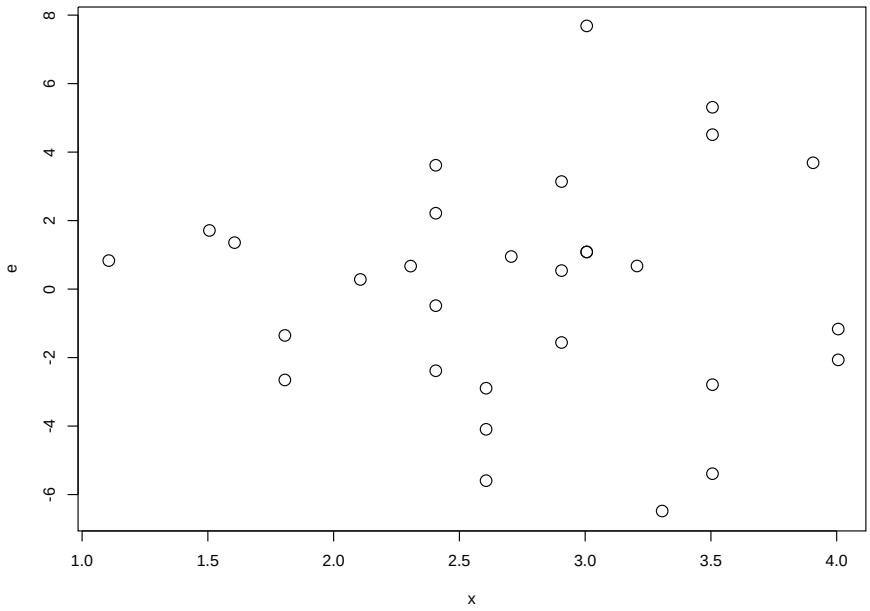

Abbildung 2.3: Kein erkennbares Muster in den Residuen

Die Muster in Abb. 2.4 wären z.B. durch Daten-Transformationen wie $a\,sin(t\,x_i + c)$ oder $a(x_i - c)^2$ zu eliminieren:

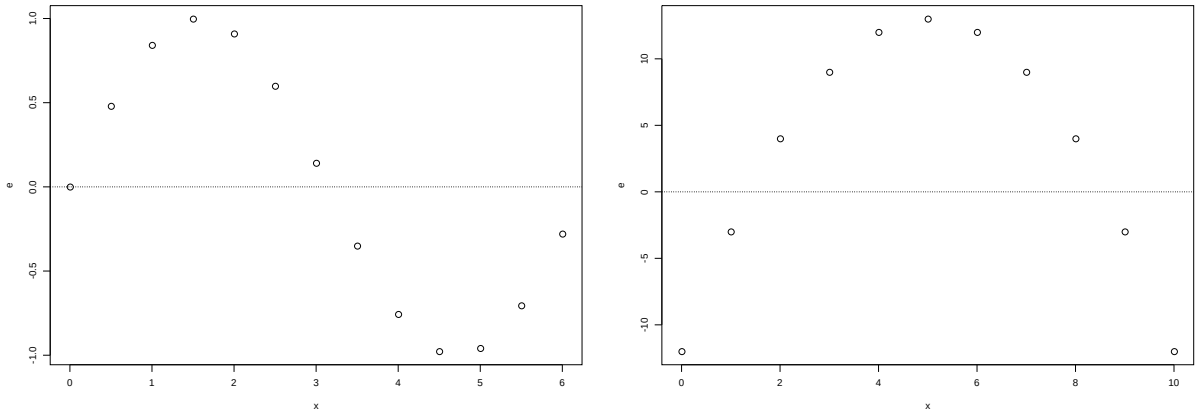

Abbildung 2.4: Funktionale Muster in den Residuen

b) Verletzung der Modellannahmen
z.B. weil die x_i den $E(Y_i \mid x_i)$ nicht gut genug erklären.
Mögliche Abhilfe, z.B.

$$Y_i = \beta_0 + \beta_1 g(x_i) + \epsilon_i,$$

wobei g nach grafischer Inspektion der Daten durch "Ausprobieren" zu finden ist.
In manchen Fällen hilft für lokale Approximationen ein kubischer Ansatz der Form

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i\,.$$

Dies ist ein multiples Regressionsmodell; siehe Kap. 3.

2.6.2 Varianzhomogenität

In der Regel gilt $Var(\epsilon_i) = \sigma_i^2,\ i = 1, \ldots, n.$
Beim Beispiel "Mietspiegel" wächst z.B. die Varianz mit der Wohnfläche.

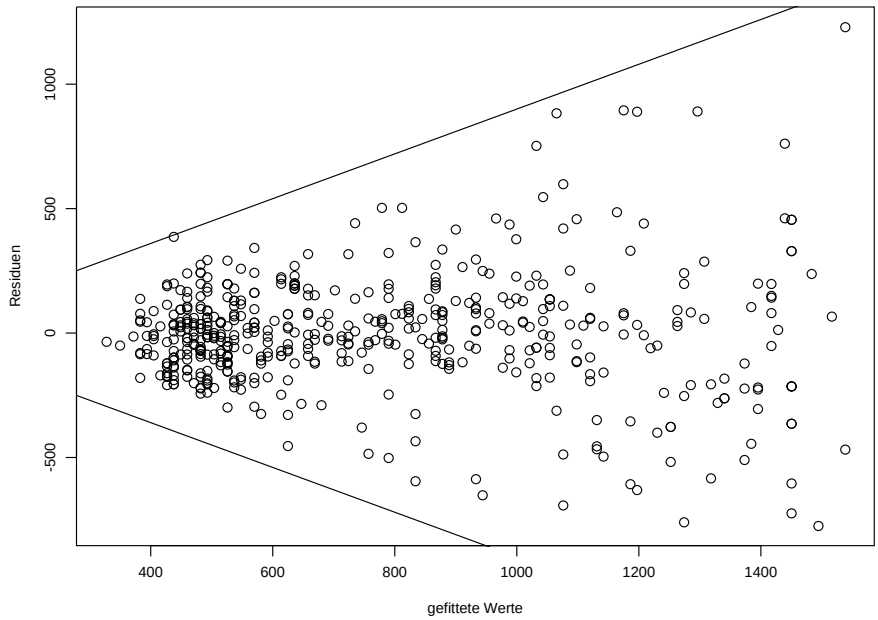

Abbildung 2.5: Inhomogene Varianz

Abhilfe: Transformationen wie $\tilde{Y}_i = log(Y_i)$ oder $\tilde{Y}_i = w_i Y_i,\ w_i > 0$, bekannte Gewichte.

2.6.3 Ausreißer

Seien $\boldsymbol{X} = (\,\mathbf{1}, \boldsymbol{x} - \mathbf{1} \cdot \bar{x})$ und $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = (h)_{ij}$, wobei in der Einfachen Linearen Regression gilt $h_{ii} = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{SS_{xx}}$ und

$$\frac{1}{n} \leq h_{ii} \leq 1,\ \sum_{i=1}^{n} h_{ii} = 2\,.$$

$Cov(Y_i, \hat{Y}_i) = \sigma^2\, h_{ii} = Var(\hat{Y}_i)$,
$Var(e_i) = Var(Y_i - \hat{Y}_i) = \sigma^2(1 - h_{ii})$, $Cov(e_i, e_j) = \sigma^2(1 - h_{ij}) \neq 0$, für $i \neq j$.
$h_{ii} = 1$ ist möglich. In diesem Fall hat man $Var(e_i) = 0$ unabhängig von der Fehlerquadratsumme SS_{Res}.
Mit $s^2 = \dfrac{1}{n-2}\sum_{i=1}^{n} e_i^2,\ se(e_i) = s\sqrt{1 - h_{ii}}$
definiert man die "standardisierten Residuen"(standardized residuals)

$$r_i := \frac{e_i}{se(e_i)} = \frac{Y_i - \hat{Y}_i}{s\sqrt{1 - h_{ii}}},\quad i = 1, \ldots, n\,.$$

Achtung: Die r_i sind weder Student-verteilt noch unkorreliert.

Abhilfe: *Jackknifed Residuals* (manchmal auch: *Studentized Residuals*)

$$r_j^* = \frac{Y_j - \hat{Y}_j^{(j)}}{se(Y_j - \hat{Y}_j^{(j)})} = r_j / \sqrt{\frac{n-2-r_j^2}{n-3}},\ j = 1, \ldots, n$$

wobei für $j = 1, \ldots, n,\ b_0^{(j)}, b_1^{(j)}$ und $s_{(j)}^2$ aus dem Modell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n,\ i \neq j$$

mit $n-1$ Beobachtungen ohne " j " berechnet sind; also $\hat{Y}_j^{(j)} = b_0^{(j)} + b_1^{(j)} x_j$.

Vorteil: Ein eventuell großes Residuum $e_j = Y_j - \hat{Y}_j$ geht nicht in die Berechnungen von $\hat{Y}_j^{(j)}, s_{(j)}^2$ und somit von r_j^* ein.
Faustregel: Beobachtung j mit $|r_j^*| > 2 \approx t_{0.975;n-3}$ ist Kandidat für *Ausreißer*.

2.6.4 Normalverteilungs-Annahme, QQ-Plots

$e_i \sim N(0, \sigma^2(1-h_{ii}))$, $r_i^* \approx N(0,1)$
geordnet: $r_{(1)}^* \leq r_{(2)}^* \leq \cdots \leq r_{(n)}^*$.

Zu $U_1, \ldots, U_n$ iid $N(0,1)$ definiert man die Ordnungsstatistiken:

$$U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}\,.$$

Aus der Statistik 1 ist bekannt: $E(U_{(k)}) = \Phi^{-1}(\frac{k}{n+1})$
Die Punktewolke $\left(\Phi^{-1}(\frac{k}{n+1}), U_{(k)}\right)_{k=1,\ldots,n}$ sollte also "eng" um die Winkelhalbierende streuen; im Fall der *jacknifed residuals* also die Punktewolke $(\Phi^{-1}(\frac{k}{n+1}), r_{(k)}^*)$, vgl. etwa den Datensatz "Korngrößen".

$n = 39:\ \Phi^{-1}(1/40) = \Phi^{-1}(0.025) = -1.96$
$n = 99:\ \Phi^{-1}(1/100) = \Phi^{-1}(0.01) = -2.33$
$n = 999:\ \Phi^{-1}(1/1000) = \Phi^{-1}(0.001) = -3.09$

Die folgenden Grafiken wurden durch Simulationen von unabhängigen normalverteilten Fehlern erzeugt; stellen also den Idealfall dar.

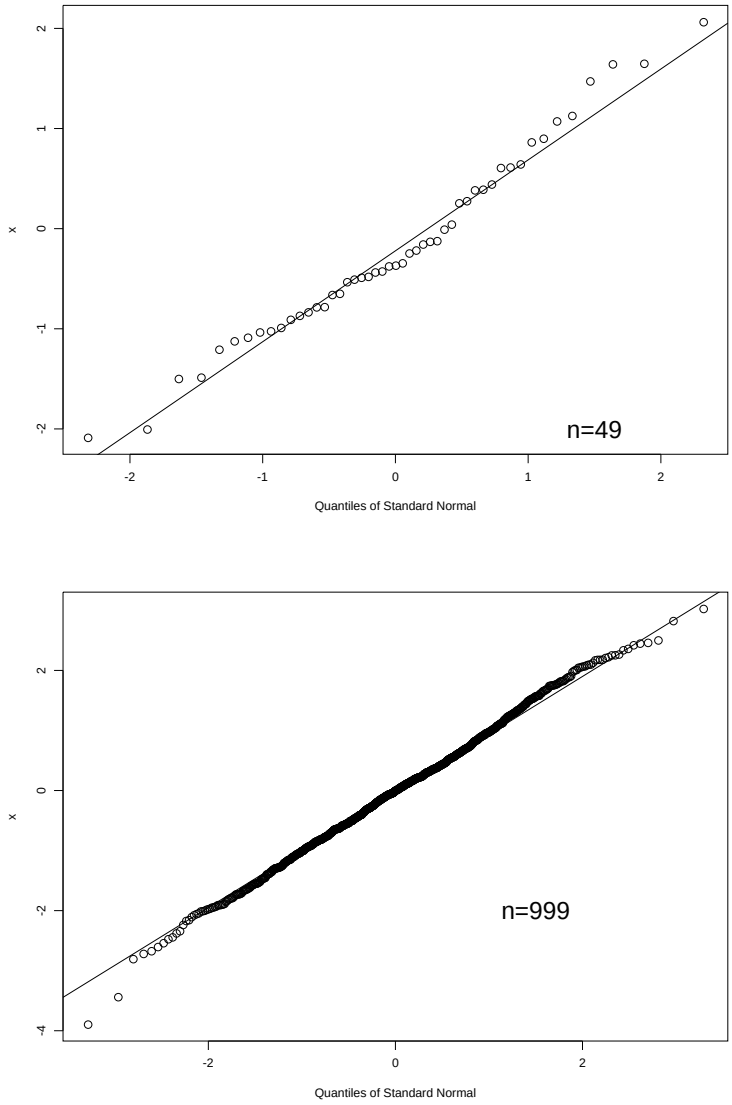

Abbildung 2.6: Ideale QQ-Plots für exakt normalverteilte Fehler

2.7 Simultane Inferenz bei Einfacher Linearer Regression

$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i,\ \epsilon_i$ iid $N(0, \sigma^2),\ i = 1, \ldots, n$
Die KQ-Schätzer b_o^* und b_1 sind unabhängig.

$$b_0^* = \bar{Y} \sim N(\beta_0^*, \sigma^2/n),\ b_1 \sim N(\beta_1, \sigma^2/SS_{xx})$$

Also ist

$$\frac{1}{\sigma^2}[n(\bar{Y} - \beta_0^*)^2 + (b_1 - \beta_1)^2 SS_{xx}] \sim \mathcal{X}_2^2$$

und unabhängig von s^2 (nach Satz 2.4), womit

$$W := \frac{(n(\bar{Y} - \beta_0^*)^2 + (b_1 - \beta_1)^2 SS_{xx})/2\sigma^2}{((n-2)s^2)/(n-2)\sigma^2} = \frac{n(\bar{Y} - \beta_0^*)^2 + (b_1 - \beta_1)^2 SS_{xx}}{2s^2} \sim F_{2,n-2}$$

also

$$P(n(\bar{Y} - \beta_0^*)^2 + (b_1 - \beta_1)^2 SS_{xx} \leq 2s^2 F_{1-\alpha;2,n-2}) = 1 - \alpha$$

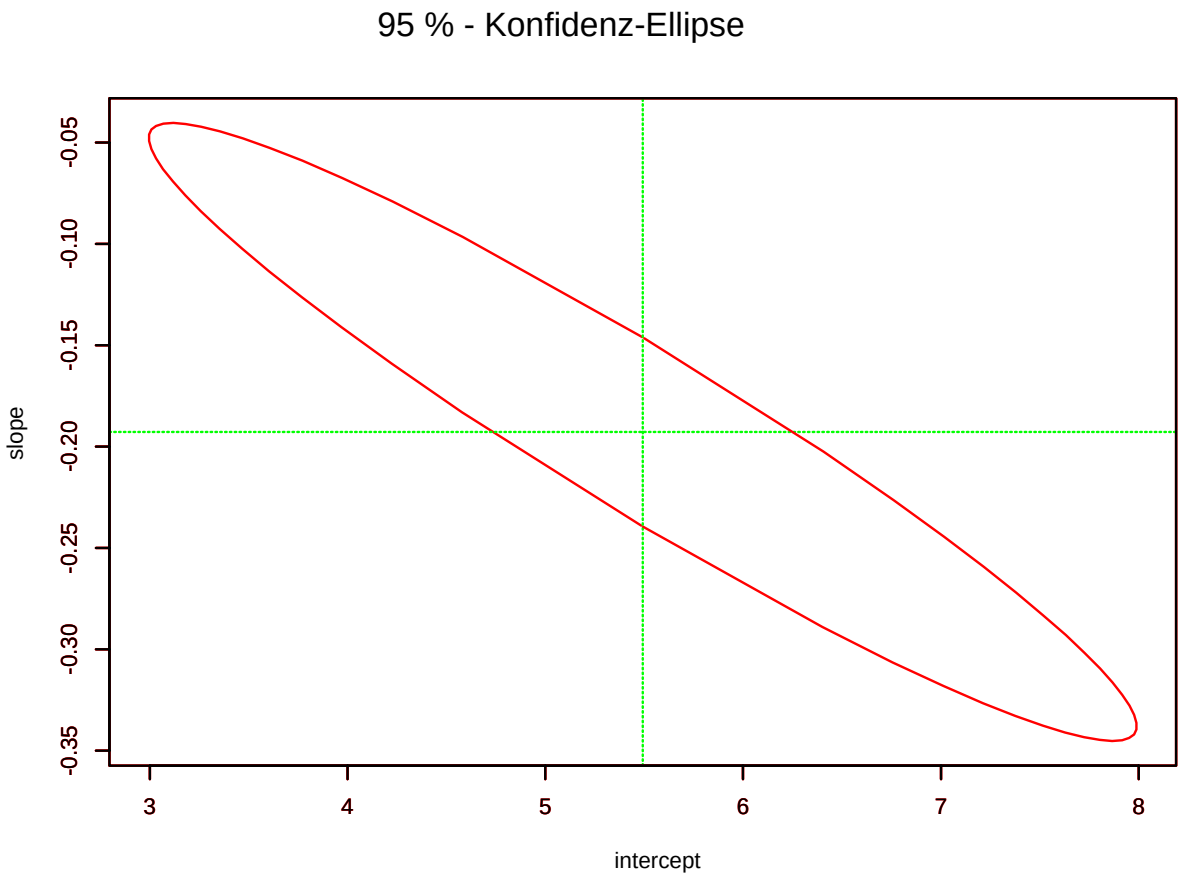

Abbildung 2.7: Simultane Inferenz

Satz 2.9 (Simultane Konfidenzintervalle)
Für ϵ_i iid $N(0, \sigma^2)$ erhält man die simultane $(1-\alpha)$-Konfidenzregion für (β_0^*, β_1) aus:

$$E := \{(\beta_0^*, \beta_1) \mid n(\bar{Y} - \beta_0^*)^2 + (b_1 - \beta_1)^2 SS_{xx} \leq 2s^2 F_{1-\alpha;2,n-2}\}$$

E ist eine Ellipse für (β_0^*, β_1) mit Mittelpunkt $(\bar{Y}, b_1)$.
Die dazu äquivalente Konfidenz-Ellipse für (β_0, β_1) um den Mittelpunkt $(\bar{Y} - b_1\bar{x}, b_1)$ ergibt sich aus dem Zusammenhang $\beta_0 = \beta_0^* - \beta_1\bar{x}$.

2.8 Gemeinsame Konfidenzbänder für den *mean response* nach Working-Hotelling

E sei der $100(1-\alpha)\%$ Konfidenzbereich von Satz 2.9. Ein $100(1-\alpha)\%$-Konfidenzband für

$$E(Y \mid x_0) = \beta_0 + \beta_1 x_0$$

ist gegeben durch

$$\begin{aligned} A(x_0) &:= \{\beta_0 + \beta_1 x_0 \mid (\beta_0, \beta_1) \in E\} \\ &= \{\hat{Y}(x_0) \pm \sqrt{2F_{1-\alpha;2,n-2}}\; s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}\;\} \end{aligned}$$

Dies ist eine andere (kleinere) Konstante verglichen mit den individuellen Konfidenzintervallen für $E(Y \mid x_0)$.

2.9 Bonferroni-Ansatz zu gemeinsamer Interferenz

Allgemein gilt:
Sei θ ein Parameter der Verteilung von Y und sei $[L(Y), U(Y)]$ ein $(1-\alpha)$-Konfidenzintervall für θ, dann gilt

$$P(\underbrace{Y : \theta \in [L(Y), U(Y)]}_{A}) = 1 - \alpha.$$

Seien $[L_1(Y), U_1(Y)]$ $(1-\alpha)$-Konfidenzintervalle für θ_1
$[L_2(Y), U_2(Y)]$ $(1-\alpha)$-Konfidenzintervalle für θ_2

und A_1, A_2 entsprechend zu A (oben) definiert.

$\Rightarrow P(A_1) = 1 - \alpha = P(A_2)$ und

$$\begin{aligned} P(A_1 \cap A_2) &= 1 - P(A_1^c \cup A_2^c) \\ &= 1 - [\underbrace{P(A_1^c)}_{\alpha} + \underbrace{P(A_2^c)}_{\alpha} - \underbrace{P(A_1^c \cap A_2^c))}_{\geq 0}] \\ &\geq 1 - 2\alpha \end{aligned}$$

Damit ist ein $100(1-2\alpha)\%$-Konfidenzbereich für (θ_1, θ_2) gegeben durch

$$\{(\theta_1, \theta_2) : \theta_1 \in [L_1(Y), U_1(Y)], \theta_2 \in [L_2(Y), U_2(Y)]\}$$

Anwendung:

$$\{(\beta_0, \beta_1) : \beta_0 \in [b_0 \pm \hat{se}(b_0)\; t_{1-\frac{\alpha}{4};n-2}], \beta_1 \in [b_1 \pm \hat{se}(b_1)\; t_{1-\frac{\alpha}{4};n-2}]\}$$

ist ein $100(1-\alpha)\%$-Konfidenzbereich für (β_0, β_1).

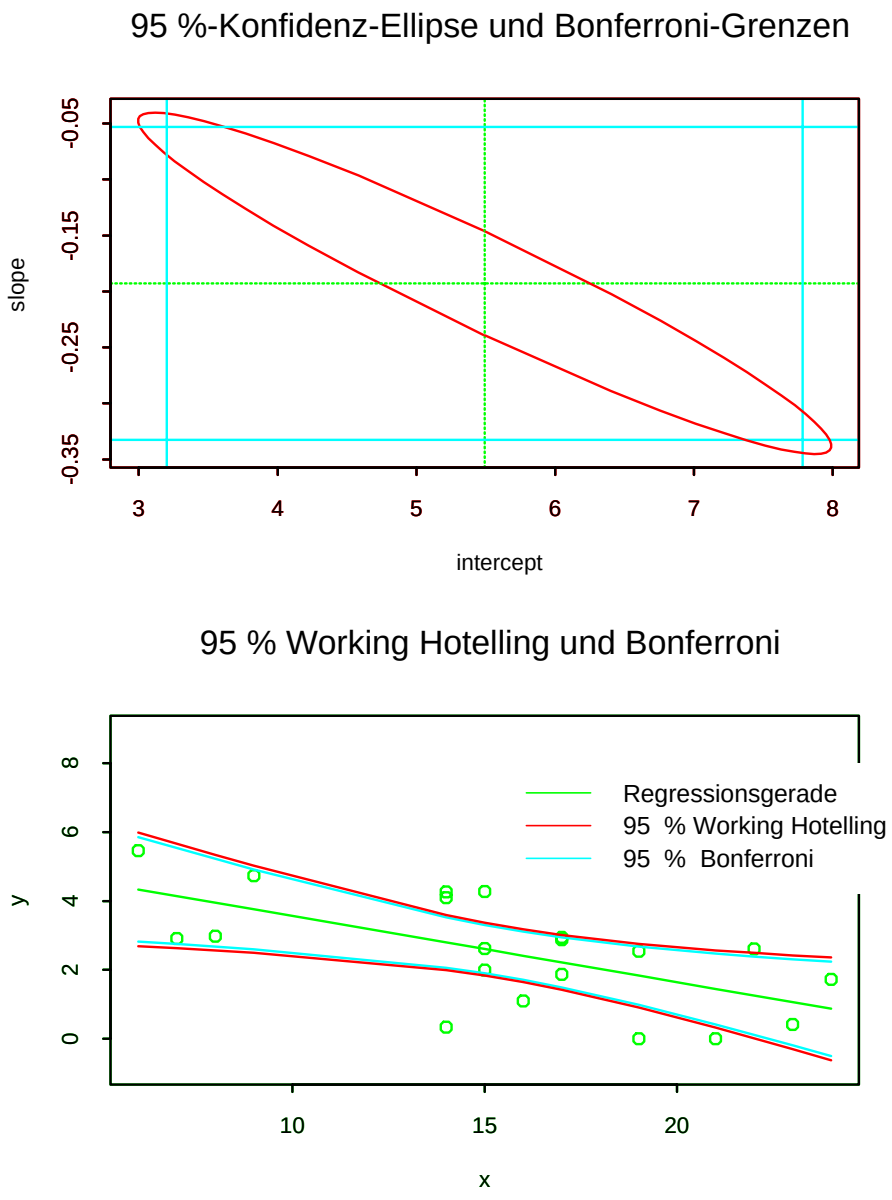

Abbildung 2.8: Vergleich der verschiedenen Methoden

Kapitel 3

Multiple Lineare Regression

3.1 Modell

Wir betrachten das Modell der Multiplen Linearen Regression für gegebene Daten $(y_i, x_{i0}, x_{i1}, ..., x_{ik})_{i=1,...,n}$. In den meisten Fällen wählt man $x_{i0} = 1$. Damit erhält man den Ansatz

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \,, \tag{3.1}$$

wobei $E(\epsilon_i) = 0$ und $Cov(\epsilon_i, \epsilon_j) = \delta_{ij}\,\sigma^2$, $i = 1, \ldots, n$.
$\beta_0, \beta_1, \ldots, \beta_k$ sind $p = k+1$ unbekannte Parameter. Damit sind die $y_i, i = 1, \ldots, n$, Zufallsvariable (ZV). In der Folge wird nicht mehr zwischen den ZV y_i und deren Realisierungen (Daten) unterschieden. Die x_{ij}, $i = 1, \ldots n$, $j = 1, \ldots, k$, werden als Ausprägungen nicht stochastischer Kovariablen angesehen.

Tabelle 1: Beispiel: College Spending 1994; vgl. [Dielman et al. (1996)], p. 202

Variable	Abk.	Erklärung
SAT	S	average SAT score
TOP10	T	freshmen in the top 10% of their high school class (percentage)
ACCRATE	A	acceptance rate (percentage)
PHD	P	faculty with PHD (percentage)
RATIO	R	student faculty ratio
SPEND	S	educational spending per full-time equivalent (FTE) student (in dollars)
GRADRATE	G	Graduation Rate (percentage)
ALUMNI	Al	alumni giving rate (percentage)

SAT = Scholastic Aptitude Test (standardisierter Test in der letzten High School Klasse (= K12 in D) für den Hochschulzugang)

Auf dieses Beispiel werden wir in diesem Abschnitt bis hin zur Variablenselektion zurückgreifen.

Seien nun $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, $\mathbf{1} = (1, \ldots, 1)^T$,

$$\boldsymbol{X} = (\,\vec{\boldsymbol{x}}_0,\, \vec{\boldsymbol{x}}_1,\, \ldots,\, \vec{\boldsymbol{x}}_k\,) \;=\; \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ . \\ . \\ \boldsymbol{x}_n^T \end{pmatrix}, \tag{3.2}$$

wobei in der Regel, aber nicht unbedingt: $\vec{\boldsymbol{x}}_0 = \mathbf{1}$.

Das Modell lässt sich also folgendermaßen schreiben:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}\,,\; \boldsymbol{Cov}(\boldsymbol{\epsilon}) = \sigma^2\,\boldsymbol{I}\,, \tag{3.3}$$

$\boldsymbol{y} \in \mathbb{R}^n$: stochastische Zielvariable (response)
$\boldsymbol{\beta} \in \mathbb{R}^p$: unbekannter Parametervektor
$\boldsymbol{X} \in \mathbb{R}^{n,p}$: Designmatrix, Regressormatrix (feste, nicht stochastische Daten)
(fast immer: $Rang(\boldsymbol{X}) = p$)
$\boldsymbol{\epsilon} \in \mathbb{R}^n$: unbeobachtbare Fehlervariable

Oft gilt zusätzlich: ϵ_i normalverteilt,
d.h. ϵ_i iid $N(0, \sigma^2)$, $\quad i = 1, \ldots, n$.

Im College-Spending-Beispiel ist man am Response $SPEND$ interessiert.

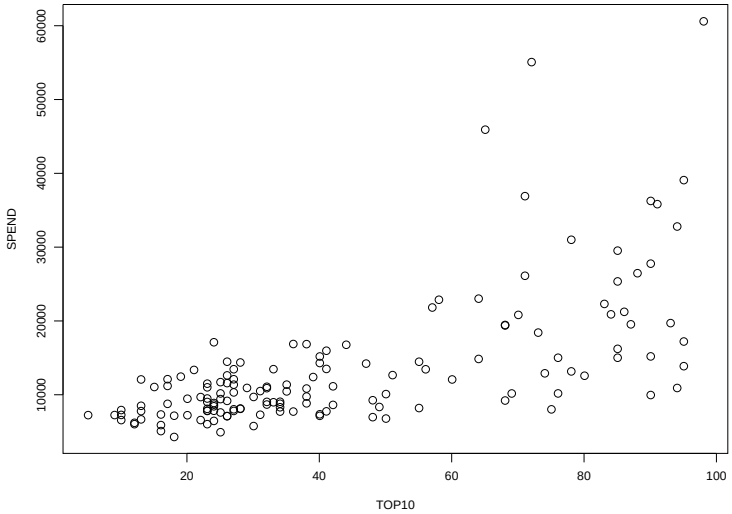

Abbildung 3.1: Scatterplot von SPEND gegen die Kovariable TOP10

An den TOP10-Werten über 60 sieht man, dass die Annnahme konstanter Varianz nicht haltbar ist. Dieses Bild verbessert sich wesentlich, wenn man die logarithmierte Zielvariable betrachtet. Wir arbeiten also fortan mit dem Response $log(SPEND)$.

Ähnlich reagieren auch die Plots gegenüber der Kovariablen SAT.

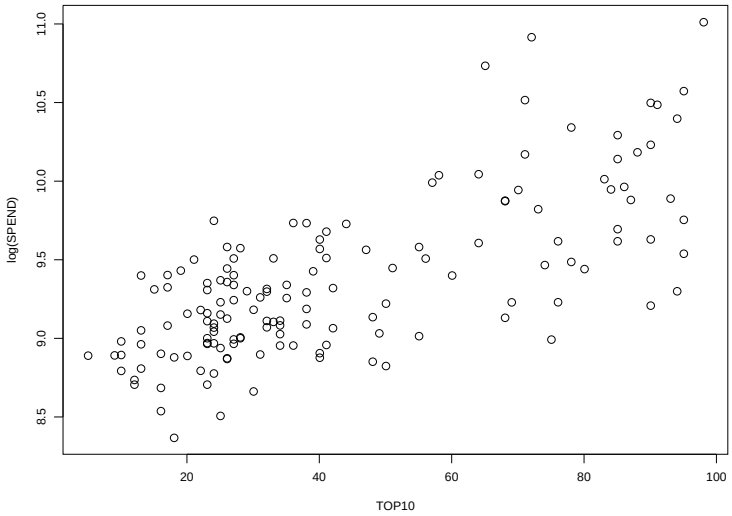

Abbildung 3.2: Scatterplot von log(SPEND) gegen die Kovariable TOP10

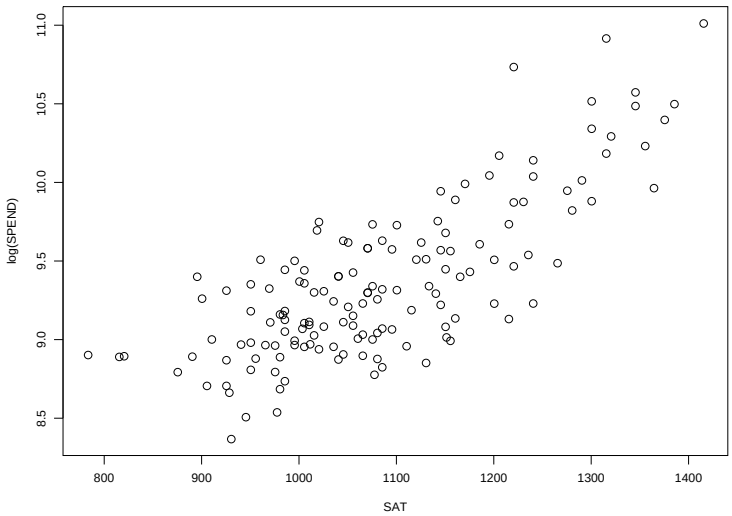

Abbildung 3.3: Scatterplot von log(SPEND) gegen die Kovariable SAT

Diese "Pairs-Plots" gegenüber potenziellen Kovariablen rechtfertigen den linearen (und additiven) Modellansatz noch nicht vollständig. Sie könnten diesen aber widerlegen. Dann müsste man noch weitere Transformationen von Kovariablen bzw. Response suchen.

Bei Prognosen und Vorhersageintervallen ist in diesem Beispiel natürlich zu berücksichtigen, dass nicht mit den Originaldaten gearbeitet wird. Rücktransformationen ergeben sich hier mit der Exponentialfunktion und allgemein aus der entsprechenden Umkehrabbildung.

3.2 KQ-Methode (LS estimation)

Wir setzen für das Folgende stets $Rang(\boldsymbol{X}) = p$ voraus.
Damit erhält man die KQ-Schätzungen $\boldsymbol{b} = (b_0, b_1, \ldots, b_k)^T$ für $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ aus den Normalgleichungen

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y} \tag{3.4}$$

durch

$$\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\,. \tag{3.5}$$

Mit der sog. "Hat-Matrix"

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \tag{3.6}$$

lautet die Schätzung $\hat{\boldsymbol{y}}$ für $\boldsymbol{y}$

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}. \tag{3.7}$$

Damit ergibt sich der Residuenvektor zu

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \tag{3.8}$$

wobei

$$SS_{Res} = \boldsymbol{e}^T\boldsymbol{e} = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.9}$$

3.3 Eigenschaften der KQ-Schätzungen

Satz 3.1 (Gauß-Markov-Theorem)
Vor.: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, Rang(\boldsymbol{X}) = p, E(\boldsymbol{\epsilon}) = \boldsymbol{0}, \mathbf{Cov}(\boldsymbol{\epsilon}) = \sigma^{\mathbf{2}}\boldsymbol{I}$
Beh.:

1. $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ *ist BLUE (best linear unbiased estimator) für* $\boldsymbol{\beta}$
2. $Cov(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
3. $s^2 = SS_{Res}/(n-p)$ *ist unverzerrter Schätzer für* σ^2

3.4 Projektionen und $\mathcal{X}^2$-Verteilung

Dieser Abschnitt dient zur Vorbereitung der Verteilungsaussagen für Schätzer.

Lemma 3.2 (Rechnen mit Kovarianzen (Wiederholung))
Wir betrachten die Zufallsvariablen V, W mit existierenden 2. Momenten und
$E(V) = \mu_V,\ E(W) = \mu_W,\ Var(V) = \sigma_V^2,\ Var(W) = \sigma_W^2,$
$cov(V,W) := E[(V-\mu_V)(W-\mu_w)] = \rho\sigma_V\sigma_W.$
Weiterhin seien
$\boldsymbol{Z} := (V,W)^T,\ \boldsymbol{\mu}_Z = (\mu_V, \mu_W)^T$
und
$Cov(\boldsymbol{Z}) := E\left[(\boldsymbol{Z}-\boldsymbol{\mu}_Z)(\boldsymbol{Z}-\boldsymbol{\mu}_Z)^T\right].$

Dann gilt mit $\boldsymbol{a} \in \mathbb{R}^2, \boldsymbol{A} \in \mathbb{R}^{2,2}$

1. $Var(V) = cov(V,V)$
2. $Cov(\boldsymbol{Z}) = \begin{pmatrix} \sigma_V^2 & \rho\sigma_V\sigma_W \\ \rho\sigma_V\sigma_W & \sigma_W^2 \end{pmatrix}$
3. $Var(\boldsymbol{a}^T\boldsymbol{Z}) = Cov(\boldsymbol{a}^T\boldsymbol{Z}) = \boldsymbol{a}^T Cov(\boldsymbol{Z})\boldsymbol{a}$
4. $Cov(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}Cov(\boldsymbol{Z})\boldsymbol{A}^T.$

Die Aussagen von 3. und 4. gelten allgemein für $\boldsymbol{Z}, \boldsymbol{a} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n,n}$.

Def. 3.3 ($\mathcal{X}^2$-Verteilung)
Seien V_i iid $N(0,1),\ i = 1,\ldots,r.$
Dann heißt

$$W^2 = \sum_{i=1}^{r} V_i^2 \sim \mathcal{X}_r^2,$$

$\mathcal{X}^2$-verteilt mit r Freiheitsgraden.

Satz 3.4 (Additionsformel)
Sind W_1^2 bzw. W_2^2 unabhängig $\mathcal{X}_{r_1}^2$ bzw. $\mathcal{X}_{r_2}^2$, dann ist

$$W^2 := W_1^2 + W_2^2 \sim \mathcal{X}_{r_1+r_2}^2$$

Satz 3.5 (Satz von Cochran)

Vor.: $\boldsymbol{w} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$
$\boldsymbol{P} \in \mathbb{R}^{n,n},\ \boldsymbol{P}^T = \boldsymbol{P},\ \boldsymbol{P}^2 = \boldsymbol{P},\ Rang(\boldsymbol{P}) = r$
Beh.: $W_1^2 := \boldsymbol{w}^T \boldsymbol{P} \boldsymbol{w}/\sigma^2 \sim \mathcal{X}_r^2$
$W_2^2 := \boldsymbol{w}^T (\boldsymbol{I} - \boldsymbol{P}) \boldsymbol{w}/\sigma^2 \sim \mathcal{X}_{n-r}^2$
W_1^2 *und* W_2^2 *sind unabhängig.*

Beweis:
Die Komponenten w_i von $\boldsymbol{w}$ sind unabhängig, also auch die Komponenten z_i von $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{w}$, für orthogonale Matrizen $\boldsymbol{A}$. Insbesondere gilt dann $\boldsymbol{z}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

Die Projektionsmatrix $\boldsymbol{P}$ ist symmetrisch und hat r Eigenwerte 1 und $n - r$ Eigenwerte 0. Insbesondere gibt es eine Orthonormalbasis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r, \boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_n$,
Matrizen $\boldsymbol{U}_r = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r),\ \boldsymbol{U}_{n-r} = (\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_n),\ \boldsymbol{U} = (\boldsymbol{U}_r, \boldsymbol{U}_{n-r}) \in \mathbb{R}^{n,n},\ \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$,
so dass gilt:

$$\boldsymbol{P} = \boldsymbol{U}_r \boldsymbol{U}_r^T,\ \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{U}_{n-r} \boldsymbol{U}_{n-r}^T,\ \boldsymbol{U}_r^T \boldsymbol{U}_{n-r} = \boldsymbol{0}.$$

Mit $\boldsymbol{A} := \boldsymbol{U}^T$ und $\boldsymbol{z} := \boldsymbol{U}^T \boldsymbol{w} = \begin{pmatrix} \boldsymbol{U}_r^T \boldsymbol{w} \\ \boldsymbol{U}_{n-r}^T \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} \boldsymbol{z}_r \\ \boldsymbol{z}_{n-r} \end{pmatrix}$
sind alle Komponenten von $\boldsymbol{z}$ unabhängig, insbesondere auch $\boldsymbol{z}_r$ und $\boldsymbol{z}_{n-r}$.
Außerdem:
$W_1^2 = \boldsymbol{w}^T \boldsymbol{P} \boldsymbol{w}/\sigma^2 = \boldsymbol{z}_r^T \boldsymbol{z}_r \sim \mathcal{X}_r^2$ ist unabhängig von
$W_2^2 = \boldsymbol{w}^T (\boldsymbol{I} - \boldsymbol{P}) \boldsymbol{w}/\sigma^2 = \boldsymbol{z}_{n-r}^T \boldsymbol{z}_{n-r} \sim \mathcal{X}_{n-r}^2$ □

Satz 3.6 (Verteilung von $\boldsymbol{b}$ und s^2)

Vor.: *Seien* $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{X} \in \mathbb{R}^{n,p},\ Rang(\boldsymbol{X}) = p,$
ϵ_i *iid* $N(0, \sigma^2),\ i = 1, \ldots, n$
$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T,\ SS_{Res} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}.$
Beh.:

1. $\boldsymbol{b} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$
2. $Q_{\boldsymbol{\beta}} := (\boldsymbol{b} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{b} - \boldsymbol{\beta})/\sigma^2 \sim \mathcal{X}_p^2$
3. $\boldsymbol{b}$ *ist unabhängig von* $s^2 = \hat{\sigma}^2 = \dfrac{SS_{Res}}{n - p}$
4. $SS_{Res}/\sigma^2 \sim \mathcal{X}_{n-p}^2$
5. $Q_{\boldsymbol{\beta}}$ *und* SS_{Res} *sind unabhängig.*

Beweis:

1. $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}$ ist als lineare Transformation von $\boldsymbol{y}$ normalverteilt. Erwartungswert und Kovarianz folgen aus dem Gauß-Markov-Theorem (Satz 3.1).

2.,4.,5. mit Satz von Cochran:
Mit $\boldsymbol{w} := \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ und $\boldsymbol{P} := \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ sind die Voraussetzungen des Satzes von Cochran (Satz 3.5) erfüllt und es gilt mit
$\boldsymbol{H}^T = \boldsymbol{H}$, $\boldsymbol{H}^2 = \boldsymbol{H}$ ($\boldsymbol{I} - \boldsymbol{H}$ analog) und $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$:

$$\begin{aligned} W_1^2 &:= \boldsymbol{w}^T\boldsymbol{P}\boldsymbol{w}/\sigma^2 = \boldsymbol{w}^T\boldsymbol{H}\boldsymbol{H}\boldsymbol{w}/\sigma^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma^2 \\ &= (\boldsymbol{b}^T - \boldsymbol{\beta}^T)\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta})/\sigma^2 \\ &= Q_{\boldsymbol{\beta}} \sim \mathcal{X}_p^2 \end{aligned}$$

ist unabhängig von

$$\begin{aligned} W_2^2 &:= \boldsymbol{w}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{w}/\sigma^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma^2 \\ &= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}/\sigma^2 \\ &= SS_{Res}/\sigma^2 \sim \mathcal{X}_{n-p}^2, \end{aligned}$$

wobei $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$ benutzt wurde.

3. Die Unabhängigkeit von $\boldsymbol{b}$ und SS_{Res} zeigt man mit Techniken analog zum Beweis des Satzes von Cochran.

□

3.5 Varianzzerlegung, Bestimmtheitsmaß

Betrachte $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_1) \in \mathbb{R}^{n,p},\ Rang(\boldsymbol{X}) = p,$
$\boldsymbol{1} = (1, \ldots, 1)^T,\ \boldsymbol{H}_0 = \boldsymbol{1} \cdot \boldsymbol{1}^T/n$

Lemma 3.7
$SS_{Total} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y}$

Beweis: Übung

Lemma 3.8
Mit $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ gilt:

1. $\boldsymbol{H}\boldsymbol{H}_0 = \boldsymbol{H}_0\boldsymbol{H} = \boldsymbol{H}_0$
2. $\boldsymbol{H} - \boldsymbol{H}_0$ *ist Orthogonalprojektion.*

Beweis: Übung

Satz 3.9
Vor.: $SS_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,\ SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Beh.: $SS_{Total} = SS_{Reg} + SS_{Res}$.

Beweisskizze:
$SS_{Res} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$
$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} \Rightarrow SS_{Reg} = \boldsymbol{y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{y}$
und
$\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{H}_0 - (\boldsymbol{H} - \boldsymbol{H}_0)$,
Rest Übung.

Def. 3.10 (Bestimmtheitsmaß, R^2)
Mit den Bezeichnungen von 3.9
$SS_{Total} = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ SS_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,\ SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
und unter der Vor. $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_1)$ *heißen*

$R^2 := \frac{SS_{Reg}}{SS_{Total}} = 1 - \frac{SS_{Res}}{SS_{Total}}$ **Bestimmtheitsmaß**,

R **multipler Korrelationskoeffizient**;

und mit $s_0^2 := SS_{Total}/(n-1)$, $s^2 = SS_{Res}/(n-p)$:
$R^2_{adj} := 1 - \frac{s^2}{s_0^2} = 1 - (1 - R^2)\frac{n-1}{n-p}$ **adjustiertes Bestimmtheitsmaß**.

Bemerkung 3.11

1. $0 \leq R^2 \leq 1$
2. $R^2 = 1 \Leftrightarrow SS_{Res} = 0$
3. $R^2(\boldsymbol{X}_0, \boldsymbol{X}_1) \geq R^2(\boldsymbol{X}_0)$ *für Zerlegungen* $\boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1)$
4. R^2_{adj} *kann für ein Modell mit mehr Variablen fallen (wenn* SS_{Reg} *kaum mehr wächst)*

Bemerkung 3.12 (Varianzzerlegung)

3.9 ist Spezialfall folgenden Prinzips:

$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2)$ *mit* $\boldsymbol{X}_0 = \mathbf{1},\ \boldsymbol{X}_1 \in \mathbb{R}^{n,p_1},\ \boldsymbol{X}_2 \in \mathbb{R}^{n,p_2}$

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},\ p = 1 + p_1 + p_2,\ Rang(\boldsymbol{X}) = p$

1. $\boldsymbol{y} = \boldsymbol{X}_0\beta_0 + \boldsymbol{\epsilon}_0 = \mathbf{1} \cdot \beta_0 + \boldsymbol{\epsilon}_0$ *(evtl. "underfitting")*
   $\boldsymbol{H}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T = \mathbf{1} \cdot \mathbf{1}^T/n$
   $SS_{Res}(\boldsymbol{X}_0) = SS_{Total} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y}$

2. $\boldsymbol{y} = \boldsymbol{X}_0\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$
   $\boldsymbol{X}_{01} = (\boldsymbol{X}_0, \boldsymbol{X}_1)$
   $\boldsymbol{H}_{01} = \boldsymbol{X}_{01}(\boldsymbol{X}_{01}^T\boldsymbol{X}_{01})^{-1}\boldsymbol{X}_{01}^T$
   $SS(\boldsymbol{X}_1 \mid \boldsymbol{X}_0) := SS_{Reg}(\boldsymbol{X}_{01}) = SS_{Res}(\boldsymbol{X}_0) - SS_{Res}(\boldsymbol{X}_{01}) =$
   $= \boldsymbol{y}^T(\boldsymbol{H}_{01} - \boldsymbol{H}_0)\boldsymbol{y}$

3. $\boldsymbol{y} = \boldsymbol{X}_0\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ *(evtl. "overfitting")*
   $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
   $SS(\boldsymbol{X}_1, \boldsymbol{X}_2 \mid \boldsymbol{X}_0) := SS_{Reg}(\boldsymbol{X})$
   $SS(\boldsymbol{X}_2 \mid \boldsymbol{X}_{01}) := SS_{Res}(\boldsymbol{X}_{01}) - SS_{Res}(\boldsymbol{X}) = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_{01})\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} =$
   $\boldsymbol{y}^T(\boldsymbol{H} - \boldsymbol{H}_{01})\boldsymbol{y}$
   $SS_{Res}(\boldsymbol{X}) = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_0 + \boldsymbol{H}_0 - \boldsymbol{H}_{01} + \boldsymbol{H}_{01} - \boldsymbol{H})\boldsymbol{y} =$
   $\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{H} - \boldsymbol{H}_{01})\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{H}_{01} - \boldsymbol{H}_0)\boldsymbol{y} =$
   $SS_{Total} - SS(\boldsymbol{X}_1 \mid \boldsymbol{X}_0) - SS(\boldsymbol{X}_2 \mid \boldsymbol{X}_{01})$

Damit hat man folgende allgemeinere Varianzzerlegung:

$SS_{Total} =$ $SS(\boldsymbol{X}_1 \mid \boldsymbol{X}_0)$ durch $\boldsymbol{X}_1$ erklärte Varianz

$+ SS(\boldsymbol{X}_2 \mid \boldsymbol{X}_{01})$ durch $\boldsymbol{X}_2$ zusätzlich erklärte Varianz

$+ SS_{Res}(\boldsymbol{X})$ von $\boldsymbol{X}$ nicht erklärte Restvarianz

Wegen $SS_{Total} = SS(\boldsymbol{X} \mid \boldsymbol{X}_0) + SS_{Res}(\boldsymbol{X})$ folgt durch Vertauschung der Rollen von $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$

$$SS(\boldsymbol{X} \mid \boldsymbol{X}_0) = SS(\boldsymbol{X}_1 \mid \boldsymbol{X}_0) + SS(\boldsymbol{X}_2 \mid \boldsymbol{X}_{01})$$
$$= SS(\boldsymbol{X}_2 \mid \boldsymbol{X}_0) + SS(\boldsymbol{X}_1 \mid \boldsymbol{X}_{02}).$$

Es kommt also auf die Reihenfolge an, in welcher die Variablen ins Modell kommen.

Das Beispiel College Spending liefert beispielsweise:

$r1 <- lm(log(SPEND\ TOP10 + ACCRATE + RATIO + GRADRATE)$
$anova(r1)$

Analysis of Variance Table
Response: log(SPEND)
Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
TOP10	1	18.78383	18.78383	274.3323	0.0000000
ACCRATE	1	1.32889	1.32889	19.4080	0.0000207
RATIO	1	6.79686	6.79686	99.2661	0.0000000
GRADRATE	1	0.15791	0.15791	2.3063	0.1310776
Residuals	142	9.72289	0.06847		

Die Spalte Sum of Sq gibt die sequential Sums of Squares. Sie hängen von der Reihenfolge der Variablen ab, mit der sie in das Modell eingehen. Dies kann man wie folgt sehen:

$r2 <- lm(log(SPEND)\ RATIO + GRADRATE + TOP10 + ACCRATE)$
$anova(r2)$

Analysis of Variance Table
Response: log(SPEND)
Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
RATIO	1	15.86151	15.86151	231.6528	0.00000000
GRADRATE	1	6.55064	6.55064	95.6702	0.00000000
TOP10	1	4.46617	4.46617	65.2271	0.00000000
ACCRATE	1	0.18915	0.18915	2.7625	0.09870084
Residuals	142	9.72289	0.06847		

Die so von einem Statistik-Programm-Paket erzeugten Resultate liefern die numerische Grundlage für wichtige Spezialfälle der folgenden Tests.

3.6 Tests linearer Hypothesen

Wichtige Hypothesen (z.B. bei der Variablenauswahl) haben etwa die Form
$H_0: \begin{matrix}\beta_2 = 0 \\ \beta_4 = 0\end{matrix}\Big\}$ zugehörige x-Variablen können aus dem Modell entfernt werden
Alternative H_1 : nicht H_0.

3.6.1 Test zur allgemeinen linearen Hypothese

Def. 3.13 (Allgemeine lineare Hypothese)

Geg.: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ Rang(\boldsymbol{X}) = p,\ \boldsymbol{X} \in \mathbb{R}^{n,p}$
$\boldsymbol{C} \in \mathbb{R}^{r,p},\ Rang(\boldsymbol{C}) = r,\ \boldsymbol{d} \in \mathbb{R}^r$

Die Restriktion

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \tag{3.10}$$

heißt allgemeine lineare Hypothese.
Die Alternative lautet stets: H_1 : *nicht* H_0.

Satz 3.14 (Restringiertes KQ-Problem)

Unter den Bedingungen von 3.13 gilt mit

$$SS_{Res} = \min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2\} = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2$$

$$SS_{H_0} = \min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \mid \underbrace{\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}}_{sprich:unterderBed.H_0}\}$$

$$SS_{H_0} = SS_{Res} + (\boldsymbol{C}\boldsymbol{b} - \boldsymbol{d})^T[\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\boldsymbol{b} - \boldsymbol{d})$$

Beweis: Lagrange-Ansatz und Differenzieren

Beispiel 3.15

$$\boldsymbol{X} = (\underbrace{\mathbf{1}, \vec{\boldsymbol{x}}_2}_{\boldsymbol{X}_1}, \underbrace{\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_3}_{\boldsymbol{X}_2})$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$$

$$H_0 : \beta_1 = \beta_3 = 0 \Leftrightarrow \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{C \in \mathbb{R}^{2,4}, r=2} \cdot \boldsymbol{\beta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \boldsymbol{d}$$

Falls H_0 *richtig ist, gilt das kleinere Modell*

$$\boldsymbol{y} = (\mathbf{1}, \vec{\boldsymbol{x}}_2)\begin{pmatrix} \beta_0 \\ \beta_2 \end{pmatrix} + \boldsymbol{\epsilon}$$

Satz 3.16 (F-Test zur linearen Hypothese)

Vor.: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \epsilon_i$ *iid* $N(0, \sigma^2)$
$Rang(\boldsymbol{X}) = p,\ \boldsymbol{X} \in \mathbb{R}^{n,p},\ \boldsymbol{C} \in \mathbb{R}^{r,p},\ Rang(\boldsymbol{C}) = r,\ \boldsymbol{d} \in \mathbb{R}^r$
SS_{H_0} *aus 3.14*

Beh: Falls die Hypothese

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$$

richtig ist, gilt

$$F = \frac{(SS_{H_0} - SS_{Res})/r}{SS_{Res}/(n-p)} \sim F_{r,n-p}.$$

Testvorschrift:
H_0 ist beim Niveau $\alpha,\ 0 < \alpha < 1$, z.B. $\alpha = 0.05$ oder $\alpha = 0.01$ abzulehnen, falls

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

Häufigste Anwendung:

3.6.2 Tests für Variablen-Subsets

$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2),\ \boldsymbol{X}_1 \in \mathbb{R}^{n,p-r},\ \boldsymbol{X}_2 \in \mathbb{R}^{n,r}$

$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},\ \boldsymbol{\beta}_1 \in \mathbb{R}^{p-r},\ \boldsymbol{\beta}_2 \in \mathbb{R}^r,\ Rang(\boldsymbol{X}) = p$

$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0},\ \boldsymbol{C} = (\boldsymbol{0}, \boldsymbol{I}_r)$

$SS_{Res} = SS_{Res}(\boldsymbol{X}) = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T,\ \boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1$

$SS_{H_0} = SS_{Res}(\boldsymbol{X}_1) = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y}$

Damit kann

$$F = \frac{(SS_{Res}(\boldsymbol{X}_1) - SS_{Res})/r}{SS_{Res}/(n-p)}$$

ohne die umständliche Formel von 3.14 leicht - insbesondere mit jedem gängigen Statistik-Programmpaket bestimmt werden.

Hilfsmittel in S-Plus: anova(r.red, r.full), wobei r.full <- lm $(y \sim x1 + \cdots + xp)$
r.red <- lm $(y \sim x1 + \cdots + x(p-r))$

Noch spezieller und wichtiger:

3.6.3 Partielle F-Tests für einzelne Parameter

Hypothese $H_j : \beta_j = 0,$ für ein $j \in \{1, \ldots, p\}$

$$F_j = \frac{(SS_{H_j} - SS_{Res})/1}{SS_{Res}/(n-p)} \sim F_{1,n-p}$$

Im Zähler: Vergrößerung der Residuenquadratsumme beim Weglassen der Variablen j.
Bekanntlich: $F_j = T_j^2$, wobei

$$T_j = \frac{b_j}{se(b_j)} \sim t_{n-p}$$

analog zur Einfachen Linearen Regression.

b_j, der zugehörige Standardfehler (standard error) $se(b_j)$ und T_j bzw. F_j finden sich im Ausdruck jedes guten Statistik-Programmpakets.

3.7 Konfidenz- und Prognoseintervalle

Seien

$\boldsymbol{x}_0 = (1, x_{01}, x_{02}, \ldots, x_{0,p-1})^T \in \mathbb{R}^p$
$y(\boldsymbol{x}_0) \sim N(\boldsymbol{x}_0^T\boldsymbol{\beta}, \sigma^2)$ unabhängig von $\boldsymbol{y} = (y_1, \ldots, y_n)^T$
$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n),\ \boldsymbol{X} \in \mathbb{R}^{n,p},\ Rang(\boldsymbol{X}) = p$
$\hat{y}(\boldsymbol{x}_0) = \boldsymbol{x}_0^T\boldsymbol{b},\ \boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
$s^2 = SS_{Res}/(n-p)$

Satz 3.17 (Konfidenz- und Prognoseintervalle)

1. $E[\hat{y}(\boldsymbol{x}_0)] = \boldsymbol{x}_0^T\boldsymbol{\beta} = E[y(\boldsymbol{x}_0)]$
   $\hat{y}(\boldsymbol{x}_0) = \boldsymbol{x}_0^T\boldsymbol{b}$*: berechenbar*
   $y(\boldsymbol{x}_0)$*: tritt erst (in der Zukunft) ein*

2. $Var[\hat{y}(\boldsymbol{x}_0)] = Var[\boldsymbol{x}_0^T\boldsymbol{b}] = \sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$

3. $Var[y(\boldsymbol{x}_0) - \hat{y}(\boldsymbol{x}_0)] = \sigma^2(1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)$

4. $(1-\alpha)$*-Konfidenz-Intervall für den "mean response"*
   $v := t_{1-\frac{\alpha}{2};n-p} \cdot s \cdot \sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0},\ 0 < \alpha < 1$
   $P(\hat{y}(\boldsymbol{x}_0) - v \leq E[y(\boldsymbol{x}_0)] = \boldsymbol{x}_0^T\boldsymbol{\beta} \leq \hat{y}(\boldsymbol{x}_0) + v) = 1 - \alpha$

5. $(1-\alpha)$*-Prognose-Intervall*
   $w := t_{1-\frac{\alpha}{2};n-p} \cdot s \cdot \sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0},\ 0 < \alpha < 1$
   $P(\hat{y}(\boldsymbol{x}_0) - w \leq y(\boldsymbol{x}_0) \leq \hat{y}(\boldsymbol{x}_0) + w) = 1 - \alpha$

6. *Für* $\boldsymbol{1} = (1, \ldots, 1)^T, \boldsymbol{x} = (x_1, \ldots, x_n)^T, \boldsymbol{x}_0 = (1, x_0)^T$ *und* $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x})$ *erhält man die Aussagen von Abschnitt 2.4.4*

3.8 Qualität des Modells, Hat-Matrix

3.8.1 Fragen

1. Ist das gewählte Modell richtig?
   Wurden die Daten so *transformiert* und die möglichen Variablen so ausgewählt, dass der lineare Modellansatz gerechtfertigt ist? Gibt es Ausreißer unter den Beobachtungen?

2. Wie lautet ein optimales Modell hinsichtlich einer möglichst kleinen Varianz des Vorhersagefehlers $Var[y(\boldsymbol{x}_0) - \hat{y}(\boldsymbol{x}_0)]$?

3.8.2 Ansätze

Mit $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ . \\ . \\ . \\ \boldsymbol{x}_n^T \end{pmatrix}$ gilt für den Vorhersagefehler

$d_i^2 = Var(y(\boldsymbol{x}_i) - \hat{y}(\boldsymbol{x}_i)) = \sigma^2(1 - \underbrace{\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-t}\boldsymbol{x}_i}_{h_{ii}})$

$\hat{d}_i^2 = s^2(1 - h_{ii})$, h_{ii} Diagonalelemente der *Hat-Matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Die Qualität der Beobachtung i hängt ab von h_{ii} und $s^2 = \dfrac{1}{n-p}\sum_{i=1}^{n}[y_i - \hat{y}(\boldsymbol{x}_i)]^2$.

Satz 3.18 (Eigenschaften von $\boldsymbol{H}$)

1. $\boldsymbol{H}$ *ist Projektionsmatrix, d.h.* $\boldsymbol{H}^T = \boldsymbol{H}$, $\boldsymbol{H}^2 = \boldsymbol{H}$ *(Orthogonalprojektion)*
2. $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$
3. $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, $SS_{Res} = \boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$
4. $0 \le h_{ii} \le 1$
5. $\dfrac{1}{n} \le h_{ii} \le 1$ *(für Modelle mit intercept)*

Konsequenzen:

Wegen $Rang(\boldsymbol{H}) = tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p$ sollten die h_{ii} idealerweise $\approx p/n$ sein.

Def. 3.19 (High Leverage Points)
Beobachtungen i mit $h_{ii} > 2p/n$ *heißen* **Hebelpunkte** *oder* **high leverage points**.

Def. 3.20 (Jackknife Methode)
Sei $\boldsymbol{b}$ *der KQ-Schätzer zum vollen Modell*
$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n$
und $b_{(j)}$ *der KQ-Schätzer zum reduzierten Modell*
$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n,\ i \neq j$
mit n-1 Beobachtungen.

Def. 3.21 (Cook's Distance)
Mit den Bezeichnungen von 3.20 heißt

$$D_j := \frac{(\boldsymbol{b} - \boldsymbol{b}_{(j)})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{b} - \boldsymbol{b}_{(j)})}{ps^2}$$

Cook's Distance *der Beobachtung j.*

Bemerkung 3.22

1. D_j *ist Standardisierung von* $\boldsymbol{b} - \boldsymbol{b}_{(j)}$, *vgl. Satz 3.6*
2. D_j *ist einfach zu berechnen. Es gilt:*
$$D_j = \frac{e_j^2}{(1 - h_{jj})^2} \cdot \frac{h_{jj}}{s^2 \cdot p}$$
3. D_j *ist groß, falls* $h_{jj} \approx 1$ *oder das j-te Residuum* $e_j = y_j - \hat{y}(\boldsymbol{x}_j)$ *betragsmäßig groß ist.*
4. D_j *misst die Verschiebung der Konfidenz-Region für* $\boldsymbol{\beta}$, *wenn die j-te Beobachtung weggelassen wird.*
5. $D_j \sim F_{p,n-p}$
6. *Mit* s^2 *anstelle von* σ^2 *in Satz 3.6 erhält man durch*
$$\{\boldsymbol{\beta} : \frac{(\boldsymbol{b} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{b} - \boldsymbol{\beta})}{ps^2} \leq F_{1-\alpha;p,n-p}\}$$
eine $(1 - \alpha)$*-Konfidenzregion für* $\boldsymbol{\beta}$.

Def. 3.23 (Influential observation)
Beobachtungen mit
$D_j > F_{0.5;p,n-p}$
heißen **einflussreiche Beobachtungen (influential observations)**.

Achtung, nicht immer gilt: "influential observation" = "Ausreißer"!

Ein Beispiel zur Demonstration: Die Beobachtungen mit den Nummern 1, 2 und 10 fallen durch ihre extreme Lage auf.

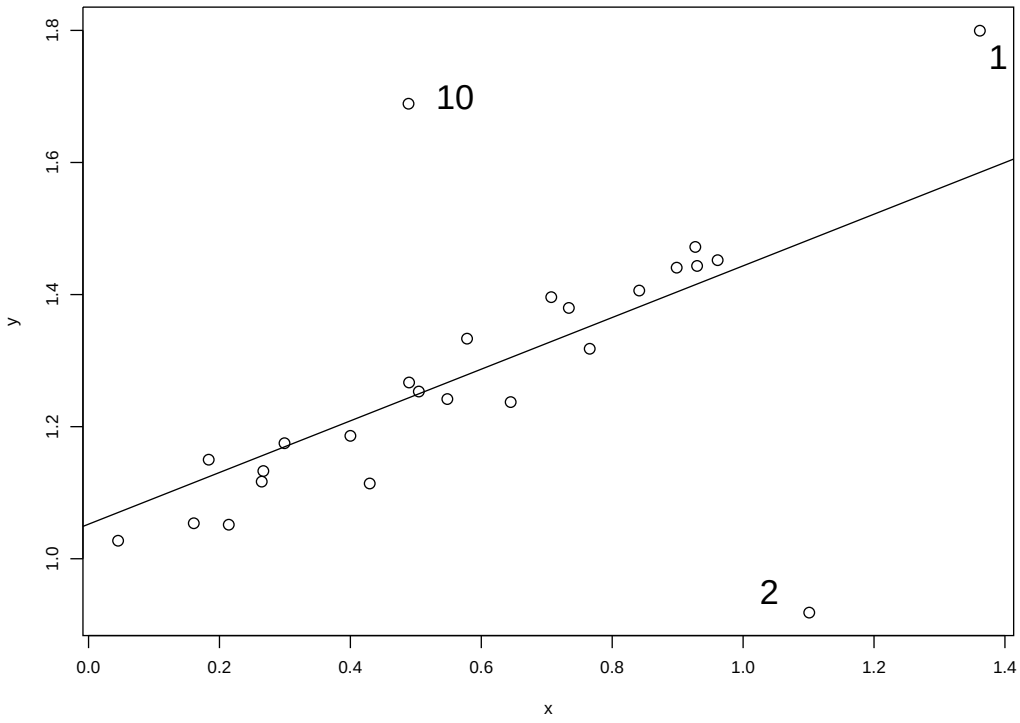

Abbildung 3.4: Plot mit auffälligen Beobachtungen

Beobachtung 1 ist wegen seiner extremen Lage am Rand der konvexen Hülle der Kovariablen-Daten ein Hebelpunkt (high leverage point): $h_{11} = 0.27 > 2p/n = 4/25 = 0.16$. Weitere Hebelpunkte existieren in diesem Datensatz nicht.
Nach dem Cook-Kriterium ist nur Beobachtung 2 eine "influential observation". Obwohl das Auge auch Beobachtung 10 als Kandidat für einen Ausreißer ansieht, schlägt das Cook-Kriterium nicht an, da die entsprechende x-Koordinate im Zentrum der anderen Beobachtungen liegt.

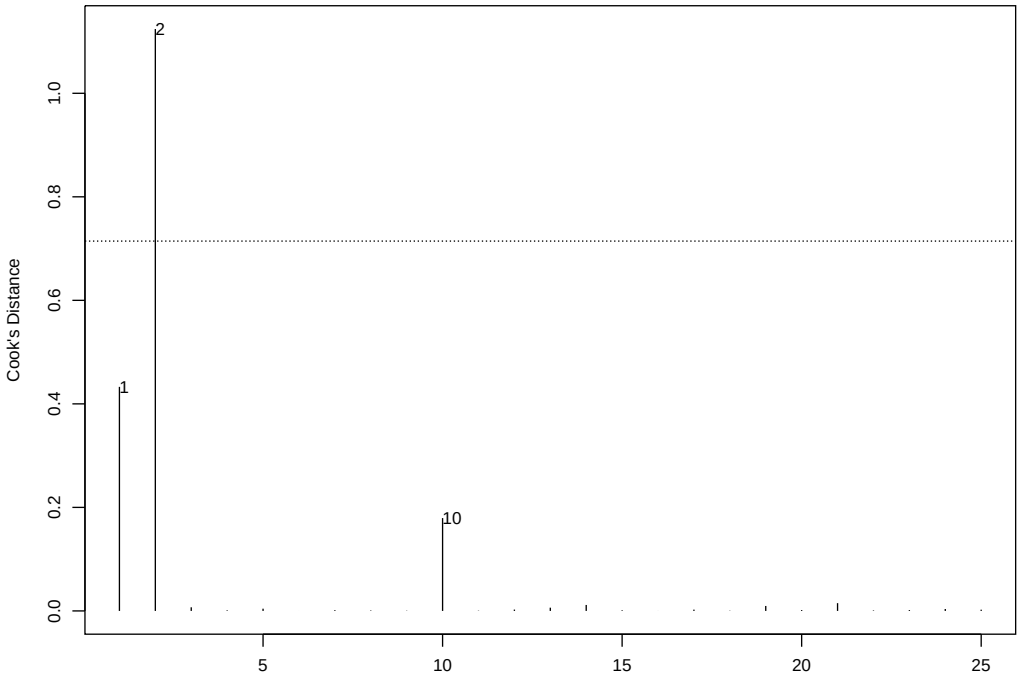

Abbildung 3.5: Cook'sche Distanz

3.9 Variablenselektion

Gegeben:

1. Daten $(y_i, x_{i1}, \ldots, x_{ik})_{i=1,\ldots,n}$
2. Zielvariable y und potenzielle Kovariablen $x_1, \ldots, x_k$
3. Modell-Terme wie $log(x_j)$, x_j^2, $\sqrt{y}, \ldots$
4. Faktoren und Interaktionen $x_j * x_l$

Ziele:

1. Welche Transformationen sind notwendig? → 2-dim. Scatter-Plots
2. Welche Variablen sollen ins Modell, welche nicht? → siehe unten

Siehe auch bsp2 in Handout 4.

Hinsichtlich Variablenauswahl gibt es zwei Fehlerarten:

- Unterspezifikation (zu wenige Variablen; "underfitting")
- Überspezifikation (zu viele Variablen; "overfitting")

Faustregel: "Lieber zu wenige als zu viele Variablen!"

Vereinbarung

Aus $m = k + 1$ Einflussgrößen (k Kovariablen, Regressoren + intercept) werden $p \leq m$ (Achtung: Hier ist $p < m$ ausdrücklich zugelassen!) ausgewählt. Seien

$$J := \{j_1, j_2, \ldots, j_p\} \subseteq \{0, 1, \ldots, k\}$$

und $\vec{\boldsymbol{x}}_{j_1}, \ldots, \vec{\boldsymbol{x}}_{j_p}$ die dem Variablensatz J entsprechenden Spalten von $\boldsymbol{X}$ mit

$$\boldsymbol{X} = (\vec{\boldsymbol{x}}_0 \equiv \mathbf{1}, \ldots, \vec{\boldsymbol{x}}_k)$$

und $\bar{J} := \{0, 1, \ldots, k\} \setminus J$.

Wir ordnen die Größen RSS_J und s_J^2 dem Modell mit Variablensatz J zu. Insgesamt o.B.d.A.

$$\boldsymbol{X} = (\boldsymbol{X}_J, \boldsymbol{X}_{\bar{J}}) \in \mathbb{R}^{n,m},\ \boldsymbol{X}_J \in \mathbb{R}^{n,p}.$$

In S-Plus und R werden oft Abkürzungen
wie $RSS_p = RSS(p) = RSS_J$
bzw. $s_p^2 = s^2(p) = s_J^2 = \dfrac{RSS_J}{n-p}$
verwendet, da $|J| = p$.

3.9.1 Unterspezifikation

Wahres Modell:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$
$$\boldsymbol{X} = (\boldsymbol{X}_J, \boldsymbol{X}_{\bar{J}}) = (\boldsymbol{X}_1, \boldsymbol{X}_2),\ \boldsymbol{X}_1 \in \mathbb{R}^{n,p_1},\ \boldsymbol{X}_2 \in \mathbb{R}^{n,p_2},\ p = p_1 + p_2,\ p_2 > 0$$

Gewähltes Modell:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1.$$

Nachteile für KQ-Schätzung $\boldsymbol{b}_1$ für $\boldsymbol{\beta}_1$:

- $\boldsymbol{b}_1$ ist verzerrt, d.h. $E(\boldsymbol{b}_1) \neq \boldsymbol{\beta}_1,\ E(\boldsymbol{y}) \neq \boldsymbol{X}_1\boldsymbol{\beta}_1$
- $E(s^2_{p_1}) > \sigma^2$
- Prognosen falsch

3.9.2 Überspezifikation

$\boldsymbol{X} = (\boldsymbol{X}_J, \boldsymbol{X}_{\bar{J}}) = (\boldsymbol{X}_1, \boldsymbol{X}_2),\ \boldsymbol{X}_1 \in \mathbb{R}^{n,p},\ \boldsymbol{X}_2 \in \mathbb{R}^{n,n-p}$

Wahres Modell: $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$
Angepasstes Modell: $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$

$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix}$ KQ-Schätzung für $\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{\beta}$.

$$\begin{aligned} \boldsymbol{H}_J = \boldsymbol{H}_1 &:= \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T \\ \boldsymbol{E} &:= \boldsymbol{X}_2^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2 \qquad \text{Orthogonalprojektion von } \boldsymbol{X}_2 \text{ auf } span(\boldsymbol{X}_1)^\perp \\ \boldsymbol{F} &:= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2 \end{aligned}$$

Es gilt:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} + \boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{F}^T & -\boldsymbol{F}\boldsymbol{E}^{-1} \\ -\boldsymbol{E}^{-1}\boldsymbol{F}^T & \boldsymbol{E}^{-1} \end{pmatrix}$$

Spezialfall: $span(\boldsymbol{X}_2) \perp span(\boldsymbol{X}_1)$
$\Rightarrow \boldsymbol{F} = \boldsymbol{0}$ und $\boldsymbol{E} = \boldsymbol{X}_2^T\boldsymbol{X}_2$

Kovarianzmatrizen von $\boldsymbol{b}_{(1)}$ und $\boldsymbol{b} = (\boldsymbol{b}_1^T, \boldsymbol{b}_2^T)^T$:

$Cov(\boldsymbol{b}_{(1)}) = \sigma^2(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}$ (kleines Modell) und
$\boldsymbol{C} := \boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{F}^T$ ist positiv semidefinit,

$\boldsymbol{C} \neq \boldsymbol{0}$, falls $\boldsymbol{X}_1^T\boldsymbol{X}_2 \neq \boldsymbol{0}$

$Cov(\boldsymbol{b}_1) = (\boldsymbol{X}^T\boldsymbol{X})^{-1} + \boldsymbol{C}$ (aus großem Modell) und
$Cov(\boldsymbol{b}_1) - Cov(\boldsymbol{b}_{(1)}) \geq 0$,
d.h. die (Ko-)Varianz von $\boldsymbol{b}_1$ wächst mit jeder hinzukommenden Variablen.

3.9.3 Kriterien zum Modellvergleich

1. Bestimmtheitsmaß R^2

$R_J^2 = 1 - \dfrac{RSS_J}{RSS_{\{0\}}}$ ist ungeeignet, da

$R_{J\cup\{k\}}^2 \geq R_J^2$ mit der Anzahl der Variablen wächst.

2. $s_J^2 = RSS_J/(n - |J|)$

Da der Nenner mit wachsender Variablenzahl kleiner wird,
kann es J_1 und J_2 geben: $s_{J_1\cup J_2}^2 > s_{J_1}^2$

3. Adjustiertes Bestimmtheitsmaß R_{adj}^2

$$R_{adj}^2(J) = 1 - [1 - R^2(J)]\frac{n-1}{n-p} = 1 - \frac{s_J^2}{s_{\{0\}}^2},$$

wobei $s_{\{0\}}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Falls s_J^2 kaum mehr wächst, kann R_{adj}^2 auch für steigende Variablenzahl fallen.

4. Mallow's C_p-Statistik

$$\hat{\sigma}^2 := s_m^2 = \frac{1}{n-m}\sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ (volles Modell)}$$

$$C_p = p + (n-p)\frac{s_J^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = C_J = \frac{RSS_J}{\hat{\sigma}^2} - n + 2p, \text{ wobei } |J| = p.$$

Ab jetzt sei $RSS_p := RSS_J$.

5. Das AIC-Kriterium

Das Akaike information criterion wurde 1973 bzw. 1974 von Akaike eingeführt.

In S-Plus:

$AIC := RSS_p + scale \cdot p \cdot k$
Eingabeparameter: *scale* und k
Voreinstellung: $k = 2$ (und meist: $scale = \hat{\sigma}^2$)
Bemerkung: $scale = 0$ liefert bei Vorwärtsselektion stets alle Variablen

In R (anders als in S-Plus):

a) $scale = 0$:
$AIC = n\ln(RSS_p/n + 2p)$
(siehe Log-Likelihoodfunktion)

b) $AIC = RSS_p - n + scale \cdot p \cdot k$, sonst

Andere Programmpakete verwenden noch:

$$AIC = \hat{\sigma}^2(C_p + n),$$

wobei wieder $\hat{\sigma}^2 = \dfrac{RSS_m}{n-m}$ aus dem vollen Modell verwendet wird.

Hier gilt dann die Identität

$$AIC = RSS_p + 2p\hat{\sigma}^2.$$

Dies ist übrigens die Voreinstellung in S-Plus bei "backward" ausgehend vom vollen Modell.

Satz 3.24 (Eigenschaften von Mallow's C_p)

1. $C_m = m$
2. *C_p schätzt den mittleren* $MSE(\boldsymbol{x}_i^T\boldsymbol{b})$ *für alle* $i = 1, \ldots, n$.
3. *Für festgehaltenes p ist C_p streng monoton fallend in s_p^2.*
4. *Die optimale Anzahl p an Einflussgrößen sollte so gewählt werden, dass C_p minimal wird.*
5. *Darstellungen:*
   $C_p = C_J = p + (n-p)\dfrac{s_J^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = \dfrac{RSS_J}{\hat{\sigma}^2} - n + 2p$
   $|J| = p$
6. *Vergleich von C_p mit Gerade $C_p = p$*

Gemäß 3.24, 2 schätzt C_p den sog. *Mean Square Error (MSE).*
Ein kleinerer MSE sichert bessere Vorhersagen.

Sei etwa $y = y(\boldsymbol{x}_0) \sim N(\underbrace{\boldsymbol{x}_0^T\boldsymbol{\beta}}_{\Theta}, \sigma^2)$

Wir suchen eine optimale Vorhersage $\hat{y}$ für y, mit $\hat{y} = \hat{\Theta} = \boldsymbol{x}_0^T\boldsymbol{b},\ \boldsymbol{b} = \hat{\boldsymbol{\beta}}$.

Def. 3.25 (MSE, Bias)
Sei $\hat{\Theta}$ mit $E(\hat{\Theta}) = \theta$ ein Schätzer für einen unbekannten Parameter Θ. Dann heißen

$MSE(\hat{\Theta}) := E[(\hat{\Theta} - \Theta)^2]$ *mean square error*

und

$Bias(\hat{\Theta}) := E[\hat{\Theta} - \Theta]$ *die Verzerrung (Bias) von $\hat{\Theta}$.*

Satz 3.26 (Zerlegung)

$$MSE(\hat{\Theta}) = Var(\hat{\Theta}) + Bias^2(\hat{\Theta}) \tag{3.11}$$

Beweis: Sei $\theta := E(\hat{\Theta})$:

$$\begin{aligned} MSE(\hat{\Theta}) &= E[(\hat{\Theta}-\Theta)^2] = E[(\hat{\Theta}-\theta+\theta-\Theta)^2] = \\ &= E[(\hat{\Theta}-\theta)^2 + 2(\hat{\Theta}-\theta)(\theta-\Theta) + (\theta-\Theta)^2] = \\ &\overset{\theta=E(\hat{\Theta})}{=} \underbrace{E(\hat{\Theta}-\theta)^2}_{Var(\hat{\Theta})} + 2(\theta-\Theta)\underbrace{E(\hat{\Theta}-\theta)}_{0} + \underbrace{(\theta-\Theta)^2}_{Bias^2(\hat{\Theta})}. \end{aligned}$$

Theoretische Grundlage des AIC ist die Log-Likelihoodfunktion für normalverteilte Fehler:

$$y_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2),\ f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)$$

$$\begin{aligned} L(\boldsymbol{\beta},\sigma^2) &:= \ln(\prod_{i=1}^{n} f(y_i)) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2 - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) \end{aligned}$$

$\hat{\boldsymbol{\beta}}_{ML}$ und $\hat{\sigma}^2_{ML}$ maximieren $L(\boldsymbol{\beta},\sigma^2)$, d.h.

$\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{KQ} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ (KQ-Schätzer)

$\hat{\sigma}^2_{ML} = \dfrac{RSS_p}{n} \neq s_p^2 = \dfrac{RSS_p}{n-p}$

$L(\hat{\boldsymbol{\beta}}_{ML},\hat{\sigma}^2_{ML}) = -\dfrac{n}{2}(1-\ln(2\pi)) - \dfrac{n}{2}\ln(RSS_p/n)$

Weglassen von Konstanten unabhängig von p:

$$AIC = -2L(\hat{\boldsymbol{\beta}}_{ML},\hat{\sigma}^2_{ML}) + \underbrace{2p}_{\text{Strafterm für zu viele Variablen}}$$

(vgl. $scale = 0$ in R)

Optimiert man die Log-Likelihoodfunktion nur nach $\boldsymbol{\beta}$ und setzt σ^2 als Parameter ein, so ergibt sich:

$$L(\boldsymbol{\beta}_{ML},\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{\sigma^2}{n}\right) - \frac{RSS_p}{2\sigma^2}$$

Weglassen der beiden ersten Konstanten liefert mit $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{KQ}$ und dem *Strafterm* $2p$:

$$-2L(\hat{\boldsymbol{\beta}}_{ML},\sigma^2) \doteq \frac{RSS_p}{\sigma^2} + 2p\,.$$

Dieser letzte Ausdruck erklärt den Gebrauch von

$$AIC = RSS_p + 2\cdot p\cdot scale$$

in S-Plus, wobei in der Regel

$$scale = \hat{\sigma}^2 = \frac{RSS_m}{n-m}$$

aus dem vollen Modell gewählt wird.

3.9.4 Algorithmisches Vorgehen

1. All subset regression mit der Prozedur "leaps" (nur in S-Plus verfügbar)
   Ausdruck der jeweils r besten Variablensätze bzgl. eines oder mehrerer der Kriterien $C_p, AIC, R^2_{adj}, s^2_p$

2. Schrittweise Regression mit der Prozedur "step" (auch in R)
   Varianten mit AIC:

   (a) Forward selection
   ausgehend vom Modell zum Kommando lm ($y \sim 1$)

   (b) Backward elimination
   ausgehend vom vollen Modell lm ($y \sim x1 + x2 + \cdots + xk$)
   m = k+1 Variablen mit intercept

3.9.5 Variablenselektion in R und S-Plus

in R verfügbar: "leaps" für R^2_{adj} und
"step" für AIC und schrittweise Regression
nur in Splus: "leaps" für C_p und Auswahl der besten Variablensätze

Splus-Befehle für Mallow's C_p am Beispiel von "College Spending"

Zusammenfügen der möglichen Variablen
x_cbind(SAT, TOP10, ACCRATE, PHD, RATIO, GRADRATE, ALUMNI)

die zu langen Variablen-Namen werden durch die "names"-Option verkürzt
name_c("S","T","A","P","R","G","A")

in der Regel will man nur die besten zwei oder drei Variablensaetze sehen; z.B. nbest=3
r.cp_leaps(x, log(SPEND), method="Cp", nbest=3, names=name)

Zusammenfassen der nach Variablensaetzen geordneten Listenelemente
list.cp_cbind(round(r.cp$Cp, 3), r.cp$size, r.cp$label)

Umformen in ein Data-Frame
list.cp_as.data.frame(list.cp)

Grafik erstellen
nur kleine Werte von Cp wegen Skalierung der Grafik

```
cp_r.cp$Cp[r.cp$Cp<15]
p_r.cp$size[r.cp$Cp<15]
plot(p,cp,ylab="Mallows Cp",xlab="Parameter im Modell")
title("Cp-Plot für log(SPEND)-Modell")
abline(0,1)
```

Zunächst zeigt sich, dass verschiedene Kriterien zu unterschiedlichen besten Variablensätzen führen.

Tabelle 2: Beste Variablensätze für "College Spending"					
p	Variable	R^2_{adj}	Rang(R^2_{adj})	C_p	Rang(C_p)
2	S	0.6005		75.554	
2	T	0.5072		126.606	
2	G	0.4657		149.300	
3	TR	0.7199		11.201	
3	SR	0.6848		30.265	
3	ST	0.6197		65.632	
4	STR	0.7339		4.589	3
4	TPR	0.7272		8.173	
4	SPR	0.6984		23.749	
5	STPR	0.7374	2	3.690	1
5	STAR	0.7342		5.430	
5	STRG	0.7325		6.297	
6	STAPR	0.7385	1	4.131	2
6	STPRG	0.7357		5.626	
6	STPRAl	0.7356		5.644	
7	STAPRAl	0.7367	3	6.072	
7	STAPRG	0.7366	4	6.098	
7	STPRGAl	0.7340		7.530	
8	STAPRGAl	0.7350	5	8.000	

Zum Schluss noch eine grafische Darstellung von Mallows' C_p. Die besten Variablensätze weisen C_p-Werte unterhalb der Geraden $C_p = p$ auf; siehe unter anderem $STPR$ für $p = 5$.

*

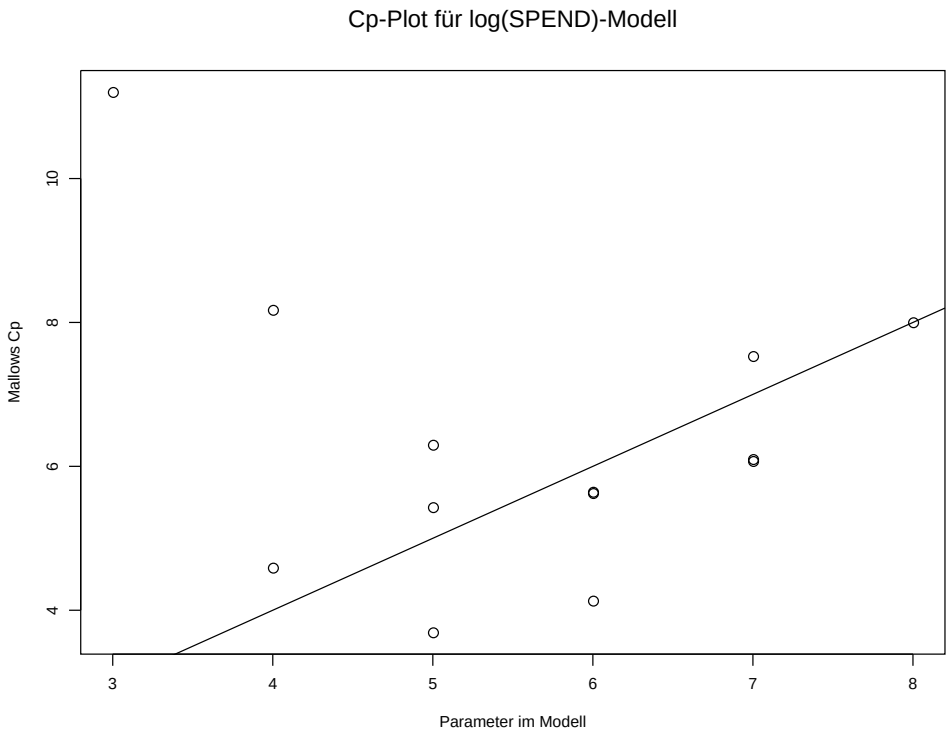

Abbildung 3.6: Mallows' C_p im Beispiel College Spending

3.10 Multikollinearität

Eines der Ziele statistischer Analysen ist die Konstruktion erwartungstreuer Schätzer mit minimaler Varianz. Wir haben bereits am Beispiel des "Overfitting" gesehen, dass man lieber mit zu wenigen als mit zu vielen Variablen arbeiten sollte.
Ein anderes Phänomen, das zur sprunghaften Vergrößerung der Varianzen der KQ-Schätzungen führen kann, ist die sog. "Multikollinearität" (M-K). Um M-K zu kontrollieren, versucht man, Größen wie

$$\sum_j Var(b_j)$$

klein zu halten. Natürlich müssen erst einmal einige Regressoren ins Modell aufgenommen werden, damit das Design den Erwartungwert von $\boldsymbol{y}$ hinreichend gut modelliert. M-K verhindert präzise statistische Aussagen. Diese Tatsache ist vielen Nutzern statistischer Software bekannt, ohne dass die meisten genau wissen, welche geometrischen und numerischen Zusammenhänge sich dahinter verbergen.

Hier können nicht alle Phänomene der M-K im Detail diskutiert werden. Die für die Praxis wichtigsten Aspekte werden jedoch angesprochen. Diese sind:

1. M-K verursacht eine "Explosion" der Varianzen der Schätzer (siehe VIF = variance inflation factor), und dementsprechend ungenau werden Vorhersagen.

2. M-K beruht nur auf Eigenschaften der Kovariablen (Regressoren) und hängt wie die leverage points in keiner Weise vom Response $\boldsymbol{y}$ ab.

3. M-K entsteht, wenn die Daten zu zwei oder mehreren Kovariablen nahezu linear abhängig sind. Wären die Kovariablen Realisierungen von Zufallsvariablen, so würde man sagen: Zwei oder mehrere Kovariablen sind hochkorreliert.

4. M-K ist der Gegenpol zur Orthogonalität. In sog. orthogonalen Designs (Spalten der Design-Matrix $\boldsymbol{X}$ stehen aufeinander senkrecht) tritt keine M-K auf.

5. Im Falle von M-K vergrößert sich die Varianz aufgrund einer geometrischen "Schiefstellung" der Kovariablen-Daten, ohne einen Gewinn bei der Modellierung von $E(\boldsymbol{y})$ zu liefern. Bei orthogonalen Designs wird dagegen die zusätzliche Erklärung von $E(\boldsymbol{y})$ durch die minimal mögliche Varianzvergrößerung erreicht.

Im Unterschied zu den übrigen Abschnitten ist es hier günstiger, den intercept nicht als erste Spalte der Design-Matrix $\boldsymbol{X}$ zu modellieren. $\boldsymbol{X}$ enthalte also in diesem Abschnitt ausnahmsweise nur echte Kovariablenspalten. Also
$\boldsymbol{X} = (\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_k),\ Rang(\mathbf{1}, \boldsymbol{X}) = k + 1 = p$
$\boldsymbol{y} = \mathbf{1} \cdot \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \mathbf{0},\ Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}.$
Ausnahmsweise: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$

3.10.1 Einfache lineare Regression

$k = 1,\ \boldsymbol{x} = (x_1, \ldots, x_n)^T \neq \mathbf{1} \cdot \bar{x}.$
Damit stehen im zentrierten Modell

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \epsilon_i,\ i = 1, \ldots, n$$

die Spalten $\mathbf{1}$ und $\boldsymbol{x} - \mathbf{1} \cdot \bar{x}$ aufeinander senkrecht. Man erkennt aber z.B. an der Formel für die Varianz von b_1, dass es beispielsweise im Fall frei wählbarer Designpunkte $x_1, \ldots, x_n$ günstiger ist, $\sum (x_i - \bar{x})^2$ möglichst groß zu halten.

3.10.2 Skalierungsinvarianz

Einige Größen wie die KQ-Schätzungen b_j und deren Varianzen hängen von der Wahl des Koordinatensystems für die Kovariablen ab. Andere wie RSS, t-Werte der b_j, R^2, R^2_{adj} oder Mallows C_p sind "skalierungsinvariant" bezüglich folgenden Transformationen der Kovariablen

$$\vec{\boldsymbol{x}}_j \to \gamma \cdot \vec{\boldsymbol{x}}_j + \delta \cdot \mathbf{1},\ \gamma \neq 0,\ \delta \in \boldsymbol{R},$$

wobei $\vec{\boldsymbol{x}}_j = (x_{1j}, \ldots, x_{nj})^T$ die Daten zur j-ten Kovariablen sind.
t-Werte, R^2 und R^2_{adj} sind sogar auch noch invariant gegen Skalierungen im Response $\boldsymbol{y}$.

Transformationen der genannten Art erhält man etwa durch den Übergang von \$ zu Euro oder von °C zu °F. Man überzeugt sich leicht, dass z.B. die Varianzen der b_j von der Wahl der Skalierung der Kovariablen abhängen. Damit ist auch unser ins Auge gefasstes Zielkriterium $\sum Var(b_j)$ skalierungsabhängig und damit so in der Praxis wertlos.
Es gibt jedoch eine eindeutige kanonische Form $\boldsymbol{X}^*$, zu der man ausgehend von beliebigen Skalierungen für die Kovariablen gelangen kann. Bezüglich dieses eindeutig bestimmten Designs ist dann $\sum Var(b_j)$ klein zu halten.

Satz 3.27 (Zentriertes, standardisiertes Modell)
Vor.:

$$\begin{aligned}
\boldsymbol{X} &= (\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_k) \in \boldsymbol{R}^{n,k},\ Rang(\mathbf{1}, \boldsymbol{X}) = p = k + 1 \\
\boldsymbol{C} &:= \mathrm{diag}\{\gamma_1, \ldots, \gamma_\mathrm{k}\} \in \boldsymbol{R}^{\mathrm{k,k}}, \\
\boldsymbol{D} &:= \mathrm{diag}\{\delta_1, \ldots, \delta_\mathrm{k}\} \in \boldsymbol{R}^{\mathrm{k,k}}, \\
\boldsymbol{Z} &:= [\boldsymbol{X} - (\mathbf{1}, \ldots, \mathbf{1}) \cdot \boldsymbol{D}] \cdot \boldsymbol{C} = (\vec{\boldsymbol{z}}_1, \ldots, \vec{\boldsymbol{z}}_k)\,.
\end{aligned}$$

Beh.:
Für beliebiges $\boldsymbol{D}$ und nichtsinguläres $\boldsymbol{C}$ führt die Zentrierung mit

$$\bar{z}_j := \frac{1}{n} \sum_{i=1}^{n} z_{ij},\ \vec{\boldsymbol{z}}_j = (z_{1j}, \ldots, z_{nj})^T$$

und Standardisierung mit

$$S_j := \sqrt{\sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2}$$

und

$$\vec{\boldsymbol{x}}^*_j := (\vec{\boldsymbol{z}}_j - \mathbf{1} \cdot \bar{z}_j)/S_j$$

zu einem eindeutigen

$$\boldsymbol{X}^* = (\vec{\boldsymbol{x}}^*_1, \ldots, \vec{\boldsymbol{x}}^*_k).$$

Gemäß diesem Satz führt also jede Ausgangsskalierung der Kovariablen zu demselben $\boldsymbol{X}^*$.

Man versucht nun im Modell

$$\boldsymbol{y} = \mathbf{1} \cdot \beta_0^* + \boldsymbol{X}^* \boldsymbol{\beta}^* + \boldsymbol{\epsilon}\,, \boldsymbol{\beta}^* \in \boldsymbol{R}^k,$$

das Kriterium

$$S^2(\boldsymbol{b}^*) = \sum_{i=1}^{k} Var(b_i^*) = E(\boldsymbol{b}^* - \boldsymbol{\beta}^*)^T(\boldsymbol{b}^* - \boldsymbol{\beta}^*)$$

zu minimieren.
$S^2(\boldsymbol{b}^*)$ ist, wie gesehen, nicht mehr von der Ausgangsskalierung abhängig.

3.10.3 Einfache Berechnung des VIF

Wegen der Zentrierung stehen alle Spalten von $\boldsymbol{X}^*$ senkrecht auf $\mathbf{1}$. Damit gilt

$$Cov \begin{pmatrix} b_0^* \\ \boldsymbol{b}^* \end{pmatrix} = \sigma^2 \begin{pmatrix} n & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{X}^{*T}\boldsymbol{X}^* \end{pmatrix}^{-1},$$

also

$$S^2(\boldsymbol{b}^*)/\sigma^2 = \mathrm{tr}[(\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{X}^*)^{-1}] =: \mathrm{w}_{11} + \cdots + \mathrm{w}_{\mathrm{kk}}\,.$$

$VIF_j := w_{jj}$ bezeichnet den *variance inflation factor*, d.h. den Anteil der j-ten Kovariablen am Zielkriterium $S^2(\boldsymbol{b}^*)$.
Dabei ist die explizite Berechnung der Inversen von $\boldsymbol{X}^{*T}\boldsymbol{X}^*$ gar nicht notwendig, denn es gilt

$$VIF_j = \frac{1}{1 - R_j^2},$$

wobei mit $\boldsymbol{X}_{(j)} := (\vec{\boldsymbol{x}}_1, \ldots, \vec{\boldsymbol{x}}_{j-1}, \vec{\boldsymbol{x}}_{j+1}, \ldots, \vec{\boldsymbol{x}}_k) \in \boldsymbol{R}^{n,k-1}$ die Design-Matrix ohne Daten zur Kovariablen j bezeichnet wird und R_j^2 das Bestimmtheitsmaß zu k fiktiven Modellen

$$\vec{\boldsymbol{x}}_j = \mathbf{1}\beta_{0j} + \boldsymbol{X}_{(j)}\boldsymbol{\beta}_{(j)} + \boldsymbol{\epsilon}_j,\ j = 1, \ldots, k$$

bezeichnet.
Wegen der Skalierungsinvarianz des Bestimmtheitsmaßes können die R_j^2 mit den ursprünglichen Daten berechnet werden. Eine Zentrierung und Skalierung ist also in der Praxis gar nicht unbedingt notwendig.

3.11 Korrelierte Fehler, gewichtete Regression

Das verallgemeinerte lineare Modell lautet:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \boldsymbol{0},\ Cov(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}\,,\quad \boldsymbol{V} \text{ positiv definit und bekannt.}$$

Wir diskutieren den wichtigsten Spezialfall:

$$V = diag\{v_{11}, \ldots, v_{nn}\},\ v_{ii} > 0,\ i = 1, \ldots, n\,.$$

Da die v_{ii} a priori bekannt sind (in der Regel Transformationen einer oder mehrerer Kovariablen), ist eine Rücktransformation zum Standardmodell möglich:
Mit

$$\boldsymbol{V}^{-1/2} = diag\{1/\sqrt{v_{11}}, \ldots, 1/\sqrt{v_{nn}}\}$$

und

$$\boldsymbol{y}_v := \boldsymbol{V}^{-1/2}\boldsymbol{y},\ \boldsymbol{X}_v := \boldsymbol{V}^{-1/2}\boldsymbol{X},\ \boldsymbol{\epsilon}_v := \boldsymbol{V}^{-1/2}\boldsymbol{\epsilon}$$

gilt:

$$\boldsymbol{y}_v = \boldsymbol{X}_v\boldsymbol{\beta}_v + \boldsymbol{\epsilon}_v,\ E(\boldsymbol{\epsilon}_v) = \boldsymbol{0},\ Cov(\boldsymbol{\epsilon}_v) = \sigma^2\boldsymbol{I}\,. \tag{3.12}$$

Der KQ-Schätzer zum transformierten Modell 3.12 lautet:

$$\boldsymbol{b}_v = (\boldsymbol{X}_v^T\boldsymbol{X}_v)^{-1}\boldsymbol{X}_v^T\boldsymbol{y}_v = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}\,.$$

Die gewichtete Residuenquadratsumme lässt sich demgemäß darstellen als

$$\|\boldsymbol{y}_v - \boldsymbol{X}_v\boldsymbol{b}_v\|^2 := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_v\|_v^2 := (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_v)^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_v).$$

In R oder S-Plus ist die Gewichtung in einfacher Weise darstellbar (siehe Beispiel bsp5).

attach(data5) Bereitstellen der Daten

1. Ohne Gewichtung (Auslassen der Beobachtungen mit fehlenden Daten)
   $r < -lm(Salary \sim Price + Work,\ na.action = na.omit)$
   $summary(r)$

	Value	Std. Error	t value	Pr(> |t|)
(Intercept)	10.0942	31.3257	0.3222	0.7488
Price	0.8688	0.1159	7.4974	0.0000
Work	-0.0167	0.0142	-1.1768	0.2458

   Multiple R-Squared: 0.6572

2. Mit Gewichtung: $1/Price$
   $r.x < -lm(Salary \sim Price + Work,\ weights = 1/Price)$
   $summary(r.x)$

	Value	Std. Error	t value	Pr(> |t|)
(Intercept)	3.1874	28.0665	0.1136	0.9101
Price	0.9373	0.1096	8.5528	0.0000
Work	-0.0156	0.0126	1.2381	0.2224

Multiple R-Squared: 0.7163

3. Zu $r.x$ entsprechendes transformiertes Regressionsmodell:
   $r.t <- lm(Salary/sqrt(Price) \sim (I(1/sqrt(Price))$
   $+ sqrt(Price) + I(Work/sqrt(Price)) - 1)$
   $summary(r.t)$

	Value	Std. Error	t value	Pr(> |t|)
I(1/sqrt(Price))	3.1874	28.0665	0.1136	0.9101
sqrt(Price)	0.9373	0.1096	8.5528	0.0000
I(Work/sqrt(Price))	-0.0156	0.0126	1.2381	0.2224

** Wert falsch ** Multiple R-Squared: 0.9012 ** wegen "-1" in lm **

4. Suche nach "optimalen" Weights:
   $r.x2 <- lm(Salary \sim Price + Work,\ weights = 1/(Price\text{^}2))$
   Multiple R-Squared: 0.7459

   $r.x3 <- lm(Salary \sim Price + Work,\ weights = 1/(Price\text{^}3))$
   Multiple R-Squared: 0.7471

   $r.x4 <- lm(Salary \sim Price + Work,\ weights = 1/(Price\text{^}4))$
   Multiple R-Squared: 0.7235

5. Fazit Nach dem R^2-Kriterium scheint das Modell $r.x3$ am besten geeignet. Das zeigt sich u. a. auch beim Vergleich der entsprechenden QQ-Plots der Residuen. Abbildung 3.7 zeigt das Ergebnis von $plot(r.x3)$.

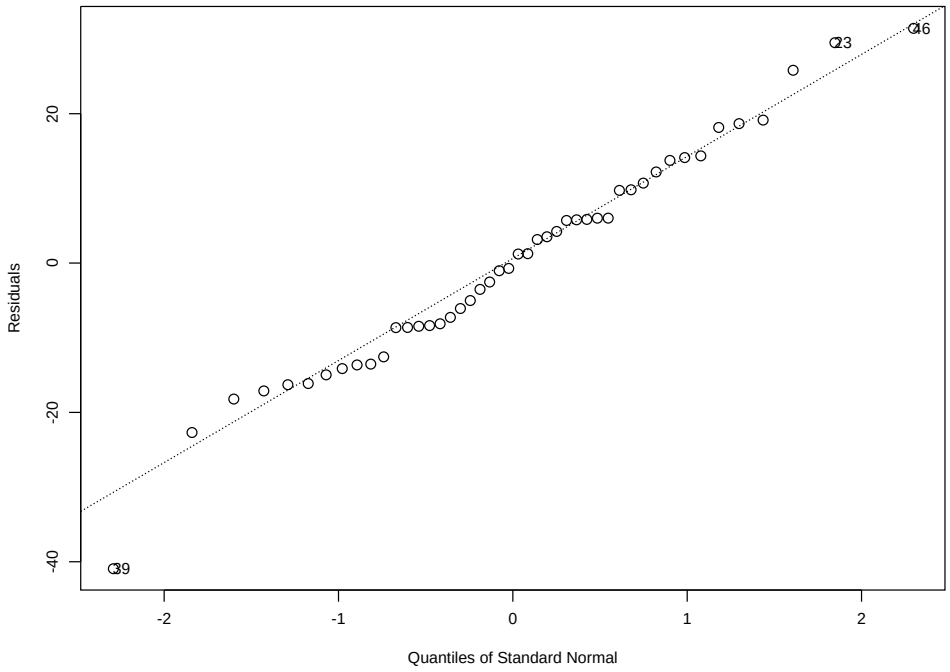

Abbildung 3.7: QQ-Plot

3.12 Autokorrelierte Fehler

Besonders bei Zeitreihen sind die Fehler oft korreliert; d.h. $E(\epsilon_i, \epsilon_j) \neq 0$. Somit ist die Unabhängigkeitsannahme verletzt. In der einfachsten Abhängigkeitsstruktur sind die Fehler *autokorreliert*.

Modell (*autoregressive Fehlerstruktur 1. Ordnung*):

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \\ \epsilon_t &= \rho\epsilon_{t-1} + U_t, \end{aligned}$$

wobei U_t *iid* $N(0, \sigma_U^2)$; $-1 < \rho < 1$.

Indizien für Fehlerkorrelation

Bei positiver Korrelation ändern sich $\epsilon_t, \epsilon_{t+1}$ kaum (vgl. Grafik 3.8)

Dagegen alternieren ϵ_t und ϵ_{t+1} bei negativer Korrelation wegen $\epsilon_t = -|\rho|\epsilon_{t-1} + U_t$, vielfach.

Formaler Test: Durbin–Watson–Test

Der formalen Einfachheit wegen betrachten wir unendlich viele t.

$$\epsilon_t = \rho\epsilon_{t-1} + U_t = \rho(\rho\epsilon_{t-2} + U_{t-1}) + U_t = \rho^2\epsilon_{t-2} + \rho U_{t-1} + U_t$$

usw. liefert

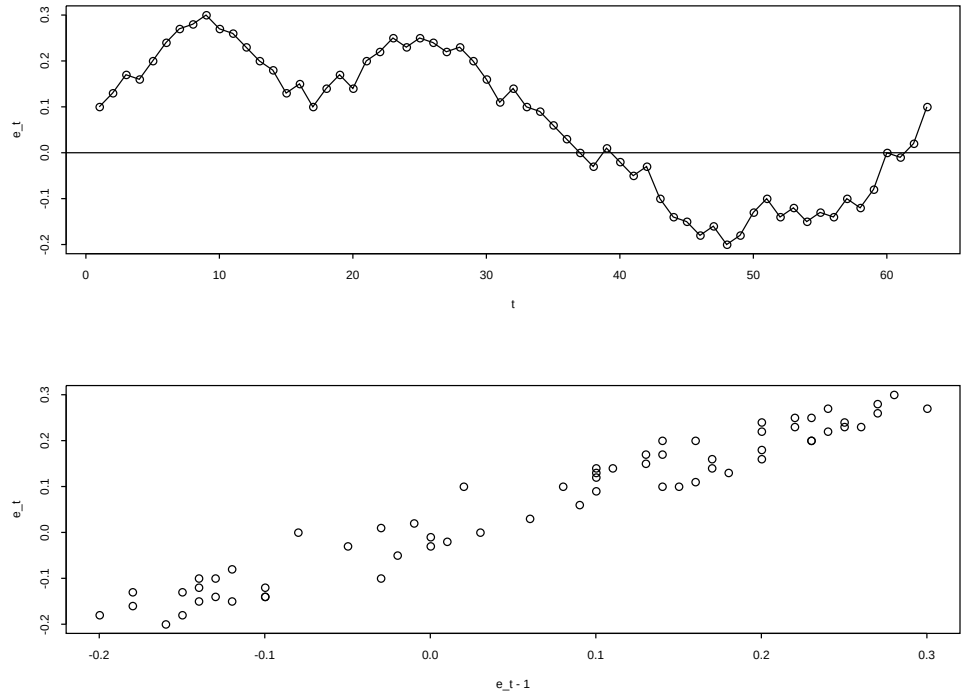

Abbildung 3.8: Autokorrelierte Fehler

$$\epsilon_t = \sum_{j=0}^{\infty} \rho^j U_{t-j}$$

woraus

$$E(\epsilon_t) = \sum_{j=0}^{\infty} \rho^j \underbrace{E(U_{t-j})}_{=0} = 0$$

und

$$Var(\epsilon_t) = \sigma_U^2 \sum_{j=0}^{\infty} (\rho^2)^j \overset{\text{geom.Reihe}}{=} \sigma_U^2 \frac{1}{1-\rho^2}$$

folgen. In ähnlicher Weise kann man leicht zeigen, dass

$$Cov(\epsilon_t, \epsilon_{t+j}) = \rho^{|j|} \sigma_U^2 \frac{1}{1-\rho^2}, \ |\rho| < 1\,.$$

Zu testen ist

$$H_0 : \rho = 0 \ \text{ gegen } \ H_1 : \rho > 0\,.$$

Mit $e_t := Y_t - \hat{Y}_t$ schlagen Durbin und Watson folgende Teststatistik vor

$$d := \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}\,.$$

Bemerkung:
Wegen $\epsilon_t - \epsilon_{t-1} = \rho\epsilon_{t-1} + U_t - \epsilon_{t-1} = (\rho - 1)\epsilon_{t-1} + U_t$ ist d für $\rho \approx 1$ klein und H_0 in

diesem Fall zu verwerfen. Andererseits erwartet man für $\rho = 0$ einen eher großen Wert, denn unabhängige Fehler liefern

$$E\left((\epsilon_t - \epsilon_{t-1})^2\right) = 2\,\sigma_U^2 \text{ und } E\left(\epsilon_t^2\right) = \sigma_U^2\,.$$

Durbin–Watson–Test
Verwerfe H_0 (*d.h.* $\rho > 0$), falls $d < d_L$.
Akzeptiere H_0 ($\rho = 0$), falls $d > d_U$.
Test ist nicht zu entscheiden, falls $d_L < d < d_U$.

d_L und d_U entnehme man geeigneten Tabellen. In der Praxis ist die Bedeutung des Durbin–Watson–Tests umstritten. Grafische Analysemethoden gewinnen immer mehr an Bedeutung. Natürlich möchte man der Einfachheit halber am liebsten die Hypothese $\rho = 0$ bestätigt haben.

Maßnahmen bei Autokorrelation

Einfachstes Modell:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 x_t + \epsilon_t \\ \epsilon_t &= \rho\epsilon_{t-1} + U_t \end{aligned}$$

Wegen $\epsilon_t - \rho\epsilon_{t-1} = U_t$ betrachtet man

$$\begin{aligned} Y_t^* := Y_t - \rho Y_{t-1} &= \beta_0 + \beta_1 x_t + \epsilon_t - \rho(\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}) \\ &= \underbrace{\beta_0(1-\rho)}_{\beta_0^*} + \underbrace{\beta_1}_{\beta_1^*}\underbrace{(x_t - \rho x_{t-1})}_{x_t^*} + \underbrace{\epsilon_t - \rho\epsilon_{t-1}}_{=U_t \sim N(0,\sigma_U^2)}\,. \end{aligned}$$

Das ist eine Einfache Lineare Regression.

Schätzung von ρ

Ausgehend von $\epsilon_t = \rho\epsilon_{t-1} + U_t$ liefert eine Einfache Lineare Regression durch den Ursprung eine KQ-Schätzung

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n-1} e_t^2}$$

für ρ.

Dies ergibt folgendes iterative Schema:

1. Berechne $\hat{\rho}$.
2. Bilde $Y_t^* = \beta_0^* + \beta_1^* X_t^* + U_t$ und berechne die KQ-Schätzer b_0^* und b_1^*.

3. Falls die Residuen $e_t^* := Y_t^* - b_0^* - b_1^* X_t$ unkorreliert sind, schätze β_0 und β_1 durch $b_0 = \dfrac{b_0^*}{1-\rho}$ und $b_1 = b_1^*$.

4. Ansonsten setze $e_t = e_t^*$ und gehe zurück zu Schritt 1.

Die Gewichtungen aus dem vorangegangenen Abschnitt sind völlig unproblematisch und tragen oft zu erheblichen Modellverbesserungen bei.

Die Behandlung abhängiger Strukturen führt aber schon im einfachsten Fall autokorrelierter Fehler nicht nur zu einem erheblichen Zusatzaufwand. Schlimmer ist noch, dass eine Interpretation der berechneten Schätzungen immer schwieriger wird. Für die Behandlung autoregressiver Fehler ist eine ausreichende Erfahrung (bei Datentransformationen und effizientem Umgang mit Programmpaketen) notwendig, die in einem Einführungskurs wie diesem nicht von allen Teilnehmern erwartet werden kann.

3.13 Kategoriale Variable; Teil 1: Regressoren

Die Behandlung qualitativer Regressoren wird in zwei Abschnitten behandelt. Dieser Abschnitt befasst sich mit linearen Modellen mit genau einer kategorialen Kovariablen.

3.13.1 Einführendes Beispiel mit einer binären Kovariablen

Beispiel: y: Geburtsgewicht
x_1: Alter (metrisch)
$\tilde{x}_2$: Geschlecht (kategorial)
mit n_1 "männlichen" und $n - n_1$ "weiblichen" Datensätzen.

Dummy-Kodierung:

$$x_{i2} = \begin{cases} 1, & \text{falls } \tilde{x}_{i2} = m \text{ (männlich)} \\ 0, & \text{sonst} \end{cases}$$

Dies entspricht der Voreinstellung in R.
In S-Plus muss vor dem lm-Kommando noch
options(contrasts = c("contr.treatment","contr.poly"))
gesetzt werden.

Man erhält folgende Design-Matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & 1 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_{n_1} & 1 \\ 1 & x_{n_1+1} & 0 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_n & 0 \end{pmatrix}$$

Effekt-Kodierung:
Die Variante
options(contrasts = c("contr.sum", "contr.poly"))
liefert die sog. Effekt-Kodierung

$$x_{i2} = \begin{cases} 1, & \text{falls } \tilde{x}_{i2} = m \\ -1, & \text{sonst} \end{cases}$$

In beiden Fällen gilt $p = 3$ und $Rang(\boldsymbol{X}) = p$.

Ein einfacher Ansatz für das Beispiel auf dem Übungsblatt liefert
out.add <- lm($y \sim$ age + sex)
$y = -1773 + 121 \cdot age + 163 \cdot sex$

Dieses rein additive Modell geht von identischen Steigungen in den beiden Gruppen m bzw. w aus. Man erhält die einfachen Regressionen
$y_m = -1610 + 121 \cdot age$

$y_w = -1773 + 121 \cdot age.$

Den allgemeinen Fall modelliert man mit einem zusätzlichen **Interaktionsterm** age : sex

out.ia <- lm ($y \sim$ age $*$ sex)
$\hat{=}$ lm ($y \sim$ age + sex + age : sex)

und der Design-Matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & 1 & x_1 \\ . & . & . & \\ . & . & . & \\ . & . & . & \\ 1 & x_{n_1} & 1 & x_{n_1} \\ 1 & x_{n_1+1} & 0 & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_n & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n,4}.$$

Die diesbezügliche KQ-Schätzung
$y = -2142 + 873 \cdot sex + 130 \cdot age - 18 \cdot sex \cdot age$
Führt zu den nun völlig unabhängigen KQ-Geraden
$y_w = -2142 + 130 \cdot age$
$y_m = -1269 + 112 \cdot age = (-2142 + 873) + (130 - 18) \cdot age$
in den verschiedenen Geschlechts-Gruppen.

Die zwei KQ-Geraden erhält man direkt über
lm ($y \sim$ sex / age -1)

Übrigens sind die Steigungen nicht signifikant voneinander verschieden, Dies zeigt sich aus einer sog. "Varianzanalyse":

out.add <- lm($y \sim$ age + sex)
out.ia <- lm ($y \sim$ age $*$ sex)
aov(out.ia,out.add)

mit dem Resultat:

	Terms	Resid. Df	RSS	Test	Df	Sum of Sq	F Value	Pr(F)
1	sex * age	20	652424.5					
2	sex + age	21	658770.7	-sex:age	-1	-6346.225	0.1945428	0.6638934

Varianzanalyse (analysis of variance, ANOVA) vergleicht ineinandergeschachtelte lineare Modelle anhand spezieller F-Tests (wir hatten die Prozedur aov() schon im Zusammenhang der Variablenselektion kennengelernt). In den nächsten Abschnitten wird noch näher auf diese Methode eingegangen.

Man nennt kategoriale Variablen wie *sex* auch *Faktoren*. Wegen der alphanumerischen Kodierung $sex \in$ {m, w} ist für R bzw. S-Plus klar, dass es sich um eine kategoriale

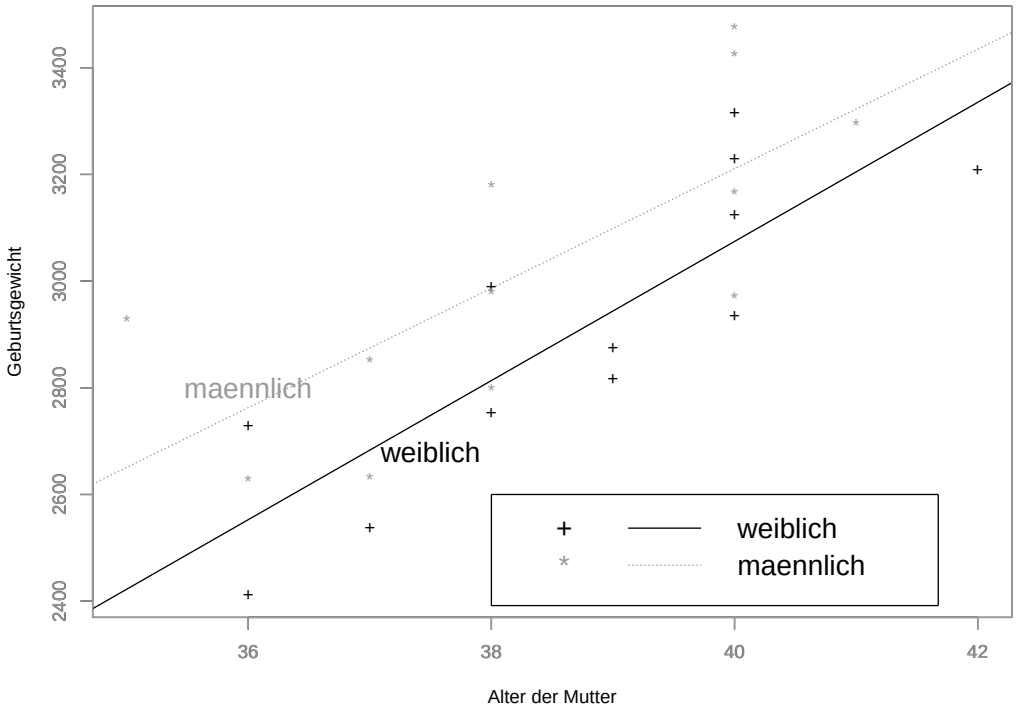

Abbildung 3.9: KQ-Geraden in den 2 Gruppen

Variable handelt.

Dies trifft nicht auf numerische Kodierungen zu wie

$$\tilde{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 4 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} 1\hat{=}\text{"ledig"} \\ \\ \\ \\ 2\hat{=}\text{"verheiratet"} \\ \\ \\ 3\hat{=}\text{"geschieden"} \\ \\ 4\hat{=}\text{"verwitwet"} \end{matrix}$$

Eine Behandlung der zunächst numerischen Variablen $\tilde{x}_2 = famstand$ als Faktor muss durch die Kommandos

f.famstand <- factor(famstand)
options(contrasts = c("contr.treatment","contr.poly"))
out.add <- lm ($y \sim$ age + f.famstand)

z.B. für y (Einkommen) erzwungen werden.

Die zugehörige Design-Matrix lautet:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & 1 & 0 & 0 \\ 1 & . & 1 & 0 & 0 \\ 1 & . & 1 & 0 & 0 \\ 1 & x_{41} & 1 & 0 & 0 \\ 1 & x_{51} & 0 & 1 & 0 \\ 1 & x_{61} & 0 & 1 & 0 \\ 1 & x_{71} & 0 & 1 & 0 \\ 1 & x_{81} & 0 & 0 & 1 \\ 1 & x_{91} & 0 & 0 & 1 \\ 1 & x_{10,1} & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n,5},\ Rang(X) = 5$$

Zur exakten Modellierung mit beliebigen Achsenabschnitten und Steigungen in den möglichen Gruppen müssen wiederum Interaktionsterme berücksichtigt werden.

out.ia <- lm ($y \sim$ age $*$ f.famstand, $x = T$)
$\hat{=}$ lm ($y \sim$ age + f.famstand + age : f.famstand, $x = T$)

Der Befehl *out.ia $ x* liefert die resultierende Design-Matrix (nur möglich, falls im vorangegangenen lm-Statement "$\ldots, x = T$" gesetzt wurde).

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & 1 & 0 & 0 & x_{11} & 0 & 0 \\ 1 & . & 1 & 0 & 0 & . & 0 & 0 \\ 1 & . & 1 & 0 & 0 & . & 0 & 0 \\ 1 & x_{41} & 1 & 0 & 0 & x_{41} & 0 & 0 \\ 1 & x_{51} & 0 & 1 & 0 & 0 & x_{51} & 0 \\ 1 & x_{61} & 0 & 1 & 0 & 0 & x_{61} & 0 \\ 1 & x_{71} & 0 & 1 & 0 & 0 & x_{71} & 0 \\ 1 & x_{81} & 0 & 0 & 1 & 0 & 0 & x_{81} \\ 1 & x_{91} & 0 & 0 & 1 & 0 & 0 & x_{91} \\ 1 & x_{10,1} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Die konkrete Behandlung mehrstufiger kategorialer Kovariablen wird nochmals an einem Beispiel erläutert.

3.13.2 Beispiel 2: Mehrstufiger Faktor

Beispiel: Y = Benzinverbrauch auf 100 km
X_1 = Autogewicht
X_2 = Automarke (VW, BMW, Mazda)

Wir beginnen wieder mit einer numerischen Kodierung, die aber kein befriedigendes Ergebnis liefert.

Erster Versuch: 1-2-3-Kodierung

$$X_2 = \begin{cases} 1 & \text{VW} \\ 2 & \text{BMW} \\ 3 & \text{Mazda} \end{cases}$$

Modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Interpretation: VW $Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon \quad = (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon$
BMW $Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 2 + \epsilon = (\beta_0 + 2\beta_2) + \beta_1 X_1 + \epsilon$
Mazda $Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 3 + \epsilon = (\beta_0 + 3\beta_2) + \beta_1 X_1 + \epsilon$

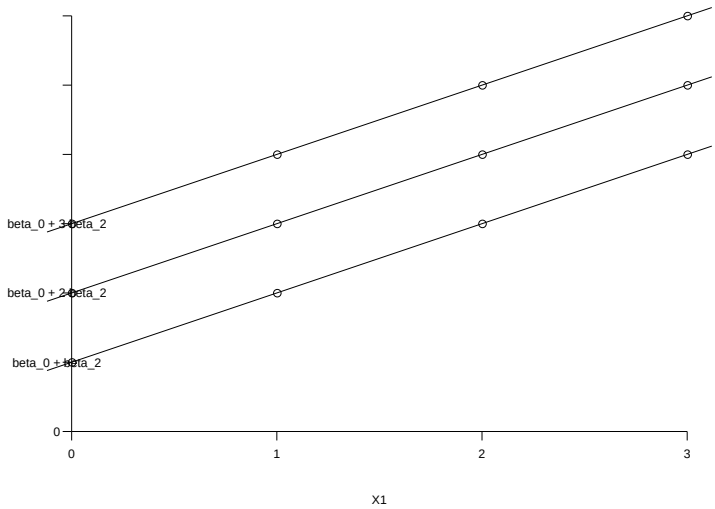

Abbildung 3.10: Unzulässige Festlegung des Intercepts; parallele KQ-Geraden

Die Grafik 3.10 zeigt, dass der Abstand der Intercepts durch die Festlegung (1,2,3) schon festgelegt ist. Außerdem impliziert der linear additive Ansatz die Parallelität der drei KQ-Geraden, die natürlich a priori nicht klar ist.

Man möchte verschiedene Intercepts und Steigungen zulassen. Dies lässt sich – wie gesehen – mit Dummy- Variablen modellieren.

$$z_{i1} = \begin{cases} 1 & \text{falls Auto i ein VW} \\ 0 & \text{sonst} \end{cases}$$
$$z_{i2} = \begin{cases} 1 & \text{falls Auto i ein BMW} \\ 0 & \text{sonst} \end{cases}$$
$$z_{i3} = \begin{cases} 1 & \text{falls Auto i ein Mazda} \\ 0 & \text{sonst} \end{cases}$$

Wegen $z_{i1} + z_{i2} + z_{i3} = 1 \; \forall \; i = 1, \ldots, n$, hätte man eine perfekte Multikollinearität. Die Entfernung der dritten Dummy-Variablen stellt keinen Informationsverlust dar, denn

$$z_{i1} = z_{i2} = 0 \Longrightarrow z_{i3} = 1 \, .$$

Zweiter Versuch:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 Z_1 + \beta_3 Z_2 + \epsilon$$

Intercepts für VW $\beta_0 + \beta_2$,
BMW $\beta_0 + \beta_3$ und
Mazda β_0.

Damit ist erreicht, dass im Allgemeinen alle Intercepts verschieden sind, die Steigungen jedoch noch nicht. Hierzu müssten analog zum einführenden Beispiel noch Interaktionen berücksichtigt werden.

3.14 Kategoriale und Indikatorvariablen; Teil 2: ANOVA–Modelle

3.14.1 Einfaktorielle Varianzanalyse; one way ANOVA

Modell

Einfache lineare Regression mit einem qualitativen Regressor Z.
Z habe k Kategorien. Definiere

$$z_{il} := \begin{cases} 1 & \text{Beobachtung i fällt in Kategorie l} \\ 0 & \text{sonst} \end{cases} \qquad l = 1, \ldots, k, \; i = 1, \ldots, n$$

Modell: $y_i = \beta_0 + \alpha_1 z_{i1} + \cdots + \alpha_{k-1} z_{i,k-1} + \epsilon_i$

Bemerkungen:

- Erwarteter Response in Kategorie l $= \beta_0 + \alpha_l, \; l = 1, \ldots, k-1$
  Erwarteter Response in Kategorie k $= \beta_0$

- H_0: keine Unterschiede zwischen den Kategorien $\Leftrightarrow H_0 : \alpha_1 = \cdots = \alpha_{k-1} = 0$

- Dies ist ein Spezialfall des allgemeinen F-Tests:

$$\begin{aligned} \alpha_1 = \cdots = \alpha_{k-1} = 0 \Leftrightarrow \boldsymbol{C}\boldsymbol{\beta} = \mathbf{0}, \; \boldsymbol{C} &= \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{(k-1),k} \\ \boldsymbol{\beta}^T &= (\beta_0, \alpha_1, \ldots, \alpha_{k-1}) \end{aligned}$$

Gemischte Modelle mit quantitativen Kovariablen und genau einem qualitativen Regressor werden analog zu obigen Beispielen behandelt.

Beispiel: one way ANOVA

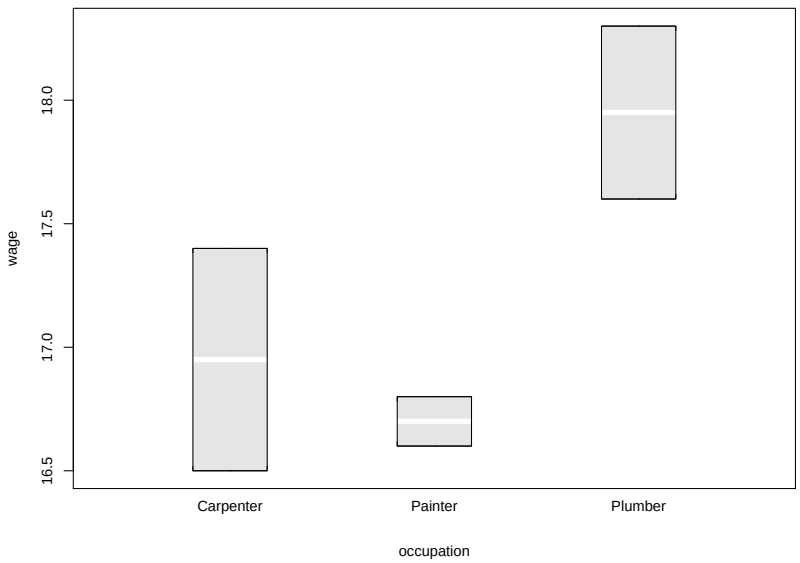

Abbildung 3.11: Boxplot: Response *wage* innerhalb der Faktorstufen

Die Daten

wage	occupation
17.6	Plumber
18.3	Plumber
16.8	Painter
16.7	Painter
16.6	Painter
17.4	Carpenter
16.5	Carpenter

geben die Einkünfte verschiedener Handwerker wieder. Die Variabilität der Zielvariablen *wage* innerhalb der Faktorstufen wird in der Regel durch Boxplots dargestellt; siehe Grafik 3.11.

Man erkennt, dass die Installateure mehr verdienen als die anderen genannten Berufe. Aber ist der im Plot sichtbare Unterschied auch statistisch signifikant?
Die S-Plus-Kommandos
$r <- aov(wage \sim occupation)$
$summary(r)$
liefern

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
occupation	2	1.964286	0.9821429	5.863539	0.06468802
Residuals	4	0.670000	0.1675000		

Der P-Wert über 5% besagt, dass die Unterschiede innerhalb der Faktorstufen **nicht signifikant** sind. Der Grund hierfür ist die geringe Fallzahl von nur zwei Beobachtungen pro Zelle.

3.14.2 Zweifaktorielles Design; two way ANOVA

Balanced Design; K Beobachtungen pro Zelle

In der Literatur ist es üblich, mit folgendem Spezialfall zu beginnen:

Daten: y_{ijk} = k-ter Response des Faktors A in Stufe (Level) i und Faktor B in Stufe j,
$i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K.$

Es gibt also in jeder Zelle bzgl. der Faktorstufen genau K Beobachtungen.

Beispiel: Aktivitätslevel von Kindern

		Drug	
		Placebo	Ritalin
		50	67
		45	60
	normal	55	58
		52	65
Group			
		70	51
	hyperaktiv	72	57
		68	48
		75	55

In Faktor A = Group und in Faktor B = Drug gibt es jeweils 2 Stufen (levels).
$K = 4$, $I = 2 = J$, z.B. $y_{121} = 67$, $y_{124} = 65$.

Explorative Graphische Analyse:

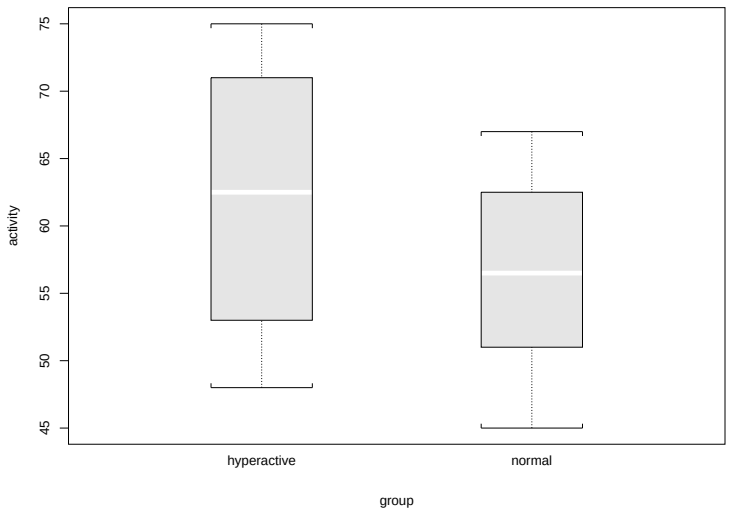

Abbildung 3.12: Boxplot: Response *activity* innerhalb der Faktorstufen von *group*

Die Bilder 3.12 und 3.13 zeigen gewisse Unterschiede in der Streuung des Response innerhalb der Faktorstufen, während die Mediane nur "wenig" voneinander abweichen. Dementsprechend liefert auch

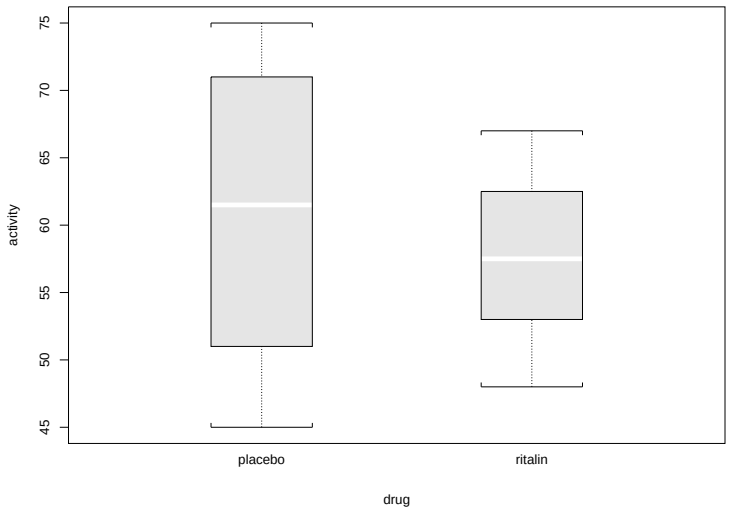

Abbildung 3.13: Boxplot: Response *activity* innerhalb der Faktorstufen von *drug*

$r <- aov(activity \sim group + drug)$
$summary(r)$
das Ergebnis

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
group	1	121.00	121.0000	1.414886	0.2555190
drug	1	42.25	42.2500	0.494041	0.4945256
Residuals	13	1111.75	85.5192		

Die P-Werte von 25% bzw. 49% scheinen darauf hinzudeuten, dass kein Zusammenhang zwischen dem Response und den Faktoren besteht. Diese Analyse ist jedoch falsch.

Da K Wiederholungen für jede Faktorkombination vorhanden sind, kann man diese benutzen, um eine Schätzung des erwarteten Response für jede Faktorkombination zu bekommen, d.h.

$$\bar{y}_{ij\cdot} := \frac{1}{K} \sum_{k=1}^{K} y_{ijk}$$

schätzt den Erwartungswert des Response in der ij-ten Zelle. Davon ausgehend dokumentiert die Grafik 3.14 einen lupenreinen Interaktionseffekt zwischen den Faktoren *group* und *drug*. Während das Placebo die hohe Aktivität bei den Hyperaktiven erwartungsgemäß nicht reduziert und bei der Kontrollgruppe die auch sonst übliche Aktivität auslöst, wirkt Ritalin umgekehrt: Die hohe Aktivität in der hyperaktiven Gruppe wird stark reduziert, dagegen steigt seltsamerweise die Aktivität in der Kontrollgruppe über das Mittelmaß hinaus, wenn Ritalin verabreicht wird.

Dies wird auch in S-Plus mit der allgemeineren Interaktions-Modellierung deutlich:
$r <- aov(activity \sim group * drug)$
$summary(r)$
liefert

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
group	1	121.00	121.000	8.00000	0.0152201
drug	1	42.25	42.250	2.79339	0.1205062
group:drug	1	930.25	930.250	61.50413	0.0000046
Residuals	12	181.50	15.125		

und damit einen hochsignifikanten Interaktionseffekt. In diesem größeren Modell verändern sich auch die Signifikanzen der Faktoren selbst im Vergleich zum rein additiven Modell.

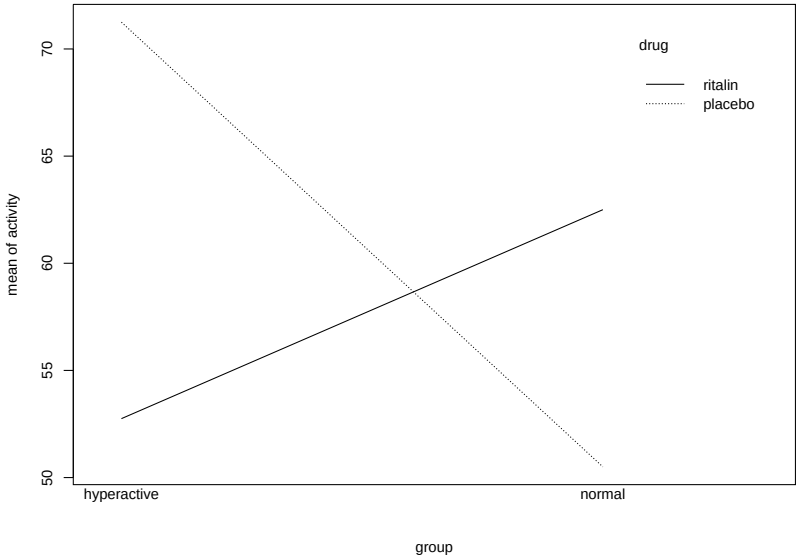

Abbildung 3.14: Interaktions-Plot: Response *activity* bzgl. der Faktoren *group* und *drug*

Modellansätze für die Parameter in ANOVA

1. ohne Interaktion

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad \begin{matrix} i = 1, \ldots, I-1 \\ j = 1, \ldots, J-1 \\ k = 1, \ldots, K \end{matrix}$$

$\alpha_I = 0,\ \beta_J = 0$

μ = overall mean
α_i = Effekt von Faktor A in Stufe (level) i
β_j = Effekt von Faktor B in Stufe (level) j

H_0 : kein Faktor A Effekt $\Rightarrow H_0 :\ \alpha_1 = \cdots = \alpha_{I-1} = 0$
H_0 : kein Faktor B Effekt $\Rightarrow H_0 :\ \beta_1 = \cdots = \beta_{J-1} = 0$

Für ein ANOVA-Modell ohne Interaktion gilt:

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j.$$

Also müssen $\bar{y}_{ij\cdot}$ parallel verlaufen, d.h. der Plot sollte im Idealfall qualitativ wie die Grafik 3.15 aussehen. Falls dies nicht hinreichend gilt, so muss man Interaktionen zwischen den Faktoren erlauben.

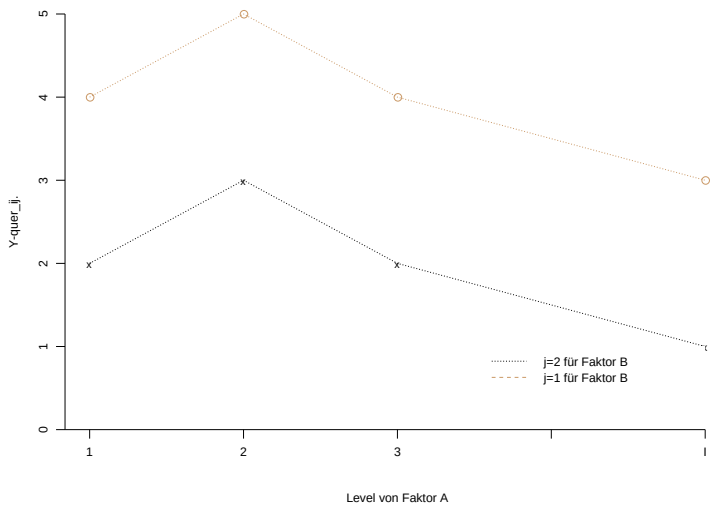

Abbildung 3.15: Two-Way-Anova ohne Interaktion

2. mit Interaktion

$$\begin{aligned} y_{ijk} = \mu + \alpha_i + \beta_j + \theta_{ij} + \epsilon_{ijk}, \; i &= 1, \ldots, I \\ j &= 1, \ldots, J \\ k &= 1, \ldots, K \end{aligned}$$

Identifizierbarkeitsbedingungen:

$$\alpha_1 + \cdots + \alpha_I = 0$$
$$\beta_1 + \cdots + \beta_J = 0$$
$$\sum_{i=1}^{I} \theta_{ij} = \sum_{j=1}^{J} \theta_{ij} = 0$$

Auch hier werden in der Regel die Gleichungsnebenbedingungen durch Einsetzen eliminiert, so dass man für obiges Beispiel (diesmal mit Effekt-Kodierung) folgende Designmatrix erhält:

$$
\begin{array}{ll}
i=1,\ j=1 & \\ & \\ & \\ & \\
i=2,\ j=1 & \\ & \\ & \\ & \\
i=1,\ j=2 & \\ & \\ & \\ & \\
i=2,\ j=2 & \\ & \\ & \\ &
\end{array}
\begin{pmatrix} 50 \\ 45 \\ 55 \\ 52 \\ 70 \\ 72 \\ 68 \\ 75 \\ 67 \\ 60 \\ 58 \\ 65 \\ 51 \\ 57 \\ 48 \\ 55 \end{pmatrix}
=
\begin{pmatrix}
1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1
\end{pmatrix}
\cdot
\begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \theta_{11} \end{pmatrix}
+ \boldsymbol{\epsilon}
$$

Einige denkbare Hypothesen sind:

$$
\begin{array}{ll}
H_0: \text{keine Interaktion} & \Leftrightarrow H_0: \theta_{11} = 0 \\
H_0: \text{keine Interaktion} & \Leftrightarrow H_0: \theta_{11} = 0 \\
\quad \text{kein Faktor A Effekt} & \qquad\quad \alpha_1 = 0 \\
H_0: \text{keine Interaktion} & \Leftrightarrow H_0: \theta_{11} = 0 \\
\quad \text{kein Faktor B Effekt} & \qquad\quad \beta_1 = 0 \\
H_0: \text{keine Interaktion} & \Leftrightarrow H_0: \theta_{11} = 0 \\
\quad \text{keine Faktor A und B Effekte} & \qquad\quad \alpha_1 = \beta_1 = 0
\end{array}
$$

Identifizierbarkeitsbedingungen:

$$
\begin{array}{ll}
\alpha_1 + \alpha_2 = 0 & \\
\beta_1 + \beta_2 = 0 & \\
\theta_{11} + \theta_{12} = 0, & \theta_{11} + \theta_{21} = 0 \\
\theta_{21} + \theta_{22} = 0, & \theta_{12} + \theta_{22} = 0
\end{array}
$$

Unbalanced Design; beliebig viele Beobachtungen pro Zelle

Beispiel: Genotypus mit zwei Faktoren

In einem Beispiel aus Scheffé (1959, pp. 139-140) wird das 28-Tage-Gewicht/Wurf von Tieren untersucht, die von Pflegemüttern aufgezogen wurden. Man unterscheidet in der betreffenden Art 4 Genotypen (Faktor A: Genotypus des Wurfes, Faktor B: Genotypus der Pflegemutter).

Man möchte anhand des Experimentes feststellen, ob die Pflegemütter Kinder des gleichen Typs besser betreuen. In diesem Fall müsste das Gewicht signifikant vom Mittelwert abweichen.

Einige Daten:

weight	A	B
61.5	1	1
68.2	1	1
64.0	1	1
65.0	1	1
59.7	1	1
55.0	1	2
42.0	1	2
60.2	1	2
52.5	1	3
61.8	1	3
49.5	1	3
52.7	1	3
42.0	1	4
...	..	..

Wir haben 5 Beobachtungen in der 1-1-Zelle, 3 Beobachtungen in der 1-2-Zelle, 4 Beobachtungen in der 1-3-Zelle usw., also keine "balanced Design" wie oben. Die Analyse kann aber mit S-Plus formal in gleicher Weise erfolgen.

Explorative Graphische Analyse:

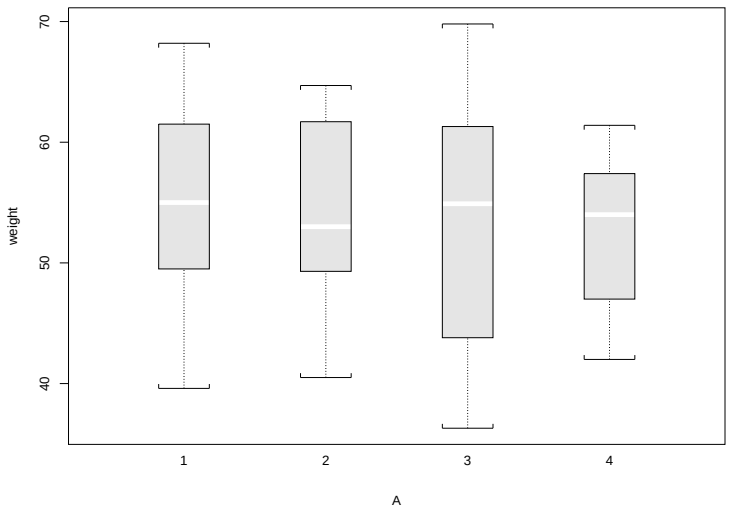

Abbildung 3.16: Response *weight* innerhalb des Faktors A: Genotypus Wurf

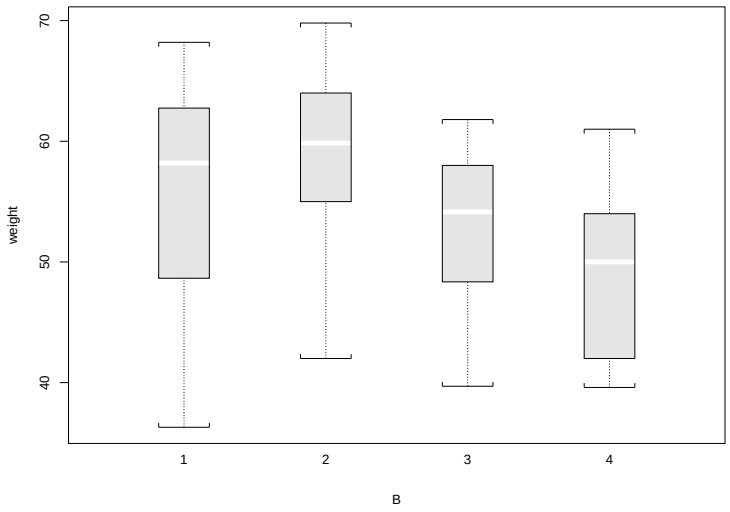

Abbildung 3.17: Response *weight* innerhalb des Faktors B: Genotypus Pflegemutter

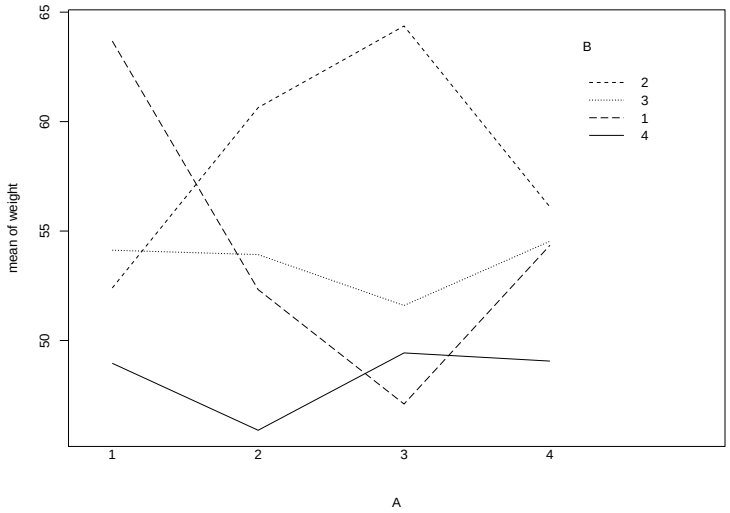

Abbildung 3.18: Interaktions-Plot: Response *weight* bzgl. der Faktoren A und B

Natürlich werden konkrete Interaktions-Plots von den Idealdarstellungen der Bilder 3.13 und 3.14 abweichen. Trotzdem deutet in der Abbbildung 3.18 nur wenig auf mögliche Interaktionen hin. Dies bestätigen auch die folgenden Rechnungen in S-Plus.

$r <- aov(weight \sim A * B)$
$summary(r)$
liefert mit Interaktionen

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
A	3	60.157	20.0524	0.369696	0.775221
B	3	775.081	258.3602	4.763246	0.005736
A:B	9	824.073	91.5636	1.688108	0.120053
Residuals	45	2440.816	54.2404		

Der Interaktionsterm ist mit einem P-Wert von 12% nicht signifikant. Also kann das rein additive Modell gewählt werden:

$r <- aov(weight \sim A + B)$
$summary(r)$
mit dem Resultat:

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
A	3	60.157	20.0524	0.331659	0.8024695
B	3	775.081	258.3602	4.273178	0.0088605
Residuals	54	3264.889	60.4609		

Dieses weist nur auf einen signifikanten Einfluss des Faktors B hin.

Fazit:

1. Die Ausgangsfragestellung des Experiments kann nicht bestätigt werden:
   Würden nämlich Pflegemütter Kinder des gleichen Typs besser betreuen, so wären signifikante Interaktionen aufgetreten. Das war nicht der Fall.

2. Der signifikante Einfluss des Faktors B besagt, dass gewisse Genotypen von Müttern besser zur Aufzucht geeignet sind. Der Boxplot 3.17 weist den Genotypus 2 als besonders gute, und den Genotypus 4 als weniger geeignete Pflegemutter aus.

Kapitel 4

Anhang

4.1 Normalgleichungen

Seien $\boldsymbol{X} \in \mathbb{R}^{n,p},\ n \geq p,\ \boldsymbol{y} \in \mathbb{R}^n$ gegebene Daten und $\boldsymbol{\beta} \in \mathbb{R}^p$ ein unbekannter Parametervektor. Dann ist jede Lösung der Normalgleichungen

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y} \tag{4.1}$$

Lösung des KQ-Problems

$$\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2\}\,. \tag{4.2}$$

Die Normalgleichungen (4.1) haben in jedem Fall mindestens eine Lösung.
Falls Rang $(\boldsymbol{X}) = p$, so lautet die eindeutige Lösung von (4.2)

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^T\boldsymbol{y}. \tag{4.3}$$

Besonders einfach stellt sich die Lösung von (4.2) dar für

$$\boldsymbol{X} = \boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p) \tag{4.4}$$

und

$$\boldsymbol{q}_i^T\boldsymbol{q}_j = \delta_{ij},$$

wenn also $\boldsymbol{Q}$ orthonormierte Spalten hat. In diesem Fall vereinfacht sich (4.3) zu:

$$\boldsymbol{b} = \boldsymbol{Q}^T\boldsymbol{y}.$$

Lemma 4.1
Lösungen des KQ-Problems

$$\min_{\boldsymbol{\beta}}\{Q(\boldsymbol{\beta})\}\,, \textit{wobei } Q(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2\,,$$

sind durch 0 nach unten beschränkt. Nach einem Satz aus der Theorie quadratischer Funktionen muss die Funktion Q ihr Minimum in mindestens einem Punkt $\boldsymbol{b} \in \mathbb{R}^p$ annehmen. Hier gilt notwendigerweise

$$\nabla Q(\boldsymbol{b}) = \mathbf{0} \quad \textit{und das ist äquivalent zu} \quad (\mathbf{4.1}).$$

Damit haben die Normalgleichungen immer eine Lösung.

4.2 Rechnen mit Erwartungswerten und Kovarianzmatrizen

4.2.1 n-dimensionale Normalverteilung

Def. 4.2 (n-dimensionale Normalverteilung $N(\boldsymbol{\mu}, \boldsymbol{C})$)

Seien $\boldsymbol{\mu} \in \mathbb{R}^n$ und $\boldsymbol{C} = (c_{ij}) \in \mathbb{R}^{n,n}$ eine positiv definite Matrix (damit ist $\boldsymbol{C}$ auch symmetrisch). $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ heißt n-dimensional normalverteilt*, wenn es eine Dichte der Form*

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \, det(\boldsymbol{C})}} \exp\left[-\tfrac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{4.5}$$

besitzt. Dabei bezeichnet $det(\boldsymbol{C})$ die Determinante von $\boldsymbol{C}$.

Def. 4.3 (2-dimensionale Normalverteilung)

Mit $(\mu_Z, \mu_Y) \in \mathbb{R}^2$, $\sigma > 0\,, \tau > 0$ und $-1 < \rho < 1$ ist die Kovarianzmatrix (vgl. Def. 4.4)

$$\boldsymbol{C} := \begin{pmatrix} \sigma^2 & \rho\,\sigma\,\tau \\ \rho\,\sigma\,\tau & \tau^2 \end{pmatrix}$$

positiv definit und $det(\boldsymbol{C}) = \sigma^2\tau^2\,(1-\rho^2)$. Die Inverse $\boldsymbol{C}^{-1}$ ist explizit berechenbar und es gilt gemäß (4.5)

$$f_{Z,Y}(z,y) = \tag{4.6}$$

$$\frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(z-\mu_Z)^2}{\sigma^2} - \frac{2\rho(z-\mu_Z)(y-\mu_Y)}{\sigma\,\tau} + \frac{(y-\mu_Y)^2}{\tau^2}\right)\right]$$

4.2.2 Erwartungswerte; n-dim.

Es sei hier nochmals an die Vereinbarung erinnert, daß $\boldsymbol{X} = (X_1, ..., X_n)^T \in \mathbb{R}^n$ und $\boldsymbol{x} = (x_1, ..., x_n)^T \in \mathbb{R}^n$ mit der Indizierung x_k, $k = 1, .., n$.

Def. 4.4 (Kovarianzen, Kovarianzmatrix)

Für $\boldsymbol{X} = (X_1, ..., X_n)^T$ *mögen alle zweiten Momente*

$$\tau_{kl} := E(X_k\, X_l)\,, \qquad 1 \le k, l \le n\,,$$

existieren. Mit $\mu_k := E(X_k)$ *nennt man*

$$\sigma_{kl} := Cov(X_k, X_l) := E\left[(X_k - \mu_k)(X_l - \mu_l)\right]\,, \qquad 1 \le k, l \le n\,,$$

die **Kovarianzen** *von* X_k *und* X_l*.*

$$\boldsymbol{C} := \boldsymbol{Cov}(\boldsymbol{X}) := (\sigma_{kl})_{1\le k,l\le n} \in \mathbb{R}^{n,n}$$

heißt **Kovarianzmatrix** *von* $\boldsymbol{X}$*.*
X_k *und* X_l *heißen* **unkorreliert**, *falls* $Cov(X_k, X_l) = 0$.
Die Existenz aller zweiten Momente von $\boldsymbol{X}$ *sichert also, dass alle Erwartungswerte* μ_k*, alle Varianzen* σ_{kk} *und alle Kovarianzen* σ_{kl} *der Komponenten* X_k, X_l, $1 \le k, l \le n$ *von* $\boldsymbol{X}$ *wohldefiniert sind.*

Folgerung 4.5 ((Ko-)Varianzen, Verschiebungsregel)

1. Für $k = l$ gilt

$$\begin{aligned} \sigma_k^2 := \sigma_{kk} = Cov(X_k, X_k) &= E\left[(X_k - \mu_k)(X_k - \mu_k)\right] \\ &= E\left[(X_k - \mu_k)^2\right] = Var(X_k)\,. \end{aligned} \tag{4.7}$$

2. Auch für Kovarianzen gibt es eine **Verschiebungsregel**, nämlich

$$\sigma_{kl} = Cov(X_k, X_l) = E\left[(X_k - \mu_k)(X_l - \mu_l)\right] = E\left(X_k\, X_l\right) - \mu_k\, \mu_l\,. \tag{4.8}$$

Def. 4.6 (Korrelationskoeffizient)

Für zwei ZV, z.B. $(Z, Y) = (X_k, X_l)$*, sollen alle zweiten Momente existieren. Ferner gelte* $\sigma_Z^2 := Var(Z) > 0$ *und* $\sigma_Y^2 := Var(Y) > 0$*. Dann heißt*

$$\rho(Z, Y) := \frac{Cov(Z, Y)}{\sigma_Z\, \sigma_Y} = \frac{Cov(Z, Y)}{D(Z)\, D(Y)} = \frac{Cov(Z, Y)}{\sqrt{Var(Z)\, Var(Y)}}$$

der **Korrelationskoeffizient** *von* Z *und* Y*.*

Satz 4.7 (Korrelationskoeffizient)

Falls alle zweiten Momente der ZV (Z, Y) *existieren und falls* $Var(Z) > 0$ *und* $Var(Y) > 0$*, so gilt:*

$$-1 \leq \rho(Z, Y) \leq 1$$

und $\rho(Z, Y) = \pm 1$ *genau dann, wenn*

$$P(Y = a + b\,Z) = P(\{\,\omega \,|\, Y(\omega) = a + b\,Z(\omega)\}\,) = 1$$

für geeignete $a \in \mathbb{R}$ *und* $b \neq 0$*. Ferner gilt für* $|\rho(Z, Y)| = 1$*:*

$$\rho(Z, Y) = 1 \iff b > 0 \quad \textit{und} \quad \rho(Z, Y) = -1 \iff b < 0\,.$$

Beispiel 4.8 (Zweidimensionale Normalverteilung)

von (Z, Y) *bzgl. Def. 4.3 mit den Parametern* $\boldsymbol{\mu} := (\mu_z, \mu_y)^T \in \mathbb{R}^2$ *sowie* $\sigma_z > 0\,, \sigma_y > 0$ *und* $-1 < \rho < 1$ *und der positiv definiten Kovarianzmatrix*

$$\boldsymbol{C} := \boldsymbol{Cov}(Z, Y) = \begin{pmatrix} \sigma_z^2 & \rho\,\sigma_z\,\sigma_y \\ \rho\,\sigma_z\,\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Dort gilt $E(Z) = \mu_z$*,* $E(Y) = \mu_y$*,* $Var(Z) = \sigma_z^2$*,* $Var(Y) = \sigma_y^2$*,* $\rho(Z, Y) = \rho$*.*

Def. 4.9 (EW von Zufallsvektoren, -matrizen)

Vorausgesetzt für $k = 1, ..., n$ *existieren die Erwartungswerte* $\mu_k := E(X_k)$*, so setzt man für* $\boldsymbol{X} = (X_1, ..., X_n)^T$

$$E(\boldsymbol{X}) := \boldsymbol{\mu} := (\mu_1, ..., \mu_n)^T\,.$$

Analog (also auch elementweise) wird der Erwartungswert von Matrizen definiert, deren Elemente ZV sind.

Satz 4.10 (EW von Zufallsvektoren)

Für $\boldsymbol{X} = (X_1, ..., X_n)^T$ *existiere* $\boldsymbol{\mu} = E(\boldsymbol{X})$*, dann gilt für (nichtstochastisches)* $\boldsymbol{A} \in \mathbb{R}^{m,n}$*,* $m \in \mathbb{N}$

$$E(\boldsymbol{A}\,\boldsymbol{X}) = \boldsymbol{A}\,E(\boldsymbol{X}) = \boldsymbol{A}\,\boldsymbol{\mu}\,. \tag{4.9}$$

Satz 4.11 (Kovarianzen von Zufallsvektoren; Varianz von Summen)

Vorausgesetzt alle zweiten Momente von $\boldsymbol{X} = (X_1, ..., X_n)^T$ *existieren, dann gilt mit* $\boldsymbol{\mu} = E(\boldsymbol{X})$, $\boldsymbol{A} \in \mathbb{R}^{m,n}$, $\boldsymbol{a} = (a_1, ..., a_n)^T \in \mathbb{R}^n$:

1. *Die Kovarianzmatrix* $\boldsymbol{C}$ *ist positiv semidefinit.*

2.
$$\begin{aligned} \boldsymbol{C} = (\sigma_{kl}) = \boldsymbol{Cov}(\boldsymbol{X}) &= E\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\right] \\ &= E\left[\boldsymbol{X}\,\boldsymbol{X}^T\right] - E\left[\boldsymbol{\mu}\,\boldsymbol{\mu}^T\right] \end{aligned}$$

3.
$$\boldsymbol{Cov}(\boldsymbol{A}\,\boldsymbol{X}) = \boldsymbol{A}\,\boldsymbol{C}\,\boldsymbol{A}^T \tag{4.10}$$

4.
$$\begin{aligned} Var(\boldsymbol{a}^T\boldsymbol{X}) = Var\left(\sum_{i=1}^{n} a_i\,X_i\right) &= Cov(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\boldsymbol{C}\,\boldsymbol{a} = \sum_{k,l} a_k\,a_l\,\sigma_{kl} \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} a_k\,a_l\,Cov(X_k, X_l) \qquad (4.11) \\ &= \sum_{k=1}^{n} a_k^2\,Var(X_k) + 2\sum_{k<l} a_k\,a_l\,Cov(X_k, X_l) \end{aligned}$$

5.
$$\begin{aligned} Var(\boldsymbol{a}^T\boldsymbol{X}) = Var(\sum_{i=1}^{n} a_i\,X_i) &= \sum_{k=1}^{n} a_k^2\,Var(X_k)\,, \qquad (4.12) \\ &\quad \textit{falls } (X_k, X_l) \textit{ unkorreliert.} \end{aligned}$$

4.3 Transformationen für Dichten und unabhängig normalverteilte Zufallsvariablen

4.3.1 Transformationssatz für Dichten

Die n-dimensionale Zufallsvariable $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ habe eine Dichte $f(\boldsymbol{x})$,

$\boldsymbol{x} = (x_1, \ldots, x_n)^T$. Weiterhin seien $B := \{\, \boldsymbol{x} \in \mathbb{R}^n \,|\, f(\boldsymbol{x}) > 0 \,\}$ und $\boldsymbol{h} : B \to D$, $D \subset \mathbb{R}^n$, differenzierbar und bijektiv. Dann ist $\boldsymbol{Y} := \boldsymbol{h}(\boldsymbol{X})$ wieder eine n-dim. ZV. Mit $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ sei die Umkehrabbildung

$$\boldsymbol{x} := \boldsymbol{h}^{-1}(\boldsymbol{y}) = (x_1(\boldsymbol{y}), \ldots, x_n(\boldsymbol{y}))^T$$

auf D (abgesehen von endlich vielen glatten Hyperflächen) stetig differenzierbar.

$$\frac{\partial(x_1,\ldots,x_n)}{\partial(y_1,\ldots,y_n)} = \frac{\partial(x_1,\ldots,x_n)}{\partial(y_1,\ldots,y_n)}(\boldsymbol{y}) := det \begin{pmatrix} \frac{\partial x_1}{\partial y_1}(\boldsymbol{y}) & \cdots & \frac{\partial x_1}{\partial y_n}(\boldsymbol{y}) \\ \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1}(\boldsymbol{y}) & \cdots & \frac{\partial x_n}{\partial y_n}(\boldsymbol{y}) \end{pmatrix}$$

sei die Funktionaldeterminante der Umkehrabbildung $\boldsymbol{h}^{-1}$. Dann hat die ZV $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^T := \boldsymbol{h}(\boldsymbol{X})$ wieder eine Dichte $g(\boldsymbol{y})$, die gegeben ist durch

$$g(\boldsymbol{y}) = g(y_1,\ldots,y_n) = f\left(x_1(\boldsymbol{y}),\ldots,x_n(\boldsymbol{y})\right) \left|\frac{\partial(x_1,\ldots,x_n)}{\partial(y_1,\ldots,y_n)}\right| , \tag{4.13}$$

wobei $|\cdot|$ hier den Betrag einer Determinante bezeichnet.

4.3.2 Lineare Transformationen

Wir betrachten nun eine weitere wichtige Klasse linearer Transformationen und folgen in der Darstellung [Stirzaker (1994)], pp. 287.

Satz 4.12 (Lineare Transformation)

Zur Matrix $\boldsymbol{A} = (a_{ij})$ existiere die Inverse $\boldsymbol{A}^{-1} = (b_{ij}) = \boldsymbol{B}$. Weiterhin betrachten wir $\boldsymbol{X} = (X_1,\ldots,X_n)^T$, $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^T$ und

$$\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{X}\,,\ \boldsymbol{X} = \boldsymbol{B}\,\boldsymbol{Y} \ \text{ bzw. für } i = 1,\ldots,n: \quad Y_i = \sum_{j=1}^{n} a_{ij}\,X_j\,,\ X_i = \sum_{j=1}^{n} b_{ij}\,Y_j\,.$$

Aus der gemeinsamen Dichte $f_X(x_1,\ldots,x_n)$ der $(X_1,..,X_n)$ kann dann wegen $det(\boldsymbol{A}) = 1/det(\boldsymbol{B}) \neq 0$ die gemeinsame Dichte $f_Y(y_1,\ldots,y_n)$ von $(Y_1,\ldots,Y_n)$ berechnet werden. Es gilt gemäß Gleichung (4.13) mit $\boldsymbol{y} = (y_1,\ldots,y_n)^T$

$$\begin{aligned} f_Y(y_1,\ldots,y_n) &= \frac{1}{|det(\boldsymbol{A})|}\, f_X(x_1(y_1,\ldots,y_n),\ldots,x_n(y_1,\ldots,y_n)) \\ &= |det(\boldsymbol{B})|\, f_X(x_1,\ldots,x_n) = \left|\, det\left(\boldsymbol{A}^{-1}\right)\right|\, f_X(\boldsymbol{A}^{-1}\,\boldsymbol{y})\,. \end{aligned}$$

Für affin lineare Transformationen $\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{X} + \boldsymbol{b}$ mit $det(\boldsymbol{A}) \neq 0$ gilt entsprechend

$$f_Y(\boldsymbol{y}) = \left|\, det\left(\boldsymbol{A}^{-1}\right)\right|\, f_X\left(\boldsymbol{A}^{-1}(\boldsymbol{y}-\boldsymbol{b})\right)\,. \tag{4.14}$$

Satz 4.13 (Orthogonale Transformation bei Normalverteilung)

Seien $(X_1, \ldots, X_n)$ $N(0,1)$-iid Zufallsvariablen und $\boldsymbol{A} = (a_{ij})$ eine orthogonale Matrix mit $det(\boldsymbol{A}) = \pm 1$ sowie $\boldsymbol{A}^{-1} = \boldsymbol{A}^T$. Weiterhin gelte mit $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ und $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ die Beziehung $\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{X}$, d. h.

$$Y_i = \sum_{j=1}^{n} a_{ij}\, X_j\,, \quad 1 \le i \le n\,. \tag{4.15}$$

Dann erhält man:

1. $(Y_1, \ldots, Y_n)$ sind unabhängige $N(0,1)$-verteilte Zufallsvariablen.
2. Das Stichprobenmittel $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ und die

   Stichprobenvarianz $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$ sind unabhängig .
3. Für $N(\mu, \sigma^2)$-iid ZV $(X_1, \ldots, X_n)$ gilt $E\,(S^2) = \sigma^2$.

Beweis:

1. Es gilt $\boldsymbol{X} = \boldsymbol{A}^T\,\boldsymbol{Y}$ und

   $$\sum_{i=1}^{n} X_i^2 = \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Y}^T\,\boldsymbol{A}\,\boldsymbol{A}^T\,\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^{n} Y_i^2\,.$$

   Die unabhängigen $N(0,1)$-verteilten X_i haben die Dichte $(2\pi)^{-n/2}\,\exp\left(-\frac{1}{2}\sum x_i^2\right)$. Somit gilt für die Dichte von $(Y_1, \ldots, Y_n)$ nach Satz 4.12

   $$f_Y(y_1, \ldots, y_n) = \frac{1}{(2\pi)^{\frac{n}{2}}}\,\exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right).$$

   Deshalb sind $(Y_1, \ldots, Y_n)$ unabhängig $N(0,1)$-verteilt.
2. Nun sei $\boldsymbol{A} = (a_{ij})$ orthogonal mit spezieller erster Zeile der Form $a_{1j} = 1/\sqrt{n}$, was

   $$Y_1 = \sum_{j=1}^{n} \frac{1}{\sqrt{n}}\, X_j = \sqrt{n}\,\bar{X}$$

   zur Folge hat. Weiterhin gilt

   $$\begin{aligned}(n-1)\,S^2 &= \sum_{i=1}^{n} X_i^2 - 2\,\bar{X}\sum_{i=1}^{n} X_i + n\,\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2 \\ &= \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2\,. \end{aligned} \tag{4.16}$$

   S^2 ist unabhängig von $\bar{X}$, da Y_1 gemäß 1. unabhängig von $(Y_2, \ldots, Y_n)$ ist.

3. Ausgehend von $N(\mu, \sigma^2)$-iid $\tilde{X}_i$ führt man zunächst eine Translation der Form $\tilde{X}_i - \mu \to X_i$, $i = 1, \ldots, n$ durch. Dann sind die X_i iid und $N(0, \sigma^2)$. Durch Einsetzen gemäß 1. verifiziert man sofort, daß die orthogonale Transformation (4.15) $N(0, \sigma^2)$ iid Y_i liefert. Damit folgt gemäß 2. und (4.16)

$$E\left[(n-1)\,S^2\right] = E\left[\sum_{i=2}^{n} Y_i^2\right] = (n-1)\,\sigma^2\,. \qquad \square$$

Lineare Transformationen normalverteilter ZV

Aus Satz 4.12 folgt noch, dass allgemeine lineare Transformationen normalverteilter ZV wieder normalverteilt sind:

Seien $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{C})$ und $\boldsymbol{A} \in \mathbb{R}^{m,n}$, $Rang(\boldsymbol{A}) = m > 0$ gegeben. Dann gilt

$$\boldsymbol{Y} := \boldsymbol{A}\,\boldsymbol{X} \sim N\left(\boldsymbol{A}\,\boldsymbol{\mu}, \boldsymbol{A}\,\boldsymbol{C}\,\boldsymbol{A}^T\right)\,. \tag{4.17}$$

Zum Beweis ergänze man $\boldsymbol{A}$ zu einer nichtsingulären Matrix.

4.4 Zentriertes Modell, orthogonale Designmatrix

Wir betrachten das zentrierte Modell der Einfachen Linearen Regresstion mit normalverteilten Fehlern. Für gegebene Daten $(x_1, y_1), \ldots, (x_n, y_n)$, $i = 1, \ldots, n$ ist dies ein auf $\bar{x}$ zentrierter Ansatz der Form

$$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i,\ \epsilon_i \text{ iid } N(0, \sigma^2),\ i = 1, \ldots, n.$$

Seien nun $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T, \mathbf{1} = (1, \ldots, 1)^T$,
$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ sowie $SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ und $SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$.
Weiterhin seien b_0^* und b_1 die KQ-Schätzungen für β_0^* und β_1. $\hat{Y}_i = b_0^* + b_1(x_i - \bar{x})$,
$e_i = Y_i - \hat{Y}_i$, $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T, \boldsymbol{e} = (e_1, \ldots, e_n)^T$ und $SS_e = \boldsymbol{e}^T\boldsymbol{e} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

Lemma 4.14

$$\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x} - \bar{x} \cdot \mathbf{1})$$

hat orthogonale Spalten und

$$\boldsymbol{b} = \begin{pmatrix} b_0^* \\ b_1 \end{pmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} \bar{Y} \\ SS_{xy}/SS_{xx} \end{pmatrix}.$$

Beweis: Einsetzen und Anwendung von Abschnitt 4.1.

Satz 4.15 *Es gilt:*

1. $E(\boldsymbol{b}) = E\begin{pmatrix} b_o^* \\ b_1 \end{pmatrix} = \begin{pmatrix} \beta_o^* \\ \beta_1 \end{pmatrix}$
2. $E(SS_e) = (n-2)\sigma^2$
3. b_0^*, b_1 *und* SS_e *sind unabhängig.*

Beweis:
Die Lösung besteht darin, den Zufallsvektor $\boldsymbol{Y} \in \mathbb{R}^n$ durch eine orthogonale Transformation $\boldsymbol{T} \in \mathbb{R}^{n,n}$ so zu "drehen", dass die beiden ersten Komponenten Z_1, Z_2 von $\boldsymbol{Z} := \boldsymbol{T}^T\boldsymbol{Y}$ in $span(\boldsymbol{X})$ liegen und die übrigen $Z_3, \ldots, Z_n$ senkrecht dazu stehen, wobei gemäß Lemma 4.14

$$\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x} - \bar{x} \cdot \boldsymbol{1}) \in \mathbb{R}^{n,2}$$
$$\boldsymbol{x} = (x_1, \ldots, x_n)^T,\ \boldsymbol{Y} = (Y_1, \ldots, Y_n)^T.$$

Dazu reparametrisiert man das zentrierte Modell

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\beta} = (\beta_0^*, \beta_1)^T$$

mit

$$\boldsymbol{q}_1 := \boldsymbol{1}/\sqrt{n},\ \boldsymbol{q}_2 := (\boldsymbol{x} - \bar{x} \cdot \boldsymbol{1})/\sqrt{SS_{xx}}$$
$$\boldsymbol{Q} := (\boldsymbol{q}_1, \boldsymbol{q}_2) \in \mathbb{R}^{n,2},\ \boldsymbol{D} = \begin{pmatrix} 1/\sqrt{n} & 0 \\ 0 & 1/\sqrt{SS_{xx}} \end{pmatrix}.$$

Nach Definition hat $\boldsymbol{Q}$ orthogonale Spalten. Weiterhin gilt

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{D}$$

und

$$\boldsymbol{X}\boldsymbol{\beta} = \underbrace{\boldsymbol{X}\boldsymbol{D}}_{\boldsymbol{Q}}\underbrace{\boldsymbol{D}^{-1}\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \boldsymbol{Q}\boldsymbol{\gamma},$$

wobei

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} \sqrt{n}\beta_0^* \\ \sqrt{SS_{xx}}\beta_1 \end{pmatrix}$$

und $span(\boldsymbol{Q}) = span(\boldsymbol{X})$. Also

$$\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \Longleftrightarrow \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Nach Lemma 4.14 gilt für die KQ-Schätzungen

$$\boldsymbol{g} = \hat{\boldsymbol{\gamma}} = (\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{Y} = \boldsymbol{Q}^T\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}^T\boldsymbol{Y} =$$
$$= \boldsymbol{D}\underbrace{(\boldsymbol{X}^T\boldsymbol{X})}_{\boldsymbol{D}^{-1}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{D}^{-1}}\underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}_{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{D}^{-2}\boldsymbol{b} = \boldsymbol{D}^{-1}\boldsymbol{b}$$

Nun ergänzen wir $\boldsymbol{Q}$ durch

$$\boldsymbol{R} := (\boldsymbol{q}_3, \ldots, \boldsymbol{q}_n) \in \mathbb{R}^{n,n-2}$$

zu einer orthogonalen Matrix $\boldsymbol{T} = (\boldsymbol{Q}, \boldsymbol{R})$.

Wegen $\boldsymbol{Y} \sim N(\boldsymbol{Q}\boldsymbol{\gamma}, \sigma^2\boldsymbol{I})$ und Satz 4.13 ist

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ - - - \\ Z_3 \\ . \\ Z_n \end{pmatrix} := \boldsymbol{T}^T\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Q}^T\boldsymbol{Y} \\ - - - \\ \boldsymbol{R}^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{g} \\ - - - \\ \boldsymbol{Z}_e \end{pmatrix} \sim N(\boldsymbol{\delta}, \sigma^2\boldsymbol{I});$$

und alle Komponenten von $\boldsymbol{Z}$ sind unabhängig.

1. $E(\boldsymbol{g}) = E(\boldsymbol{Q}^T\boldsymbol{Y}) = \boldsymbol{Q}^T E(\boldsymbol{Y}) = \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{\gamma} = \boldsymbol{\gamma} = \begin{pmatrix} \sqrt{n}\beta_0^* \\ \sqrt{SS_{xx}}\beta_1 \end{pmatrix}$.

Insbesondere
$Z_1 = g_0 \sim N(\sqrt{n}\beta_0^*, \sigma^2)$,
$Z_2 = g_1 \sim N(\sqrt{SS_{xx}}\beta_1, \sigma^2)$ und
$SS_Z = \sum_{i=3}^{n} Z_i^2 = \boldsymbol{Z}_e^T\boldsymbol{Z}_e$.

Es gilt $SS_Z/\sigma^2 \sim \mathcal{X}^2_{n-2}$, da $E(\boldsymbol{Z_e}) = \boldsymbol{R}^T\boldsymbol{Q}\boldsymbol{\gamma} = \boldsymbol{0}$.

Wegen $\boldsymbol{b} = \boldsymbol{D}\boldsymbol{g}$ folgt b_0^* und b_1 sind unabhängig mit
$b_0^* = g_0/\sqrt{n} \sim N(\beta_0^*, \sigma^2/\sqrt{n})$ und
$b_1^* = g_1/\sqrt{SS_{xx}} \sim N(\beta_1, \sigma^2/\sqrt{SS_{xx}})$.

Es bleibt noch zu zeigen:
$SS_Z = SS_e = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$,
wobei

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{Q}\boldsymbol{D}^{-1}(\underbrace{\boldsymbol{D}^{-1}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{D}^{-1}}_{\boldsymbol{D}^{-2}})^{-1}\boldsymbol{D}^{-1}\boldsymbol{Q}^T = \\ &= \boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{D}^2\boldsymbol{D}^{-1}\boldsymbol{Q}^T = \boldsymbol{Q}\boldsymbol{Q}^T\,. \end{aligned}$$

Mit
$$\boldsymbol{I} = \boldsymbol{T}\boldsymbol{T}^T = (\boldsymbol{Q}, \boldsymbol{R}) \begin{pmatrix} \boldsymbol{Q}^T \\ \boldsymbol{R}^T \end{pmatrix} = \boldsymbol{Q}\boldsymbol{Q}^T + \boldsymbol{R}\boldsymbol{R}^T$$
folgt schließlich mit $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{b}$:

$$\begin{aligned} SS_e &= \boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{X}\,\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\,\boldsymbol{b} = \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \\ &= \boldsymbol{Y}^T(\boldsymbol{Q}\boldsymbol{Q}^T + \boldsymbol{R}\boldsymbol{R}^T - \boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{Y} = \\ &= \boldsymbol{Y}^T\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{Y} = \boldsymbol{Z}_e^T\boldsymbol{Z}_e = SS_Z\,. \end{aligned}$$

Insbesondere sind b_0^*, b_1 und SS_e unabhängig. Damit ist 3. gezeigt.

2. $E(SS_e) = E(SS_Z) = (n-2)\sigma^2$, da $SS_Z/\sigma^2 \sim \mathcal{X}^2_{n-2}$.

Literaturverzeichnis

[Dielman et al. (1996)] T.E.: Applied Regression Analysis for Business and Economics (2nd ed). Wadsworth Publishing Company

[Falk et al. (1995)] Falk M., Becker R. und Mahrohn F.: Angewandte Statistik mit SAS. Springer, Berlin.

[Fahrmeir et al. (1996)] Fahrmeir. L, Hamerle A. und Tutz G. (Hrsg.): Multivariate statistische Verfahren, 2., erweiterte Auflage. De Gruyter, Berlin.

[Kredler (1999)] Ch.: Einführung in die Wahrscheinlichkeitsrechnung und Statistik. Ausgearbeitetes Vorlesungsskript zur Stochastik 1. Zentrum Mathematik, TU München.

[Seber (1977)] G.A.F.: Linear Regression Analysis. Wiley, New York.

[Myers (1990)] R.H. (1990): Classical and Modern Regression with Applications. Duxbury Press, Belmont, CA.

[Draper/Smith (1998)] Draper N.R. and Smith H. (1998): Applied Regression Analysis (3rd ed), Wiley, N.Y.

[Cook/Weisberg (1999)] Cook, R.D. and Weisberg S.D. (1999): Applied Regression Including Computing and Graphics, Wiley, N.Y.

[Kleinbaum/Kupper (1998)] Kleinbaum D.G. and Kupper L.L.: Applied Regression Analysis and Other Multivariate Methods, Duxbury Press, Belmont, CA.

[Toutenburg (1992)] H. : Lineare Modelle, Physica-Verlag, Heidelberg

[S-PLUS (2000)] S-PLUS 2000 - Guide to Statistics, Vol.1, MathSoft, Seattle