Studiengang Medical Biometry: Statistische Modellierung 2

Prof. Dr. Vanessa Didelez

BIPS, Universität Bremen

SoSe 2019

Vorüberlegung

Ziel:
Es gibt verschiedene Eigenschaften (i.d.R. genau zwei, 'Hypothesen') eines zufälligen Prozesses, über die Unsicherheit herrscht; Daten sollen dahin gehend untersucht werden, mit welcher der beiden Hypothesen sie eher übereinstimmen. Das Ergebnis wird manchmal zu einer Entscheidung zwischen den Hypothesen benutzt.

Beispiel: Bei einem Produktionsprozess werden Bauteile hergestellt, die 10cm lang sein sollen. Hypothesen sind (1) der Prozess ist unter Kontrolle und die Länge beträgt (im Durchschnitt?) 10cm; (2) der Prozess ist nicht unter Kontrolle.

Ausganssituation:
$X_1, ..., X_n$ i.i.d. mit Dichte $f_\vartheta(x)$, $\vartheta \in \Theta \subset \mathbb{R}^s$ beliebig.

Definition

(a) Eine Hypothese H ist eine Teilmenge des Parameterraums Θ mit $\vartheta \in H :\Leftrightarrow H$ trifft auf ϑ zu.

(b) H heißt einfach, falls $H = \{\vartheta'\}$, H heißt zusammengesetzt, falls sie nicht einelementig ist.

(c) Ein Testproblem besteht aus einer Nullhypothese H_0 und einer Alternativhypothese H_1 mit $H_0 \cap H_1 = \emptyset$, $H_0 \cup H_1 \subset \Theta$.

(d) Eine Entscheidungsfunktion für H_0 vs H_1 ist eine (messbare) Funktion $d : \mathbb{R}^n \to \{0, 1\}$ mit

$$d(\mathbf{x}) = 1 \Leftrightarrow H_0 \text{ wird zugunsten } H_1 \text{ abgelehnt.}$$

(e) Die Menge

$$C = \{\mathbf{x}; d(\mathbf{x}) = 1\}$$

heißt kritischer oder Ablehnungsbereich, sein Komplement

$$A = \{\mathbf{x}; d(\mathbf{x}) = 0\}$$

Annahmebereich.

Anmerkung

(a) Beispiele für zusamengesetzte Hypothesen:
- $H : \vartheta \leq \vartheta_0$ bzw. $H = \{\vartheta \in \Theta; \vartheta \leq \vartheta_0\}$, $\vartheta_0 \in \Theta$ fest
- $H : \vartheta \neq \vartheta_0$
- $H : \vartheta_1 \neq \vartheta_2$

(b) Im Allgemeinen gilt: $H_0 \dot{\cup} H_1 = \Theta$

(c) Ein Testproblem (H_0, H_1) zusammen mit d heißt statistischer Test. Hier ist die Entscheidung $D = d(\mathbf{X})$ als Funktion von $\mathbf{X}$ selbst zufällig!

(d) Fehlentscheidungen können immer auftreten. Die Qualität eines statistischen Tests bemisst sich im Wesentlichen daran, die Wahrscheinlichkeit für Fehlentscheidungen zu minimieren – das ist aber immer nur unter Annahmen möglich. Und solange Zufälligkeit im Spiel ist, können Fehler nie ausgeschlossen werden.

Definition: Fehler 1./2. Art

Falls $d(\mathbf{x}) = 1$, obwohl $\vartheta \in H_0$, spricht man von einem Fehler 1. Art (falsch-positive Entscheidung);
falls $d(\mathbf{x}) = 0$, obwohl $\vartheta \in H_1$, spricht man von einem Fehler 2. Art (falsch-negative Entscheidung).

Zu Illustration: Mögliche Entscheidungen eines statistischen Tests

Testentscheidung	H_0 wahr	H_1 wahr
H_0 wird nicht abgelehnt: $D = 0$	richtig	falsch-negative Entscheidung: Fehler 2. Art (β-Fehler)
H_0 wird abgelehnt: $D = 1$	falsch-positive Entscheidung: Fehler 1. Art (α-Fehler)	richtig

Definition: Gütefunktion

Zur Beurteilung der Qualität statistischer Tests werden Kriterien benötigt wie etwa die Wahrscheinlichkeit für den Fehler 1. Art bzw. 2. Art. Allgemein betrachtet man die sogenannte Gütefunktion.

Die Gütefunktion $g_d(\vartheta)$ eines Tests d ist die Wahrscheinlichkeit für die Ablehnung von H_0, d.h.

$$g_d(\vartheta) : \Theta \to [0,1] \text{ mit } g_d(\vartheta) := P_\vartheta\{D = 1\} = E_\vartheta\{d(\mathbf{X})\} \quad (0.1)$$

Die Gütefunktion stellt die Wahrscheinlichkeiten für Fehler 1. Art und 2. Art dar.

Definition: Fehlerrisiko 1. Art

Das Fehlerrisiko 1. Art $\alpha(\vartheta)$ bezeichnet die Wahrscheinlichkeit für den Fehler 1. Art mit

$$\alpha(\vartheta) := g_d(\vartheta), \vartheta \in H_0. \tag{0.2}$$

Das maximale Fehlerrisiko 1. Art α ist definiert als

$$\alpha := \max\{g_d(\vartheta); \vartheta \in H_0\}. \tag{0.3}$$

α heißt auch (Signifikanz-)Niveau des statistischen Tests.

Definition: Fehlerrisiko 2. Art

Das Fehlerrisiko 2. Art $\beta(\vartheta)$ ist entsprechend definiert als

$$\beta(\vartheta) := 1 - g_d(\vartheta), \vartheta \in H_1. \tag{0.4}$$

Man bezeichnet $1 - \beta(\vartheta)$, $\vartheta \in H_1$, als Power (Macht) des Tests.

Das maximale Fehlerrisiko 2. Art β ist definiert als

$$\beta := \max\{\beta(\vartheta); \vartheta \in H_1\} = 1 - \min\{g_d(\vartheta); \vartheta \in H_1\}. \tag{0.5}$$

Wünschenswert: Ein statistischer Test, bei dem beide maximalen Fehlerrisiken minimiert werden.

Problem: Dies ist i.A. nicht möglich, da die Verringerung von α i.d.R. eine Vergrößerung von β bewirkt und umgekehrt.

Traditionelle Vorgehensweise: Gebe Siknifikanzniveau α vor, meist $\alpha = 1\%$, 5% oder 10%, und wähle unter allen Test denjenigen mit dem geringsten β, d.h. dem Fehler 1. Art wird größere Bedeutung zugemessen.
Diese unsymmetrische Behandlung von Fehler 1. und 2. Art, muss bei der Formulierung von H_0 und H_1 bedacht werden.

Beispiel:
Ist es 'schlimmer', wenn ein nicht-effektives Medikament für den Markt zugelassen wird, oder wenn ein wirksames Medikament nicht zugelassen wird?

Beachte: nicht nur die Wahl des Tests, sondern auch der Stichprobenumfang wirken sich auf β aus — bei allen 'sinnvollen' Tests, wird β (für festes ϑ) kleiner je größer n.

Die Entscheidung über H_0 wird anhand von Daten $\mathbf{x}$ getroffen, indem diese Information in eine sogenannte Teststatistik $T : \mathbb{R}^n \to \mathbb{R}$ mit $t = T(\mathbf{x})$ komprimiert wird.
Oft: überschreitet die Teststatistik einen vorgegebenen Wert t_0 (kritischer Wert) wird H_0 abgelehnt, d.h.:

$$D = 1 \Leftrightarrow T(\mathbf{x}) > t_0. \tag{0.6}$$

Damit ist der kritische Bereich gegeben als

$$C = \{\mathbf{x}; T(\mathbf{x}) > t_0\}. \tag{0.7}$$

Mächtigste Tests

Wie bereits erwähnt, dient die Gütefunktion zur Festlegung von Gütekriterien für Tests.

Frage: Wann ist ein Test besser als ein anderer?

Beispiel: t-Test im Vergleich zu Wilcoxon-Vorzeichen-Rang-Test.

Definition: Bester Test

Gegeben sei ein einfaches Testproblem mit $H_0 = \{\vartheta_0\}$ vs $H_1 = \{\vartheta_1\}$, $\vartheta_0 \neq \vartheta_1 \in \Theta$.

Ein Test d heißt mächstigster bzw. bester (engl.: most powerful) Test zum Niveau α, falls

$$g_d(\vartheta_0) = \alpha \text{ (Niveau } \alpha\text{-Test)}$$

und jeder andere Test d^* zum Niveau $\leq \alpha$ keine größere Macht auf H_1 besitzt, d.h.

$$g_{d^*}(\vartheta_0) \leq \alpha \quad \Rightarrow \quad g_{d^*}(\vartheta_1) \leq g_d(\vartheta_1).$$

Definition: Gleichmäßig bester Test

Gegeben sei ein Testproblem mit H_0 vs H_1, H_0, H_1 beliebig, dann heißt ein Test d gleichmäßig bester (engl.: uniformly most powerful, UMP) Test zum Niveau α, wenn es ein $\vartheta_0 \in H_0$ gibt, mit

$$\alpha = g_d(\vartheta_0) = \max\{g_d(\vartheta); \vartheta \in H_0\}$$

und für jedes $\vartheta_1 \in H_1$ der Test d ein bester Test zum Niveau α für die einfachen Hypothesen $H_0' = \{\vartheta_0\}$ vs $H_1' = \{\vartheta_1\}$ ist, d.h. für jeden anderen Test d^* gilt:

$$g_{d^*}(\vartheta_0) \leq \alpha \quad \Rightarrow \quad g_{d^*}(\vartheta_1) \leq g_d(\vartheta_1) \;\; \forall \vartheta_1 \in H_1.$$

Unverfälschtheit

Eine weitere Eigenschaft eines Test betrifft seine Unverfälschtheit, vergleichbar zur Unverzerrheit eines Schätzers. Bei einem Test bedeutet diese Eigenschaft, dass die Wahrscheinlichkeit dafür, H_0 abzulehnen, für $\vartheta \in H_1$ größer ist als für $\vartheta \in H_0$.

Definition: Unverfälschtheit

Gegeben sei ein Testproblem mit H_0 vs H_1, mit H_0, H_1 beliebig, dann heißt ein Test d unverfälscht (engl. unbiased), falls

$$g_d(\vartheta_0) \leq g_d(\vartheta_1) \ \forall \vartheta_0 \in H_0, \vartheta_1 \in H_1.$$

Ist d ein Test zum Niveau α, so ist er unverfälscht, falls

$$g_d(\vartheta_1) \geq \alpha \ \forall \vartheta_1 \in H_1.$$

Ein Test, der gleichmäßig bester unverfälschter Test ist, heißt UMPU-Test. Er ist gleichmäßig bester unter allen unverfälschten.

Regressionsanalyse I

Modellierung von *asymmetrischen* stochastischen Zusammenhängen zwischen Y und x oder $(x_1, \ldots, x_p)$.

Z.B.:
1) Cholesterin in Abhängigkeit vom Alter.
2) Holzvolumen in Abh. von Durchmesser und Höhe eines Baums.
3) Blutdruck in Abhängigkeit von Dosierung eines Medikaments.
4) Online-Kaufverhalten in Abhängigkeit von geschalteter Werbung.
5) Zähne mit Karies in Abh. von Putzhäufigkeit.

Regressionsanalyse II

Asymmetrisch?

- x kann zeitlich vor Y erhoben werden
- x kann einfacher / billiger erhoben werden
- x ist möglicherw. ursächlich (kausal) für Y
- oder x ist in einem anderen Sinne Y vorgeordnet

Unterscheidungen I

Modellierung kann verschiedenen Zielen dienen:

Wissenschaftlich (scientific): 'wahre' Zusammenhänge erkennen, verstehen, ausnutzen, z.B. physikalische / biologische 'Gesetze'; kausale Beziehungen finden und quantifizieren etc.
Beispiele: Temperatur bei dem Produktionsprozess und Eigenschaften der produzierten Komponenten? Führt eine Zuckersteuer zu reduziertem Zuckerkonsum bei Kindern?

Technologisch (technological): gute Beschreibung vergangener Daten, gute Vorhersagen zukünftiger Daten, wobei irrelevant ist, ob das Modell 'richtig' ist.
Beispiel: Wettervorhersage (?), Sprach-/Bilderkennung (machin. Lernen).

$\Rightarrow$ Unterschied relevant bei Kriterien der Modellanpassung/Güte.

Unterscheidungen II

Experimentelle Daten: z.B. verschiedene Dosierungen werden an Gruppen vergleichbarer Patienten unter vergleichbaren Bedingungen verabreicht. Wenn 'Vergleichbarkeit' nicht kontrolliert werden kann, wird Zufall (Rondomisierung) benutzt, um systematische Unterschiede zu verhindern.

$\Rightarrow$ x-Werte i.d.R. fest vorgegeben.

Beobachtungsdaten: z.B. Befragungen nach life-style; Fall-Kontroll Studien oder Kohortenstudien; vorhandene Datenquellen/banken (Krankenkassendaten etc.).

$\Rightarrow$ x-Werte zufällig (ZV X).

$\Rightarrow$ Unterschied relevant bei Interpretation

Unterscheidungen III

(z.B. kausal oder prädiktiv).

Unterscheidungen IV

Kausale Zusammenhänge: können zur 'Steuerung' genutzt werden. Z.B. Zuckersteuer zur Steuerung des *durchschnittlichen* Zuckerkonsums; Aktivität zur Steuerung des Herz-Kreislauf-Erkrankungs*risikos* etc.

$\Rightarrow$ eher bei experimentellen Daten, sonst: gute Begründung & geeignete Methoden.

Assoziative / prädiktive Zusammenhänge: Für präzise Vorhersagen vor allem wichtig, gute Prädiktoren zu finden, egal ob diese kausal verantwortlich sind. Z.B. Postleitzahl des Wohnorts als Prädiktor für Kreditwürdigkeit.

Unterscheidungen V

Hoher Stichprobenumfang & niedrige Dimension: Es können komplexe Modelle benutzt werden, die oft eine hohe Modellanpassung erreichen; auch schwache 'Signale' können gefunden werden; es kann zwischen gut passenden und schlecht passenden Modellen unterschieden werden; der Datensatz kann in 'Trainingdaten' (zur Modellanpassung) und 'Testdaten' (zur Bewertung der Modellgüte) aufgeteilt werden.

Geringer Stichprobenumfang / hohe Dimension: Es können nur schlichte Modelle (wenig Parameter) benutzt werden; Modellwahl sollte durch inhaltliche Gesichtspunkte (Biologie, Physik etc.) begründet werden.

Standardmodell der linearen Einfachregression

Es gilt

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.1}$$

$Y_1, \ldots, Y_n$	beobachtbare metrische Zufallsvariablen,
$x_1, \ldots, x_n$	gegebene deterministische Werte oder Realisierungen einer metrischen Zufallsvariable X,
$\epsilon_1, \ldots, \epsilon_n$	unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $\mathbb{E}(\epsilon_i) = 0$ und $\mathbb{V}(\epsilon_i) = \sigma^2$. (2.2)

Die Regressionskoeffizienten α, β und die Varianz σ^2 sind unbekannte Parameter, die aus den Daten (y_i, x_i), $i = 1, \ldots, n$, zu schätzen sind.

Normalverteilungsannahme

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ bzw. } Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2), i = 1, \ldots, n. \quad (2.3)$$

Die Normalverteilungsannahme wird für die Konsistenz der KQ-Methode (s.u.) nicht benötigt; zum Herleiten von Konfidenzintervallen, Standardfehlern bzw. Tests wird eine Verteilungsannahme oder großes n benötigt.

Unter Normalverteilung ist die KQ-Methode äquivalent mit der ML-Schätzung.

Kleinste-Quadrate-Schätzer I

Für das Standardmodell der linearen Regression wird gewöhnlich die *KQ (Kleinste-Quadrate-) Methode* eingesetzt.
KQ-Prinzip: Bestimme die Schätzer $\widehat{\alpha}$ und $\widehat{\beta}$ für α und β so, dass

$$\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2 \to \min_{\alpha,\beta}$$

d.h. die Summe der quadratischen Abweichungen wird durch $\widehat{\alpha}$, $\widehat{\beta}$ minimiert.

Andere Ansätze?

Kleinste-Quadrate-Schätzer II

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{2.4}$$

mit $S_{xx} = \sum(x_i - \overline{x})^2$ und $S_{xy} = \sum(x_i - \overline{x})(Y_i - \overline{Y})$,

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x}, \tag{2.5}$$

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2, \tag{2.6}$$

Kleinste-Quadrate-Schätzer III

Seien

$$\begin{aligned} \widehat{Y}_i &= \widehat{\alpha} + \widehat{\beta} x_i \quad \text{gefittete Werte} \\ \widehat{\epsilon}_i &= Y_i - \widehat{Y}_i \quad \text{Residuen} \end{aligned} \tag{2.7}$$

$\widehat{\alpha}$, $\widehat{\beta}$ sind Lösungen der Normalengleichungen (2.8):

$$\begin{aligned} \sum \widehat{\epsilon}_i &= 0 \\ \sum \widehat{\epsilon}_i x_i &= 0 \end{aligned} \tag{2.8}$$

Kleinste-Quadrate-Schätzer IV

Es gilt:

$$\mathbb{E}(\widehat{\alpha}) = \alpha, \quad \mathbb{E}(\widehat{\beta}) = \beta, \quad \mathbb{E}(\widehat{\sigma}^2) = \sigma^2, \tag{2.9}$$

$$\mathbb{V}(\widehat{\alpha}) = \sigma^2_{\widehat{\alpha}} = \sigma^2 \frac{\sum x_i^2}{n\, S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right) \tag{2.10}$$

$$\mathbb{V}(\widehat{\beta}) = \sigma^2_{\widehat{\beta}} = \frac{\sigma^2}{S_{xx}}.$$

Somit sind $\widehat{\alpha}$, $\widehat{\beta}$ und $\widehat{\sigma}^2$ erwartungstreue Schätzer.

Kleinste-Quadrate-Schätzer V

- Gilt für $n \to \infty$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \to \infty,$$

so sind sie auch konsistent.

- Die Konsistenzbedingung $\sum(x_i - \overline{x})^2 \to \infty$ ist erfüllt, da mit Wahrscheinlichkeit 1 gilt:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \to \mathbb{V}(X) = \sigma_X^2.$$

Verteilungsaussagen über die Schätzer

Unter

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{bzw.} \quad Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2), \quad i = 1, \ldots, n,$$

erhält man wegen $\widehat{\alpha} = \sum a_i Y_i$, $\widehat{\beta} = \sum b_i Y_i$ mit

$$a_i = \frac{1}{n} - \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \overline{x}, \quad b_i = \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

sofort, dass

$$\widehat{\alpha} \sim \mathcal{N}(\alpha, \sigma^2_{\widehat{\alpha}}), \quad \widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2_{\widehat{\beta}}). \tag{2.11}$$

Außerdem gilt: $(\widehat{\alpha}, \widehat{\beta})$ und $\widehat{\sigma}^2$ sind stochastisch unabhängig.

Verteilung der standardisierten Schätzfunktionen

Unter der Normalverteilungsannahme gilt

$$\frac{\widehat{\alpha}-\alpha}{\widehat{\sigma}_{\widehat{\alpha}}} \sim t_{n-2}, \quad \frac{\widehat{\beta}-\beta}{\widehat{\sigma}_{\widehat{\beta}}} \sim t_{n-2} \tag{2.12}$$

mit

$$\widehat{\sigma}_{\widehat{\alpha}} = \widehat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}}, \quad \widehat{\sigma}_{\widehat{\beta}} = \frac{\widehat{\sigma}}{\sqrt{S_{xx}}}.$$

Konfidenzintervalle für α und β

$$\begin{aligned} \widehat{\alpha} &\pm \widehat{\sigma}_{\widehat{\alpha}} \cdot t_{n-2;1-\alpha/2}, \\ \widehat{\beta} &\pm \widehat{\sigma}_{\widehat{\beta}} \cdot t_{n-2;1-\alpha/2} \end{aligned} \tag{2.13}$$

Für $n > 30$: Quantile der t_{n-2}-Verteilung durch Quantile der $\mathcal{N}(0,1)$-Verteilung ersetzen.

Teststatistiken und Ablehnbereiche

$$T_{\alpha_0} = \frac{\widehat{\alpha} - \alpha_0}{\widehat{\sigma}_{\widehat{\alpha}}} \quad \text{bzw.} \quad T_{\beta_0} = \frac{\widehat{\beta} - \beta_0}{\widehat{\sigma}_{\widehat{\beta}}}$$

Ablehnbereiche zu den Hypothesen (a), (b), (c)

(a) $|T_{\alpha_0}| > t_{n-2;1-\alpha/2}$ bzw. $|T_{\beta_0}| > t_{n-2;1-\alpha/2}$

(b) $T_{\alpha_0} < -t_{n-2;1-\alpha}$ bzw. $T_{\beta_0} < -t_{n-2;1-\alpha}$

(c) $T_{\alpha_0} > t_{n-2;1-\alpha}$ bzw. $T_{\beta_0} > t_{n-2;1-\alpha}$

Für $n > 30$: Quantile der t_{n-2}-Verteilung durch Quantile der $\mathcal{N}(0, 1)$-Verteilung ersetzen.

Streuungszerlegung I

$$SST = SSM + SSE$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{2.14}$$

$SST \mathrel{\hat{=}}$ Sum of Squares Total

$SSM \mathrel{\hat{=}}$ Sum of Squares Model

$SSE \mathrel{\hat{=}}$ Sum of Squares Error

Streuungszerlegung II

R^2: *Bestimmtheitsmaß* bzw. Determinationskoeffizient

Maßzahl für die Güte der Modellanpassung, definiert als Anteil der Gesamtvarianz der y_i, der durch die Regression von X auf Y erklärt wird:

$$R^2 = \frac{SSM}{SST} = \frac{\sum(\widehat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{\sum(y_i - \overline{y})^2 - \sum(y_i - \widehat{y})^2}{\sum(y_i - \overline{y})^2}$$
$$= 1 - \frac{\sum(y_i - \widehat{y})^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{SSE}{SST}. \qquad (2.15)$$

Es gilt: $0 \leq R^2 \leq 1$.

Viele Aspekte der KQ-Methode sind sehr empfindlich gegenüber Ausreißern, so auch das Bestimmtheitsmaß R^2.

Streuungszerlegung III

Varianzanalysetabelle

Streuung	FG	mittlerer quadratischer Fehler	Prüfgröße
SSM	1	$MSM = SSM/1$	$F = MSM/MSE$
SSE	$n-2$	$MSE = SSE/(n-2)$	
SST	$n-1$		

Zur Prüfung von $H_0 : \beta = 0$ kann der F-Wert verwendet werden. Falls die Normalverteilungsannahme zutrifft, gilt unter H_0 $F \sim F_{1,n-2}$.

Streuungszerlegung IV

Außerdem gilt die Beziehung:

$$F = \frac{R^2}{1 - R^2}(n - 2)$$

Beziehung zwischen $R^2, \widehat{\rho}_{xy}$ und $\widehat{\beta}$

Sei $\widehat{\rho}_{xy}$ der Korrelationskoeffizient nach Bravais-Pearson.
Es gilt:

$$R^2 = \widehat{\rho}_{xy}^2 \tag{2.16}$$

$$\widehat{\beta} = \widehat{\rho}_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}} \tag{2.17}$$

Konfidenz- und Prognoseintervall

1. Konfidenzintervall für $\mu_i = \alpha + \beta x_i$

$$\widehat{\alpha} + \widehat{\beta} x_i \pm t_{n-2;1-\alpha/2} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}} \tag{2.18}$$

2. Betrachte weitere Beobachtung x_0 mit unbekanntem Y_0. Der Prognosewert für Y_0 ist

$$\widehat{Y}_0 = \widehat{\alpha} + \widehat{\beta} x_0$$

Prognoseintervall für Y_0

$$\widehat{Y}_0 \pm t_{n-2;1-\alpha/2} \cdot \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \tag{2.19}$$

Schärfebetrachtungen für Steigungsparameter I

Test: $H_0 : \beta \leq 0 \quad vs \quad H_1 : \beta > 0$
H_0 wird nicht abgelehnt, falls

$$\widehat{\beta}/\widehat{\sigma}_{\widehat{\beta}} \leq t_{n-2;\,1-\alpha} =: t_\alpha$$

entsprechend wird H_0 abgelehnt, falls

$$\widehat{\beta}/\widehat{\sigma}_{\widehat{\beta}} > t_\alpha.$$

Sei

$$\beta_{\beta_1} = \mathrm{P}_{\beta_1}(\text{„}H_0 \text{ nicht ablehnen“}),$$

i.e. β_{β_1} ist Fehler 2.Art unter Vorliegen von $\beta = \beta_1$, entsprechend ist

$$\text{Power}_{\beta_1} = 1 - \beta_{\beta_1} = \mathrm{P}_{\beta_1}(\text{„}H_0 \text{ ablehnen“}).$$

Schärfebetrachtungen für Steigungsparameter II

Damit ergibt sich

$$\begin{aligned}
\text{Power}_{\beta_1} &= 1 - \beta_{\beta_1} \\
&= \mathrm{P}_{\beta_1}(\widehat{\beta}/\widehat{\sigma}_\beta > t_\alpha) \\
&= \mathrm{P}_{\beta_1}\left(\frac{\widehat{\beta} - \beta_1}{\widehat{\sigma}_\beta} > t_\alpha - \frac{\beta_1}{\widehat{\sigma}_\beta}\right) \\
&\overset{*}{\approx} 1 - \Phi\left(t_\alpha - \frac{\beta_1}{\widehat{\sigma}_{\widehat{\beta}}}\right) = \Phi\left(\frac{\beta_1}{\widehat{\sigma}_{\widehat{\beta}}} - t_\alpha\right) \qquad (2.20)
\end{aligned}$$

mit $\widehat{\sigma}_{\widehat{\beta}} = \widehat{\sigma}/\sqrt{S_{xx}}$

* Hier wird die t-Verteilung durch die Normalverteilung approximiert (Grund: t-Verteilung hängt von gesuchter Fallzahl ab).

Schärfebetrachtungen für Steigungsparameter III

- erforderliche Quadratsumme S_{xx}, um Regressionskoeffizienten β_1 mit Power$_{\beta_1}$=$1 - \beta_{\beta_1}$ zu entdecken:

$$S_{xx} \overset{*}{\approx} \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\beta_1} \cdot \sigma \right)^2 \tag{2.21}$$

- nötiger Stichprobenumfang für obige Fragestellung:

$$n \geq 1 + \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\widehat{\sigma}_x \cdot \beta_1} \cdot \sigma \right)^2 \tag{2.22}$$

$$\text{mit } \widehat{\sigma}_x^2 = S_{xx}/(n-1)$$

* Hier werden Quantilwerte aus der Normalverteilung statt der t-Verteilung verwendet (Grund: t-Verteilung hängt von gesuchter Fallzahl ab).

Residualanalyse

Abweichungen von der Normalverteilung lassen sich anhand eines Normal-Quantil-Plots für die Residuen überprüfen. Dabei werden statt der $\widehat{\epsilon}_i$ oft sogenannte *standardisiert Residuen*

$$r_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}}}$$

verwendet. Man kann zeigen, dass sie den Eigenschaften der Fehler ϵ_i noch näher kommen als die Residuen $\widehat{\epsilon}_i$.

Inverse Vorhersage (Kalibrierung) - nicht prüfungsrelevant

- Ausgangspunkt ist das einfache lineare Modell

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$$
$$\Rightarrow \widehat{Y} = \widehat{\alpha} + \widehat{\beta} x$$

- Man beobachtet einen neuen Wert y_{neu}
- gesucht: x_{neu}, der Wert, der y_{neu} zugrunde liegt
- $\widehat{x}_{neu} = \frac{y_{neu} - \widehat{\alpha}}{\widehat{\beta}}$ für $\widehat{\beta} \neq 0$
- Approximatives Vorhersage-Intervall für x_{neu}

$$\widehat{x}_{neu} \pm t_{n-2;1-\alpha/2} \cdot \widehat{\sigma}_{\widehat{x}_{neu}}$$

$$\text{und} \quad \widehat{\sigma}^2_{\widehat{x}_{neu}} = \frac{\widehat{\sigma}^2}{\widehat{\beta}^2} \cdot \left[1 + \frac{1}{n} + \frac{(\widehat{x}_{neu} - \overline{x})^2}{S_{xx}} \right]$$

Standardmodell der multiplen linearen Regression

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n. \tag{3.1}$$

mit

$Y_1, \ldots, Y_n$	beobachtbare metrische Zufallsvariablen,
$x_{1j}, \ldots, x_{nj}$	deterministische Werte der Variablen X_j oder Realisierungen von Zufallsvariablen X_j, $j = 1, \ldots, p$.
$\epsilon_1, \ldots, \epsilon_n$	unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $\mathbb{E}(\epsilon_i) = 0$ und $\mathbb{V}(\epsilon_i) = \sigma^2$. (3.2)
$\beta_1, \ldots, \beta_p$	zu schätzende **partielle Regressionskoeffizienten**

Die Bemerkungen im Anschluss an das Standardmodell der einfachen linearen Regression bleiben, entsprechend modifiziert, gültig, beispielsweise gilt bei gegebenen Regressorwerten, dass $Y_1, \ldots, Y_n$ unabhängig sind und

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \quad \mathbb{V}(Y_i) = \sigma^2, \quad i = 1, \ldots, n.$$

Aus

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n,$$

folgt

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma^2), \\ \mu_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots, n. \end{aligned}$$

Multiple lineare Regression in Matrixnotation

Wir fassen die Zielvariablen Y_i und die Werte $x_{i1}, \ldots, x_{ip}$, $i = 1, \ldots, n$, in einem $n \times 1$-Vektor $\boldsymbol{Y}$ und in einer $n \times (p+1)$-Matrix $\boldsymbol{X}$ zusammen:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}.$$

Die Matrix $\boldsymbol{X}$ enthält in der ersten Spalte die Werte der künstlichen Variable $\boldsymbol{X}_0 \equiv 1$.

Definiere den $(p+1) \times 1$-Vektor $\boldsymbol{\beta}$ der Regressionskoeffizienten und den $n \times 1$-Vektor $\boldsymbol{\epsilon}$ der Fehlervariablen:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Grundmodell der multiplen linearen Regression in Matrixnotation:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{mit} \\ \mathbb{E}(\boldsymbol{\epsilon}) &= \boldsymbol{0}, \quad \mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}. \end{aligned} \tag{3.3}$$

Die folgenden dargestellten Schätz- und Testverfahren arbeiten dann besonders gut, wenn die Fehlervariablen und damit die Zielvariablen zumindest approximativ normalverteilt sind.

Wie im einfachen Fall, sind die Schätzer aber auch ohne NV erwartungstreu.

KQ-Methode

Bestimme $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ so, dass die Summe der quadratischen Abweichungen bezüglich $\beta_0, \beta_1, \ldots, \beta_p$ minimiert wird:

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \rightarrow \min_{\beta_0, \beta_1, \ldots, \beta_p}$$

Damit das Minimierungsproblem eine eindeutige Lösung $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ besitzt, muss gelten:

1. $n \geq p + 1$
   Besser noch: $n \gg p + 1$, damit Schätzfehler möglichst klein!
2. Die Spaltenvektoren $\boldsymbol{X}_j$ von $\boldsymbol{X}$, $j = 0, \ldots, p$, mit $\boldsymbol{X}_0 \equiv 1$ müssen linear unabhängig sein, d. h. es darf für kein $j = 0, \ldots, p$ gelten:

$$\boldsymbol{X}_j = \sum_{k \neq j} a_k \boldsymbol{X}_k + b$$

Das KQ-Prinzip

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \to \min_{\boldsymbol{\beta}},$$

Nullsetzen der ersten Ableitung nach $\boldsymbol{\beta}$ liefert das $(p+1)$-dimensionale Gleichungssystem der „Normalengleichungen“

$$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^T\widehat{\boldsymbol{\epsilon}} = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}.$$

Geometrische Interpretation: Der geschätzte Fehlervektor $\widehat{\boldsymbol{\epsilon}}$ ist orthogonal zu dem von den Spalten von $\boldsymbol{X}$ erzeugten Modellraum. Unter den getroffenen Annahmen ist die Matrix $\boldsymbol{X}^T\boldsymbol{X}$ invertierbar, sodass sich

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{3.4}$$

als KQ-Schätzer ergibt.

Produktsummen-Matrix

Die Matrix $\boldsymbol{X}^T\boldsymbol{X}$ heißt Produktsummen-Matrix.

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & \sum x_{i1} & \dots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \dots & \sum x_{i1}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ip} & \dots & \dots & \sum x_{ip}^2 \end{pmatrix}$$

Mit $\boldsymbol{x}_i^T = (1, x_{i1}, \dots, x_{ip})$ hat man auch die Darstellung:

$$\boldsymbol{X}^T\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T$$

Unter den Voraussetzungen gilt

- Der KQ-Schätzer ist **erwartungstreu:**

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{3.5}$$

- Für die Kovarianzmatrix von $\widehat{\boldsymbol{\beta}}$ gilt:

$$\Sigma_{\widehat{\boldsymbol{\beta}}} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{3.6}$$

- Unter Normalverteilungsannahmen gilt:

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \tag{3.7}$$

- wobei

$$\widehat{\sigma}^2 := \frac{1}{n-(p+1)}\sum(Y_i - \widehat{Y}_i)^2; \quad \mathbb{E}(\widehat{\sigma}^2) = \sigma^2. \tag{3.8}$$

Vergleich: Interpretation einfach- vs. multiple Regression I

Einfachregresion:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$$

Hierbei ist α der Erwartungswert für Y_i für ein i mit $x_i = 0$.
Beachte: ist $x_i = 0$ sinnvoll gewählt? (evtl. umskalieren.)

Wenn die x_i **experimentell gesetzt** und andere Einflussgrößen experimentell kontrolliert wurden (bzw. x_i randomisiert), dann ist β die erwartete (mittlere) Veränderung in Y, wenn x um eine Einheit vergrößert wird. Beachte: ist 'eine Einheit' sinnvoll gewählt?

Beachte auch: β ist dann der **Gesamteffekt** von x, der möglicherweise durch andere Faktoren vermittelt wird.

Vergleich: Interpretation einfach- vs. multiple Regression II

Wenn die x_i **nur beobachtet** wurden, dann beschreibt β den zu erwartenden (mittlere) Unterschied zwischen Y_i und Y_j für den Fall, dass x_i eine Einheit größer ist als x_j.

Da die Objekte / Personen i und j aufgrund fehlender experimentelle Kontrolle nicht unbedingt in aller anderer Hinsicht vergleichbar sind, kann man diesen Unterschied i.d.R. nicht kausal auf den Unterschied in der X Variable zurückführen;

aber X ist potenziell immer noch wichtig / nützlich als Prädiktor für Y.

Vergleich: Interpretation einfach- vs. multiple Regression III

Multiple Regression

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n.$$

Hierbei ist β_0 der Erwartungswert für Y_i für den Fall, dass **alle** $x_{i1} = 0, \ldots, x_{ip} = 0$.

Seien alle $x_{i1}, \ldots, x_{ip}$ **experimentell** gesetzt (oder randomisiert) worden, und zwar so, dass sie **unkorreliert** sind (orthogonale Spalten der Designmatrix (mehr: später)), und andere Einflußfaktoren wurden kontrolliert. Dann sind die β_k, $k = 1, \ldots, p$, einzeln fast genauso zu interpretieren wie bei der Einfachregression im experimentellen Fall, aber als **direkte** (nicht durch die anderen $x_{il}, l \neq k$, vermittelte) Effekte.

Vergleich: Interpretation einfach- vs. multiple Regression IV

Seien alle $x_{i1}, \ldots, x_{ip}$ **nur beobachtet** worden.

Dann ist ein β_k, $k = 1, \ldots, p$, der zu erwartende (mittlere) Unterschied zwischen Y_i und Y_j für den Fall, dass x_{ik} eine Einheit größer ist als x_{jk} und **alle anderen** $x_{il} = x_{jl}, l \neq k$.

Mit demselben Grund wie oben, ist β_k typischerweise nicht kausal zu interpretieren. Manchmal kann man plausibel machen, dass die anderen Variablen $\boldsymbol{X}_l, l \neq k$, ausreichen, um alle Störfaktoren ('confounder') darzustellen; dann wäre β_k als **direkter** Effekt zu interpretieren.

Illustration: an der Tafel.

Vergleich: Interpretation einfach- vs. multiple Regression V

Vorhersage / Prädiktion: in manchen Anwendungen ist man nicht an der Interpretation der β_k's selbst interessiert, sondern nur daran, ob die Menge der $X_1, \ldots, X_p$ zusammen gute Vorhersagen für Y liefern.

$\Rightarrow$ Maße für Vorhersagegüte — später.

In **komplexen Modellen:** mit vielen nicht-Linearitäten / Interaktionen wird es auch schwieriger, jeden Koeffizienten einzeln zu interpretieren; stattdessen könnte man für 'typische Probanden' den vorhergesagten Wert $\hat{y}$ berechnen und vergleichen, z.B. Mann versus Frau beide im Durchschnittsalter mit Durchschnittsgewicht.

Hut-Matrix und Residual-Matrix I

$\boldsymbol{X}$ habe vollen Rang. Definiere

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\boldsymbol{P} := \underbrace{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T}_{n\times n} \qquad (\text{„Hut-Matrix“})$$

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

$$= \underbrace{(\boldsymbol{I}_n - \boldsymbol{P})}_{n\times n}\boldsymbol{Y}$$

$$\boldsymbol{Q} := \boldsymbol{I}_n - \boldsymbol{P} \qquad (\text{Residual-Matrix})$$

Hut-Matrix und Residual-Matrix II

$\boldsymbol{P}$ und $\boldsymbol{Q}$ sind zueinander orthogonale Projektionsmatrizen:

$$\boldsymbol{P}^T = \boldsymbol{P}, \quad \boldsymbol{P}^2 = \boldsymbol{P}$$
$$\boldsymbol{Q}^T = \boldsymbol{Q}, \quad \boldsymbol{Q}^2 = \boldsymbol{Q}$$
$$\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{P} = \boldsymbol{0}$$

Für die Kovarianz-Matrizen gilt:

$$\mathbb{V}(\widehat{\boldsymbol{Y}}) = \sigma^2 \boldsymbol{P}$$
$$\mathbb{V}(\widehat{\boldsymbol{\epsilon}}) = \sigma^2 \boldsymbol{Q}$$

Insbesondere sind die Komponenten $\widehat{\epsilon}_i$ des Residuenvektors nicht unabhängig und nicht identisch verteilt. Des Weiteren sind $\widehat{\boldsymbol{Y}}$ und $\widehat{\boldsymbol{\epsilon}}$ unkorreliert:

$$\mathrm{Cov}(\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{\epsilon}}) = \boldsymbol{0}.$$

Hut-Matrix und Residual-Matrix III

Projektionen in Worten:

Modellraum (= Spaltenraum von $\boldsymbol{X}$): kann durch die Parameter des Modells vollständig beschrieben werden, ist also $(p+1)$-dimensional; insbesondere ist $\widehat{\boldsymbol{Y}}$ im Modellraum; da $\widehat{\boldsymbol{Y}} = \boldsymbol{PY}$ kann man sagen, dass $\boldsymbol{P}$ den n-dimensionalen Datenvektor $\boldsymbol{Y}$ auf den $(p+1)$-dimensionalen Modellraum projiziert.

Residuenraum: da $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{QY}$ kann man $\boldsymbol{Q}$ analog als Projektion von $\boldsymbol{Y}$ auf den 'Rest', den $(n-(p+1))$-dimensionalen Residuenraum, interpretieren.

Streuungszerlegung I

Gegeben sei das lineare Modell (3.3) mit Design-Matrix $\boldsymbol{X}$, wobei $\text{Rang}(\boldsymbol{X}) = p + 1 =: p'$. Dann gilt:

$$\underbrace{(\boldsymbol{Y} - \overline{\boldsymbol{Y}})^T(\boldsymbol{Y} - \overline{\boldsymbol{Y}})}_{SST} = \underbrace{(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})}_{SSE} + \underbrace{(\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}})}_{SSM} \tag{3.9}$$

$SST \hat{=}$ Sum of Squares Total -(korrigierte) Gesamtquadratsumme

$SSM \hat{=}$ Sum of Squares Model -Modell-Quadratsumme

$SSE \hat{=}$ Sum of Squares Error -Fehler-Quadratsumme

Obige Zerlegung setzt ein Absolutglied in der Regression voraus.

Streuungszerlegung II

Folgende Zerlegung setzt nicht notwendig ein Absolutglied voraus (wird selten so gemacht – nicht Prüfungsrelevant):

$$\underbrace{\boldsymbol{Y}^T\boldsymbol{Y}}_{SST^*} = \underbrace{(\boldsymbol{Y}-\widehat{\boldsymbol{Y}})^T(\boldsymbol{Y}-\widehat{\boldsymbol{Y}})}_{SSE} + \underbrace{\widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}}}_{SSM^*}$$

SST^* – nicht korrigierte Gesamtquadratsumme

SSE – Fehler-Quadratsumme

SSM^* – nicht korrigierte Modellquadratsumme

Beachte: die beiden Streuungszerlegungen werden benötigt, um (wie im einfachen Fall) auch für das multiple Regressionsmodell einen F-Test herzuleiten, bzw. eine R^2 Statistik.

Erwartungswerte der Quadratsummen I

Betrachte wieder das Modell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$
$$\epsilon_i \quad iid, \quad \mathbb{E}(\epsilon_i) = 0, \quad \mathbb{V}(\epsilon_i) = \sigma^2$$
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{Rang}(\boldsymbol{X}) = p + 1 = p'$$

Sei $\boldsymbol{e} \in \mathbb{R}^n$, $\boldsymbol{e}^T = (1, \ldots, 1)$.
Dann bezeichne

$$\boldsymbol{P_e} := \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T = \boldsymbol{e}(\boldsymbol{e}^T\boldsymbol{e})^{-1}\boldsymbol{e}^T \qquad \text{(Projektionsoperator auf MW)}$$
$$\boldsymbol{Q_e} := \boldsymbol{I}_n - \boldsymbol{P_e} \qquad \text{(„Zentrier-“ Operator)}$$

Erwartungswerte der Quadratsummen II

Für die Erwartungswerte der Quadratsummen gilt:

$$\mathbb{E}(SST^*) = \mathbb{E}(\boldsymbol{Y}^T\boldsymbol{Y}) = n \cdot \sigma^2 + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

$$\begin{aligned}\mathbb{E}(SST) &= \mathbb{E}((\boldsymbol{Y} - \overline{\boldsymbol{Y}})^T(\boldsymbol{Y} - \overline{\boldsymbol{Y}})) \\ &= (n-1) \cdot \sigma^2 + \boldsymbol{\beta}^T(\boldsymbol{Q_e}\boldsymbol{X})^T(\boldsymbol{Q_e}\boldsymbol{X})\boldsymbol{\beta}\end{aligned}$$

$$\mathbb{E}(SSE) = \mathbb{E}(\widehat{\boldsymbol{\epsilon}}^T\widehat{\boldsymbol{\epsilon}}) = (n - p') \cdot \sigma^2$$

$$\mathbb{E}(SSM^*) = \mathbb{E}(\widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}}) = p' \cdot \sigma^2 + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

$$\begin{aligned}\mathbb{E}(SSM) &= \mathbb{E}((\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}})) \\ &= p \cdot \sigma^2 + \boldsymbol{\beta}^T(\boldsymbol{Q_e}\boldsymbol{X})^T(\boldsymbol{Q_e}\boldsymbol{X})\boldsymbol{\beta}\end{aligned}$$

Grundlegende Statistiken unter Normalverteilung I

Regressionsmodell $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$,
$\mathrm{Rang}(\boldsymbol{X}) = p + 1 =: p'$.

Seien $\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2$ KQ-Schätzer. Dann gilt:

a) $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$
$\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}} := \widehat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

b) $(n - p')\dfrac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p'}$

c) $\widehat{\sigma}^2$ und $\widehat{\boldsymbol{\beta}}$ sind stochastisch unabhängig.

Grundlegende Statistiken unter Normalverteilung II

d) $$\frac{\widehat{\beta}_k - \beta_k}{\widehat{\sigma}_k} \sim t_{n-p'},$$

wobei $\widehat{\sigma}_k = \widehat{\sigma}_{\widehat{\boldsymbol{\beta}}_k} = \sqrt{c_{kk}}\widehat{\sigma}$ mit

$$c_{kk} = \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{kk}$$

e) $$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p'\widehat{\sigma}^2} \sim F_{p',n-p'}$$

f) für eine Linearkombination $\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{\beta}$, wobei $\boldsymbol{B}$ eine $q \times p'$ – Matrix ist:

$$\frac{(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^T \boldsymbol{V}^{-1} (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})}{q \cdot \widehat{\sigma}^2} \sim F_{q,n-p'}$$

mit $\boldsymbol{V} = \boldsymbol{B}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{B}^T$

Vertrauens- und Prognose-Intervalle

g) für den E-Wert der i-ten Beobachtung gilt

$$\frac{\widehat{y}_i - \mathbb{E}(y_i)}{\widehat{\sigma}\sqrt{p_{ii}}} \sim t_{n-p'}, \quad \text{wobei } p_{ii} = \boldsymbol{P}_{ii}$$

h) für den E-Wert einer neuen Beobachtung bei beliebiger Versuchsbedingung $\boldsymbol{x}_0$:

$$\frac{\widehat{y}(\boldsymbol{x}_0) - \mathbb{E}(y(\boldsymbol{x}_0))}{\widehat{\sigma}\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-p'}$$

i) für eine neue Beobachtung bei beliebiger Versuchsbedingung $\boldsymbol{x}_0$ (Prognose):

$$\frac{y(\boldsymbol{x}_0) - \widehat{y}(\boldsymbol{x}_0)}{\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-p'}$$

Globale Tests über β I

Betrachte folgende Nullhypothesen:

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$
$$H_0^* : \beta_0 = \ldots = \beta_p = 0$$

Bezeichne

$$MSM = \frac{SSM}{p}$$
$$MSM^* = \frac{SSM^*}{p'}$$
$$MSE = \frac{SSE}{n - p'}$$

Globale Tests über $\boldsymbol{\beta}$ II

Prüfgrößen für obige Tests sind

$$F_0 = \frac{MSM}{MSE} \sim F_{p,n-p',\delta_0}$$

$$\text{mit } \delta_0 = \frac{1}{\sigma^2}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{Q_e} \boldsymbol{X}\boldsymbol{\beta}$$

$$\text{unter } H_0 : \delta_0 = 0$$

$$F_0^* = \frac{MSM^*}{MSE} \sim F_{p',n-p',\delta_0^*}$$

$$\text{mit } \delta_0^* = \frac{1}{\sigma^2}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta}$$

$$\text{unter } H_0^* : \delta_0^* = 0$$

Globale Tests über β III

ANOVA-Tabelle

	Quadrat-summe	FG	Quadrat-mittel	$\mathbb{E}$ (Quadratmittel)
Regression	SSM	p	SSM$/p$	$\sigma^2 + \frac{\|\mathbb{E}(\boldsymbol{Y}) - \mathbb{E}(\overline{\boldsymbol{Y}})\|^2}{p}$
Fehler	SSE	$n - p'$	SSE$/(n-p')$	σ^2
Total	SST	$n - 1$		

Bestimmtheitsmaß R^2

$$R^2 := \frac{\text{SSM}}{\text{SST}}$$

$R^2 = \rho^2_{Y|X}$ (multipler Korrelationskoeffizient)
Es gilt:

$$F := \frac{R^2}{1-R^2} \cdot \frac{(n-p')}{p} = \frac{SSM}{SSE} \cdot \frac{(n-p')}{p} \tag{3.10}$$

Unter $H_0 : \beta_1 = \ldots = \beta_p = 0$ ist F verteilt wie $F_{p,n-p'}$.
Lehne H_0 ab, falls $F > F_{p,n-p';1-\alpha}$.

Tests für lineare Hypothesen (geschachtelte Modelle) I

Betrachte Hypothesen, die sich via linearer Transformation des Parametervektors beschreiben lassen:

$$\boldsymbol{B} \in \mathbb{R}^{q \times p'}, \boldsymbol{b} \in \mathbb{R}^{q}$$
$$\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b} \text{ und } \operatorname{Rang}(\boldsymbol{B}) = q < p$$

'Geschachtelt': ein Untermodell $\mathcal{M}_0$ läßt sich aus dem Obermodell $\mathcal{M}$ erhalten, indem die Parameter von $\mathcal{M}$ eingeschränkt werden; $\mathcal{M}_0$ darf keine anderen / neuen Parameter enthalten.

Tests für lineare Hypothesen (geschachtelte Modelle) II

Beispiel:

$$\boldsymbol{B} = \begin{pmatrix} 1 & \dots & \dots & 0 & 0 & \dots & 0 \\ \vdots & 1 & & \vdots & \vdots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 1 & 0 & \dots & 0 \end{pmatrix}, \quad \boldsymbol{b} = \mathbf{0} \in \mathbb{R}^q$$

Hypothese: Die ersten q Regressionskoeffizienten sind $= 0$
Test: Die ersten q Variablen sind überflüssig

Tests für lineare Hypothesen (geschachtelte Modelle) III

$$\begin{aligned} \text{Modell:} \quad & \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n, \boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \text{Rang } \boldsymbol{X} = p' \\ \text{Zusätzlich:} \quad & \boldsymbol{B} \in \mathbb{R}^{q \times p'}, \boldsymbol{b} \in \mathbb{R}^q \\ \Rightarrow \quad & \boldsymbol{B}\widehat{\boldsymbol{\beta}} - \boldsymbol{b} \sim \mathcal{N}_q(\boldsymbol{B}\boldsymbol{\beta} - \boldsymbol{b}, \boldsymbol{B}\Sigma_{\widehat{\boldsymbol{\beta}}}\boldsymbol{B}^T) \end{aligned}$$

- Nullhypothese H_0: $\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}$

Tests für lineare Hypothesen (geschachtelte Modelle) IV

$$SSH = (\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_0)^T(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_0)$$

ist Quadratsumme, die die Abweichung der Schätzung unter dem vollen Modell von der unter $H_0 : \boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}$ beschreibt, algebraisch:

$$\text{SSH} := (\boldsymbol{B}\widehat{\boldsymbol{\beta}} - \boldsymbol{b})^T[\boldsymbol{B}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{B}^T]^{-1}((\boldsymbol{B}\widehat{\boldsymbol{\beta}} - \boldsymbol{b}))$$

Tests für lineare Hypothesen (geschachtelte Modelle) V

- $\frac{\text{SSH}}{\sigma^2} \sim \chi^2_{q,\delta_H}$
  mit $\delta_H = \sigma^{-2} \cdot (\boldsymbol{B\beta} - \boldsymbol{b})^T [\boldsymbol{B}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{B}^T]^{-1}(\boldsymbol{B\beta} - \boldsymbol{b})$
- $\text{MSH} := \frac{\text{SSH}}{q}$
- $\frac{\text{MSH}}{\text{MSE}} \sim F_{q,n-p';\delta_H}$
- Wald-Test der Hypothese H_0: $\boldsymbol{B\beta} = \boldsymbol{b}$:
  Lehne H_0 ab, falls

$$\frac{\text{MSH}}{\text{MSE}} > F_{q,n-p';1-\alpha} \tag{3.11}$$

Tests für lineare Hypothesen (geschachtelte Modelle) VI

- Die globalen Tests auf

$$H_0: \quad \beta_1 = \ldots = \beta_p = 0$$
$$H_0*: \quad \beta_0 = \beta_1 = \ldots = \beta_p = 0$$

sind Wald-Tests.

Tests für lineare Hypothesen (geschachtelte Modelle) VII

Der KQ-Schätzer $\widehat{\boldsymbol{\beta}}_0$ unter der Nebenbedingung $\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}$ ist gleich

$$\widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{B}[\boldsymbol{B}(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^{-1}(\boldsymbol{B}\widehat{\boldsymbol{\beta}} - \boldsymbol{b}).$$

Ferner gilt

$$\underbrace{(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_0)^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_0)}_{\text{SSE}_0} = \underbrace{(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})}_{\text{SSE}} + \underbrace{(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_0)^T(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_0)}_{\text{SSH}},$$

also

$$\text{SSH} = \text{SSE}_0 - \text{SSE}. \tag{3.12}$$

Partielle Quadratsummen

- Betrachte Teilmodelle, die durch Nullrestriktionen von Komponenten von $\boldsymbol{\beta}$ entstehen.
- $\mathcal{M}$: volles Modell $\quad (\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p))$
- $\mathcal{M}_0$: Teilmodell $\quad (\boldsymbol{\beta}^T = (\underbrace{0, \ldots, 0}_{J}, \beta_{J+1}, \ldots, \beta_p))$

$$R(\beta_1, \ldots, \beta_J \mid \beta_{J+1}, \ldots, \beta_p) := \text{SSH} = \text{SSE}(M_0) - \text{SSE}(M)$$

(Hypothesen-Quadratsumme zur Hypothese $\beta_1 = \ldots = \beta_J = 0$ in Modell M.)

Speziell heißen die zur Hypothese $\beta_j = 0$ gehörigen Quadratsummen des Gesamtmodells partielle Quadratsummen:

$$R(\beta_j \mid \beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p) = \text{SSE}(M_{-j}) - \text{SSE}$$

M_{-j} : Modell mit $\beta_j = 0$.

Simultane Konfidenzintervalle I

'Simultan': wollen für alle/mehrere β's 'gleichzeitig' KIs haben – was heißt das? W'keit, dass eins oder mehr KIs den jeweiligen Parameter nicht enthalten soll begrenzt werden.

Voraussetzung: Multiples lineares Modell mit NV-Annahme.

Simultane Konfidenzintervalle II

- Bonferroni-Konfidenzintervalle
  Für die K Parameter $\beta_{j1}, \ldots, \beta_{jK}$ sind

$$\widehat{\beta}_{jk} \pm \widehat{\sigma}_{\widehat{\beta}_{jk}} \cdot t_{n-p';1-\alpha/(2K)}, \quad k = 1, \ldots, K, \tag{3.13}$$

  simultane Konfidenzintervalle zum Niveau $(1-\alpha)$:

$$\mathrm{P}\{\text{es gibt } k \mid |\beta_{jk} - \widehat{\beta}_{jk}| > \widehat{\sigma}_{\widehat{\beta}_{jk}} \cdot t_{n-p';1-\alpha/(2K)}\} \leq \alpha.$$

  Man kann die Parameter durch Linearkombinationen $\theta_k = \boldsymbol{b}_k^T \boldsymbol{\beta}$ mit den entsprechenden Standardabw. ersetzen.

Simultane Konfidenzintervalle III

- Konfidenzintervalle nach Scheffé
  Es sind

$$\widehat{\beta}_j \pm \widehat{\sigma}_{\widehat{\beta}_j} \cdot \sqrt{p' F_{p',n-p';1-\alpha}} \tag{3.14}$$

und

$$\widehat{\theta} \pm \widehat{\sigma}_{\widehat{\theta}} \cdot \sqrt{p' F_{p',n-p';1-\alpha}} \tag{3.15}$$

simultane Konfidenzintervalle für alle Parameter β_j und beliebige Linearkombinationen $\theta = \boldsymbol{b}^T \boldsymbol{\beta}$.

Simultane Konfidenzintervalle IV

Beispiel (Konfidenzband für Regressionsgerade)

- Für einzelne Werte $x \in \mathbb{R}$ ist $\mathbb{E}(Y \mid X = x) = \alpha + \beta x$ mit Wahrscheinlichkeit $(1 - \alpha)$ im Intervall

$$\widehat{y} \pm \widehat{\sigma}_{\widehat{y}(x)} \cdot t_{n-2;1-\alpha/2}$$

  enthalten.

- Simultan für alle $x \in \mathbb{R}$ ist $\alpha + \beta x$ mit Wahrscheinlichkeit $(1 - \alpha)$ im Bereich

$$\widehat{y}(x) \pm \widehat{\sigma}_{\widehat{y}(x)} \cdot \sqrt{2F_{2,n-2;1-\alpha}}$$

  enthalten.

Simultane Konfidenzintervalle V

- Konfidenz-Ellipsoide
  Es ist

$$\{\boldsymbol{\beta} \mid (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq p' \widehat{\sigma}^2 F_{p',n-p',1-\alpha}\} \qquad (3.16)$$

  ein Konfidenz-Ellipsoid von $\boldsymbol{\beta}$ zum Niveau $1-\alpha$.
  Für lineare Transformationen $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\beta}$ mit $\text{Rang}(\boldsymbol{B}) = q$ gilt:

$$\{\boldsymbol{\theta} \mid (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T M_{\widehat{\boldsymbol{\theta}}}^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \leq q \widehat{\sigma}^2 F_{q,n-p';1-\alpha}\}$$

  ist ein $(1-\alpha)$ Konfidenz-Ellipsoid für $\boldsymbol{\theta}$,
  $M_{\widehat{\boldsymbol{\theta}}} = \boldsymbol{B}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{B}^T$.

Gauß-Markov-Theorem

Sei das Modell

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{Rang}(\boldsymbol{X}) = p + 1 = p'$$
$$\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad \mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \cdot \boldsymbol{I}_n \quad \text{gegeben}$$

Dann hat der KQ-Schätzer $\widehat{\boldsymbol{\beta}}$ die kleinste Varianz unter allen erwartungstreuen, linearen Schätzern von $\boldsymbol{\beta}$
($\widehat{\boldsymbol{\beta}}$ ist BLUE: best linear unbiased estimate).
Ist $\widetilde{\boldsymbol{\beta}}$ ein weiterer unverzerrter, linearer Schätzer, so gilt

$$\mathbb{V}(\widetilde{\boldsymbol{\beta}}) \geq \mathbb{V}(\widehat{\boldsymbol{\beta}}).$$

Asymptotische Normalität I

Sei das Modell $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ gegeben mit $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.
Wir betrachten Modelle mit steigendem Stichprobenumfang n (gegebene Folge $\boldsymbol{x}_n$ von Einflussgrößen). Zu jedem $n > p + 1$ sei

$\boldsymbol{X}_n$: Design-Matrix der ersten n Beobachtungen

$\widehat{\boldsymbol{\beta}}^{(n)}$: KQ-Schätzer der ersten n Beobachtungen

Sei $\mathrm{Rang}(\boldsymbol{X}_n) = p + 1$ für alle $n \geq p + 1$.

Asymptotische Normalität II

a) Konsistenz

$$\text{aus } \lim_{n\to\infty} (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} = 0$$

$$\Rightarrow \quad \widehat{\boldsymbol{\beta}}^{(n)} \xrightarrow[n\to\infty]{\mathrm{P}} \boldsymbol{\beta}$$

b) Asymptotische Normalität

Gilt die „Vernachlässigbarkeits-Bedingung“

$$\max_{i\leq n} \boldsymbol{x}_i^T (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{-1} \boldsymbol{x}_i \xrightarrow[n\to\infty]{} 0$$

$$\Rightarrow \quad (\boldsymbol{X}_n^T \boldsymbol{X}_n)^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{p+1})$$

Polynomiale Regression

Quadratische Regression

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{4.1}$$

ist unter anderem angebracht, wenn $\mathbb{E}(Y)$ ein Maximum oder Minimum im Wertebereich des Prediktors hat

$$\widehat{X}_M = \frac{-\widehat{\beta}_1}{2\widehat{\beta}_2}$$

Quadratische Regression kann gleichfalls angebracht sein, wenn die Mittelwertfunktion gekrümmt ist.

Polynomiale Regression

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d \qquad (4.2)$$

- jede glatte Funktion kann durch ein Polynom von genügend hohem Grad approximiert werden (i.A. repräsentiert eine solche Darstellung kein „physikalisches“ Modell)
- um numerische Instabilität zu vermeiden, sollte man die Kovariablen zentrieren $(\tilde{x}_k := (x_k - \overline{x}_k))$
- Alternative: orthogonale Polynome
- Ergänzung: Polynome mit rationalem Exponenten (fractional polynomials)

Polynome mit mehreren Faktoren

Beispiel mit zwei Prediktoren:

$$\begin{aligned}\mathbb{E}(Y \mid X_1 = x_1, X_2 = x_2) =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &+ \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2\end{aligned} \tag{4.3}$$

$x_1 \cdot x_2$ heißt Wechselwirkung (interaction).

Mit Interaktion (d.h. $\beta_{12} \neq 0$) hängt der prädiktive Effekt einer Kovariable auch vom Wert der anderen Kovariable ab!

Andere ("lokale") Parametrisierungen I

Bei einer metrischen Variable X kann man, wenn bekannte *"Bruchpunkte"* $c_1, \ldots, c_K$ vorliegen, u. U. folgende Parametrisierungen vornehmen:

- *Abschnittweise konstante Funktion*

$$X_k = \begin{cases} 1, & c_k \leq X < c_{k+1} \\ 0, & \text{sonst} \end{cases} \quad k = 1, \ldots, K-1,$$

zusätzlich $X_0 = I\{X < c_1\}$ und $X_K = I\{X \geq c_K\}$. Dies entspricht einer Kategorisierung von X. Die Regressionsgleichung wäre dann

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \sum_{k=1}^{K} \beta_k \cdot x_k.$$

Andere ("lokale") Parametrisierungen II

- *Abschnittweise lineare Funktion*
  Definiere Funktionen

$$f_k(x) = (x - c_k)_+ := \max(0, x - c_k), \quad k = 1, \ldots, K.$$

Dann beschreibt

$$\mathbb{E}(Y \mid X = x) = \alpha + \beta_0 \cdot x + \sum_{k=1}^{K} \beta_k \cdot f_k$$

eine stückweise lineare Funktion.

Andere ("lokale") Parametrisierungen III

- *Regressions-Spline*
  Definiere Funktionen

$$g_k(x) = (x - c_k)_+^3 := (\max(0, x - c_k))^3, \quad k = 1, \ldots, K.$$

Dann beschreibt

$$\mathbb{E}(Y \mid X = x) = \alpha + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3 + \sum_{k=1}^{K} \gamma_k \cdot g_k$$

einen Regressions-Spline 3.ten Grades.

Die Verwendung von Splines fällt in das Thema "Glättungsverfahren" und damit in das Kapitel *nichtparametrische Regression*.

Faktoren I

Faktoren erlauben das Einbeziehen qualitativer oder kategorialer Prediktoren in ein multiples Regressionsmodell. Betrachte eine diskrete Variable D mit J Ausprägungen.

- *Dummy-Kodierung*

$$D_j = \begin{cases} 1, & D = j \\ 0, & \text{sonst} \end{cases} \quad j = 1, \ldots, J$$

- *Effekt Kodierung*

$$E_j = \begin{cases} 1, & D = j \\ 0, & \text{für } D \neq j,\ D \neq J, \\ -1, & \text{für } D = J \end{cases} \quad j = 1, \ldots, J-1$$

Faktoren II

Beispiel: Sleep-Data: Zusammenhang von Schlafdauer und Gefährdungs-Index $D, D \in \{1, \ldots, d\}$ bei Säugetieren.

Definiere Dummy-Variable

$$D_j = \begin{cases} 1, & D = j \\ 0, & \text{sonst} \end{cases} \quad j = 1, \ldots, d$$

Regressionsmodelle:

$$\mathbb{E}(TS \mid D) = \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_d D_d \tag{4.4}$$

$$\mathbb{E}(TS \mid D) = \alpha + \beta_2 D_2 + \ldots + \beta_d D_d \tag{4.5}$$

Obacht! Die Regressionskoeffizienten β_j in (4.4) und (4.5) haben verschiedene Bedeutung! Die F-Tests aus den zwei Varianzanalyse-Tabellen testen verschiedene Hypothesen! In der eigentlichen Varianzanalyse wird wiederum anders parametrisiert.

Faktoren und stetige Regression

Vergleich von Regressionsgeraden
Beispiel: Sleep-Daten (Regressoren sind D und $\log$(Körpergewicht))

(M1) das „volle“ Modell

$$\mathbb{E}(TS \mid D = j, \log bw = x) = \sum_{j=1}^{d} (\beta_{0j} + \beta_{1j} x) \cdot D_j \qquad (4.6)$$

in anderer Parametrisierung:

$$\mathbb{E}(TS \mid D = j, \log bw = x) = \alpha + \beta x + \sum_{j \geq 2} (\beta_{0j} + \beta_{1j} x) \cdot D_j \qquad (4.7)$$

Dieses Modell hat insgesamt $2 \cdot d$ Parameter.

In Softwares werden solche Modelle realisiert durch Definition sogenannter Wechselwirkungen. Für die Steigungsparameter definiere

$$\log bw_j = \log bw \cdot D_j = \begin{cases} \log bw, & \text{falls } D = j \\ 0, & \text{sonst} \end{cases}$$

(M2) parallele Regression:

$$\beta_{11} = \ldots = \beta_{1d} \quad \text{in (4.6)}$$

Dieses Modell postuliert einen separaten Intercept, aber gleiche Steigung ($d+1$ Parameter).

(M3) gemeinsamer Intercept:

$$\beta_{01} = \ldots = \beta_{0d} \quad \text{in (4.6)}$$

Dieses Modell impliziert einen gemeinsamen Intercept, aber verschiedene Steigungsparameter ($d+1$ Parameter).

(M4) Übereinstimmende Regressionsgeraden

Dieses Modell ist

$$\mathbb{E}(TS \mid D = j, \log bw = x) = \alpha + \beta x, \tag{4.8}$$

postuliert also identische Parameter für alle Gruppen (2 Parameter).

$\Rightarrow$ Vergleich der Modelle (geschachtelte Hypothesen) via F-Tests, in diesem Beispiel:

	Modell	FG	SSE	F	$Pr(>F)$
M1:	allgemein	44	495,19	–	
M2:	parallel	48	510,27	0,33	0,853
M3:	gemeins. Intercept	48	728,35	5,18	0,002
M4:	eine Gerade	52	832,33	3,74	0,002

- $M2, M3, M4$ werden mit $M1$ verglichen.
- Schlussfolgerung?

- Die Teststatistiken F werden gebildet als

$$\begin{aligned} F_k &= \frac{(SSE_k - SSE_1)/(FG_k - FG_1)}{SSE_1/FG_1} \\ &= \frac{MSH_k}{MSE_1}, \quad k = 2, 3, 4, \end{aligned} \tag{4.9}$$

vergleiche mit den Formeln (3.12) und (3.11) aus dem dritten Kapitel.

Test auf Parallelität von zwei Regressionsgeraden

Alternative (Stratifizierung):
Seien $\widehat{\beta}_j, \widehat{\sigma}_j^2, n_j, S_{xx_j}$ die geschätzte Steigung, Fehlervarianz, Stichprobenumfang und korrigierte Regressor-Quadratsumme aus zwei Regressionsanalysen. Es ist

$$\mathbb{E}(Y \mid X = x, \text{ Gruppe } j) = \alpha_j + \beta_j x, \quad j = 1, 2.$$

Ein t-Test auf die Gleichheit $\beta_1 = \beta_2$ wird berechnet mit

$$T = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\sigma}(1/S_{xx_1} + 1/S_{xx_2})^{\frac{1}{2}}}$$

mit

$$\widehat{\sigma}^2 = \frac{(n_1 - 2)\widehat{\sigma}_1^2 + (n_2 - 2)\widehat{\sigma}_2^2}{n_1 + n_2 - 4}$$

T wird verglichen mit dem entsprechenden Quantil einer $t_{n_1+n_2-4}$-Verteilung.

Mehrere Faktoren

Die Analyse des Einflusses mehrerer diskreter Faktoren auf eine stetige Zielvariable Y ist Gegenstand der Varianzanalyse, die ihre eigene Nomenklatur hat.

Beispiel: 3-faktorieller Versuch mit Faktoren A, B und C mit jeweils drei Ausprägungen

$$Y_{ijkl} = \mu_{ijk} + \epsilon_{ijkl}, \quad i, j, k = 1, 2, 3;\ l = 1, \ldots, n_{ijk}$$

Verschiedene Möglichkeiten, μ_{ijk} zu modellieren:

- Haupteffekt-Modell
- Modelle mit Wechselwirkungen

Transformationen I

Der Hauptzweck von Transformationen ist, eine Funktion $\mathbb{E}(Y \mid X)$ zu erhalten, die auf der transformierten Skala Modellannahmen plausibler macht, z.B. Linearität.
Eine häufig verwandte Transformation T ist die logarithmische für positive Zufallsvariablen U:

- $T(U) = \log(U)$
- Logarithmische Transformation mit Nullpunkt-Erhaltung:

$$T(U) = \log(U + 1),$$

z.B. bei der Quantifizierung von Zigarettenrauchen:
$LogPY = \log(\text{Packyear} + 1)$.

Transformationen II

Allgemeiner definiert man

- *Power-Transformationen*
  Für Variable $U > 0$:

$$T(U, \lambda) = U^{\lambda} \tag{4.10}$$

heißt Powertransformation. Für $\lambda = 0$ setzt man

$$T(U, \lambda = 0) = \log(U).$$

In der Regel betrachtet man λ im Wertebereich $-1 \leq \lambda \leq +1$; andere Ansätze existieren.

Transformationen III

Zwei empirische Regeln

- Log-Regel: Falls der Wertebereich einer Variable mehr als eine Größenordnung umfasst und die Variable streng positiv ist, dann ist Logarithmieren wahrscheinlich hilfreich.

- Bereichs-Regel: Falls der Wertebereich einer Variable erheblich kleiner als eine Größenordnung ist, dann hilft wahrscheinlich keine Transformation dieser Variable.

Log-Transformation: ist oft hilfreich (linearisierend), wenn Y im weitesten Sinne 'Wachstum' darstellt; dabei wird auch die Varianz homogenisiert.

Box-Cox Transformation

Ziel der Box-Cox Transformation ist es, die abhängige Variable Y, die größer null sein muss, so zu transformieren, dass deren Dichte *normalverteilt* ist.

Die Box-Cox Transformation

$$U_\lambda = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \text{für } \lambda \neq 0 \\ \log(Y) & \text{für } \lambda = 0 \end{cases}$$

(Für $\lambda \neq 0$ arbeitet man oft mit $U_\lambda = Y^\lambda$ statt mit $\frac{Y^\lambda - 1}{\lambda}$. Auch hier wird gewöhnlich $\lambda \in [-1, +1]$ gewählt.)

Wahl von λ (hier: ML-Schätzung; auch andere Methoden möglich.)

- für festes λ ist $U := Y^\lambda$ normalverteilt, also

$$f_U(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) \tag{4.11}$$

- nach dem Transformationssatz für Dichten hat Y dann die Dichte

$$\begin{aligned} f_Y(y) &= f_U(u) \cdot \left|\frac{du}{dy}\right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^\lambda-\mu)^2\right) \cdot y^{\lambda-1}|\lambda| \end{aligned}$$

- für n Beobachtungen ist dann die Likelihood gegeben als

$$L(Y;\mu,\sigma^2,\lambda) = \left(\frac{1}{\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum(y_i^\lambda - \mu_i)^2\right) \cdot |\lambda|^n \cdot \prod_1^n y_i^{\lambda-1}$$

- die Loglikelihood ist entsprechend

$$\begin{aligned} l = & -n\,\log\sigma - \frac{1}{2\sigma^2}\sum(y_i^\lambda - \mu_i)^2 \\ & + n\,\ln|\lambda| + (\lambda-1)\sum\log y_i \end{aligned} \tag{4.12}$$

- Maximieren von l in einem 2-Schrittverfahren
  1. fixiere λ und finde KQ-Schätzer $\widehat{\mu}_i$ und

$$\widehat{\sigma}^2_\lambda := \frac{1}{n}\sum_{1}^{n}(y_i^\lambda - \widehat{\mu}_i)^2$$

  2. Setze entsprechendes λ in (4.12) ein, d.h. berechne die sogen. Profile-Loglikelihood.
     $\Rightarrow$ finde optimales λ

Alternativ: betrachte einige diskrete Werte für λ und vergleiche Modellgüte der resultierenden Modelle.

Regressionsdiagnostik: Motivation

Sind die Modellannahmen plausibel? Könnte irgendwas die Ergebnisse verfälschen (Ausreißer)?

– inhaltliche / substanzwissenschaftliche Überlegungen

– statistische explorative Verfahren

– manche 'Tests' (diese besser auf einer separaten Teilstichprobe)

Ziel: grobe Verletzungen der Modellannahmen u.a. Probleme auszuschliessen.

Dann: Modell verbessern, z.B. durch Transformationen, Ausreißer eliminieren, robuste / andere Schätzverfahren verwenden, die weniger Annahmen benötigen (siehe später).

Residuenplots

Annahmen:

– $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{V}(\epsilon_i) = \sigma^2$ (homogene Varianz), unabhängig $\forall\, i$

– zusätzlich: Normalverteilung

Die Residuen $\widehat{\epsilon}_i = y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}$ des linearen Modells sind zwar weder unabhängig noch haben sie die gleiche Varianz (s.u.), aber als 'schnelle' Diagnostik werden sie trotzdem oft verwendet: man 'plotte'

- $\widehat{\epsilon}_i$ gegen $\widehat{y}_i$ (Tukey-Anscombe-Diagramm);
- $\widehat{\epsilon}_i$ gegen x_{ij} (separat für jede j-te Variable);
- evtl. $\widehat{\epsilon}_i$ gegen i, wenn i systematisch (z.B. Reihenfolge der Messungen);
- QQ-Plot: ungefähre Normalverteilung?

Besser: man verwende die *standardisierten* Residuen!

Für die Residuen $\widehat{\epsilon}_i$ des linearen Modells gilt (s. oben) mit $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$:

$$\widehat{\boldsymbol{\epsilon}} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \cdot (\boldsymbol{I} - \boldsymbol{P})).$$

Daher definiert man

- standardisierte Residuen

$$r_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}\sqrt{1 - p_{ii}}} \tag{5.1}$$

und verwende r_i statt $\widehat{\epsilon}_i$ in obigen Plots.
(Meistens sind die Unterschiede zwischen r_i und $\widehat{\epsilon}_i$ in der Praxis recht klein.)

Vorbemerkungen

Von *Kollinearität* spricht man, wenn die Spalten von $\boldsymbol{X}$ linear abhängig sind; in der Praxis auch wenn sie überhaupt korreliert sind.

Problem? Wenn zwei Spalten von $\boldsymbol{X}$ sehr 'ähnliche' Information enthalten, wird es schwierig, die jeweiligen Einflüsse auf Y auseinander zu halten; der KQ-Schätzer wird 'instabil' (s.u.).

Beachte (1) Wenn $\boldsymbol{X}$ durch das Design der Studie deterministisch festgelegt ist, kann und soll Kollinearität vermieden werden.
(2) Problem vor allem dann, wenn Einflussgrößen $X_1, \ldots, X_p$ selbst zufällig und mehr oder weniger korreliert sind.
In Epidemiologie: oft $X_1 =$ Exposition, die untersucht werden soll; $X_2, \ldots, X_p$ Kovariablen, die 'kontrolliert' werden sollen, *weil* sie mit X_1 korreliert sind, um kausale Interpretation von X_1 zu ermöglichen.
$\Rightarrow$ Inhaltliche Gründe für Modell wichtiger als statistische.

Zunächst: Interpretation; z.B. zwei Einflussgrößen X_1, X_2,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{5.2}$$

$$Y = \alpha_0 + \alpha_1 X_1 + u \tag{5.3}$$

Was sind α_0, α_1, u? Annahme: (5.2) korrekt, (X_1, X_2) bivariat normalverteilt mit $(\mu_1, \mu_2, \rho, \sigma_1^2, \sigma_2^2)$; unabh. von ϵ. Dann

$$\alpha_0 = \mathbb{E}(Y|X_1 = 0) = \beta_0 + \beta_2(\mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1)$$

$$\alpha_1 = \beta_1 + \beta_2\rho\frac{\sigma_2}{\sigma_1}$$

$$\mathbb{V}(u) = \mathbb{V}(Y|X_1 = x_1) = \beta_2^2\sigma_2^2(1-\rho^2) + \sigma_\epsilon^2$$

$\Rightarrow \alpha_1 = \beta_1$ identisch wenn $\beta_2 = 0$ oder $\rho = 0$;
sonst: Modell (5.3) zwar korrekt, α_1 andere Interpretation als β_1, andere (größere) Residuenvarianz. *('Pfadmodelle')*

Kollinearität – formal: Auswirkung auf KQ-Schätzer

Das lineare Modell (3.3) sei gegeben.

Zwei Variablen X_1, X_2 heißen kollinear (linear abhängig), wenn eine lineare Gleichung

$$c_1 X_1 + c_2 X_2 = c_0$$

mit Konstanten $c_1, c_2 \neq 0$ gilt.

Kollinearität zweier Faktoren wird gemessen durch r^2, den quadrierten Korrelationskoeffizienten:

$$r^2 = 1 \quad - \quad \text{exakte Kollinearität}$$
$$r^2 = 0 \quad - \quad \text{nicht kollinear}$$

Die Diskussion bezieht sich in der Regel auf approximative Kollinearität, d.h. $r^2 \approx 1$. Die Definition erweitert sich auf $p > 2$ Variablen: $X_1, \ldots, X_p$ heißen approximativ kollinear, falls es Zahlen $c_0, \ldots, c_p$ gibt mit

$$c_1 X_1 + \ldots + c_p X_p \approx c_0$$

mit mindestens zwei $c_j \neq 0$.

Ein einfaches diagnostisches Analogon zu r^2 ist der quadrierte multiple Korrelationskoeffizient zwischen X_j und den anderen Variablen, bezeichnet mit R_j^2.

Man kann zeigen: $R_j^2 =$ Bestimmtheitsmaß aus linearer Regression von X_j auf alle anderen Variablen.

Man kann zeigen, dass für die Varianz des j-ten Regressionskoeffizienten $\widehat{\beta}_j$ in Modell (3.3) gilt:

$$\mathbb{V}(\widehat{\beta}_j) = \frac{1}{1 - R_j^2} \cdot \frac{\sigma^2}{S_{x_j x_j}} \tag{5.4}$$

mit

$$S_{x_j x_j} = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2.$$

$VIF_j := \frac{1}{1 - R_j^2}$ heißt j-ter Varianz-Inflations Faktor.

$VIF_j^{-1} \approx$ "Faktor um den sich die Varianz von $\widehat{\beta}_j$ verkleinern würde, wäre X_j von allen anderen Einflussvariablen unabhängig."

Je größer VIF_j bzw. R_j^2, umso ungenauer ('unstabiler') wird β_j geschätzt.
Dagegen hängt die Genauigkeit, mit der $\widehat{\boldsymbol{Y}}$ geschätzt wird, wegen $\mathbb{V}(\widehat{\boldsymbol{Y}}) = \sigma^2 \boldsymbol{P}$, nur vom 'Modellraum', nicht von dessen Parametrisierung ab. (Allerdings enthält $\boldsymbol{P}$ ein Inverse, die bei nahezu exakter Kollinearität numerisch instabil wird.)

Große Wert von VIF_j können bedeuten, dass kleine Änderungen in den Daten zu stark veränderten Schätzern $\widehat{\beta}_j$ führen.

Wenn alle VIF_j groß sind, ist es möglich, dass kein einzelnes $\widehat{\beta}_j$ signifikant ist, auch wenn der F–Test signifikant ist.

Kollinearität – Was tun?

- Starke Kollinearität sollte immer angegeben werden.
- Falls kollineare Variablen ähnliche Phänomene messen (z.B. syst./diast. Blutdruck; zwei versch. aber ähnliche Intelligenztests etc.) sollten diese Variablen zusammengefaßt werden.
- Aus inhaltlichen Gründen kann es trotzdem nötig sein, kollineare Variablen im Modell zu behalten, z.B. Interpretation oder 'Kontrolle'.

Beobachtungen mit Hebelkraft

Beobachtungen, deren x-Werte „am Rande“ der Kovariablen-Menge liegen, können einen großen Einfluss auf das Ergebnis der Regressionsanalyse ausüben. Man spricht von „Beobachtungen mit Hebelkraft“. Dabei spielen die Diagonalelemente p_{ii} der Hut-Matrix $\boldsymbol{P}$ des linearen Modells $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ eine große Rolle, $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$

- $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$, insbesondere

$$\widehat{y}_i = \sum_{j=1}^{n} p_{ij}y_j = p_{ii}y_i + \sum_{j \neq i} p_{ij}y_j \tag{5.5}$$

$\Rightarrow p_{ii} =$ „Gewicht“, mit dem y_i in die Schätzung des eigenen Erwartungswertes eingeht

- $\widehat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon} \Rightarrow \widehat{\epsilon}_i = (1 - p_{ii})\epsilon_i - \sum_{j \neq i} p_{ij}\epsilon_j$

Hut-Matrix und Hebelkraft

Sei $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ die Hut-Matrix des linearen Modells (3.3) mit $\text{Rang}(\boldsymbol{P}) = p' = p + 1$. Dann gilt für die Diagonalelemente p_{ii} von $\boldsymbol{P}$:

(i) $0 \leq p_{ii} \leq 1$;

(ii) falls $\boldsymbol{X}$ einen konstanten Term enthält:

$$\frac{1}{n} \leq p_{ii};$$

(iii) $\text{Spur}(\boldsymbol{P}) = \sum_{i=1}^{n} p_{ii} = p'$.

p_{ii} heißen Hebel (leverage) der Beobachtung i.

- Faustregel: p_{ii} ist kritisch hoch wenn :

$$p_{ii} \geq \frac{2 \cdot p'}{n} \quad \text{oder} \quad p_{ii} \geq \frac{3 \cdot p'}{n}$$

Einflussreiche Beobachtungen - Influenz-Diagnostik

Ziel: Untersuchung der Robustheit der Regressionsanalyse gegenüber leichten „Störungen“ der Daten.

Definition: Man nennt eine Beobachtung einflussreich, wenn ihre Streichung aus den Daten zu „größeren Änderungen“ in der Analyse führt.

Dies kann eine Beobachtung mit großer Hebelkraft sein, wenn sie einen ‘ungewöhnlichen’ y-Wert aufweist. Es kann auch eine Beobachtung ohne Hebelkraft sein, die einen ‘ungewöhnlichen’ y-Wert aufweist. Insbesondere bei mehr als zwei Dimensionen lassen sich einflussreiche oder ungewöhnliche Beobachtungen nicht mehr leicht visuell identifizieren.

Notation für das 'Weglassen' einer Beobachtung:

$$\begin{aligned} \text{Modell:} \quad & \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \\ & \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \\ & \boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T, \quad \text{Rang}(\boldsymbol{X}) = p' = p + 1 \end{aligned}$$

In obigem Modell soll die i-te Beobachtung gestrichen werden, im so reduzierten Modell habe die resultierende Designmatrix auch Rang p'.

- Vektoren und Matrizen, in denen die i-te Zeile gestrichen ist, werden mit Index$_{(i)}$ gekennzeichnet, also $\boldsymbol{Y}_{(i)}$, $\boldsymbol{X}_{(i)}$.
- Kenngrößen des so reduzierten Modells, die ohne die i-te Beobachtung geschätzt werden, werden mit Index$_{(i)}$ bezeichnet, z.B. $\widehat{\sigma}^2_{(i)}$.

Satz 5.1

Bezeichnen $\boldsymbol{w}_i$ die i-te Spalte von $\boldsymbol{W} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\boldsymbol{p}_i$ die i-te Spalte von $\boldsymbol{P}$ und sei $\widehat{\boldsymbol{\mu}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)}$.

Dann gilt:

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{(i)} + \boldsymbol{w}_i \cdot \frac{\widehat{\epsilon}_i}{1 - p_{ii}}$$

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_{(i)} + \boldsymbol{p}_i \cdot \frac{\widehat{\epsilon}_i}{1 - p_{ii}}$$

insbesondere

$$\widehat{\mu}_i = \widehat{\mu}_{i(i)} + p_{ii} \cdot \frac{\widehat{\epsilon}_i}{1 - p_{ii}}$$

$$\boldsymbol{Y} = \widehat{\boldsymbol{\mu}}_{(i)} + \boldsymbol{p}_i \cdot \frac{\widehat{\epsilon}_i}{1 - p_{ii}} + \widehat{\boldsymbol{\epsilon}}$$

insbesondere

$$Y_i = \widehat{\mu}_{i(i)} + \frac{\widehat{\epsilon}_i}{1 - p_{ii}}$$

$$\|\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_{(i)}\|^2 = p_{ii} \cdot \left(\frac{\widehat{\epsilon}_i}{1 - p_{ii}}\right)^2$$

Für die Varianzen der Schätzer gilt:

$$SSE = SSE_{(i)} + \frac{\widehat{\epsilon}_i^2}{1 - p_{ii}}$$

$$\widehat{\sigma}^2 = \frac{n - p'}{n - p} \cdot \widehat{\sigma}^2_{(i)} + \frac{\widehat{\epsilon}_i^2}{(n - p)(1 - p_{ii})}$$

$$Cov(\widehat{\boldsymbol{\beta}}) = Cov(\widehat{\boldsymbol{\beta}}_{(i)}) - \frac{\boldsymbol{w}_i \boldsymbol{w}_i^T}{1 - p_{ii}} \cdot \sigma^2$$

insbesondere $$\mathbb{V}(\widehat{\beta}_j) = \mathbb{V}(\widehat{\beta}_{j(i)}) + \frac{w_{ji}^2}{1 - p_{ii}} \cdot \sigma^2$$

Studentisierte Residuen

- standardisierte Residuen r_i siehe (5.1).
- studentisierte Residuen

$$t_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}_{(i)}\sqrt{1-p_{ii}}} \tag{5.5}$$

($\widehat{\sigma}_{(i)}$ ist die Schätzung von σ ohne die i-te Beobachtung)

- Für die studentisierten Residuen gilt:
  $\widehat{\epsilon}_i$ und $\widehat{\sigma}_{(i)}$ sind stochastisch unabhängig, $t_i \sim t_{n-p'-1}$
- Umrechnungsformel:

$$t_i = r_i \cdot \left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{\frac{1}{2}}$$

Bemerkung: r_i und t_i werden manchmal (z.B. bei SAS) als studentisierte und R-studentisierte Residuen bezeichnet.

t_i als Prüfgröße eines Ausreißertests

Betrachte das folgende Modell M_1 mit einem 'Ausreißer' in der i-ten Beobachtung

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \delta + \epsilon_i \tag{5.6}$$

$$y_{i'} = \boldsymbol{x}_{i'}^T \boldsymbol{\beta} + \epsilon_{i'} \quad \text{für } i' \neq i \tag{5.7}$$

Die Nullhypothese 'kein Ausreißer' ist parametrisiert als $H_0 : \delta = 0$.

Test: Im Ausreißermodell (5.6) und (5.7) wird der „Sprungparameter“ δ geschätzt durch

$$\widehat{\delta} = \frac{\widehat{\epsilon}_i}{1 - p_{ii}}$$

H_0 wird getestet mit dem t-Test und der Prüfgröße

$$T = \frac{\widehat{\delta}}{\widehat{\sigma}_{\widehat{\delta}}} = t_i. \tag{5.8}$$

Cook's Distanz C_i

- Zusammenfassendes Maß (über alle Komponenten von $\boldsymbol{\beta}$)

$$C_i := \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p' \widehat{\sigma}^2}$$

$$= \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p' \widehat{\sigma}^2}$$

$$\text{mit } \widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)}) \boldsymbol{X}_{(i)}^T \boldsymbol{Y}_{(i)}$$

- Berechnung von C_i:

$$C_i = \frac{1}{p'} r_i^2 \frac{p_{ii}}{1 - p_{ii}}$$

- „Kritische Werte“ für C_i.

  Es wird vorgeschlagen, Quantile der $F_{p',n-p'}$-Verteilung als „Warnschranken“ für C_i zu benutzen (C_i ist allerdings nicht F-verteilt!).

- Faustregel:

  $C_i > 0.5 \rightarrow$ auffällig
  $C_i > 1 \rightarrow$ Beobachtung sollte näher untersucht werden

Bemerkung:

Obige Checks funktionieren unter der Annahme, dass Ausreisser isoliert auftreten; wenn eine mehrere Beobachtungen (auf die gleiche Art) 'seltsam' sind, kann dies durch das Streichen einer einzelnen Beobachtung nicht immer entdeckt werden ('Masking' Effekt).

Beispiel: Hirngewicht in Abhängigkeit vom Körpergewicht bei Tieren und Dinosauriern.

Alternative: man verwende *robuste* Regressionsmethoden statt KQ-Schätzung ($\rightarrow$ eigenes Thema.)

Korreliertheit der Residuen

Eine *Annahme* im linearen Modell ist die der unkorrelierten (unabhängigen) Residuen ϵ_i. Wenn dies nicht plausibel ist, kann man versuchen, die Anhängigkeit mit zu modellieren.

Autoregressiver Prozess 1. Ordnung

Seien ω_i eine Folge unabhängiger, $\mathcal{N}(0, \tau^2)$-verteilter ZVen, $|\rho| < 1$. Dann bezeichnet

$$\epsilon_i = \rho\epsilon_{i-1} + \omega_i, \quad i = 1, 2, \ldots$$

einen autoregressiven Prozess 1. Ordnung.
$\Rightarrow$ Plausibel, wenn $i =$ zeitliche Reihenfolge o.ä.

Für die Abhängigkeit der Residuen gilt dann:

$$\epsilon_1 = \omega_1$$
$$\epsilon_2 = \rho\epsilon_1 + \omega_2 = \rho\omega_1 + \omega_2$$
$$\epsilon_3 = \rho\epsilon_2 + \omega_3 = \rho^2\omega_1 + \rho\omega_2 + \omega_3$$
$$\vdots$$
$$\epsilon_i = \sum_{k=1}^{i} \rho^{i-k}\omega_k$$

Sei

$$\sigma^2 := \frac{\tau^2}{1-\rho^2}.$$

Es gilt:

$$\begin{aligned}
\mathrm{Cov}(\epsilon_i, \epsilon_{i+s}) &= \rho^s \tau^2 \cdot \frac{1-\rho^{2i}}{1-\rho^2} = \rho^s \sigma^2 \cdot \left(1-\rho^{2i}\right) \\
\mathbb{V}(\epsilon_i) &= \sigma^2 \\
\mathrm{Cov}(\epsilon_i, \epsilon_{i+s}) &= \rho^s \sigma^2 \qquad \text{(für große } i\text{)} \\
\rho(\epsilon_i, \epsilon_{i+s}) &= \rho^s \qquad \text{(für große } i\text{)}
\end{aligned}$$

Für einen Zeitabschnitt von n aufeinanderfolgenden Beobachtungen erhält man

$$\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{C})$$

mit

$$\boldsymbol{C} = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & & \ddots & \vdots & \vdots \\ \rho^{n-1} & & \dots & \rho & 1 \end{pmatrix}$$

und

$$\boldsymbol{C}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & \ddots & & \vdots \\ 0 & -\rho & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}$$

Wird die Autokorrelation der Beobachtungen verkannt und irrtümlich $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ angenommen, so wird $\sum_{\widehat{\boldsymbol{\beta}}}$ falsch geschätzt.

(Ist ρ bekannt, so ist der BLUE-Schätzer von $\boldsymbol{\beta}$ gemäß Kapitel 6 gegeben als

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{Y})$$

Durbin-Watson Statistik

Die am häufigsten verwendete Kenngröße zur Überprüfung der Autokorrelation der Residuen $\widehat{\epsilon}_i$ ist die Durbin-Watson Statistik DW, die für den Durbin Watson Test eingesetzt wird. Dieser prüft $H_0: \rho = 0$ versus $H_1: \rho \neq 0$.

$$DW := \frac{\sum_{i=2}^{n} (\widehat{\epsilon}_i - \widehat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n} \widehat{\epsilon}_i^2} \approx 2 \cdot (1 - \widehat{\rho}), \tag{5.9}$$

damit gilt: $0 \leq DW \leq 4$.
Kritische Werte für den Test hängen von n und p und werden von Programmpaketen berechnet.

Das allgemeine lineare Modell I

Das *klassische lineare Regressionsmodell* ist spezifiziert durch (siehe (3.3))

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad \mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I} \tag{6.1}$$

mit *unabhängigen, homoskedastischen* Fehlern ϵ_i, $i = 1. . . . , n$.
Im *allgemeinen linearen Modell* ersetzt man die Kovarianzmatrix des Fehlerterms durch

$$\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{W}, \tag{6.2}$$

wobei $\boldsymbol{W}$ eine positiv definite Matrix sei. Diese Formulierung lässt auch den Fall korrelierter Fehlerterme zu.

Das allgemeine lineare Modell II

Für den Fall weiterhin unkorrelierter, aber heteroskedastischer Fehler hat $\boldsymbol{W}$ die Gestalt

$$\boldsymbol{W} = \mathrm{Diag}(w_1, \ldots, w_n), \tag{6.3}$$

so dass sich für die Fehlervarianzen $\mathbb{V}(\epsilon_i) = \sigma_i^2 = \sigma^2 w_i$ ergibt.

Das allgemeine lineare Modell III

Verwendet man im Falle des allgemeinen linearen Modells (6.2) weiterhin den gewöhnlichen KQ-Schätzer $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, so zeigt sich, dass

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$
$$\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Das bedeutet, dass der gewöhnliche KQ-Schätzer zwar erwartungstreu ist, aber die Varianz falsch abschätzt. Damit wären alle aus einer solchen Anwendung stammenden Tests und Konfidenzintervalle falsch. Man braucht daher andere Methoden der Schätzung. Im Folgenden wird die *gewichtete lineare Regression* vorgestellt, die zum Einsatz kommt, wenn man die Matrix $\boldsymbol{W}$ kennt.

Gewichtete Kleinste Quadrate Methode I

Wir betrachten zunächst den Fall unkorrelierter, heteroskedastischer Fehler, d.h. $\mathbb{V}(\epsilon_i) = \sigma_i^2 = \sigma^2 w_i$. Transformiere alle Variablen mit $1/\sqrt{w_i}$, sodass $y_i^\star = y_i/\sqrt{w_i}$, $x_{ik}^\star = x_{ik}/\sqrt{w_i}$ und $\epsilon_i^\star = \epsilon_i/\sqrt{w_i}$. Damit kommt man zu Regressionsgleichungen

$$y_i^\star = \sum_k \beta_k x_{ik}^\star + \epsilon_i^\star$$

mit homoskedastischen Fehlern $\epsilon_i^\star$, die man mit der klassischen KQ-Methode lösen kann. In Matrix-Notation schreibt sich dies als

$$\boldsymbol{y}^\star = \boldsymbol{X}^\star \boldsymbol{\beta} + \boldsymbol{\epsilon}^\star, \quad \text{wobei} \tag{6.4}$$

$$\begin{aligned} \boldsymbol{y}^\star &= \boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{y}, \\ \boldsymbol{X}^\star &= \boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{X} \text{ und } \boldsymbol{\epsilon}^\star = \boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{\epsilon}. \end{aligned} \tag{6.5}$$

Gewichtete Kleinste Quadrate Methode II

Der KQ-Schätzer $\widehat{\boldsymbol{\beta}}$ für Modell (6.4) ergibt sich dann via $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star})^{-1}\boldsymbol{X}^{\star T}\boldsymbol{y}^{\star}$ zu

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{y}. \tag{6.6}$$

Dies ist auch der sogenannte *Aitken Schätzer*, der sich ergibt, wenn man die *gewichtete* Summe der Abstandsquadrate GSAQ minimiert:

$$\begin{aligned} GSAQ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \sum_i \frac{1}{w_i}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2. \end{aligned}$$

Daher der Name *gewichtete Regression:* Beobachtungen mit großen Varianzen (große w_i) erhalten ein kleines Gewicht.

Gewichtete Kleinste Quadrate Methode III

Ein erwartungstreuer Schätzer der Fehlervarianz ergeibt sich als

$$\widehat{\sigma}^2 = \frac{1}{n-p'}\widehat{\boldsymbol{\epsilon}}^T \boldsymbol{W}^{-1}\widehat{\boldsymbol{\epsilon}} = \frac{1}{n-p'}\sum_i \frac{1}{w_i}(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})^2. \qquad (6.7)$$

Die für das klassische lineare Modell entwickelten Tests und Konfidenzintervalle übertragen sich problemlos auf die gewichtete lineare Regression.

Gewichtete Kleinste Quadrate Methode IV

Beispiel: Gruppierte Daten

Falls mehrere Kovariablenvektoren bzw. Zeilen der Kovariablen-Datenmatrix identisch sind, können die Daten gruppiert werden, sodass die resultierende Designmatrix nur noch G verschiedene „Kovariablenmuster“ enthält. Die zugehörigen Responses $\overline{y}_g$, $g = 1, \ldots, G$, sind dann die entsprechenden arithmetischen Mittel der Original-Responses. Beinhaltet das g-te Kovariablenmuster n_g Einzelmessungen, so ist $\mathbb{V}(\overline{y}_g) = \sigma^2/n_g$, entsprechend ist hier $\boldsymbol{W} = \mathrm{Diag}(1/n_1, \ldots, 1/n_G)$.

Allgemeiner Fall

Die eben für den Fall heteroskedastischer, aber unkorrelierter Fehler beschriebene Vorgehensweise lässt sich auf den Fall übertragen, bei dem die Kovarianzmatrix die allgemeine Gestalt $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{W}$ mit positiv definiter Matrix $\boldsymbol{W}$ hat.
Man macht wiederum Transformationen (6.5), um zu einem Modell (6.4) zu gelangen. Dies beinhaltet das „Wurzel-Zehen“ aus der Matrix $\boldsymbol{W}$. Eine Möglichkeit ist die Benutzung der Spektraldarstellung. Ist $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, wobei $\boldsymbol{\Lambda}$ die Diagonalmatrix der Eigenwerte λ_i von $\boldsymbol{W}$ darstellt, so wäre $\boldsymbol{W}^{-\frac{1}{2}} = \boldsymbol{U}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ die gesuchte Transformationsmatrix.

$\boldsymbol{W}$ unbekannt

In vielen Fällen ist $\boldsymbol{W}$ nicht vollständig bekannt, aber die Struktur ist gegeben. Z.B. kann man einen Autoregressiven Prozess 1. Ordnung annehmen (s.o.); unbekannte Parameter für die Varianz (z.B. ρ) können dann mit geschätzt werden, etwa in einem zweistufigem Verfahren oder durch ML-Schätzung.

Modellwahl — Motivation I

Oft stehen bei Anwendungen eine Vielzahl von Variablen zur Verfügung, von denen nicht klar ist, wie und in welcher Form sie in das Modell aufgenommen werden sollen.

- *Welche* Variablen: vor allem *inhaltliche* Überlegungen (z.B. 'Alter', 'Geschlecht', 'Blutdruck' etc.)
  $\Rightarrow$ Interpretation der (partiellen) Regressionskoeffizienten, 'Kontrollvariable' für interessierende Exposition (kausale Interpretation) etc.
- In welcher *Form*: nur Haupteffekte? quadratisch? Polynome? logarithmisch? Wechselwirkungen (welcher Ordnung)? Stufenfunktionen / diskretisiert?
  $\Rightarrow$ schon bei einer einzelnen stetigen Variable gibt es unendlich viele Formen, wie diese in ein Modell eingehen könnte.
- In der Regel sind **viele Modelle mit den Daten kompatibel**.

Modellwahl — Motivation II

- Modellselektion kann verschiedene Ziele verfolgen und hat je nachdem unterschiedliche Ergebnisse:
  – Modellgüte zu 'optimieren' (versch. Kriterien);
  – Parameterschätzung zu 'optimieren' (i.d.R. ein spezieller Expositionseffekt);
  – Prognosen zu 'optimieren';
  – das 'wahre' Modell zu finden;
  – etc.
- **Achtung:** wenn mit statistischen Verfahren ein Modell unter vielen ausgewählt wird, und dieses anschliessend mit *den selben* Daten angepasst wird, sind (unadjustierte) Standardfehler, p-Werte und Konfidenzintervalle **ungültig**, i.d.R anti-konservativ. $\Rightarrow$ Vorsicht bei mehrstufigen Verfahren!

Modellwahl — Motivation III

- Aus interpretations- und philosophischen Gründen werden traditionell oft 'schlichte' Modelle bevorzugt, e.g. nur Haupteffekte.
- Neuerdings, im Zeitalter von Big-Data und Machine-learning, können und werden auch extrem komplexe Modelle routinemäßig benutzt, wenn sie hohe Prognosegüte haben und nicht interpretierbar zu sein brauchen.

Effekt von Modellspezifikation auf Bias & Varianz I

Hinsichtlich 'Optimierung' der Parameterschätzung kann man folgende Fragen untersuchen:

(1) *Irrelevante Kovariablen / Terme:* Welchen Effekt haben diese auf Bias und Varianz?

(2) *Fehlende (relevante) Kovariablen / Terme:* Welchen Effekt haben diese auf Bias und Varianz?

Effekt von Modellspezifikation auf Bias & Varianz II

Zum Beispiel:
Sei korrekt spezifiziert ($\beta_3 \neq 0$, (x_1, x_2) nicht orthogonal):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i2}^2 + \epsilon_i$$

$\Rightarrow$ KQ-Schätzer von β_1 in einem Modell ohne x_{i2}^2 *verzerrt* (bzw. andere Bedeutung).

Aber auch: die geschätzte Residuenvarianz $\hat{\sigma}^2$ wird durch zusätzliche Variablen / Terme im Modell i.d.R. kleiner, Bestimmtheitsmass R^2 wird immer größer, auch wenn diese irrelevant sind $\Rightarrow$ Modellgüte kann besser erscheinen, aber Prognosegüte wird dennoch ungenauer.

Effekt von Modellspezifikation auf Bias & Varianz III

Zerlege dazu den $(p+1)-$dimensionalen Kovariablenvektor $\boldsymbol{x}^T = (x_0, x_1, \ldots, x_p)$ in zwei Teile,

$$\boldsymbol{x}_1^T = (x_0, \ldots, x_{p_1}) \text{ und } \boldsymbol{x}_2^T = (x_{p_1+1}, \ldots, x_p).$$

Betrachte dazu die beiden Schätzmodelle (vgl. (5.2) und (5.3))

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{7.1}$$

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{u}. \tag{7.2}$$

Die KQ-Schätzer ergeben sich zu

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \text{ und } \widetilde{\boldsymbol{\beta}}_1^* = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}.$$

Effekt von Modellspezifikation auf Bias & Varianz IV

Unter Annahme von (7.1): Für $\widetilde{\boldsymbol{\beta}}_1^*$, den Schätzer des Teilmodells, ergeben sich dann für Erwartungswert

$$\begin{aligned}
\mathbb{E}(\widetilde{\boldsymbol{\beta}}_1^*) &= \mathbb{E}((\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}) \\
&= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\,\mathbb{E}(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}) \\
&= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2) \\
&= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2.
\end{aligned}$$

Effekt von Modellspezifikation auf Bias & Varianz V

Wir untersuchen nun zwei Situationen:

(1) *Irrelevante Kovariablen*, das heißt, das Teilmodell (7.2) wäre korrekt ($\boldsymbol{\beta}_2 = \mathbf{0}$), die anderen Variablen ($\boldsymbol{x}_2$) überflüssig; Es ergibt sich

- $\widehat{\boldsymbol{\beta}}$ ist, wie auch $\widetilde{\boldsymbol{\beta}}_1^*$, unverzerrt für $\boldsymbol{\beta}_1$;

(2) *Fehlende Kovariablen*, das heißt, das volle Modell (7.1) wäre korrekt, wir berücksichtigen aber nicht die eigentlich relevanten Variablen aus $\boldsymbol{X}_2$, d.h. zunächst i.A. $\boldsymbol{\beta}_1^* \neq \boldsymbol{\beta}_1$; und es ergibt sich

- $\widetilde{\boldsymbol{\beta}}_1^*$ ist *verzerrt* für $\boldsymbol{\beta}_1$, außer es gilt $\boldsymbol{X}_1^T \boldsymbol{X}_2 = \mathbf{0}$.
  Die Bedingung $\boldsymbol{X}_1^T \boldsymbol{X}_2 = \mathbf{0}$ bedeutet, dass jede Variable aus $\boldsymbol{X}_1$ zu jeder aus $\boldsymbol{X}_2$ unkorreliert ist.
- Wenn $\boldsymbol{X}_1^T \boldsymbol{X}_2 \neq \mathbf{0}$, ist $\widetilde{\boldsymbol{\beta}}_1^*$ vor allem deshalb verzerrt für $\boldsymbol{\beta}_1$, weil es $\boldsymbol{\beta}_1^*$ schätzt und $\boldsymbol{\beta}_1^*$ eine inhaltlich und mathematisch andere Größe als $\boldsymbol{\beta}_1$ ist (siehe Kollinearität).

Effekt von Modellspezifikation auf Bias & Varianz VI

Varianz?

$$\Sigma_{\widetilde{\boldsymbol{\beta}}_1^*} = \sigma_\epsilon^2 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \quad \text{oder} \quad \Sigma_{\widetilde{\boldsymbol{\beta}}_1^*} = \sigma_u^2 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1}$$

deshalb unterscheide:

(a) $\boldsymbol{X}$ durch experimentelles Design gegeben (also *fest*)
$\Rightarrow$ in (7.1/7.2) $\mathbb{V}(u_i) = \mathbb{V}(\epsilon_i) = \sigma^2$;

(b) Spalten von $\boldsymbol{X}$ *zufällig* (Beobachtungsstudie)
$\Rightarrow \sigma_u^2 = \mathbb{V}(u_i) \geq \mathbb{V}(\epsilon_i) = \sigma_\epsilon^2$ (Gleichheit, wenn $\boldsymbol{\beta}_2 = \mathbf{0}$);

- Im Falle (a): wenn (7.2) korrekt, dann $\mathbb{V}(\widetilde{\boldsymbol{\beta}}_1^*) \leq \mathbb{V}(\widehat{\boldsymbol{\beta}}_1)$; sonst, wenn $\boldsymbol{X}_1^T \boldsymbol{X}_2 = \mathbf{0}$, dann auch $\mathbb{V}(\widetilde{\boldsymbol{\beta}}_1^*) \leq \mathbb{V}(\widehat{\boldsymbol{\beta}}_1)$;
also: in diesen zwei Fällen höhere Präzision ohne $\boldsymbol{X}_2$.

Effekt von Modellspezifikation auf Bias & Varianz VII

- Im Falle (b): wenn $\boldsymbol{\beta}_2 = \mathbf{0}$, dann $\mathbb{V}(\widetilde{\boldsymbol{\beta}}_1^*) \leq \mathbb{V}(\widehat{\boldsymbol{\beta}}_1)$; sonst, wenn $\boldsymbol{X}_1^T \boldsymbol{X}_2 = \mathbf{0}$, dann i.A. $\mathbb{V}(\widetilde{\boldsymbol{\beta}}_1^*) \geq \mathbb{V}(\widehat{\boldsymbol{\beta}}_1)$!
  $\Rightarrow$ zusätzliche Prädiktoren von Y im Modell erhöhen die Präzision (weil sie die Residuenvarianz verringern).
- Achtung! 'Empfehlung' hängt also entscheidend von der Ausgangssituation ab.
  Dies wird in vielen Lehrbüchern nicht klar dargestellt.

Effekt von Modellspezifikation auf Prognosegüte I

Im Folgenden wird nicht notwendigerweise vorausgesetzt, dass das spezifizierte Modell korrekt ist, d.h. ein Bias ist 'erlaubt'.
Wir betrachten das folgende Szenario:

- Unabhängige Beobachtungen Y_i, $i = 1, \ldots, n$, mit $\mathbb{E}(Y_i) = \mu_i$, $\mathbb{V}(Y_i) = \sigma^2$;
- Potenzielle Regressoren $x_0 = 1, x_1, \ldots, x_p$;
- Es werde eine Teilmenge $M \subset \{x_0, x_1, \ldots, x_p\}$ zur Modellierung verwendet, die zugehörige Designmatrix wird mit $\boldsymbol{X}_M$ bezeichnet, entsprechende Schätzer mit $\widehat{\boldsymbol{\beta}}_M$ und $\widehat{\sigma}_M$;
- Man erhält den KQ-Schätzer $\widehat{\boldsymbol{\beta}}_M$ sowie eine Schätzung $\widehat{\boldsymbol{y}}_M$ auf $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$ via

$$\widehat{\boldsymbol{\beta}}_M = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \boldsymbol{Y}, \quad \widehat{\boldsymbol{Y}}_M = \boldsymbol{X}_M \widehat{\boldsymbol{\beta}}_M.$$

Effekt von Modellspezifikation auf Prognosegüte II

Fasse die Schätzungen $\widehat{Y}_{iM}$ als Prognose für zukünftige Beobachtungen $Y_{n+i} = \mu_i + \epsilon_{n+i}$, $i = 1, \ldots, n$, bei gegebenen selben Regressoren $x_{i1}, \ldots, x_{ip}$ auf.
Für den erwarteten quadrierten Prognosefehler SPSE (sum of prediction squared errors), definiert als

$$\mathrm{SPSE} = \sum_{i=1}^{n} \mathbb{E}(Y_{n+i} - \widehat{Y}_{iM})^2, \tag{7.3}$$

gilt folgende wichtige Formel:

$$\mathrm{SPSE} = n \cdot \sigma^2 + \mathrm{Rang}(M) \cdot \sigma^2 + \sum_{i=1}^{n} (\mu_{iM} - \mu_i)^2. \tag{7.4}$$

Effekt von Modellspezifikation auf Prognosegüte III

Damit ist SPSE zerlegt in drei Terme:

1. *irreduzibler Prognosefehler:* $n \cdot \sigma^2$;
   hängt von Varianz des Fehlerterms ϵ ab, lässt sich nicht verkleinern;
2. *Varianz*: $\text{Rang}(M) \cdot \sigma^2$;
   dieser Term stellt $\sum_i \mathbb{V}(\widehat{Y}_{iM})$ dar und ist durch Modellwahl zu beeinflussen: $\text{Rang}(M) \cdot \sigma^2$ ist umso kleiner, je weniger Variablen das Modell umfasst;
3. *Quadrierter Bias*: $\sum_i (\mu_{iM} - \mu_i)^2$;
   erfasst den quadrierten Bias der Schätzungen $\widehat{Y}_{iM}$ für den Erwartungswert μ_i; er wird umso kleiner, je komplexer das Modell ist.

Fazit: Je komplexer das Modell, umso kleiner ist der Bias, aber umso größer ist die Varianz.
Es handelt sich hier um das typische **Varianz-Bias-Dilemma**, *das bei allen Regressionsmodellen, nicht allein bei den linearen, auftritt.*

Effekt von Modellspezifikation auf Prognosegüte IV

Bemerkungen:

- Wenn M das 'wahre' Modell wäre, ist der Bias null und $\frac{1}{n}\text{SPSE} = FPE = \sigma^2(1 + \frac{p_M+1}{n})$, wobei p_M die Anzahl der Regressoren in Modell M ist.
  'FPE' steht für 'Final Prediction Error' — den erwarteten Prognosefehler im wahren Modell.
- Beachte: SPSE und die obigen Eigenschaften beziehen sich speziell auf den (erwarteten) *quadratischen* Prognosefehler; die Prognosegüte kann auch mit anderen Verlustfunktionen bewertet werden, z.B. dem *absoluten* Prognosefehler (große Abweichungen fallen weniger ins Gewischt), oder unsymmetrischen Verlustfunktionen (z.B. wenn ein Unterschätzen von Y_{n+i} 'schlimmer' ist als ein Überschätzen, oder wenn Y binär ist) etc.

Beweis von (7.4):

Bezeichne mit

$$\boldsymbol{P}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T$$

die Hut-Matrix für Modell M. Dann gilt

$$\begin{aligned}
&\mathbb{E}(\widehat{\boldsymbol{Y}}_M) = \mathbb{E}(\boldsymbol{P}_M\boldsymbol{Y}) = \boldsymbol{P}_M\boldsymbol{\mu} =: \boldsymbol{\mu}_M \\
&\mathbb{V}(\widehat{\boldsymbol{Y}}_M) = \sigma^2 \cdot \boldsymbol{P}_M \\
&\sum_i \mathbb{V}(\widehat{Y}_{iM}) = \sigma^2 \cdot \mathrm{Rang}(M)
\end{aligned}$$

Damit erhält man für den erwarteten quadratischen Prognosefehler

$$
\begin{aligned}
\text{SPSE} &= \sum_{i=1}^{n} \mathbb{E}(Y_{n+i} - \widehat{Y}_{iM})^2 \\
&= \sum \mathbb{E}((Y_{n+i} - \mu_i) + (\mu_i - \widehat{Y}_{iM}))^2 \\
&= \sum \Big\{ \mathbb{E}(Y_{n+i} - \mu_i)^2 + 2 \underbrace{\mathbb{E}[(Y_{n+i} - \mu_i)(\mu_i - \widehat{Y}_{iM})]}_{=0} \\
&\quad + \mathbb{E}(\mu_i - \widehat{Y}_{iM})^2 \Big\} \\
&= \sum \mathbb{E}(Y_{n+i} - \mu_i)^2 + \sum \mathbb{E}(\mu_i - \widehat{Y}_{iM})^2 \\
&= n \cdot \sigma^2 + \underbrace{\sum \Big\{ \mathbb{E}[(\mu_i - \mu_{iM}) + (\mu_{iM} - \widehat{Y}_{iM})]^2 \Big\}}_{\sum(\mu_i - \mu_{iM})^2 + \text{Rang}(M) \cdot \sigma^2}
\end{aligned}
$$

Modellwahlkriterien I

Bias, Varianz und Prognosegüte und ihre Relevanz in einer gegebenen Anwendung sollten sich in der Wahl des Kriteriums wiederspiegeln, welches zum Vergleich verschiedener Modelle benutzt wird.
Beliebte Kriterien:

- SPSE — kann aber nur angenähert / geschätzt werden;
- korrigiertes Bestimmtheitsmass R^2_{adj} (da R^2 nicht geeignet);
- Mallow's C_p;
- AIC — Akaike Information Criterion;
- BIC — Bayesian Information Criterion;
- statistische Tests — wird heutzutage nicht mehr empfohlen (Hypothesentests sind nicht auf Modellselektion ausgerichtet und haben i.A. keine entsprechenden Optimalitätseigenschaften; zudem: multiples-Testen-Problem).

SPSE als Kriterium I

Leider kann man SPSE nicht direkt berechnen, da in der Regel σ^2 und die μ_i unbekannt sind, daher benötigt man eine Schätzung für SPSE. Zwei Strategien sind vorstellbar:

1. *Schätze SPSE mithilfe neuer, unabhängiger Daten*
   Hat man tatsächlich einen unabhängigen Validierungsdatensatz mit Werten vorliegen, so kann man SPSE einfach schätzen durch

$$\widehat{\mathrm{SPSE}} = \sum_{i=1}^{n} (Y_{n+i} - \widehat{Y}_{iM})^2$$

SPSE als Kriterium II

Da man solche Validierungsdaten in den seltensten Fällen vorliegen hat, wäre ein gangbarer Ausweg:

- Teile den Datensatz zufällig in Test- und Validierungsdatensatz;
- Benutze den Testdatensatz, um die $\boldsymbol{\beta}_M$ und $\boldsymbol{\mu}_M$ für die verschiedenen Modelle M zuschätzen;
- Benutze den Validierungsdatensatz, um SPSE zu berechnen.
- Beachte: dabei wird es selten möglich sein, dass die $\boldsymbol{x}_i$–Werte im Test- und Validierungsdatensatz genau gleich sind. D.h. SPSE wird dann etwas überschätzt. Allerdings gilt dies für alle Modelle, die verglichen werden, und fällt deshalb nicht sehr ins Gewicht.

Diese Methode ist im Prinzip wünschenswert, wird aber oft nur dann benutzt, wenn sehr große Datensätze zur Verfügung stehen.

SPSE als Kriterium III

2. *Schätze SPSE mithilfe der vorhandenen Daten*
   Es zeigt sich, dass man die Residuenquadratsumme $\sum(Y_i - \widehat{Y}_{iM})^2$ verwenden kann, es gilt nämlich

$$\mathbb{E}\left(\sum_{i=1}^{n}(Y_i - \widehat{Y}_{iM})^2\right) = \mathrm{SPSE} - 2 \cdot \mathrm{Rang}(M) \cdot \sigma^2.$$

   Damit erhält man als Schätzung für SPSE

$$\widehat{\mathrm{SPSE}} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_{iM})^2 + 2 \cdot \mathrm{Rang}(M) \cdot \widehat{\sigma}^2_{voll}, \qquad (7.5)$$

SPSE als Kriterium IV

- Beachte: $\widehat{\sigma}^2_{voll}$ soll die Residuenvarianz im 'wahren' Modell schätzen, aber das ist ja unbekannt; wenn es in der Menge der betrachteten Modelle so etwas wie ein 'volles' Modell gibt, sollte dies hier verwendet werden; ansonsten sollte aber immer dasselbe $\widehat{\sigma}^2_{voll}$ aus einem 'großen' (wenig Bias) Modell gewählt werden.
- Wenn für $\widehat{\mathrm{SPSE}}$ stattdessen $\widehat{\sigma}^2_M$ benutzt wird, also die geschätzte Residuenvarianz unter dem Modell M, dann gilt $\frac{1}{n}\widehat{\mathrm{SPSE}} = \widehat{FPE}$ — ein etwas anderes mögliches Kriterium.

Adjustiertes R^2

Das *korrigierte Bestimmtheitsmaß* R^2_{adj} ist definiert als

$$R^2_{adj} = 1 - \frac{n-1}{n-p'_M}(1-R^2),$$

- Modellwahl mit R^2_{adj} is äquivalent zur Benutzung von $\widehat{\sigma}^2_M$.
- R^2_{adj} bzw. $\hat{\sigma}^2_M$ messen die Güte des *angepassten* Modells, aber $\widehat{\sigma}^2_M$ unterschätzt SPSE auch wenn das betrachtete Modell korrekt ist.
- Man kann daher der Ansicht sein, dass R^2_{adj} zusätzlich aufgenommener Terme (größeres p_M) zu wenig "bestraft".

Mallow's C_p

Mallow's C_p ("Complexity Parameter") ist definiert als

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_{iM})^2}{\widehat{\sigma}^2_{voll}} - n + 2 \cdot \mathrm{Rang}(M),$$

seine Verwendung liefert dasselbe Modell wie die Verwendung von $\widehat{\mathrm{SPSE}}$ (bei gleicher Wahl von $\widehat{\sigma}^2_{voll}$).

Informationskriterien I

Während obige Kriterien mehr oder weniger direkt versuchen die Prognosegüte zu optimieren, orientieren sich die folgenden 'Informationskriterien' eher an der (maximalen) Likelihood, also Anpassungsgüte.

Akaike's Information Criterion AIC ist ein im Rahmen der ML-Inferenz sehr häufig verwendetes Kriterium, es ist definiert als

$$AIC = -2 \cdot l(\widehat{\boldsymbol{\beta}}_M, \tilde{\sigma}^2_M) + 2 \cdot (\text{Rang}(M) + 1), \qquad (7.6)$$

$l(\widehat{\boldsymbol{\beta}}_M, \tilde{\sigma}^2_M)$ ist das Maximum der Log-Likelihood, wobei man die ML-Schätzer $\widehat{\boldsymbol{\beta}}_M$ und $\tilde{\sigma}^2_M$ eingesetzt hat.

Optimal: Modelle mit kleinerem AIC sind laut diesem Kriterium zu bevorzugen.

Informationskriterien II

Im linearen Modell mit normalverteilten Fehlern hat man

$$\begin{aligned} -2 \cdot l(\widehat{\boldsymbol{\beta}}_M, \tilde{\sigma}^2_M) &= n \cdot \log(\tilde{\sigma}^2_M) + \frac{1}{\tilde{\sigma}^2_M}(\boldsymbol{y} - \boldsymbol{X}_M \widehat{\boldsymbol{\beta}}_M)^T(\boldsymbol{y} - \boldsymbol{X}_M \widehat{\boldsymbol{\beta}}_M) \\ &= n \cdot \log(\tilde{\sigma}^2_M) + \frac{n\tilde{\sigma}^2_M}{\tilde{\sigma}^2_M} \\ &= n \cdot \log(\tilde{\sigma}^2_M) + n. \end{aligned}$$

Damit ergibt sich AIC, unter Weglassen von Konstanten, in diesem Fall zu

$$AIC = n \cdot \log(\tilde{\sigma}^2_M) + 2 \cdot (\text{Rang}(M) + 1).$$

Beachte, dass hier $\tilde{\sigma}^2_M = \widehat{\boldsymbol{\epsilon}}^T \widehat{\boldsymbol{\epsilon}}/n$ der ML-Schätzer von σ^2 ist, nicht der sonst übliche (unter M erwartungstreue) Schätzer $\widehat{\sigma}^2_M$.

Informationskriterien III

Das Bayesianische Informationskriterium BIC ist definiert als

$$BIC = -2 \cdot l(\widehat{\boldsymbol{\beta}}_M, \tilde{\sigma}^2_M) + \log(n) \cdot \text{Rang}(M),$$

und hat damit eine ähnliche Struktur wie AIC, wird allerdings unter anderen Prinzipien hergeleitet. Im Vergleich zu AIC werden bei Verwendung von BIC komplexe Modelle stärker bestraft.

Im linearen Modell mit normalverteilten Fehlern hat man

$$BIC = n \cdot \log(\tilde{\sigma}^2_M) + \log(n) \cdot \text{Rang}(M).$$

Optimal: Modelle mit kleinerem BIC sind laut diesem Kriterium zu bevorzugen.

Informationskriterien IV

Bemerkungen:

- Man sieht bei beiden Kriterium das wiederum die Residuenvarianz von entscheidender Bedeutung ist.
- AIC, Mallow's C_p bzw. $\widehat{\mathrm{SPSE}}$ sind asymptotisch äquivalent.
- Optimierung des AIC minimiert in gewissem Sinne den 'Abstand' zwischen dem 'wahren' Modell und dem gewählten Modell;
- obwohl es also scheint, dass AIC das 'wahre' Modell finden möchte, kann man zeigen, dass AIC sehr gut im Sinne der Prognosegüte ist, aber *nicht konsistent* hinsichtlich Modellwahl. Die gewählten Modelle sind 'zu komplex'.

Informationskriterien V

- Optimierung des BIC versucht in gewissem Sinne das 'wahrscheinlichste' Modell zu finden;
- dabei kann man zeigen, dass BIC suboptimal im Sinne der Prognosegüte ist, dafür aber *konsistent* hinsichtlich Modellwahl.
- Das Phänomen ist als AIC–BIC Dilemma bekannt: Prognosegüte versus Konsistenz.

Kreuzvalidierung I

Bei der *"leave-one-out"*-Methode (LOO) der Kreuzvalidierung lässt man sukzessive jede Beobachtung i bei der Schätzung von $\boldsymbol{\beta}_M$ weg. Bezeichne $\widehat{y}_{iM}^{-i}$ die Vorhersage von y_i aus einem solchen Modell. Als Kriterium verwendet man die z.B. gerne*Kreuzvalidierungsfunktion* CV,

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{iM}^{-i})^2.$$

CV ist tatsächlich ohne großen Aufwand zu berechnen, man zeigt, dass

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_{iM}}{1 - p_{iiM}}\right)^2,$$

p_{iiM} sind die Diagonalelemente der Hut-Matrix $\boldsymbol{P}_M$.

Kreuzvalidierung II

Bemerkungen:

- Die vereinfachte Berechnung von CV ergibt sich nur für die quadratischen Verlustfunktion.
- LOO-CV stellt allerdings ein ganz flexibles Prinzip dar, welches sich mit beliebigen Verlustfunktionen kombinieren läßt, z.B. $CV_{abs} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_{iM}^{-i}|$.
- Der Prognosefehler, der durch CV geschätzt wird, ist etwas anders als $SPSE/n$, da die Prognosen nicht für dieselben x-Werte sind.
- Im linearen Regressionsmodell mit Normalverteilungsannahme sind CV und AIC äquivalent.
- AIC und BIC sind an eine Likelihood gebunden, während LOO-CV 'verteilungsfrei' ist.

Praktische Verwendung der Kriterien I

- Im Prinzip würde man für alle zu vergleichenden Modelle das gewünschte Kriterium berechnen und dann das Modell wählen, welches am besten abschneidet (dies können auch mehrere gleich gute Modelle sein).
  Z.B. wenn maximal p Regressoren betrachtet werden (das 'volle' Modell), würde man alle Modelle vergleichen, die eine Teilmenge dieser Regressoren enthalten $\Rightarrow$ 'best subset selection'.
- Aber es ist oft rechentechnisch unmöglich 'alle' Modelle zu vergleichen, da dies zu viele sind (z.B. 'best subset selection' $\Rightarrow 2^p$ Modelle). Deswegen werden oft *heuristische* Suchstrategien verwendet, von denen bekannt ist, dass sie nicht garantiert das/ein optimale(s) Modelle finden.

Praktische Verwendung der Kriterien II

Folgende Strategie wäre wünschenswert:

- Durch substanzwissenschaftliche Überlegungen bzw. Vorwissen aus vorangegangenen Experimenten wird eine Vorauswahl potenzieller Modelle getroffen. Dadurch sollte versucht werden, die Gesamtzahl der zu vergleichenden Modelle gering zu halten;
- Auf die ausgewählten Modelle können obige Kriterien angewandt werden. Bei der Beschreibung der Ergebnisse sollte man sich nicht schlicht auf ein einziges "bestes" Modell beschränken, da in der Regel verschiedene Modelle zu ähnlichen Ergebnissen führen.

Praktische Verwendung der Kriterien III

Zu den gängigen Heuristiken gehören die folgenden *Selektionsverfahren*

- *Vorwärts-Selektion (Forward-Selection)*

  Ausgehend von einem Startmodell wird bei jedem Schritt des Verfahrens diejenige Kovariable in das Modell zugefügt, deren Aufnahme die größte Reduktion eines der Kriterien (C_p, AIC, BIC, CV) bewirkt. Der Algorithmus bricht ab, wenn keine Verbesserung mehr stattfindet.

- *Rückwärts-Selektion (Backward-Selection)*

  Hier wird mit dem vollen Modell, das alle Kovariablen umfasst, gestartet. Bei jeder Iteration wird diejenige Variable aus dem Modell entfernt, sodass die größte Reduktion eines der Kriterien erfolgt. Der Algorithmus stoppt, wenn keine Reduktion mehr möglich ist.

Praktische Verwendung der Kriterien IV

- *Schrittweise-Selektion (Stepwise-Selection)*

  Hier handelt es sich um eine Kombination von Vorwärts- und Rückwärts-Selektion. Mit jedem Schritt können sowohl Variablen aufgenommen wie entfernt werden.

Bemerkungen:

- Diese drei Selektionsverfahren führen in der Regel nicht zu denselben Modellen.
- Da sie nur wenige Modellsequenzen durchlaufen, gibt es keine Garantie, dass ein optimales Modell gefunden wird.
- Bei der Realisierung muss man bei dummy-kodierten kategoriellen Kovariablen aufpassen (sie sollten als "Block", d.h. zusammen, betrachtet werden.)

Praktische Verwendung der Kriterien V

- 'Forward' kann auch bei kleinen Stichprobenumfängen benutzt werden; hat aber den Nachteil, dass die ersten Modelle i.d.R. unrealistisch sind.
- 'Backward' kann nur benutzt werden, wenn der Stichprobenumfang gross genug ist, das volle Modelle anzupassen ($n > p$); hat aber den Vorteil , dass das 'volle' Modelle am ehesten 'nicht falsch' ist.
- Andere Strategien könnten z.B. mit verschiedenen Modellen mittlerer Größe anfangen, und sich dann schrittweise vortasten.
- Wiederum andere Strategien (Bayesianisch) vergeben a-priori Wahrscheinlichkeiten an die zu vergleichenden Modelle und betrachten zunächst die wahrscheinlichsten Modelle.

Praktische Verwendung der Kriterien VI

- Es ist auch möglich und gängig, Prognosen auf der Basis mehrerer plausibler Modelle zu erstellen (z.B. durch Gewichtung).

Achtung: Eine auf demselben Datensatz vorgeschaltete Modellselektion 'verfälscht' die anschließende Inferenz (z.B. sind Konfidenzintervalle zu schmal, p-Werte zu klein / antikonservativ – siehe Folie 7-2)
$\longrightarrow$ das sogenannte 'post-selection-inference' (PoSI) Problem.
$\longrightarrow$ Aufteilen des Datensatzes in einen Teil, der für Modellselektion benutzt wird, und einen anderen, der für die Inferenz benutzt wird, ist am sichersten (und ehrlichsten).

Lasso und Ridge Regression I

Zwei Beispiele für Abwandlungen des KQ-Prinzips, die auch für sehr großes p (viele Variablen oder Terme) inspesondere $p \gg n$ funktionieren, sind *Lasso* und *Ridge* Regression. Ähnlich wie schon oben eingeführt beruhen sie auf der Idee einer "Bestrafung" ('penalisation' oder 'regularisation') komplexer Modelle.

Lasso und Ridge Regression II

Lasso (least absolute shrinkage and selection operator):
KQ minimiert $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$. Hier ist $||\cdot||_2$ die L_2-Norm.

Das Lasso Prinzip zielt auch darauf ab, möchte aber die Absolutwerte $|\beta_j|$ nicht 'zu groß' werden lassen: $\widehat{\boldsymbol{\beta}}_{lasso}$ ist die Lösung von

$$||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \to \min_{\boldsymbol{\beta}},$$

wobei $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$ (die L_1-Norm) und $\lambda \geq 0$.

- Bestrafung mit der L_1-Norm hat zur Folge, dass viele Einträge des Lasso Schätzers $\widehat{\boldsymbol{\beta}}_{lasso}$ *genau gleich null* sind (selection), während die anderen Einträge in Richtung Null 'geschrumpft' werden (shrinkage).

Lasso und Ridge Regression III

- Wenn $\widehat{\beta}_{j,lasso} = 0$ ist X_j nicht mehr im Modell — Lasso Schätzung ist also gleichzeitig ein Schätz- und Modellselektionsverfahren.
- λ ist ein 'tuning' Parameter; je größer, desto mehr Nullen und desto größer die Schrumpfung in $\widehat{\boldsymbol{\beta}}_{lasso}$.
- λ wird oft mit Hilfe einer Kreuzvalidierung gewählt.
- Die Lasso Schätzung hat asymptotisch, wenn n und $p \to \infty$ (in einem gewissen Verhältnis) und unter gewissen Annahmen Optimalitätseigenschaften hinsichtlich Modellwahl.
- Lasso ist unstabil, wenn Gruppen der X_j *stark* korreliert sind; andere Verfahren sind besser hinsichtlich Prognosegüte.

Lasso und Ridge Regression IV

Ridge Regression:
Das Ridge Prinzip bestraft die Absolutwerte $|\beta_j|$ mit der L_2-Norm:
$\widehat{\boldsymbol{\beta}}_{ridge}$ ist die Lösung von (für $\lambda \geq 0$)

$$||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 \rightarrow \min_{\boldsymbol{\beta}}.$$

- Es gilt

$$\widehat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad (7.7)$$

  wobei $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})$ für $\lambda > 0$ immer invertierbar ist
- Bestrafung mit der L_2-Norm hat zur Folge, dass alle Einträge des Ridge Schätzers $\widehat{\boldsymbol{\beta}}_{ridge}$ Richtung Null 'geschrumpft' werden (shrinkage), aber keine *genau gleich null* gesetzt werden.

Lasso und Ridge Regression V

- Im Spezialfall $n < p$ und orthogonale Spalten von $\boldsymbol{X}$ gilt: $\widehat{\beta}_{j,ridge} = \widehat{\beta}_{j,KQ}/(1+\lambda)$.
- Auch hier ist λ der 'tuning' Parameter; je größer, desto größer die Schrumpfung in $\widehat{\boldsymbol{\beta}}_{ridge}$.
- λ wird oft mit Hilfe einer Kreuzvalidierung gewählt.
- Die Ridge Schätzung ist stabiler als Lasso, wenn Gruppen der X_j *stark* korreliert sind.
- Beide, Lasso und Ridge Regression, vermeiden "overfitting" und haben deshalb für großes p niedrigeren MSE als KQ (wenn existent) und bessere Prognosegüte, sind aber (wegen Schrumpfung) verzerrt ('regularisation bias').
  $\longrightarrow$ Wieder ein Beispiel für den Varianz vs. Bias 'trade off'.

Lasso und Ridge Regression VI

- Beide, Lasso und Ridge Regression, lassen sich abwandeln, so dass unterschiedliche Parameter unterschiedlich stark bestraft werden. Z.B. wird der Achsenabschnit (β_0) nie bestraft (bzw: man zentriere Y_i und alle X_{ij} zunächst); desweiteren könnte man Haupteffekte weniger bestrafen als Interaktionen oder höhere Polynomialterme; auch sollte die Bestrafung z.B. für binäre X_j anders sein als für stetige etc.

Longitudinal- und Clusterdaten 纵向和集群数据

Von Longitudinaldaten spricht man, wenn an Untersuchungseinheiten über Zeitpunkte hinweg Wiederholungsmessungen vorgenommen werden. In der Regel sind solche Messungen innerhalb einer Einheit korreliert. Ähnliches gilt für Clusterdaten, etwa Mäusen aus einem Wurf: die Nachkommen einer Maus sind einander ähnlicher (wegen der gleichen Mutter) als Nachkommen anderer Muttertiere. Daher sind Messungen aus einem Cluster oft korreliert. Solche Daten sind also keine "unabhängige" Stichprobe.
Um systematischen Unterschieden zwischen Individuen bzw. Clustern Rechnung zu tragen, verwendet man gerne *gemischte Modelle*: in ihnen wird der lineare Prädiktor um sogenannte zufällige Effekteerweitert. Ein solcher zufälliger Effekt repräsentiert z.B. eine unbeobachtbare Gemeinsamkeit (latente Variable) der Mäuse desselben Muttertiers.
当在一段时间内重复对调查单位的测量时，会参考纵向数据。 通常，这种测量在一个单元内是相关的。 这同样适用于群集数据，例如来自垃圾的小鼠：小鼠的后代彼此更相似（因为同一个母亲）作为其他水坝的后代。 因此，来自群集的测量通常是相关的。 因此，这些数据不是一个独立的样本。
为了解释个体或群集之间的系统差异，经常使用混合模型：在其中，线性预测因子通过所谓的随机效应扩展。 这种随机效应表示例如 同一母亲的老鼠的一个不可观察的社区（潜在变量）。

Beispiel: Hormontherapie bei Ratten[1]

睾酮对大鼠生长的影响

- Die Wirkung von Testosteron auf das Wachstum von Ratten wurde untersucht: insgesamt 50 Ratten wurden zufällig einer Kontrollgruppe und zwei Therapiegruppen (niedrige oder hohe Dosis von Decapeptyl, das die Testosteronsynthese hemmt) zugeteilt.抑制睾酮
- Die Behandlung begann im Alter von 45 Tagen, beginnend mit dem 50. Tag wurden alle 10 Tage das Wachstum des Kopfes via Röntgenmessunggen ermittelt. Zielvariable war der Abstand (in Pixeln) zwischen zwei wohldefinierten Punkten im Röntgenbild.

每10天通过X射线测量基因确定头部的生长。目标变量是X射线图像中两个明确定义的点之间的距离（以像素为单位）

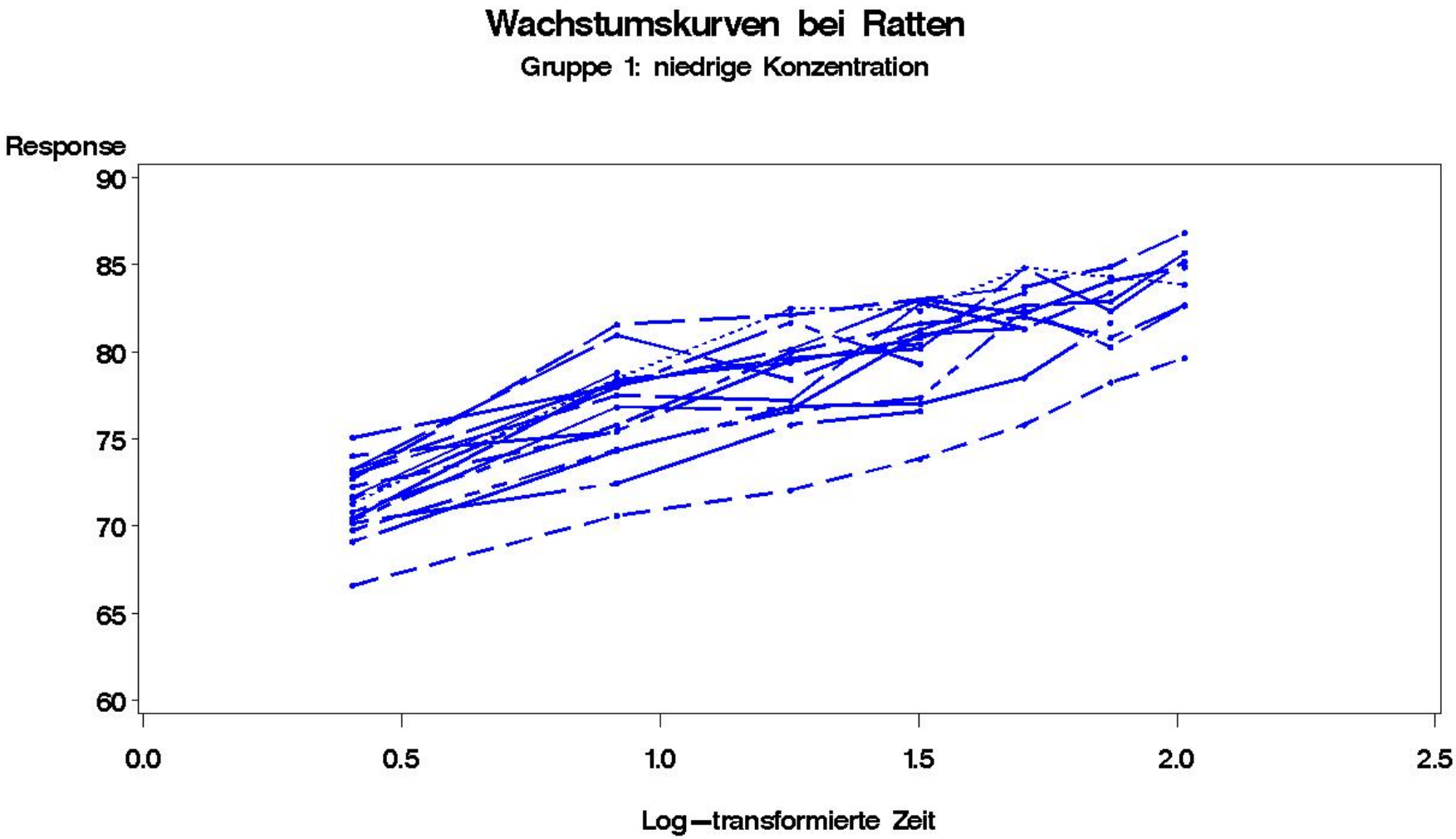
- Die Anzahl n_i von wiederholten Messungen y_{ij}, $j = 1, \ldots, n_i,$, $i = 1, \ldots, 50$, war für die Ratten unterschiedlich (an vier Ratten wurde nur eine Messung vorgenommen, an 22 Ratten sieben Messungen).
- Die folgende Tabelle beschreibt das so entstandene Untersuchungsdesign, die folgenden Grafiken die Wachstumskurven der Gruppen.

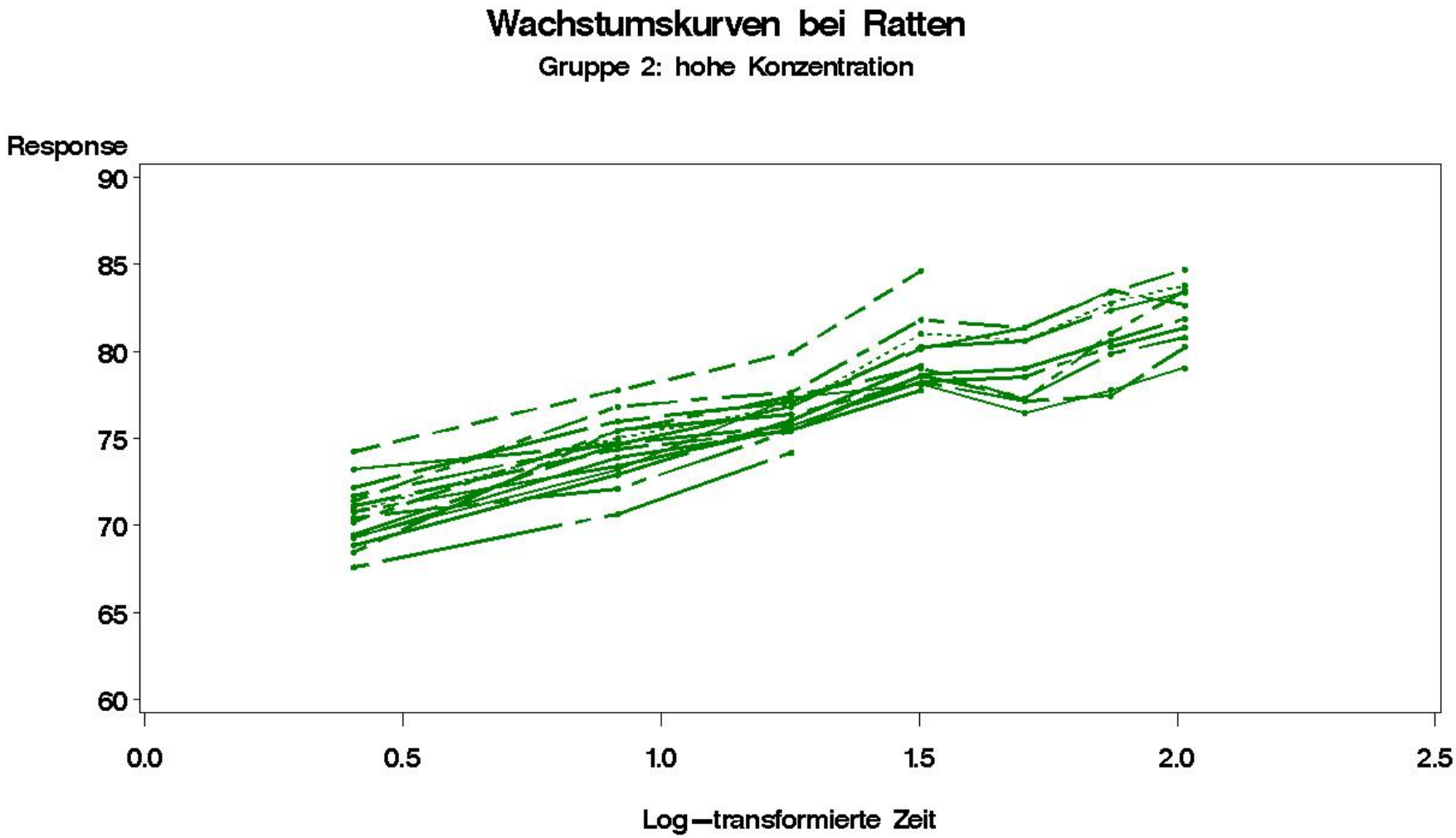
[1] Quelle: Fahrmeir et al. (2007), S. 35; Verbeke & Molenberghs (2000)

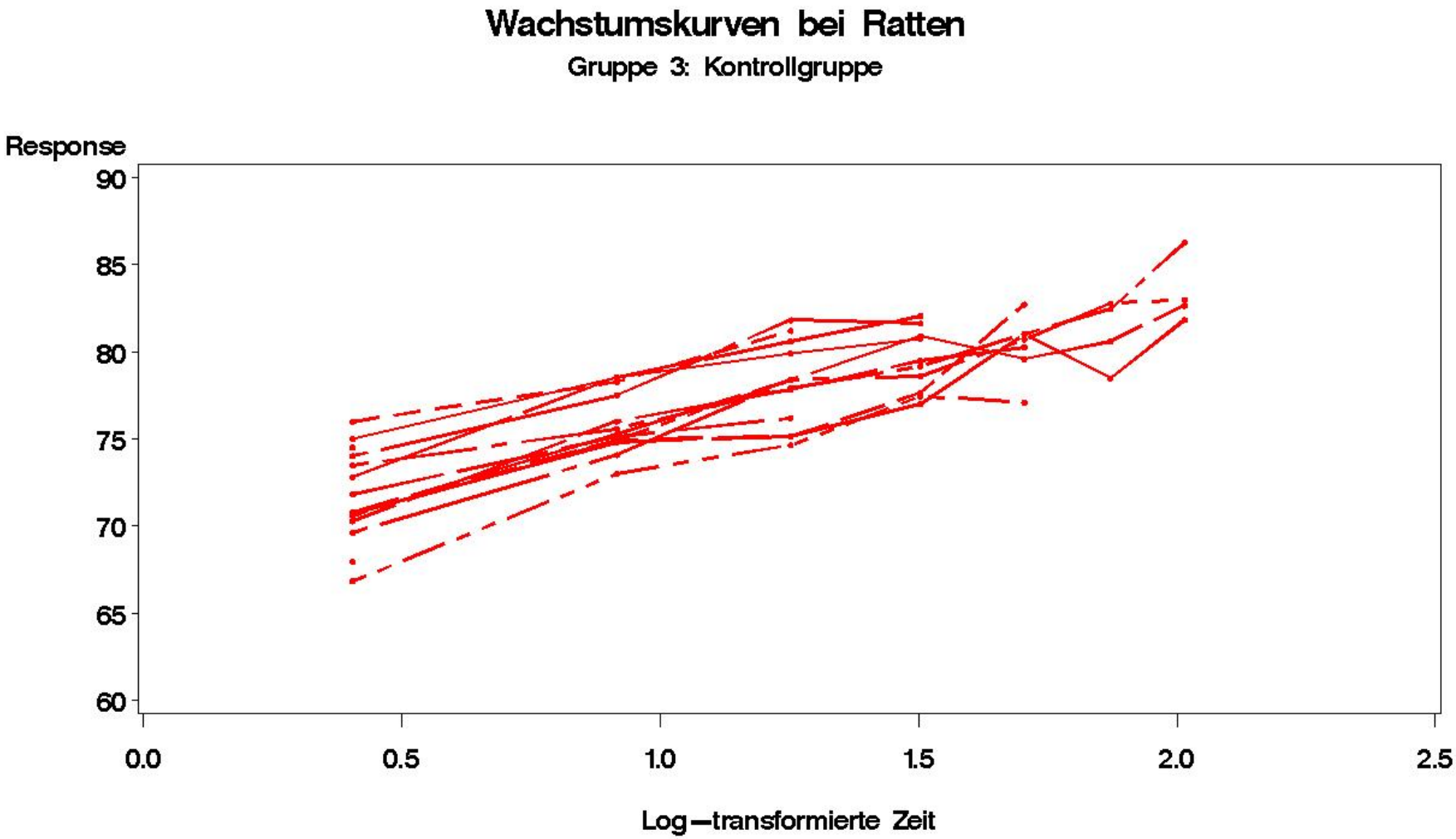
Tabelle: Anzahl Beobachtungen pro Zeitpunkt und Dosierungsgruppe

Alter (in Tagen)	Kontrolle	Niedrig	Hoch	Gesamt
50	15	18	17	50
60	13	17	16	46
70	13	15	15	43
80	10	15	13	38
90	7	12	10	29
100	4	10	10	24
110	4	8	10	22

Obs	SUBJECT	GROUP	RESPONSE	TIME	transf_ time
1	8	niedrig	71.7008	50	0.40547
2	8	niedrig	78.8162	60	0.91629
3	8	niedrig	.	70	1.25276
4	8	niedrig	.	80	1.50408
5	8	niedrig	.	90	1.70475
6	8	niedrig	.	100	1.87180
7	8	niedrig	.	110	2.01490
8	10	niedrig	71.2811	50	0.40547
9	10	niedrig	78.4920	60	0.91629


Wachstumskurven bei Ratten
Gruppe 1: niedrige Konzentration
Response
90
85
80
75
70
65
60
0.0
0.5
1.0
1.5
2.0
2.5
Log−transformierte Zeit


Wachstumskurven bei Ratten
Gruppe 2: hohe Konzentration
Response
90
85
80
75
70
65
60
0.0
0.5
1.0
1.5
2.0
2.5
Log−transformierte Zeit


Wachstumskurven bei Ratten
Gruppe 3: Kontrollgruppe
Response
90
85
80
75
70
65
60
0.0
0.5
1.0
1.5
2.0
2.5
Log−transformierte Zeit

Formulierung des Modells:

- transformiertes Alter:

$$t = \log(1 + (\text{Alter-45})/10)$$

- Dummy-Variable für die Behandlungsgruppen:
  $C \in \{0, 1\}$ für Kontrolle, $N \in \{0, 1\}$ für niedrige Dosis und
  $H \in \{0, 1\}$ für hohe Dosis
- Modell auf **Populationsebene**:

$$y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} & \text{für } i \text{ in Niedrigdosisgruppe} \\ \beta_0 + \beta_2 t_{ij} + \varepsilon_{ij} & \text{für } i \text{ in Hochdosisgruppe} \\ \beta_0 + \beta_3 t_{ij} + \varepsilon_{ij} & \text{für } i \text{ in Kontrollgruppe} \end{cases}$$

$$\Longleftrightarrow$$

$$y_{ij} = \beta_0 + \beta_1 N_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} + \varepsilon_{ij} \qquad (8.1)$$

- Die Parameter β_i von Modell (8.1) beschreiben Effekte auf Populationsebene, nicht auf individueller Ebene
- Man sieht an den Grafiken, dass die Messungen für einzelne Ratten abhängig sind.
- Um diese Effekte abzubilden, wird ein individueller Regressionsansatz gewählt:

$$y_{ij} = \begin{cases} \beta_0 + \gamma_{0i} + (\beta_1 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in Niedrigdosisgruppe} \\ \beta_0 + \gamma_{0i} + (\beta_2 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in Hochdosisgruppe} \\ \beta_0 + \gamma_{0i} + (\beta_3 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in Kontrollgruppe} \end{cases}$$

$$\Longleftrightarrow$$

$$\begin{aligned} y_{ij} =& (\beta_0 + \gamma_{0i}) + \beta_1 N_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} \\ & + \gamma_{1i} \cdot t_{ij} + \varepsilon_{ij} \end{aligned} \tag{8.2}$$

- γ_{0i} stellen die individuellen Abweichungen vom Populationsmittel β_0, γ_{1i} die individuellen Abweichungen von den Populationsparametern β_1, β_2 und β_3 dar.

- Im Gegensatz zu den „fixen“ Effekten $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \beta_3)$ werden die individuen-spezifischen Effekte $\boldsymbol{\gamma}_i^T = (\gamma_{0i}, \gamma_{1j})$ als zufällige Größen (‘random effects’) angesehen, da die Ratten eine Zufallsauswahl aus einer Population darstellen.
- Man trifft die spezifische Annahme, dass die zufälligen Effekte unabhängig normalverteilt sind:

$$\gamma_{0i} \sim \mathcal{N}(0, \tau_0^2), \qquad \gamma_{1i} \sim \mathcal{N}(0, \tau_1^2) \tag{8.3}$$

- Für die Messfehler ε_{ij} wird in diesem Beispiel angenommen, dass sie $iid\, \mathcal{N}(0, \sigma^2)$ sind.
- Da das Modell (8.2) neben den Parametern aus Modell (8.1) auch die zufälligen Effekte $\boldsymbol{\gamma}_i$ enthält, spricht man von einem *gemischten linearen Modell* oder einem *linearen Modell mit zufälligen Effekten.*

Lineares gemischtes Modell für Longitudinal- oder Clusterdaten

Daten: Für $i = m$ Individuen bzw. Cluster werden jeweils n_i zeitlich oder pro Cluster wiederholte Daten $(y_{ij}, \boldsymbol{x}_{ij})$ für eine metrische Zielvariable y mit Kovariablen $x_1, \ldots, x_k$ erhoben.

Modell: Für ein lineares Modell wird angenommen:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} \\ &\quad + \gamma_{0i} + \gamma_{1i} u_{ij1} + \cdots + \gamma_{li} u_{ijl} + \varepsilon_{ij} \\ &\quad \text{mit } i = 1, \ldots, m \text{ und } j = 1, \ldots, n_i. \end{aligned} \tag{8.4}$$

- $\beta_0, \ldots, \beta_k$: feste Populationsparameter
- $\gamma_{0i}, \ldots, \gamma_{li}$: individuen- bzw. clusterspezifische Effekte
- die zufälligen Effekte werden als unabhängig und normalverteilt angenommen.

Für Wiederholungsmessungen an Individuum i bzw. Messungen am Objekt j im Cluster i hat man in der Regel

- Zeitpunkte $t_{i1}, \ldots, t_{ij}, \ldots, t_{in_i}$
- die zusätzlichen Designvariablen $u_{ij1}, \ldots, u_{ijl}$ bestehen oft aus einem Teil der Kovariablen $x_{ij1}, \ldots, x_{ijk}$ wie t_{ij} in dem Beispiel.
- Standardmäßig wird für die Fehlervariablen ε_{ij} angenommen, dass sie $iid\, \mathcal{N}(0, \sigma^2)$ seien. Man kann jedoch auch Korrelationen modellieren.
- Analoges gilt für die zufälligen Effekte同样适用于随机效应

Vorteile gemischter Modelle für die Analyse von Longitudinaldaten:

- Sie sind oft plausibler.
- Die Benutzung der individuenspezifischen Information kann im Vergleich zur Schätzung eines einfachen linearen Modells zu einer verbesserten Schätzgenauigkeit führen.
- Individuenspezifische Effekte können als Surrogat für die Effekte von Kovariablen dienen, die in den vorliegenden Daten nicht oder unzureichend gemessen wurden. Man spricht dann von *unbeobachteter Heterogenität* $\rightarrow$ latente Variablen. 潜在变量
- Die geschätzten individuellen Verlaufskurven erlauben auch *individuelle Prognosen*, die in einem herkömmlichen Regressionsmodell nicht möglich wären.
- Information über inter-individuelle Variabilität τ kann an sich von Interesse sein.

Beispiel (contd.)

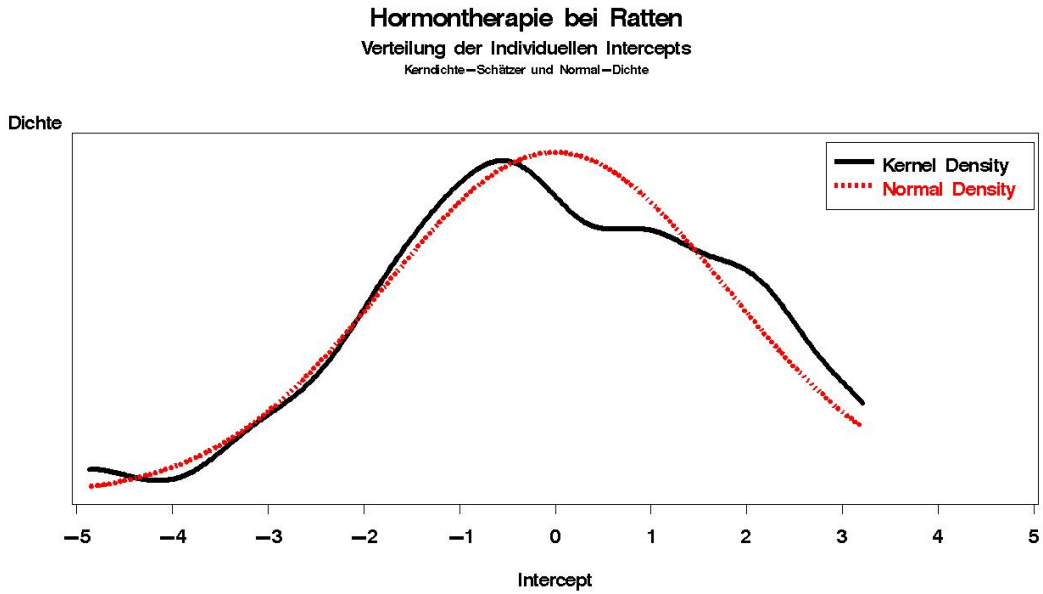
Modelle:

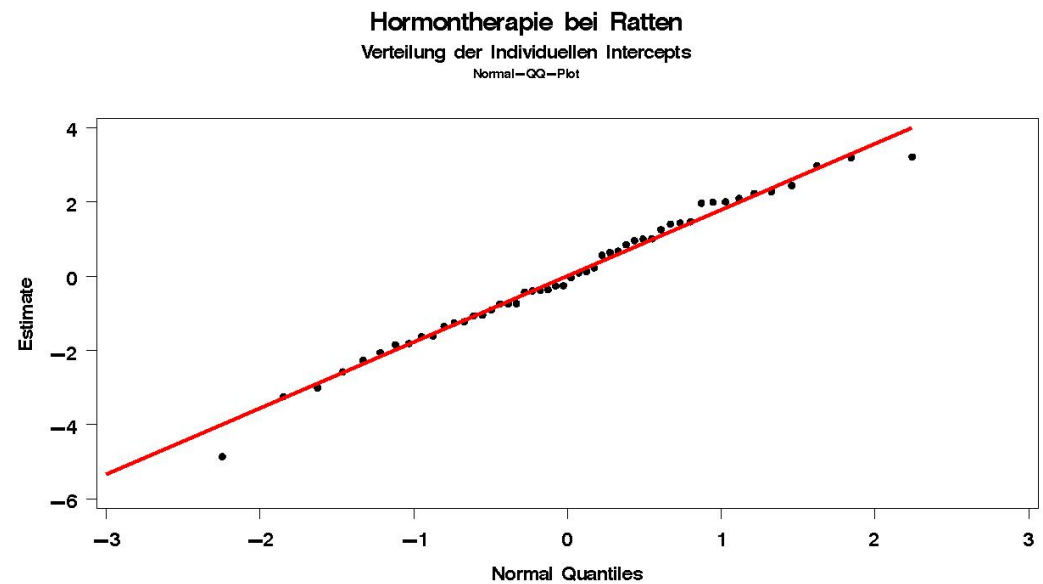
- Modell (8.2): hat individuenspezifische Intercepts (Abweichung von β_0) und individuenspezifische Steigungen (Abweichungen von β_i). Es werden die fixen Effekte $\beta_0 - \beta_3$, die Varianzparameter $\sigma^2, \tau_0^2, \tau_1^2$ (siehe Tabelle) und die zufälligen Effekte geschätzt.
- Da sich zeigt, dass die individuellen Steigungen kaum von den Populationswerten abweichen, wird ein vereinfachtes Modell geschätzt, das keine Terme $\gamma_{1i} t_{ij}$ enthält.
- Für das vereinfachte Modell zeigen die folgenden Grafiken die Verteilung der $\hat{\gamma}_{0i}$ (Kerndichteschätzer und Normal-Quantilplot).

	Parameter	Modell (8.2) Schätzwert	(SE)	vereinfachtes Modell Schätzwert	(SE)
Intercept	β_0	68.606	(0.325)	68.607	(0.331)
Niedrigdosis	β_1	7.503	(0.228)	7.507	(0.225)
Hochdosis	β_2	6.877	(0.231)	6.871	(0.228)
Kontrolle	β_3	7.319	(0.285)	7.314	(0.281)
$\mathbb{V}(\gamma_{0i})$	τ_0^2	3.564		3.565	
$\mathbb{V}(\gamma_{1i})$	τ_1^2	<.001			
$\mathbb{V}(\varepsilon_{ij})$	σ^2	1.445		1.445	

```
proc mixed data=rats01;
     class group ;
     model response= transf_time(group)/ s;
     random int transf_time  /type=un subject=subject  ;
run;

*Vereinfachtes Modell (nur Random Intercept);
    ods output SolutionR=solr;
proc mixed data=rats01;
     class group;
     model response= transf_time(group)/ s;
     random int   /type=un subject=subject s;
run;
```


Hormontherapie bei Ratten
Verteilung der Individuellen Intercepts
Kerndichte−Schätzer und Normal−Dichte
Dichte
Kernel Density
Normal Density
−5
−4
−3
−2
−1
0
1
2
3
4
5
Intercept


Hormontherapie bei Ratten
Verteilung der Individuellen Intercepts
Normal−QQ−Plot
Estimate
4
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3
Normal Quantiles

Lineares gemischtes Modell in Matrixnotation

Wir fassen die Zielvariablen y_{ij} einer Person / eines Clusters i und die zugehörigen Kovariablenvektoren $\boldsymbol{x}_{ij}^T$ und $\boldsymbol{u}_{ij}^T$, $i = 1, \ldots, m$, $j = 1, \ldots n_i$, in Vektoren bzw. Matrizen zusammen:

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1}^T \\ \boldsymbol{x}_{i2}^T \\ \vdots \\ \boldsymbol{x}_{in_i}^T \end{pmatrix}. \quad \boldsymbol{U}_i = \begin{pmatrix} \boldsymbol{u}_{i1}^T \\ \boldsymbol{u}_{i2}^T \\ \vdots \\ \boldsymbol{u}_{in_i}^T \end{pmatrix}, \quad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix}.$$

Dabei kann die Dimension n_i über einzelne Personen bzw. Cluster variieren. $\boldsymbol{X}_i$ und $\boldsymbol{U}_i$ sind $n_i \times p-$ bzw. $n_i \times q-$dimensionale Designmatrizen mit bekannten Kovariablen.

Lineares gemischtes Modell für Longitudinal- oder Clusterdaten

Ein lineares, gemischtes Modell (LMM) für Longitudinal- und Clusterdaten hat die Form

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, m. \tag{8.5}$$

$\boldsymbol{\beta}$ ist ein $p-$dimensionaler Vektor fixer Effekte, $\boldsymbol{\gamma}_i$ ein $q-$dimensionaler Vektor individueller bzw. clusterspezifischer Effekte und $\boldsymbol{\varepsilon}_i$ ein n_i-dimensionaler Fehlerterm. Für $\boldsymbol{\gamma}_i$ und $\boldsymbol{\varepsilon}_i$ gelten die Verteilungsannahmen:

$$\begin{gathered} \boldsymbol{\gamma}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}), \\ \text{und unabhängig davon} \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i). \end{gathered} \tag{8.6}$$

Formulierung als allgemeines lineares gemischtes Modell

Um das obige Modell (8.5) als allgemeines lineares Modell zu formulieren, fasst man die Beobachtungen noch weiter zusammen. Man definiert dazu die Vektoren

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_m \end{pmatrix}$$

und Designmatrizen

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_2 & \cdots & \boldsymbol{0} \\ \vdots & & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{U}_m \end{pmatrix}.$$

Lineares gemischtes Modell (LMM)

Allgemeine Form des *linearen gemischten Modells*:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{8.7}$$

mit

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \right) \tag{8.8}$$

und den blockdiagonalen Kovarianzmatrizen

$$\boldsymbol{G} = \mathrm{Diag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_m),\ \boldsymbol{R} = \mathrm{Diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m). \tag{8.9}$$

Dabei sind $\boldsymbol{X}$ und $\boldsymbol{U}$ Designmatrizen, $\boldsymbol{\beta}$ ein Vektor von fixen Effekten und $\boldsymbol{\gamma}$ ein Vektor von zufälligen Effekten. $\boldsymbol{\gamma}$ und $\boldsymbol{\varepsilon}$ werden als unabhängig angenommen, ihre Kovarianzmatrizen seien positiv definit.

Formulierung als zweistufiges hierarchiches Modell:
Die bedingte Verteilung von $\boldsymbol{y}$, gegeben $\boldsymbol{\gamma}$, ist gegeben durch

$$\boldsymbol{y} \mid \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{U\gamma}, \boldsymbol{R}), \tag{8.10}$$

wobei

$$\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}). \tag{8.11}$$

- Aufgrund der obigen hierarchichen Struktur, bietet sich gerade bei gemischten Modellen ein *Bayesianischer* Ansatz besonders gut an.
- Das Prinzip hierarchicher Modelle läßt sich direkt erweitern, z.B. bei Studien mit Patienten in verschiedenen Krankenhäusern (unterste Ebene), Krankenhäusern in verschiedenen Ländern (zweite Ebene), Länder mit verschiedenen Gesundheitssystemen (dritte Ebene) etc.

Marginales Modell

Das *marginale Modell* für $\boldsymbol{y}$ ergibt sich durch (vgl. Kapitel 6)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*, \quad \boldsymbol{\varepsilon}^* = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Man erhält das *allgemeine lineare Modell*

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{R} + \boldsymbol{U}\boldsymbol{G}\boldsymbol{U}^T) \tag{8.12}$$

bzw.

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}) \\ \boldsymbol{V} &= \boldsymbol{R} + \boldsymbol{U}\boldsymbol{G}\boldsymbol{U}^T. \end{aligned} \tag{8.13}$$

- Man kann zwar aus der Definition des hierarchischen Modells (8.10) das marginale Modell (8.12) ableiten, nicht aber umgekehrt (d.h. aus der marginalen Verteilung von $\boldsymbol{y}$ kann man weder die bedingte Verteilung $\boldsymbol{y} \mid \boldsymbol{\gamma}$ noch die Verteilung von $\boldsymbol{\gamma}$ ableiten.
- Ist man allein an den fixen Effekten $\boldsymbol{\beta}$ interessiert, kann man das marginale Modell benutzen (ähnlich zu Kapitel 6). Dabei können auch andere Kovarianzmatrizen $\boldsymbol{V}$ als solche mit der Struktur aus (8.13) zum Einsatz kommen.
- Ist man aber auch an der Verteilung der individuellen Parametern aus $\boldsymbol{\gamma}$ interessiert, braucht man die zweistufige Darstellung.

Parameterschätzung - Likelihood-Inferenz

Es wird die auf der Likelihood basierende Schätzung von $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ sowie die der Parameter aus $\boldsymbol{G}$ und $\boldsymbol{R}$ behandelt. Bei der Schätzung der zufälligen Effekte spricht man von *Prädiktion*.

(1) Man geht zunächst davon aus, dass man $\boldsymbol{G}$ und $\boldsymbol{R}$ kennt.

- Dann ergibt sich für die Schätzung von $\boldsymbol{\beta}$:

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y} \tag{8.14}$$

- Eine Möglichkeit, $\boldsymbol{\gamma}$ zu schätzen, beruht auf der gemeinsamen Verteilung von $\boldsymbol{\gamma}$ und $\boldsymbol{y}$:

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{\gamma}\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{V} & \boldsymbol{U}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{U}^T & \boldsymbol{G}\end{pmatrix}\right). \tag{8.15}$$

- Der „beste“ Schätzer von γ ergibt sich dann als der auf y bedingte Erwartungswert von γ, der sich berechnet als

$$\mathbb{E}(\gamma \mid y) = GU^T V^{-1}(y - X\beta) \tag{8.16}$$

- Ersetzt man in der letzten Formel β durch $\widetilde{\beta}$, so ergibt sich $\widetilde{\gamma}$ als

$$\widetilde{\gamma} = GU^T V^{-1}(y - X\widetilde{\beta}). \tag{8.17}$$

- Den gleichen Schätzer erhält man auch nach anderen Prinzipien.

(2) Schätzung der Varianzstruktur (G und R)

Zur Schätzung der Varianzparameter wird Maximum Likelihood (ML) oder restringierte Maximum Lieklihood (REML) verwendet. Die unbekannten Parameter aus $\boldsymbol{G}$ und $\boldsymbol{R}$ fassen wir in einen Vektor $\boldsymbol{\delta}$ zusammen und schreiben

$$\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\delta}) = \boldsymbol{U}\boldsymbol{G}(\boldsymbol{\delta})\boldsymbol{U}^T + \boldsymbol{R}(\boldsymbol{\delta}).$$

ML-Schätzung:

- Ausgangspunkt ist Likelihood des marginalen Modells:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}(\boldsymbol{\delta})).$$

- Die Loglikelihood für $\boldsymbol{\beta}$ und $\boldsymbol{\delta}$ ist proportional zu

$$l(\boldsymbol{\beta}, \boldsymbol{\delta}) = -\frac{1}{2}\{\log|\boldsymbol{V}(\boldsymbol{\delta})| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}(\boldsymbol{\delta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \tag{8.18}$$

- Maximieren von l bzgl. $\boldsymbol{\beta}$ für festes $\boldsymbol{\delta}$ liefert

$$\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = (\boldsymbol{X}^T \boldsymbol{V}(\boldsymbol{\delta})^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}(\boldsymbol{\delta})^{-1} \boldsymbol{y}$$

- Einsetzen von $\widetilde{\boldsymbol{\beta}}$ in l liefert den Wert der *Profil-Loglikelihood*:

$$l_P(\boldsymbol{\delta}) = -\frac{1}{2}\{\log|\boldsymbol{V}(\boldsymbol{\delta})| + (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta}))^T \boldsymbol{V}(\boldsymbol{\delta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta}))\} \tag{8.19}$$

- Maximieren von $l_P(\boldsymbol{\delta})$ bzgl. $\boldsymbol{\delta}$ liefert den ML-Schätzer $\widehat{\boldsymbol{\delta}}_{ML}$

REML-Schätzung:

- Als Alternative zu $l_P(\boldsymbol{\delta})$ kann man zur Schätzung von $\boldsymbol{\delta}$ auch die marginale Loglikelihood $l_R(\boldsymbol{\delta})$ benutzen (restricted maximum likelihood estimation)

$$l_R(\boldsymbol{\delta}) = \log\left(\int L(\boldsymbol{\beta}, \boldsymbol{\delta}) d\boldsymbol{\beta}\right). \tag{8.20}$$

- Man zeigt, dass

$$l_R(\boldsymbol{\delta}) = l_P(\boldsymbol{\delta}) - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\delta})^{-1}\boldsymbol{X}|. \tag{8.21}$$

- Maximieren von $l_R(\boldsymbol{\delta})$ liefert den *restringierten ML-Schätzer* $\widehat{\boldsymbol{\delta}}_{REML}$.

Bei beiden Ansätzen liefert Einsetzen der Schätzer für $\boldsymbol{\delta}$ Schätzungen der Kovarianzmatrizen

$$\widehat{\boldsymbol{R}} = \boldsymbol{R}(\widehat{\boldsymbol{\delta}}),\ \widehat{\boldsymbol{G}} = \boldsymbol{G}(\widehat{\boldsymbol{\delta}}) \quad \text{und damit } \widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\delta}})$$

(3) Zusammenfassen:

- Den Schätzer für den Varianzparameter $\boldsymbol{\delta}$ erhält man via ML oder REML, dabei hat REML bei kleinem Stichprobenumfang oft bessere Eigenschaften.
- Die Schätzer für die fixen und zufälligen Effekte erhält man, indem man die geschätzten Kovarianzmatrizen einsetzt:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{y} \\ \widehat{\boldsymbol{\gamma}} &= \widehat{\boldsymbol{G}} \boldsymbol{U}^T \widehat{\boldsymbol{V}}^{-1} (\boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}}). \end{aligned} \tag{8.22}$$

- Beim Einsetzen der Schätzungen für $\boldsymbol{G}$, $\boldsymbol{R}$ und $\boldsymbol{V}$ wird deren Variabilität nicht berücksichtigt. Deshalb werden die Kovarianzmatrizen von $\widehat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\gamma}}$ tendenziell unterschätzt.
- Zur Konstruktion von Konfidenzintervallen benutzt man, dass $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ *asymptotisch* normalverteilt ist.
- Hypothesentests über Parameter führt man über Wald-Tests oder über LQ-Tests durch.

Beispiel (contd):

Die folgende Tabelle zeigt neben den Schätzungen des vereinfachten linearen gemischten Modells auch die Schätzungen unter einem parametrischen Modell ohne Berücksichtigung zufälliger Effekte:

	Parameter	vereinfachtes Modell Schätzwert	(SE)	parametrisches Modell Schätzwert	(SE)
Intercept	β_0	68.607	(0.331)	68.687	(0.348)
Niedrigdosis	β_1	7.507	(0.225)	7.677	(0.286)
Hochdosis	β_2	6.871	(0.228)	6.529	(0.284)
Kontrolle	β_3	7.314	(0.281)	7.212	(0.326)
$\mathbb{V}(\gamma_{0i})$	τ_0^2	3.565			
$\mathbb{V}(\varepsilon_{ij})$	σ^2	1.445		4.730	

Während die Parameterschätzungen für die fixen Effekte sich kaum unterscheiden, ist der Unterschied in ihren Standardabweichungen beträchtlich. Dies liegt an der dreimal so großen Fehlervarianz des parametrischen Modells (das muss nicht immer so sein!).