Lineare Modelle

Vorlesung: Prof. Dr. Helmut Küchenhoff
Skriptbearbeitung: Juliane Manitz

Stand: 22. September 2009

Vorbemerkungen

Das ist ein Skript der Vorlesung "Lineare Modelle", das auf dem LaTeX-File von Prof. Küchenhoff basiert. Es ist für mich zur besseren Übersicht entstanden und hat sich durch die Vorbereitung auf die Vordiploms-Prüfungen stark erweitert. Zusaetzlich sollte es mir helfen LaTeXdurch Anwendung zu erlernen.

Dabei habe ich auch meine eigenen Gedanken eingebracht, der sich in Form von Kommentaren und Beispielen, die ich aus der Vorlesung mit nach Hause genommen habe, äußern. Weiterhin habe ich für die Vordiplomsvorbereitung hilfreiche Beweise und Erklärungen aus dem Buch von Prof. Toutenburg[1] ergänzt. Zusätzlich erschien es mir sinnvoll, Grafiken einzufügen, um für einzelne Sachverhalte eine bessere Vorstellung erhalten zu können.
Eigene Kommentare sind durch solche kleinen Dreiecke gekennzeichnet: ▷.

Derzeit sind Kapitel 1 bis 9 von Prof. Küchenhoff durchgesehen, korrigiert und ergänzt. Druckt euch auch bitte erstmal nur diese Teile aus, da im Laufe des Sommersemesters 2009 weitere Korrekturen stattfinden werden.

Sollten euch Fehler, weitere hilfreiche Kommentare oder andere Hinweise, Verbesserungsvorschläge etc. einfallen, bin ich euch sehr dankbar. (Bitte an jule@statistiker-wg.de oder persönlich) Die anderen Nutzer werden euch auch sehr dankbar sein.

Also viel Erfolg damit. Ich hoffe es hilft euch etwas.

Viele Grüße, Jule

[1] Toutenburg, Helge: Lineare Modelle. Theorie und Anwendungen. Heidelberg: 2., neu bearbeitete und erweiterte Auflage, Physica-Verlag, 2003.

Contents

1 Das einfache lineare Regressionsmodell

Annahmen:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i;\ i = 1, \ldots, n && (1.1)\\ E(\epsilon_i) &= 0 && (1.2)\\ V(\epsilon_i) &= \sigma^2 && (1.3)\\ \{\epsilon_i \,|\, i = 1,\ \ldots,\ n\} &\quad \text{stoch. unabhängig} && (1.4)\\ \epsilon_i &\sim N(0,\, \sigma^2) && (1.5) \end{aligned}$$

Y_i:	Zielgröße (Zufallsgröße), abhängige Variable
x_i:	**feste** bekannte Einflussgröße, unabhängige Variable
ϵ_i:	Zufallsfehler
β_0, β_1, σ^2:	unbekannte Parameter
n:	Anzahl der Beobachtungen

▷ (1.2) $\Rightarrow E(Y|x) = \beta_0 + \beta_1 x$ und $\Rightarrow E(Y_i|x_i) = \beta_0 + \beta_1 x_i$. Man betrachtet also die bedingte Verteilung $Y|X = x$.
▷ (1.5) ist interessant für kleine Stichproben. Bei großen Stichproben greift der zentrale Grenzwertsatz.

1.1 KQ-Schätzung

Wir betrachten Modell (1.1). Dann der **KQ-Schätzer** (Schätzer nach der Methode der kleinsten Quadrate)

$$(\hat\beta_0,\, \hat\beta_1) = \arg\min_{\beta_0,\,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{1.6}$$

▷ Er berechnet für welche β_0, β_1 die Summe minimal ist. D.h. die Abstände der tatsächlichen Werte zu der Regressionsgerade sollen minimal werden. (Dabei kann der Abstand in y-Richtung, x-Richtung oder der geometrische Abstand relevant sein. Hier wird der Abstand in y-Richtung betrachtet.)

$$\hat\epsilon_i := Y_i - \hat\beta_0 - \hat\beta_1 x_i \quad \text{heißen } \textbf{Residuen.} \tag{1.7}$$

Der KQ-Schätzer **existiert** und ist **eindeutig**, (falls $\sum_{i=1}^n (x_i - \bar x)^2 \neq 0$):

$$\begin{aligned} \hat\beta_1 &= \frac{S_{xY}}{S_x^2} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar x)(Y_i - \bar Y)}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar x)^2} && (1.8)\\ \hat\beta_0 &= \bar Y - \hat\beta_1 \bar x. && (1.9) \end{aligned}$$

Beweis:

Eine notwendige Bedingung für die Existenz eines Minimums der quadratischen Funktion $(\hat\beta_0,\, \hat\beta_1) = \arg\min_{\beta_0,\,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$ ist das Vorliegen einer Nullstelle der partiellen Ableitungen 1. Ordnung nach β_0 und β_1.

1. Bestimmung der partiellen Ableitung 1. Ordnung von (1.6):

$$(I) \qquad \frac{d}{d\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \frac{d}{d\beta_0} (Y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^{n} Y_i - \beta_0 - \beta_1 x_i$$

$$(II) \qquad \frac{d}{d\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \frac{d}{d\beta_1} (Y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i$$

2. Normalengleichung durch Nullsetzen:

$$(I) \quad \Rightarrow \sum_{i=1}^{n} Y_i - \beta_0 - \beta_1 x_i = 0 \Rightarrow \; n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$(II) \quad \Rightarrow \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i = 0 \Rightarrow \; \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i$$

3. Berechnung von $\hat{\beta}_0$ durch Multiplikation von (I) mit $\frac{1}{n}$:

$$\Rightarrow \; \beta_0 + \beta_1 \bar{x} = \bar{y} \Rightarrow \; \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

4. Einsetzen von $\hat{\beta}_0$ in (II) um $\hat{\beta}_1$ zu bestimmen:

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Wegen $\sum_{i=1}^{n} (x_i - \bar{x})^2 = S_{XX}$ und $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = S_{XY}$ folgt: $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$

□

1.1.1 Eigenschaften des KQ-Schätzers

Durch Differenzieren von $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$ erhält man: $(\hat{\beta}_0, \hat{\beta}_1)$ sind Lösung der **Normalgleichungen**

$$\sum_{i=1}^{n} \hat{\epsilon}_i \;\; = \;\; 0 \tag{1.10}$$

$$\sum_{i=1}^{n} \hat{\epsilon}_i x_i \;\; = \;\; 0 \tag{1.11}$$

▷ Der Mittelwert der Residuen (geschätzte Abweichung) ist also immer Null. Es ist also ein Trugschluss nach der Schätzung anzunehmen, wenn der EW Null ist, dass die Regressionsgerade stimme. Sie ist und bleibt eine Schaetzung
▷ Also sind die $\hat{\epsilon}_i$ nicht stochastisch unabhängig.
▷ Jede Regressionsgerade läuft wegen $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$ durch den Punkt $(\bar{x}, \bar{y})$.

Gegeben sei Modell (1.1) mit Annahme (1.2).

a. Dann ist $(\hat{\beta}_0, \hat{\beta}_1)$ ein **erwartungstreuer** Schätzer für (β_0, β_1):

$$E(\hat{\beta}_0, \hat{\beta}_1) = (\beta_0, \beta_1). \tag{1.12}$$

b. Für die **Varianzen** von $(\hat{\beta}_0, \hat{\beta}_1)$ gilt unter den Annahmen (1.3), (1.4):

$$V(\hat{\beta}_1) \quad = \quad \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \;\; = \;\; \frac{\sigma^2}{n S_x^2} \tag{1.13}$$

$$V(\hat{\beta}_0) \quad = \quad \sigma^2 \left[\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{n S_x^2} \right] \tag{1.14}$$

$$\text{mit} \quad S_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

▷ Je größer die Streuung S_x^2 der x-Werte ist, desto genauer ist die Schätzung von β_1. ▷ Je größer der Stichprobenumfang ist, desto genauer ist die Schätzung von β_0 und β_1.

c. Unter der NV-Annahme (1.5) ist der KQ-Schätzer $(\hat{\beta}_0, \hat{\beta}_1)$ **ML-Schätzer**.

Beweis:

(a) Es ist zu beachten, dass die x_i fest sind und dass die einzige stochastische Komponente des Modells ε_i ist.

$$\begin{aligned}
E(\hat{\beta}_1) &= E\left[\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n} E[(\beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1\bar{x} - \bar{\varepsilon})(x_i - \bar{x})] \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + E[\varepsilon_i] - \beta_0 - \beta_1\bar{x} - E[\bar{\varepsilon}])(x_i - \bar{x})] \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}\beta_1(x_i - \bar{x})(x_i - \bar{x}) = \beta_1 \\
E(\hat{\beta}_0) &= E[\bar{Y} - \hat{\beta}_1\bar{x}] = E[\beta_0 + \beta_1\bar{x} + \bar{\varepsilon}] - \beta_1\bar{x} = \beta_0
\end{aligned}$$

(b) Übung: Aufgabe 1a

(c) Die Likelihood von Beobachtungen (y_i, x_i) lautet:

$$\begin{aligned}
L(y_i, x_i) &= \prod_{i=1}^{n}\left(\sqrt{2\cdot\pi\cdot\sigma^2}\right)^{-1}\exp\left[-\frac{[\varepsilon_i(\beta_0,\beta_1)]^2}{2\sigma^2}\right] \\
\ln L(y_i, x_i) &= -n/2\cdot\ln(\sigma^2\cdot 2\cdot\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2
\end{aligned}$$

Da die Parameter β_0 und β_1 nur in $\sum_{i=1}^{n}\varepsilon_i^2$ vorkommen, entspricht die Maximierung von $\ln L(y_i, x_i)$ der Minimierung von $\sum_{i=1}^{n}\varepsilon_i^2$. Damit entspricht die KQ-Methode der ML-Methode.

□

1.1.2 Schätzung von σ^2 und Konfidenzintervalle für β_0 und β_1

Gegeben sei das Modell (1.1) bis (1.4).

1. Dann ist
$$\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \tag{1.15}$$
ein **erwartungstreuer** Schätzer für σ^2

▷ Als ML-Schätzung von σ^2 ergibt sich:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

▷ Dieser Schätzer ist nicht erwartungstreu und wird er selten verwendet. Statt dessen werden 2 FG beachtet, da zwei geschätzte Parameter β_0 *und* β_1 genutzt werden. Das ergibt dann die Korrektur (n-2)

2. Unter der Normalverteilungsannahme (1.5) gilt:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 \quad \sim \quad \chi^2_{n-2} \tag{1.16}$$

$$(\hat{\beta}_0,\, \hat{\beta}_1) \text{ und } \hat{\sigma}^2 \qquad \textbf{stochastisch unabhängig} \tag{1.17}$$

▷ Aus $\sum_{i=1}^{n}\hat{\epsilon}_i = 0$ und $\sum_{i=1}^{n}\hat{\epsilon}_i x_i = 0$ ergeben sich die $(n-2)$ Freiheitsgrade der χ^2-Verteilung

3. Unter (1.5) gilt für die Schätzer $\hat{\beta}_1$ und $\hat{\beta}_0$:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \quad \sim \quad t_{n-2} \text{ mit } \hat{\sigma}_{\hat{\beta}_1} := \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1.18}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \quad \sim \quad t_{n-2} \text{ mit } \hat{\sigma}_{\hat{\beta}_0} := \sqrt{\hat{\sigma}^2\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \tag{1.19}$$

4. Konfidenzintervalle zum Niveau $1-\alpha$ für β_1 und β_0 unter Normalverteilungsannahme (1.5):

$$[\hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} t_{1-\alpha/2}(n-2); \hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} t_{1-\alpha/2}(n-2)] \tag{1.20}$$

$$[\hat{\beta}_0 - \hat{\sigma}_{\hat{\beta}_0} t_{1-\alpha/2}(n-2); \hat{\beta}_0 + \hat{\sigma}_{\hat{\beta}_0} t_{1-\alpha/2}(n-2)] \tag{1.21}$$

$t_{1-\alpha/2}(n-2)$: $1-\alpha/2$-Quantil der $t(n-2)$-Verteilung.

Beweis:

Teil 1 und 2 später als Spezialfall im multiplen Regressionsmodell

Teil 3:

$$\left.\begin{array}{ll} \text{Def. t-Verteilung:} & X_1 \sim N(0;1) \\ & X_2 \sim \chi^2_n \\ & X_1, X_2 \text{ unabh.} \end{array}\right\} \frac{X_1}{\sqrt{\frac{X_2}{n}}} \sim t_n$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right), \text{ da } \quad \hat{\beta}_1 = \sum \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} y_i \sim NV$$

(Summe von unabh. NV ZG)

Aus der Def. der t-Vert. und Teil 2 $\Rightarrow$ Behauptung

Teil 4: Standardkonstruktion von Konfidenzintervallen □

1.1.3 Quadratsummenzerlegung

Gegeben sei das Modell (1.1). Dann gilt:

1.

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSM}} \tag{1.22}$$

mit den **angepassten Größen** $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$

SST:	Sum of Squares Total	Gesamtstreuung von Y
SSE:	Sum of Squares Errors	Streuung der Residuen
SSM:	Sum of Squares Model	Streuung, die das Modell erklärt

▷ SST $\hat{=}$ Streuung auf der y-Achse um $\bar{y}$
▷ SSE $\hat{=}$ Reststreuung; Abweichungen VON der Regressionsgerade
▷ SSM $\hat{=}$ Abweichungen AUF der Regressionsgerade

2.
$$R^2 = \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{1.23}$$
heißt **Bestimmtheitsmaß**. Es gilt
$$R^2 = r_{xY}^2. \tag{1.24}$$
r_{xY} := Korrelationskoeffizient nach Bravais-Pearson.

▷ $R^2 \in [0, 1]$
▷ R^2 beschreibt den Anteil der Varianz, die durch das Regressionsmodell erklärt werden kann, was (1 – den Anteil der nicht erklärten Variabilität) entspricht.

Beweis:

1. Mit dem Nulltrick ergibt sich
$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

2. Man quadriert und summiert beide Seiten:
$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

3. Für das gemischte Glied erhält man dann mit den Normalengleichungen (1.10) und (1.11):
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{\beta}_1(x_i - \bar{x}) = \hat{\beta}_1 \sum_{i=1}^{n} \hat{\varepsilon}_i(x_i - \bar{x}) = \hat{\beta}_1 \left(\sum_{i=1}^{n} \hat{\varepsilon}_i x_i - \bar{x} \sum_{i=1}^{n} \hat{\varepsilon}_i \right) = 0$$

4. Damit ergibt sich:
$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \Leftrightarrow \quad SST = SSE + SSM$$

Merke: Nulltrick als Technik für Quadratsummen-Beweise. □

Bemerkung: Freiheitsgrade
Zu den obigen Quadratsummen wird üblicherweise die Zahl der Freiheitsgrade angegeben. Sie bezeichnet die Anzahl der frei bestimmbaren Summanden der obigen Quadratsummen (bei geg. x_i). Die anderen Summanden ergeben sich aus diesen.

bei SST: $\sum(y_i - \bar{y}) = 0 \Rightarrow df = n - 1$
bei SSE: NGL liefern 2 Restriktionen $\Rightarrow df = n - 2$
$(\sum \hat{\varepsilon}_i = 0; \sum \varepsilon_i x_i = 0)$
bei SSM: $\sum(\hat{y}_i - \bar{y})^2 = \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum(\hat{\beta}_1(x_i - \bar{x}))^2$
$\Rightarrow df = 1$ (durch Wählen von 1 y-Wert liegt $\hat{\beta}_1$ fest)

1.1.4 Prognose

Neben (1.2) - (1.4) betrachten wir eine weitere Beobachtung x_{n+1} mit zugehörigem unbekanntem Y_{n+1}. Der Prognosewert von Y_{n+1} ist gegeben durch

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \tag{1.25}$$

Für den Erwartungswert und die Varianz des Prognosefehlers gilt:

$$E\left(\hat{Y}_{n+1} - Y_{n+1}\right) = 0 \tag{1.26}$$

$$V\left(\hat{Y}_{n+1} - Y_{n+1}\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \tag{1.27}$$

Prognoseintervall für y_{n+1} zum Niveau $1 - \alpha$:

$$[\hat{Y}_{n+1} - \hat{\sigma}_{\hat{Y}_{n+1}} t_{1-\alpha/2}(n-2); \hat{Y}_{n+1} + \hat{\sigma}_{\hat{Y}_{n+1}} t_{1-\alpha/2}(n-2)] \tag{1.28}$$

$$\text{mit} \quad \hat{\sigma}^2_{\hat{Y}_{n+1}} = \hat{\sigma}^2 \left[1 + 1/n + (x_{n+1} - \bar{x})^2 / \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right] \tag{1.29}$$

Beweis:

- Erwartungswert des Prognosefehlers:

$$\begin{aligned} E(\hat{Y}_{n+1} - Y_{n+1}) &= E(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} - (\beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1})) \\ &= \underbrace{E(\hat{\beta}_0 - \beta_0)}_{=0 \ ,da \ E(\hat{\beta}_0)=0} + \underbrace{E((\hat{\beta}_1 - \beta_1)x_{n+1})}_{=0 \ ,da \ E(\hat{\beta}_1)=0} - \underbrace{E(\varepsilon_{n+1})}_{=0} \end{aligned}$$

- Varianz des Prognosefehlers:

$$\begin{aligned} Var(\hat{Y}_{n+1} - Y_{n+1}) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} - (\beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1})) \\ &= Var((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_{n+1}) + Var(\varepsilon_{n+1}) \\ &= Var(\bar{y} - \hat{\beta}_1(x_{n+1} - \bar{x} + \sigma^2 \\ &= \sigma^2 + \frac{1}{n}\sigma^2 + Var(\hat{\beta}_1(x_{n+1} - \bar{x})^2 \end{aligned}$$

- Prognoseintervall: Standardkonstruktion von Konfidenzintervallen.

□

Bemerkungen:

1. Alle Aussagen gelten nur unter der zentralen Modellannahme des linearen Zusammenhangs von E(Y) und x. (▷ Beispiel des quadratischen Modells)

2. Transformationen sind grundsätzlich möglich. Zu beachten sind dann die geänderte Interpretation der Modellparameter und der Modellannahmen. Insbesondere ist

$$E\left[g(Y)\right] \neq g\left[E(Y)\right]$$

   ▷ Zum Beispiel ist bei $Y_i = \beta_0 + \beta_1 {x_i}^2 + \varepsilon_i \Rightarrow \ ln Y_i = ln\beta_0 + \beta_1 x_i + ln\varepsilon_i$ zu beachten, dass $E(\varepsilon_i) = 0$ nicht mehr gilt.

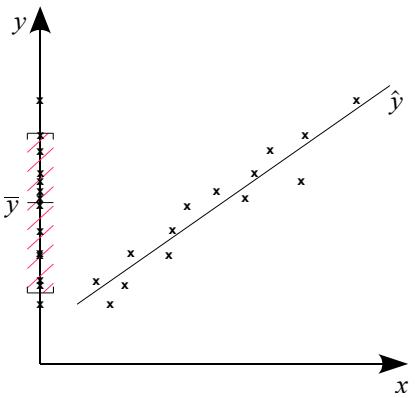
3. Das lineare Modell ist in vielen Beispielen eine sinnvolle Näherung. Die Zusammenhänge in der Realität sind komplexer. (Bsp.: Quadratischer Zusammenhang wird durch ein lineares Modell versucht zu erklären, siehe 1.2.3)

1.2 Beispiele

Dieser Abschnitt wurde von mir mit Beispielen aus der Vorlesung hinzugefuegt.

1.2.1 Erläuterung der Quadratsummenzerlegung:

Angenommen man kennt nur die y-Werte und man möchte damit ein Konfidenzintervall für Normwerte festlegen (z.B. Herzfrequenz). Dann wählt man den Durchschitt $\overline{y}$ als Mitte und legt anhand dessen die $1 - \frac{\alpha}{2}$-Grenzen fest. Damit ergibt sich der SST.

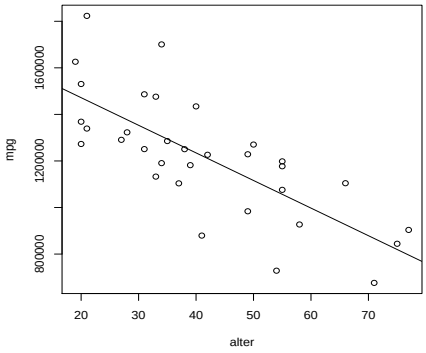


Bekommt man zu jedem y-Wert zusätzlich auch einen x-Wert (z.B. Gewicht), so werden die Normwerte nicht nur als KI angegeben, sondern die Punkte auf der Regressionsgerade. Hier ergibt sich dann also der SSM.

Der SSE ergibt sich dann durch die Streuung der wahren Punkte um die Regressionsgerade.

1.2.2 Beipiel: Osteoporose

Um zu untersuchen, ob Osteoporose eine „moderne“ Krankheit ist, wurden Daten aus Reihengräbern analysiert.



Die zentrale Fragestellung ist also der Zusammenhang zwischen Alter und Knochenbälkchendicke des 4. Lendenwirbels. Dabei sollen Außreißer gefunden werden, die ein Zeichen für Osteoporose sein können.

Im Zentrum der Interpretation stehen die Parameterschätzung von β_1, die Schätzung der durchschnittlichen Abweichung von der Regressionsgerade und R^2.

Probleme können entstehen durch:

- Messfehler (Das Alter kann nur ungenau gemessen werden)
- Die sehr kleine Stichprobe kann eine starke Verschiebung der Gerade durch nur 2-3 Außreißer bewirken.
- x_i ist nicht zufällig

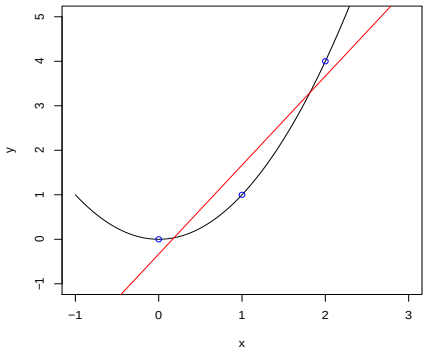
1.2.3 Lineares Modell als sinnvolle Annäherung

Gegeben sei ein quadratisches Modell:

$$Y_i = {x_i}^2 + \varepsilon \text{ mit } E(\varepsilon) = 0, \quad x_1 = 0, x_2 = 1, x_3 = 2 \text{ und } y_1, y_2, y_3$$

Angenommen wird aber ein lineares Modell: **KQ-Schätzung**

$$
\begin{aligned}
E(\hat{\beta}_1) &= \frac{\sum_{i=1}^{3}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{3}(x_i-\bar{x})^2} \\
&= \frac{1}{2}\left(-1E(y_1-\bar{y})+1E(y_3-\bar{y})\right) \\
&= \frac{1}{2}(-0^2+4)=2 \\
E(\hat{\beta}_0) &= E(\bar{y}-\hat{\beta}_1\bar{x}) = \frac{5}{3}\cdot 2\cdot 1 = -\frac{1}{3}
\end{aligned}
$$



▷ Falscher Zusammenhang, aber KQ-Schätzer existiert.

Je nachdem, wie die x-Werte (unter Annahme der Richtigkeit des quadratischen Modells) liegen, bekommt man eine andere Regressionsgerade (bei Annahme eines linearen Zusammenhangs).

- Näherung im Bereich $[0, 2]$ ist akzeptabel (abhängig von der Größe des Störterms).
- Prognose außerhalb von $[0, 2]$ ergibt falsche Ergebnisse. Auch das Prognose-Intervall ist vollkommen falsch: z.B. $x = -1$ ergibt eine Prognose von $\hat{y} = -1$, obwohl $y = 1$ richtig ist.

2 Das multiple lineare Regressionsmodell

2.1 Darstellung

$$\begin{aligned} Y_i &= \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}_{x_i'\beta} + \epsilon_i \; i = 1, \ldots n \\ Y &= X\beta + \epsilon \qquad\qquad x_i' = (1\, x_{i1} \ldots x_{ip}) \end{aligned} \tag{2.1}$$

mit

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{11}, & \cdots & x_{1p} \\ \vdots & & \vdots & \\ 1 & x_{n1} & & x_{np} \end{pmatrix} \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

▷ Interpretation von β_1 *in* $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$:
Y steigt um β_1-Einheiten, falls x_1 um eine Einheit steigt und x_2 festgehalten wird. x_1 muss also vom Einfluss von x_2 bereinigt werden.
▷ Es sind jetzt $(p+1)\beta$-Parameter und σ^2 zu schätzen.

2.2 Modellannahmen

$$\begin{aligned} E(\epsilon_i) &= 0 \\ E(\epsilon) &= \mathbf{0} && (2.2) \\ V(\epsilon_i) &= \sigma^2 && (2.3) \\ \{\epsilon_i \,|\, i = 1, \ldots, n\} & \quad \text{unabh.} && (2.4) \\ \text{Aus (2.3), (2.4) folgt: } V(\epsilon) &= \sigma^2 I \\ \text{Aus } \epsilon_i &\sim N(0, \sigma^2) \text{ und (2.4) folgt:} \\ \epsilon &\sim N(\mathbf{0}, \sigma^2 I) && (2.5) \end{aligned}$$

X: feste Design-Matrix (Matrix der Einflussgrößen)
β: Vektor der Regressionsparameter
Y: Zufallsvektor der Zielgröße
ϵ: Störgrößen

▷ Wenn (2.2) erfuellt ist, ist das Modell als richtig anzusehen.
▷ (2.3) und (2.4) $\Rightarrow \epsilon_i$ sind unkorreliert und haben die gleiche Varianz.

2.3 KQ-Schätzer

Wir betrachten Modell (2.1). Dann heißt

$$\hat{\beta} = \arg\min_{\beta} \underbrace{(Y - X\beta)'(Y - X\beta)}_{\sum_{i=1}^{n}(Y_i - x_i\beta)^2} \tag{2.6}$$

KQ-Schätzer.

$$\hat{\epsilon}_i = Y_i - x_i'\beta \tag{2.7}$$

Es gilt für $(X'X)$ invertierbar: $\hat{\beta}$ **existiert**, ist **eindeutig** und

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{2.8}$$

Der KQ-Schätzer erfüllt die Normalgleichungen:

$$X'\hat{\epsilon} = \mathbf{0} \tag{2.9}$$

Dabei heißt die Matrix $X'X$ **Produktsummenmatrix**. Es gilt:

$$X'X = \begin{pmatrix} n & \sum x_{i1} & \cdots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ip} & \cdots & \cdots & \sum x_{ip}^2 \end{pmatrix} \tag{2.10}$$

2.3.1 Eigenschaften des KQ-Schätzers

Sei das Modell (2.1) mit (2.2) gegeben.

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

$$E(\hat{\beta}) = \beta \tag{2.11}$$

2. Für die Varianz-Kovarianz-Matrix von $\hat{\beta}$ gilt unter (2.3) und (2.4):

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \tag{2.12}$$

3. Unter (2.5) gilt:

$$\hat{\beta} \sim N\left(\beta,\ \sigma^2 (X'X)^{-1}\right) \tag{2.13}$$

Beweis:

1.

$$E(\hat{\beta}) = E((X'X)^{-1}X'Y) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$$

2.

$$Var(\hat{\beta}) = Var((X'X)^{-1}X'Y) = (X'X)^{-1}X'Var(Y)X(X'X)^{-1} = (X'X)^{-1}\sigma^2$$

□

2.4 Hat-Matrix P und Residualmatrix Q

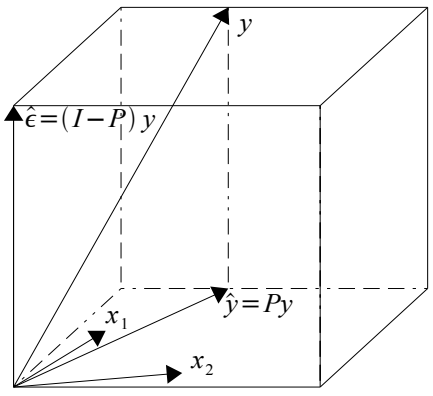
Sei das Modell (2.1) mit einer Designmatrix X mit $rg(X) = p + 1$ gegeben. Es gilt:

$$\hat{Y} = X(X'X)^{-1}X'Y = X\hat{\beta} \tag{2.14}$$

$$P := \underbrace{X(X'X)^{-1}X'}_{n \times n} \tag{2.15}$$

$$\hat{\epsilon} = Y - \hat{Y} = (I - P)Y = QY \tag{2.16}$$

$$Q := I - P \tag{2.17}$$



▷ $\hat{Y} = PY$: P wird Hat-Matrix genannt, weil sie Y den „Hut“ aufsetzt.
▷ $\hat{\epsilon} = QY$, da $QY = (I - P)Y = Y - PY = Y - \hat{Y} = \hat{\epsilon}$
▷ Die Residualmatrix Q bietet also die Abbildung von Y auf $\hat{\epsilon}$, die Residuen.
▷ $\hat{Y}$ kann geometrisch als Projektion auf den Unterraum interpretiert werden, der von x aufgespannt wird (siehe Grafik).

2.4.1 Eigenschaften von P und Q

P heißt **Hat-Matrix** ($\hat{Y} = PY$), Q **Residualmatrix**.

$$P' = P,\ P^2 = P \tag{2.18}$$

$$Q' = Q,\ Q^2 = Q \tag{2.19}$$

$$PQ = QP = \mathbf{0} \tag{2.20}$$

▷ $P' = P \Leftrightarrow$ P sind symmetrisch.
▷ $P^2 = P \Leftrightarrow$ zweimaliges Anwenden der Regression führt zum gleichen Ergebnis, d.h. P ist idempotent.

▷ (2.20) $\Rightarrow$ P und Q sind orthogonal, d.h. sie sind Projektionsmatrizen $\Leftrightarrow$ Anwendung der Regression auf die Residuen liefert $\hat{y} = 0$.

Für die Varianz-Kovarianz-Matrizen von $\hat{Y}$ bzw. $\hat{\epsilon}$ gilt:

$$V(\hat{Y}) = \sigma^2 P \tag{2.21}$$
$$V(\hat{\epsilon}) = \sigma^2 Q \tag{2.22}$$
$$\text{da } \hat{\epsilon} = Q\epsilon \tag{2.23}$$

2.5 Erwartungstreue Schätzung von σ^2

Gegeben sei das Modell (2.1) mit (2.2) bis (2.4). Dann ist:

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n-(p+1)}\sum \hat{\varepsilon}_i^2 \tag{2.24}$$

ein erwartungstreuer Schätzer für σ^2.

▷ (p+1) Freiheitsgrade, da (p+1) Parameter geschätzt wurden.

Bemerkung: Für Projektionsmatrizen gilt allgemein:

$$Sp(P) = rg(P)$$

3 Quadratsummenzerlegung und statistische Inferenz im multiplen linearen Regressionsmodell

3.1 Quadratsummenzerlegung

Gegeben sei das Modell (2.1) mit Design-Matrix X und

$$rg(X) = p + 1 =: p'. \tag{3.1}$$

▷ Da X $(p+1)$ Spalten hat: $X = \begin{pmatrix} 1 & x_{11}, & \cdots & x_{1p} \\ \vdots & & \vdots & \\ 1 & x_{n1} & & x_{np} \end{pmatrix}$

Dann gilt:

$$\underbrace{(Y-\bar{Y})'(Y-\bar{Y})}_{\text{SST}} = \underbrace{(Y-\hat{Y})'(Y-\hat{Y})}_{\text{SSE}} + \underbrace{(\hat{Y}-\bar{Y})'(\hat{Y}-\bar{Y})}_{\text{SSM}} \tag{3.2}$$

▷ SST, SSE und SSM haben die Dimension 1, sind also Skalare.
▷ Die Quadratsummen haben folgende Freiheitsgrade: $SST : n-1,\ SSE : n-p',\ SSM : p.$

Interpretation:

SST	:	Gesamt-Streuung, (korrigierte) Gesamt-Quadratsumme, "Total"
SSE	:	Fehler-Quadratsumme, "Error"
SSM	:	Modell-Quadratsumme, "Model"

Die Zerlegung (3.2) setzt ein Absolutglied (▷ β_0 als Konstante) in der Regression voraus. Eine weitere Zerlegung setzt nicht notwendig ein Absolutglied in der Regression voraus:

$$\underbrace{Y'Y}_{\text{SST*}} = \underbrace{(Y-\hat{Y})'(Y-\hat{Y})}_{\text{SSE}} + \underbrace{\hat{Y}'\hat{Y}}_{\text{SSM*}} \tag{3.3}$$

SST^*	:	nicht korrigierte Gesamt-Quadratsumme Erfasst auch Abweichungen von $Y = 0$ und nicht nur von $\bar{Y}$.
SSE	:	Fehler-Quadratsumme, wie bei (3.2)
SSM^*	:	nicht korrigierte Modell-Quadratsumme

3.1.1 Erwartungswerte der Quadratsummen

Wir betrachten das multiple Regressionsmodell (2.1) mit (2.2) bis (2.4) und

$$P_e = e(e'e)^{-1}e',\ \ Q_e = I - P_e \text{ mit } \ e = (1,1,\ldots,1)'. \tag{3.4}$$

Dann gilt für die Erwartungswerte der Quadratsummen:

$$E(SST^*) = E(Y'Y) = \sigma^2 n + \beta' X'X\beta \tag{3.5}$$
$$E(SST) = E(Y-\bar{Y})'(Y-\bar{Y}) = \sigma^2(n-1) + \beta'(Q_eX)'(Q_eX)\beta \tag{3.6}$$
$$E(SSE) = E(\hat{\varepsilon}'\hat{\varepsilon}) = \sigma^2(n-p') \tag{3.7}$$
$$E(SSM^*) = E(\hat{Y}'\hat{Y}) = \sigma^2 p' + \beta' X'X\beta \tag{3.8}$$
$$E(SSM) = E(\hat{Y}-\bar{Y})'(\hat{Y}-\bar{Y}) = \sigma^2 p + \beta'(Q_eX)'(Q_eX)\beta \tag{3.9}$$

▷ mit $\hat{Y} = PY$ und $(Y-\hat{Y}) = QY$
▷ Die Erwartungswerte der Quadratsummen sind Hauptinstrumente für die Inferenz im Regressionsmodell.

Beweis:
Allgemein berechnet man den Erwartungswert von Quadratischen Formen wie folgt:

$$\begin{aligned} E(Y'AY) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}Y_iY_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}E(Y_iY_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}E(Y_i)E(Y_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}Cov(Y_i, Y_j) \\ &= E(Y)'AE(Y) + Sp(AV(Y)) \\ V(Y) &:= \text{Varianz-Kovarianzmatrix von Y} \end{aligned}$$

Unter Benutzung von $Sp(AV(Y)) = sp(A)\cdot\sigma^2 = rang(A)*\sigma^2$ für Projektionsmatrizen erhält man obige Identitäten. □

Bemerkungen:

1. Die stochastischen Eigenschaften der Quadratsummen werden zur Konstruktion von Tests bezüglich β genutzt
2. $\beta = 0 \Longrightarrow E(Y'Y) = n\sigma^2$, da der hintere Teil wegfällt (3.5).
3. $\beta_1, ...\beta_p = 0 \Longrightarrow E(SSM) = p\sigma^2$, wegen (3.9).
4. Zum Nachweis von 3. benutzt man, daß Q_e der "Mittelwertsbereinigungs- Operator" ist:

   $$Q_e x = (I - P_e)x = x - \bar{x}$$

   Die erste Spalte von Q_eX ist also der Nullvektor:

   $$Q_eX = \begin{pmatrix} 0 & x_{11} - \bar{x}_{\cdot 1} & \cdots & x_{1p} - \bar{x}_{\cdot p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n1} - \bar{x}_{\cdot 1} & \cdots & x_{np} - \bar{x}_{\cdot p} \end{pmatrix}$$

3.1.2 Mittlere Quadratsummen

Wir definieren entsprechend der Zahl der Freiheitsgrade die **mittleren Quadratsummen**:

$$MST^* := \frac{SST^*}{n} \tag{3.10}$$

$$MST := \frac{SST}{n-1} \tag{3.11}$$

$$MSE := \frac{SSE}{n-p'} \tag{3.12}$$

$$MSM^* := \frac{SSM^*}{p'} \tag{3.13}$$

$$MSM := \frac{SSM}{p} \tag{3.14}$$

3.2 Verteilungsdefinitionen

3.2.1 Normalverteilung

Ein n-dimensionaler Zufallsvektor Z heißt multivariat normalverteilt, falls für seine Dichtefunktion gilt:

$$f_z(z) = \frac{1}{\sqrt{\det\Sigma}\sqrt{(2\pi)^n}} \exp\left[-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right] \tag{3.15}$$

mit positiv definiter, symmetrischer Matrix Σ.

Bezeichnung: $Z \sim N_n(\mu, \Sigma)$

Eigenschaften: Ist $Z \sim N_n(\mu, \Sigma)$, so gilt:

1. Momente:

$$\begin{aligned} E(Z) &= \mu \\ V(Z) &= \Sigma \end{aligned}$$

2. Lineare Transformationen:
Ist $A : R^n \longrightarrow R^m$ eine lineare Tranformation mit $rg(A) = m$

$$\Longrightarrow AZ \sim N_m(A\mu, A\Sigma A') \tag{3.16}$$

▷ Beweis über Dichtetransformationssatz

3. Orthogonale Transformation in unabhängige Komponenten
Es existiert eine Matrix $T \in R^{n\times n}$ mit $T'T = I$ und $T\Sigma T' =$ diag $(\lambda_1, \ldots, \lambda_n)$, sodass

$$TZ \sim N_n(T\mu, \text{ diag } (\lambda_1, \ldots, \lambda_n)) \text{ gilt.} \tag{3.17}$$

▷ Σ ist symmetrisch und positiv definit, also auch diagonalisierbar.
▷ T ist orthogonal.
▷ Wenn normalverteilte Zufallsvariablen/-vektoren unkorreliert sind, sind sie auch unabhängig!!

3.2.2 Chi-Quadrat-Verteilung

Ist $Z \sim N_n(\mu, I)$ so heißt $X = Z'Z = \sum_{i=1}^{n} {Z_i}^2$ (nicht-zentral) Chi-Quadrat-verteilt.

Bezeichnung: $X \sim \chi^2(n, \delta)$

n:	Zahl der Freiheitsgrade
$\delta := \mu'\mu = \sum_{i=1}^{n} {\mu_i}^2$:	Nichtzentralitätsparameter

Im Fall $\delta = 0$ erhält man die zentrale $\chi^2(n)$-Verteilung.

Eigenschaften: Ist $X \sim \chi^2(n, \delta)$, so gilt:

1. Momente:

$$\begin{aligned} E(X) &= n + \delta \\ V(X) &= 2n + 4\delta \end{aligned}$$

2. Allgemeiner Bezug zur Normalverteilung

$$Z \sim N_n(\mu, \Sigma) \qquad \Longrightarrow Z'\Sigma^{-1}Z \sim \chi^2(n, \mu'\Sigma^{-1}\mu) \tag{3.18}$$

▷ entspricht im normierten Fall: $X_i \sim N(\mu, \sigma^2) \Rightarrow \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$
▷ Beweis durch Normierung, Dividieren durch die Varianz und orthogonaler Transformation.

3.2.3 t-Verteilung

Seien Z und W voneinander unabhängige Zufallsgrößen mit

$$\begin{aligned} Z &\sim N(\delta, 1), \\ W &\sim \chi^2(n). \end{aligned}$$

Dann heißt $X = \frac{Z}{\sqrt{\frac{W}{n}}}$ (nicht-zentral) t-verteilt.

Bezeichnung: $X \sim t(n, \delta)$

n: Zahl der Freiheitsgrade,
δ: Nicht-Zentralitätsparameter
Im Fall $\delta = 0$ erhält man die zentrale $\chi^2(n)$-Verteilung.

Erwartungswert: Ist $X \sim t(n, \delta)$, so gilt:

$$E(X) \quad = \quad \delta\sqrt{\frac{n}{2}}\,\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \text{ für } n > 1. \tag{3.19}$$

3.2.4 F-Verteilung

Seien W_1 und W_2 voneinander unabhängige Zufallsgrößen mit

$$\begin{aligned} W_1 &\sim \chi^2(n_1, \delta) \\ W_2 &\sim \chi^2(n_2) \end{aligned}$$

Dann heißt $X = \frac{W_1/n_1}{W_2/n_2}$ (nicht-zentral) F-verteilt.

Bezeichnung: $X \sim F(n_1, n_2, \delta)$

n_1: Zählerfreiheitsgrade
n_2: Nennerfreiheitsgrade
δ: Nichtzentralitätsparameter

Erwartungswert: Ist $X \sim F(n_1, n_2, \delta)$, so gilt:

$$E(X) = \frac{n_2}{n_2 - 2}(1 + \delta/n_1) \text{ für } n_2 > 2. \tag{3.20}$$

3.3 Statistische Inferenz im multiplen Regressionsmodell

3.3.1 Satz von Cochran

Sei $Z \sim N(\mu, \Sigma), dim Z = n, A \in R^{n\times n}$, $A' = A$ und $rg(A) = r$, $B \in R^{n\times n}$.
Dann gilt:

$$\mathbf{\Sigma = I, A^2 = A} \implies \mathbf{Z'AZ} \sim \chi^{\mathbf{2}}(\mathbf{r}, \mu'\mathbf{A}\mu) \tag{3.21}$$

$$\mathbf{A, B \in R^{n\times n}, \Sigma = I, AB = 0} \implies \mathbf{Z'AZ} \text{ und } \mathbf{Z'BZ} \text{ sind unabh.} \tag{3.22}$$

$$\mathbf{B \in R^{n\times n}, \Sigma = I, BA = 0} \implies \mathbf{BZ} \text{ und } \mathbf{Z'AZ} \text{ sind unabh.} \tag{3.23}$$

▷ (3.21) $\Rightarrow$ $A = A^k$, also ist A idempotent.

Allgemeiner Fall (brauchen wir später):

$$A\Sigma A = A \implies Z'AZ \sim \chi^2(r, \mu'A\mu) \tag{3.24}$$

$$A, B \in R^{n\times n}, A\Sigma B = 0 \implies Z'AZ \text{ und } Z'BZ \text{ sind unabh.} \tag{3.25}$$

$$B \in R^{n\times n}, B\Sigma A = 0 \implies BZ \text{ und } Z'AZ \text{ sind unabh.} \tag{3.26}$$

3.3.2 Verteilung des KQ-Schätzers unter Normalverteilung

Sei multiple Regressionsmodell (2.1) mit (2.5) und $rg(X) = p'$ gegeben. Für den KQ-Schätzer $\hat{\beta}$ gilt:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \quad (3.27)$$
$$\hat{\Sigma}_{\hat{\beta}} := \hat{\sigma}^2(X'X)^{-1} \quad (3.28)$$
$$(n-p')\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p') \quad (3.29)$$
$$\hat{\sigma} \text{ und } \hat{\beta} \text{ sind unabhängig} \quad (3.30)$$
$$\hat{\sigma}_{\hat{\beta}_k} := \sqrt{c_{kk}}\hat{\sigma}, \ (c_{kk} \text{ entspr. Diagonalelement von} (X'X)^{-1}) \quad (3.31)$$
$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \sim t(n-p', 0) \quad (3.32)$$

▷ mit $\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-p'}$
▷ Normalverteilung des KQ-Schätzers ist besonders für kleine Stichproben interessant. Bei großen Stichproben greift der Zentrale Grenzwertsatz.

Beweis:
(3.27) Normalverteilung von $\hat{\beta}$ folgt aus Eigenschaften der NV
(3.29) folgt aus Theorem von Cochran (1.Teil), da $\hat{\varepsilon} = QY$
(3.30) folgt aus Theorem von Cochran (3.Teil): $B \in R^{n\times n}, \Sigma = I, BA = 0 \Longrightarrow BZ$ und $Z'AZ$ sind unabh.
mit

$$B := (X'X)^{-1}X', A := Q, Z := Y$$
$$BA = (X'X)^{-1}X'Q = (X'X)^{-1}X'(I-P) =$$
$$(X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X' = 0.$$

□

3.3.3 Overall-Tests

▷ Der Overall-Test wird oft zu Beginn der Regression angewendet.
▷ Allgemeine Fragestellung: Ist das Modell überhaupt hilfreich? Ist das Modell geeignet, den Sachverhalt zu beschreiben?

Sei Modell (2.1) mit (2.5) und $rg(X) = p'$ gegeben. Dann gilt für die mittleren Quadratsummen:

$$F_O = \frac{MSM}{MSE} \sim F(p, n-p', \sigma^{-2}\beta'(Q_eX)'(Q_eX)\beta) \quad (3.33)$$
$$F_O^* = \frac{MSM^*}{MSE} \sim F(p', n-p', \sigma^{-2}\beta'(X'X)\beta) \quad (3.34)$$

Die Verteilungen werden zu der Konstruktion folgender Tests benutzt:

Lehne $H_0^O : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ ab, falls

$$F_0 > F_{1-\alpha}(p, n-p') \quad (3.35)$$

▷ entspricht dem Test von $H_0 : Y_i = \beta_0$ mit $\hat{\beta}_0 = \bar{y}$ gegen H_1 : volles Modell mit $y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Overall Test mit $\beta_0 = 0$:

Lehne $H_0^{O*} : \beta_0 = \beta_1 = \cdots = \beta_p = 0$ ab, falls

$$F_0^* > F_{1-\alpha}(p', n-p') \quad (3.36)$$

Dieser Test wird nur in Ausnahmefällen angewendet.

$F_{1-\alpha}(p, n-p')$:$(1-\alpha)$-Quantil der zentralen $F(p, n-p')$–Verteilung.

3.3.4 Wald-Test

Als nächstes Ziel sollen Hypothesen, die sich mit Hilfe linearer Transformationen von β darstellen lassen, betrachtet werden:

$$A \in R^{a\times(p+1)} \quad c \in R^a$$
$$A\beta = c \text{ mit } rg(A) = a$$

Beispiele:

1. Haben x_1, x_2 den gleichen Einfluss?
   $p = 2 \;\; \beta_1 = \beta_2 \leftrightarrow A = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}, c = 0$
2. Haben x_3 *und* x_4 überhaupt einen Einfluss auf Y?
   $p = 4 \;\; \beta_3 = \beta_4 = 0 \leftrightarrow A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
3. Direkter Test: Haben die Parameter bestimmte Werte?
   $p = 4 \;\; \beta_2 = 3 \;\; \beta_3 + \beta_4 = 1 \leftrightarrow A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}, c = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$

Allgemeine lineare Hypothese

Sei das Modell (2.1) mit (2.5) und $A \in R^{a\times p'}, rg(A) = a, c \in R^a$ gegeben.

$$V(A\hat{\beta} - c) = \sigma^2 A(X'X)^{-1}A' \tag{3.37}$$
$$SSH := (A\hat{\beta} - c)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - c) \tag{3.38}$$
$$\sigma^{-2}SSH \sim \chi^2(a, \sigma^{-2}(A\beta - c)'(A(X'X)^{-1}A')^{-1}(A\beta - c)) \tag{3.39}$$
$$MSH := \frac{SSH}{a} \tag{3.40}$$
$$\frac{MSH}{MSE} \sim F(a, n - p', \sigma^{-2}(A\beta - c)'(A(X'X)^{-1}A')^{-1}(A\beta - c)) \tag{3.41}$$

SSH: Quadratsumme, die die Abweichung von der Hypothese $A\beta = c$ beschreibt.

Test nach Wald: $H_0 : A\beta = c$. Lehne H_0 ab, falls:

$$\frac{MSH}{MSE} > F_{1-\alpha}(a, n - p') \tag{3.42}$$

▷ mit $MSH = \frac{SSH}{a} = \frac{\hat{\varepsilon}'\hat{\varepsilon} - \varepsilon'\varepsilon}{a}$ und $MSE = \frac{\varepsilon'\varepsilon}{n-p'}$
▷ F-Verteilung (keine χ^2-Verteilung) wegen der Normierung mit $\frac{1}{\sigma^2}$

Overall-Test und zweiseitiger t-Test
Die Overall-Tests aus (3.35) und (3.36) sind spezielle Wald-Tests.
Der mit Hilfe von (3.32) konstruierte zweiseitige Test auf $\beta_k = \beta_k^0$ ist ebenfalls ein Wald-Test.

3.3.5 Likelihood–Quotienten-Test

Grundidee des Likelihood - Quotienten Tests:
Vergleiche (Bilde den Quotienten) maximierte Likelihood des Modells unter H_0 mit maximierter Likelihood ohne H_0.

ML-Schätzung: Die maximierte Likelihood des Modells ist

$$(2\pi)^{-n/2} \cdot \hat{\sigma}^{-n} \cdot exp(-\sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{2\hat{\sigma}^2})$$

Es folgt (siehe Nachweis ML = KQ aus Kapitel 1)

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n} \tag{3.43}$$
$$MaxL = C \cdot \hat{\sigma}^n = C \cdot (SSE/n)^{(n/2)} \tag{3.44}$$

Wir betrachten also den ML - Schätzer mit und ohne die Restriktion $A\beta = c$:

$\hat{\hat{\varepsilon}}$: Residuen unter dem Modell mit H_0

$\hat{\varepsilon}$: Residuen unter dem Modell ohne Einschränkung

Die LQ- Teststatistik lautet dann:

$$\tau_{LQ} = \left(\frac{\hat{\hat{\sigma}}}{\hat{\sigma}}\right)^{-n} = \left(\frac{\hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}}}{\hat{\varepsilon}'\hat{\varepsilon}}\right)^{-n/2} \sim \chi^2(a) \tag{3.45}$$

Wald-Test ist Likelihood–Quotienten-Test
Sei das Modell (2.1) mit (2.5) und $A \in R^{a \times p'}, rg(A) = a, c \in R^a$ gegeben.

$$H_0 : A\beta = c \text{ gegen } H_1 : A\beta \neq c$$

Dann ist der Wald-Test zu dem Likelihood-Quotiententest äquivalent, d. h. :
Die Testgröße des LQ–Tests ist eine streng monotone Funktion der Testgröße des Wald - Tests:

$$\tau_{LQ} = g\left[\frac{MSH}{MSE}\right] \tag{3.46}$$

g streng monoton

▷ Hier wird mit den Modellabweichungen die Modellanpassung verglichen: SSE mit SSE_{H_0}.

Beweis:
Wir behalten das lineare Modell mit der linearen Restriktion $A\beta = c$.
Wir lösen das Minimierungsproblem

$$(Y - X\beta)'(Y - X\beta) \to \min \quad \text{unter} \quad A\beta = c$$

mit der Lagrange-Methode:

$$S(\beta, \lambda) = (Y - X\beta)'(Y - X\beta) + 2\lambda'(A\beta - c)$$

λ ist der Vektor der Lagrange-Multiplikationen.

$$\begin{aligned}
\frac{\partial S}{\partial \beta} &= 0 \Leftrightarrow -2X'Y + 2X'X\hat{\hat{\beta}} + 2A'\lambda = 0 \\
\frac{\partial S}{\partial \lambda} &= 0 \Leftrightarrow A\hat{\hat{\beta}} = c \\
A'\lambda &= X'Y - X'X\hat{\hat{\beta}} \quad | \cdot (X'X)^{-1} \\
(X'X)^{-1}A'\lambda &= \hat{\beta} - \hat{\hat{\beta}} \quad | \cdot A \; (*) \\
\Rightarrow A(X'X)^{-1}A'\lambda &= A'\hat{\beta} - c \quad (A\hat{\hat{\beta}} = c) \\
\Rightarrow \lambda &= (A(X'X)^{-1}A')^{-1}(A\hat{\beta} - c) \\
(*) \Rightarrow \hat{\hat{\beta}} &= \hat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c) \\
w &:= (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c)
\end{aligned}$$

Darstellung von SSH

$$\begin{aligned}
\hat{\hat{Y}} &= X\hat{\hat{\beta}} = \hat{Y} - Xw \\
\hat{\hat{\varepsilon}} &= Y - \hat{\hat{Y}} = \hat{\varepsilon} + Xw \\
\Rightarrow \hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}} &= \hat{\varepsilon}'\hat{\varepsilon} + w'X'Xw \;\; (\text{da } X'\hat{\varepsilon} = 0) \\
&= \hat{\varepsilon}'\hat{\varepsilon} + (A\hat{\beta} - c)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c) \\
\Rightarrow SSH &= \hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}} - \hat{\varepsilon}'\hat{\varepsilon}
\end{aligned}$$

Damit haben wir eine andere Möglichkeit, SSH aus den Residuenquadratsummen zu berechnen. Als Spezialfall liefert obige Gleichung die schon besprochene Quadratsummenzerlegung.

Damit ist τ_{LQ} eine monotone Funktion der Testgröße des Wald-Tests:

$$\frac{MSH}{MSE} = \frac{(\hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}} - \hat{\varepsilon}'\hat{\varepsilon})/a}{\hat{\varepsilon}'\hat{\varepsilon}/(n-p')} =: \tau_W$$

$$\Rightarrow \left[\tau_W \cdot \frac{a}{(n-p)'} + 1\right]^{-n/2} = \tau_{LQ}$$

□

3.3.6 Reparametrisierung des Modells unter linearer Restriktion

Sei das Modell (2.1) mit (2.5) und $A \in R^{a \times p'}, rg(A) = a, c \in R^a$ gegeben. ($\triangleright\ Y = \beta X + \epsilon$)
Dann gibt es eine **Reparametrisierung** des Modells (2.1)

$$V = Z\gamma + \epsilon \tag{3.47}$$
$$V = Y - Xd \tag{3.48}$$
$$Z = XB \tag{3.49}$$

Das Modell ist das reparametrisierte Modell mit

Zielgröße:	$V = Y - Xd \in R^n$
Design-Matrix:	$Z,\ Z \in R^{n \times (p'-a)};\ \ rgZ = p' - a$
Parameter:	$\gamma \in R^{p'-a}$
Störterm :	ε stimmt mit dem aus dem Grundmodell überein!

Zusammenhang Reparametrisierung und Modell unter linearer Restriktion
Es gilt:

$$\hat{\hat{\beta}} = B\hat{\gamma} + d \tag{3.50}$$
$$SSH = \hat{\hat{\varepsilon}}'\hat{\hat{\varepsilon}} - \hat{\varepsilon}'\hat{\varepsilon} \tag{3.51}$$

mit

$\hat{\hat{\beta}}$:	KQ-Schätzer unter Restriktion $A\beta = c$
$\hat{\gamma}$:	KQ-Schätzer aus Modell (3.47)
$\hat{\hat{\varepsilon}}$:	Residuenvektor aus Modell (3.47)= Residuenvektor aus KQ-Schätzung unter Restriktion.

▷ **Beispiel:**
Angenommen wird das Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Mit der Hypothese $H_0 : \beta_1 = \beta_2$ ergibt sich ein neues einfaches lineares Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \varepsilon = \beta_0 + \beta_1(x_1 + x_2) + \varepsilon$$

mit nur noch einer Einflussgröße $(x_1 + x_2)$ und den Parametern β_0 und β_1.

3.4 Sequentielle und Partielle Quadratsummen

Gegeben sei das Modell (2.1). Wir betrachten nun Teilmodelle, die durch Nullrestriktionen von Komponenten des Vektors β entstehen und deren Residuenquadratsummen.

$$R(\beta_{i1}, \ldots, \beta_{ik} | \beta_{j1}, \ldots, \beta_{jl}) = SSH = SSE(M1) - SSE(M2) \tag{3.52}$$

M2:	Modell, das die Parameter $\beta_{i1}, \ldots, \beta_{ik}, \beta_{j1}, \ldots, \beta_{jl}$ enthält.
M1:	Modell, das die Parameter $\beta_{j1}, \ldots, \beta_{jl}$ enthält.
SSH:	Hypothesenquadratsumme zur Hypothese $(\beta_{i1} = \ldots = \beta_{ik} = 0)$ im Modell M2.

▷ Hierbei geht es um den Vergleich von zwei Modellen M_1 und M_2, wobei M_1 Untermodell von M_2 ist.
▷ Die Gesamtabweichung der beiden Modelle bleibt gleich. Es gilt also:

$$\begin{aligned} M_1 : SST &= SSM(M_1) + SSE(M_1) \\ M_2 : SST &= SSM(M_2) + SSE(M_2) \end{aligned}$$

3.4.1 Partielle Quadratsummen

Die zu der Hypothese $\beta_i = 0$ gehörigen Quadratsummen bzgl. des Gesamtmodells heißen **partielle Quadratsummen**:

$$R(\beta_i|\beta_0, \ldots \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p) = SSE(M_{-i}) - SSE \;\; (= SSM - SSM(M_{-i})) \tag{3.53}$$

M_{-i}: Modell mit $H_0 : \beta_i = 0$.

▷ Hierbei wird das volle Modell mit dem um ein x_i reduziertes Modell verglichen.
▷ Die partiellen Quadratsummen ist der Zähler der F-Statistik zum Testen von $\beta_i = 0$.
▷ entsprechen den Typ III-Quadratsummen in SAS

3.4.2 Sequentielle Quadratsummen

Wir betrachten die Folge von Modellen:

$$M_0 : \qquad Y = \beta_0 + \epsilon \tag{3.54}$$
$$M_1 : \qquad Y = \beta_0 + \beta_1 x_1 + \epsilon \tag{3.55}$$
$$\ldots$$
$$M_p : \qquad Y = X\beta + \epsilon \tag{3.56}$$

$$R(\beta_k|\beta_0, \ldots, \beta_{k-1}) = SSE(M_{k-1}) - SSE(M_k) \tag{3.57}$$

heißen **sequentielle Quadratsummen** und es gilt:

$$SST = \sum_{k=1}^{p} R(\beta_k|\beta_0, \ldots, \beta_{k-1}) + SSE \tag{3.58}$$

▷ Die sequentiellen Quadratsummen messen die Verbesserung von SSM nach Hinzunahme einer weiteren Variable x_k zu den bereits vorliegenden $x_1, \ldots, x_{k-1}$
▷ Sie beantworten die Frage nach dem Einfluss der Variable von x_k.
▷ entsprechen den Typ I-Quadratsummen in SAS und SPSS

3.4.3 ▷ Beispiel

$$\begin{aligned} y_A &= \beta_0 + \beta_1 x_1 + \epsilon_A \\ y_B &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_B \\ y_C &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_C \\ \epsilon_A \geq \epsilon_B \geq \epsilon_C \;\; &\Rightarrow \;\; SSE_A \geq SSE_B \geq SSE_C \;\Leftrightarrow\; SSM_A \leq SSM_B \leq SSM_C \end{aligned}$$

Die sequentiellen Quadratsummen messen wie groß nun der Unterschied zwischen den Fehler-Quadratsummen ist. Je kleiner der quadratische Fehler, desto größer der Teil, der durch das Modell erklärt wird.

3.5 Bereinigung

Wir betrachten folgende Zerlegung des Modells (2.1):

$$\begin{aligned} Y &= X\beta + \epsilon && (3.59)\\ Y &= (X_1X_2)(\beta_1'\beta_2')' + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon && (3.60)\\ \hat{\beta} &= (X'X)^{-1}X'Y = (\hat{\beta}_1', \hat{\beta}_2')' && (3.61)\\ Q_2 &:= I - X_2(X_2'X_2)^{-1}X_2' && (3.62)\\ Y^* &:= Q_2Y && (3.63)\\ X_1^* &:= Q_2X_1 && (3.64)\\ \Longrightarrow \quad \hat{\beta}_1 &= (X_1^{*\prime}X_1^*)^{-1}X_1^{*\prime}Y^* && (3.65) \end{aligned}$$

Y^*, X_1^* : von X_2 bereinigte Variablen

▷ Q_2 ist hierbei der Bereinigungsoperator

Beispiele zur Bereinigung Mittelwerts-Bereinigung

$$\begin{aligned} X_2 &= (1 \dots 1)'\\ X_1 &= (x_1 \dots x_n)'\\ Y^* &= Y - \bar{y}\\ X_1^* &= X_1 - \bar{x}\\ (X'^*X^*)^{-1}X'^*Y^* &= S_x^{-2}S_{xy} \end{aligned}$$

Geschlechtseffekt, Trendbereinigung

Beipiel: Analyse von Fehleranzahlen in Tests zu starken Verben

Zielgröße :	Anzahl der Fehler bei Test
Einflussgrößen:	Geschlecht Alter Leseverhalten, Fernsehverhalten, etc.

Regression auf binäre Variable Geschlecht entspricht Mittelwertsschätzung
Bereinigung nach Geschlecht: Abziehen des jeweiligen Gruppenmittelwertes

3.6 Simultane Konfidenzintervalle

▷ **Bemerkung:**

Die Konfidenzintervalle für die geschätzten Parameter β_i werden aus der t-Verteilung hergeleitet (3.32)

$$\beta_i \pm \hat{\sigma}_{\hat{\beta}_i} t_{1-\frac{\alpha}{2}}(n-p')$$

Vielleicht hat man auch Interesse an Linearkombinationen $\gamma_j = \alpha'\beta$ der β_i oder man möchte mehrere β_i oder γ_j betrachten.
Dann halten die obrigen Konfidenzintervalle das Sicherheitsniveau α nicht gleichzeitig ein. Deshalb verwendet man silmutane Konfidenzintervalle.

3.6.1 Bonferroni-Konfidenzintervalle

Gegeben sei das Modell (2.1) mit der Normalverteilungsannahme (2.5) Dann sind für die Parameter $\beta_{i1}, \dots, \beta_{ik}$

$$\hat{\beta}_{il} \pm \hat{\sigma}_{\hat{\beta}_{il}} t_{1-\frac{\alpha}{2k}}(n-p'),\ l = 1, \dots, k$$

simultane Konfidenzintervalle zum Sicherheitsniveau $1-\alpha$:

$$P(\text{es gibt l} : |\beta_{il} - \hat{\beta}_{il}| > \hat{\sigma}_{\beta_{il}} t_{1-\frac{\alpha}{2k}}(n-p')) \leq \alpha$$

Bemerkung:

Die Parameter β_{il} können durch Linearkombinationen $\gamma_l = a'\beta$ mit den entsprechenden geschätzten Standardabweichungen ersetzt werden.

▷ Aus (3.66) folgt $P(\beta_i \in [\hat{\beta}_i \pm \hat{\sigma}_{\hat{\beta}_i} t_{1-\frac{\alpha}{2k}}(n-p')]) \leq \frac{\alpha}{k}$. Das entspricht der Ungleichung von Bonferroni (3.67). Auf Tests übertragen heißt das, dass ein multipler Test, der k Mittelwertsvergleiche zum Niveau von $\frac{\alpha}{k}$ durchführt.
▷ einfache Adjustierung von α
▷ benutzt man bei wenigen Variablen β_i oder γ_i
▷ sehr konserativ (d.h. bei der Verwendung ist man immer auf der richtigen Seite, aber es ist sehr unflexibel)
▷ berücksichtigt nicht, dass $\hat{\beta}_i$ oder $\hat{\gamma}_i$ korreliert sind.

3.6.2 Konfidenzintervalle nach Scheffé

Sei Modell (2.1) mit NV-Annahme (2.5) gegeben.
Dann sind

$$\hat{\beta}_j \pm \sqrt{p' F_{1-\alpha}(p', n-p')}\ \hat{\sigma}_{\hat{\beta}_j} \tag{3.66}$$

für die Parameter β_j und

$$\hat{\gamma} \pm \sqrt{p' F_{1-\alpha}(p', n-p')}\ \hat{\sigma}_{\hat{\gamma}}$$

für beliebige Linearkombinationen $\gamma = a'\beta$ simultane Konfidenzintervalle.

Bemerkungen:

Die KIs nach Scheffé sind z.B zur Bestimmung von simultanen Konfidenzregionen für Y geeignet

▷ Die KIs nach Scheffé sind bei einer Vielzahl von Parametern sinnvoll, da sie für alle möglichen β_i und γ_i gleichzeitig gelten
▷ Es gibt in Analogie zu den KI's einen entsprechenden multiplen Test über die F-Verteilung
▷ berücksichtigt, dass $\hat{\beta}_i$ oder $\hat{\gamma}_i$ nicht unabhängig sind.
▷ ist ein Bayesianischer Ansatz.

3.6.3 Konfidenzellipsoide

Sei das Modell (2.1) mit NV-Annahme (2.5) gegeben.
Dann ist

$$\left\{\beta \,|\, (\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta}) \leq p'\hat{\sigma}^2 F_{1-\alpha}(p', n-p')\right\} \tag{3.67}$$

eine Konfidenzregion für β.
Entsprechendes gilt für lineare Transformationen $\gamma = A\beta$:

$$\left\{\gamma \,|\, (\gamma-\hat{\gamma})'\widehat{V(\hat{\gamma})}^{-1}(\gamma-\hat{\gamma}) < \dim\gamma \cdot F_{1-\alpha}(\dim(\gamma), n-p'))\right\}$$

ist Konfidenzregion zum Sicherheitsniveau $1-\alpha$.

▷ Angenommen Y und X sind die Einheitsmatrizen, dann ergibt sich der Einheitskreis als Konfidenzellipsoid. Ist dies nicht der Fall, sonst ergibt sich eine Ellipse.

3.7 Eigenschaften des KQ-Schätzers

3.7.1 Das Gauss-Markov-Theorem

Sei das Modell

$$\begin{aligned} Y &= X\beta+\epsilon, \quad \text{rg } X = p' \\ E(\epsilon) &= 0 \\ V(\epsilon) &= \sigma^2 I \qquad\qquad \triangleright \text{Unabhängigkeit der Störterme} \end{aligned}$$

gegeben.

▷ Dieses Modell hat keine Normalverteilungsannahme.

Dann ist der KQ-Schätzer $\hat{\beta}$ unter den erwartungstreuen linearen Schätzern derjenige mit der kleinsten Varianz: $\hat{\beta}$ ist BLUE-Schätzer (**b**est **l**inear **u**nbiased **e**stimator).

Ist $\tilde{\beta}$ ein weiterer Schätzer von β mit $E(\tilde{\beta}) = \beta$ und $\tilde{\beta} = CY$, so gilt:

$$V(\tilde{\beta}) \geq V(\hat{\beta})$$

$V(\tilde{\beta}) \geq V(\hat{\beta}) \Leftrightarrow V(\tilde{\beta}) = V(\hat{\beta}) + M$ mit M positiv semidefinit (d.h. $\forall a \neq 0 : a'Ma \geq 0$).

▷ Der Schätzer $\hat{\beta}$, der durch die KQ-Schätzung entsteht, ist also besser als jeder Schätzer $\tilde{\beta}$, der die Vorraussetzungen der Erwartungstreue und Linearität erfüllt.

3.7.2 Konsistenz des KQ–Schätzers

Sei das Modell (2.1) - (2.4) gegeben (die Normalverteilungsannahme ist also nicht notwendig). Wir betrachten nun das Modell mit steigendem Stichprobenumfang n. Da die Einflussgrößen fest sind, gehen wir von einer gegebenen Folge x_n der Einflussgrößen aus. Sei zu jedem $n > p'$

X_n : Designmatrix, die aus den ersten n Beobachtungen besteht
$\hat{\beta}^{(n)}$: KQ- Schätzer aus den ersten n Beobachtungen

Vor.: X_n hat vollen Rang für alle $n \geq p$

$$\lim_{n\to\infty} (X_n'X_n)^{-1} = 0.$$

▷ Es wird also vorausgesetzt, dass der Schätzer mit steigenden n immer mehr am wahren Wert liegt.

Dann folgt die schwache Konsistenz des KQ-Schätzers (Konvergenz in Wahrscheinlichkeit):

$$\hat{\beta}^{(n)} \xrightarrow{P} \beta. \tag{3.68}$$

Sind die Störgrößen zusätzlich identisch verteilt, so folgt die starke Konsistenz (fast sichere Konvergenz)

$$\hat{\beta}^{(n)} \xrightarrow{\text{f.s.}} \beta. \tag{3.69}$$

3.7.3 Asymptotische Normalität des KQ–Schätzers

Sei das Modell (2.1) - (2.4) gegeben. Sei zu jedem $n > p'$

X_n : Designmatrix, die aus den ersten n Beobachtungen besteht.
$\hat{\beta}^{(n)}$: KQ- Schätzer aus den ersten n Beobachtungen
Vor.: X_n hat vollen Rang für alle $n \geq p$

$$\lim_{n\to\infty} \max x_i'(X_n'X_n)^{-1}x_i = 0.$$

▷ Vorraussetzung ist also, dass das Gewicht von den Einzelwerten in der Schätzung mit steigenden n gegen Null geht.

Dann folgt die asymptotische Normalität von β

$$(X_n'X_n)^{1/2}(\hat{\beta}^{(n)} - \beta) \xrightarrow{d} N(0, I). \tag{3.70}$$

▷ Wenn man also große Stichprobenumfänge hat, kann auf die Normalverteilungsannahme verzichtet werden (ähnlich dem Gesetz der großen Zahlen).

4 Modelle mit diskreten Einflussgrößen

▷ Die kodierten Merkmalsausprägungen (z.B. Geschlecht: männlich 1, weiblich 2) können nicht wie reelle Zahlen in die Berechnung der Parameterschätzungen einbezogen werden, da diese nicht unbedingt einer Ordnung unterliegen und die Abstände nicht definiert sind. Deshalb müssen kategoriale Regressoren umkodiert werden.

▷ typische Beispiele für kategoriale Regressoren sind:

- Geschlecht: weiblich, männlich
- Familienstand: ledig, verheiratet, geschieden, verwitwet
- Prädikat für Diplomzeugnisse: sehr gut, gut, befriedigend, ausreichend
- Standort der Börse: New York, Tokio, Frankfurt
- Aktientyp: Standard, New Economy

Wir betrachten ein nominales Merkmal C mit K Ausprägungen.

a) **Einfache Dummy Codierung**

$$Z_k(C) = \begin{cases} 1 \text{ für } C = k; \\ 0 \text{ für } C \neq k; \end{cases} \quad k = 1, \ldots, K \tag{4.1}$$

b) **Effekt-Codierung**

$$Z_k^e(C) = \begin{cases} 1 & \text{für } C = k; \\ 0 & \text{für } C \neq k, C \neq K \\ -1 & \text{für } C = K. \end{cases} \quad k = 1, \ldots, K-1; \tag{4.2}$$

▷ mit K als Referenzkategorie

4.1 Einfache Varianzanalyse

▷ Ziel der Varianzanalyse ist die Untersuchung, ob es Unterschiede in den Erwartungswerten der einzelnen Gruppen gibt. Dies lässt sich auch als „Einfluss“ des diskreten Merkmals auf die Zielgröße interpretieren.

Gegeben sei eine nominale Einflussgröße C mit K Ausprägungen (Gruppen).
Der Zielgrößenvektor Y wird in die K Gruppen mit jeweils n_k Beobachtungen aufgeteilt:

$$Y = (Y_{11}, \ldots Y_{1n_1}, Y_{21}, \ldots, Y_{Kn_K})$$

▷ Im Weiteren verwenden wir folgendes **Beispiel**:

$$Y = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \varepsilon$$

▷ Modell mit einem Faktor mit $K = 3$ Ausprägungen und je $n_i = 2$ Wiederholungen

▷μ Gesamtmittelwert (Over-all-mean)
▷τ_k Abweichung des Mittelwertes der k-ten Gruppe vom Gesamtmittelwert ($\mu_k = \mu + \tau_k$).

▷ Für dieses Modell können keine Parameter geschätzt werden, da $X'X$ nicht vollen Rang hat und somit nicht invertierbar ist. Die Parameter werden aber mit $\hat{\beta} = (X'X)^{-1}X'Y$ geschätzt.

4.1.1 Mittelwertsmodell

$$Y_{kl} = \mu_k + \varepsilon_{kl} \qquad l = 1, \ldots, n_k;\ k = 1, \ldots, K$$

$$Y = (Z_1(C) \ldots Z_K(C)) \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} + \varepsilon \tag{4.3}$$

▷ Y_{kl} ist die Gruppenangehörigkeit.
▷ k ist der Laufindex für die Gruppenzugehörigkeit.
▷ l ist der Laufindex für die Wiederholungen der Beobachtungen für festen Faktor.

Beispiel

Design-Matrix X für $K = 3$ Gruppen mit je $n_k = 2$ Beobachtungen pro Gruppe:

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Gruppenmittelwert μ_k.

▷ Hier wird die erste Spalte weggelassen
$\Rightarrow$ lineare Unabhängigkeit der Spalten $\Rightarrow$ $(X'X)$ ist invertierbar $\Rightarrow$ Eindeutigkeit der KQ-Schätzung.

▷ Man schätzt für jede Faktorstufe einen eigenen

4.1.2 Effektkodierung

$$Y_{kl} = \mu + \tau_k + \varepsilon_{kl}; \qquad \sum_{k=1}^{K} \tau_k = 0$$

$$Y = (e \ \ Z_1^e(C) \ldots Z_{K-1}^e(C)) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K-1} \end{pmatrix} + \varepsilon \tag{4.4}$$

Beispiel

Design-Matrix X für $K = 3$ Gruppen mit je $n_k = 2$ Beobachtungen pro Gruppe.

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}.$$

▷ μ ist der ungewichtete Gesamtmittelwert über die Gruppen.
▷ τ_k ist die Abweichung des Gruppenmittelwertes vom Gesamtmittelwert.
▷ τ_K ergibt sich durch die anderen τ's:
$\tau_K = -\sum_{k=1}^{K-1} \tau_k$

▷ Eindeutigkeit der Schätzung erreicht man, indem man die τ's durch $\sum_{k=1}^{K} \tau_k = 0$ einschränkt. Dadurch sind nur noch $(n-1)$ τ's "frei wählbar".

Bemerkung:

- Eine Alternative Effektkodierung entsteht durch die Einschränkung der τ's durch $\sum_{k=1}^{K} n_k \tau_k = 0$.
- n_k: Stichprobenumfänge der einzelnen Gruppen
- Hier werden die Gruppen entsprechend den Stichprobenumfängen in den Gruppen gewichtet. Es ergibt sich μ als gewichteter Gesamtmittelwert über die Gruppen.

4.1.3 Modell mit Referenzkategorie K:

$$\begin{aligned} Y_{kl} &= \mu_K + \tau_k + \varepsilon_{kl} \;\; \tau_K = 0; \\ Y &= (e \;\; Z_1(C) \dots Z_{K-1}(C)) \begin{pmatrix} \mu_K \\ \tau_1 \\ \vdots \\ \tau_{K-1} \end{pmatrix} + \varepsilon \end{aligned} \tag{4.5}$$

▷ Als Referenzgruppe K wird im Allgemeinen die letzte Gruppe gewählt.

Beispiel

Design-Matrix X für 3 Gruppen mit je 2 Beobachtungen pro Gruppe:

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

▷ τ_k ist hier der Unterschied des Gruppenmittelwerts der k-ten Gruppe zum Gruppenmittelwert der Referenzgruppe K.
▷ μ_K entspricht somit dem Gruppenmittelwert der Referenz K.

4.1.4 Nullhypothesen zum Test auf "Effekt von C":

Mittelwertsmodell:	$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$
Effektkodierung:	$H_0 : \tau_1 = \tau_2 = \ldots = \tau_{K-1} = 0$
Referenzmodell:	$H_0 : \tau_1 = \tau_2 = \ldots = \tau_{K-1} = 0$

▷ Hat die Gruppenzugehöhrigkeit C einen Einfluss?
▷ H_0 : C hat keinen Effekt. Wird H_0 abgelehnt, so ist der Effekt von C signifkant.

4.2 Modell der zweifaktoriellen Varianzanalyse

Wir betrachten zwei diskrete Einflussgrößen C und D mit K_1 bzw. K_2 Ausprägungen. Man spricht dann von einer zweifaktoriellen Varianzanalyse mit einem K_1-stufigen und einem K_2-stufigen Faktor

4.2.1 Modell mit einfachen Effekten (Effektdarstellung)

$$Y = (e \;\; Z_1^e(C) \dots Z_{K_1-1}^e(C) \;\; Z_1^e(D) \dots Z_{K_2-1}^e(D)) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K_1-1} \\ \gamma_1 \\ \vdots \\ \gamma_{K_2-1} \end{pmatrix} \tag{4.6}$$

Test auf Effekt von C:

$$H_0 : \tau_1 = \ldots = \tau_{K_1-1} = 0$$

Test auf Effekt von D:

$$H_0 : \gamma_1 = \ldots = \gamma_{K_2-1} = 0$$

Designmatrix X für Modell mit einem zweistufigen und einem dreistufigen Faktor und jeweils zwei Beobachtungen pro Faktorkombination

$$X = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}.$$

4.2.2 Modell mit Interaktion

Interaktionen lassen sich durch Aufnahme aller Produktterme $Z_k^e(C)Z_l^e(D)$ modellieren:

$$Y = (e, Z_1^e(C) \dots Z_{K_2-1}(D)Z_1^e(C) \cdot Z_1^e(D) \dots Z_{K_1-1}(C)Z_{K_2-1}(D)) \cdot \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \gamma_{K_2-1} \\ (\tau\gamma)_{11} \\ \vdots \\ (\tau\gamma)_{K_1-1,K_2-1} \end{pmatrix} \quad (4.7)$$

▷ Interpretation einer Interaktion: Die Wirkung von C ist abhängig von dem Wert des Faktor B.
▷ Umsetzung, indem man die Faktoren jeweils miteinander multipliziert und entsprechend eine neue Variablen ins Modell einfügt.

Test auf Interaktion:

$$H_0 : (\tau\gamma)_{11} = \dots = (\tau\gamma)_{K_1-1,K_2-1} = 0$$

▷ Beeinflussen sich die Gruppen wirklich gegenseitig?
▷ Wird H_0 abgelehnt, so interagieren die Gruppen.

Beispiel

Design-Matrix X für 2-Faktor Modell mit einem zweistufigen und einem dreistufigen Faktor: (jeweils eine Beobachtung pro Merkmalskombination).

$$X\beta = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \tau_1 \\ \gamma_1 \\ \gamma_2 \\ (\tau\gamma)_{11} \\ (\tau\gamma_{12} \end{pmatrix}$$

▷ 2. Spalte · 3. Spalte = 5. Spalte
▷ 2. Spalte · 4. Spalte = 6. Spalte

4.2.3 Zwei-Faktor-Modell mit Referenz-Kategorie

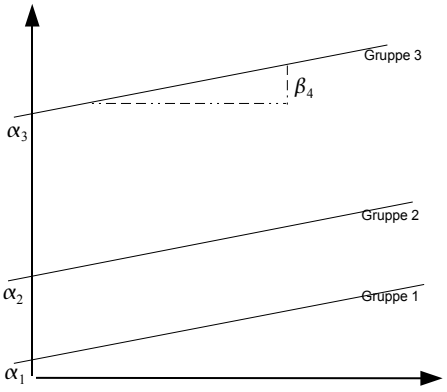
$$X\beta = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$$

μ: Der Mittelwert in Kategorie (2,3).
τ_1: Unterschied zwischen beiden Gruppen des ersten Merkmals.
γ_1, γ_2 Unterschied zwischen zwei verschiedenen Gruppen zu Referenzgruppe des zweiten Merkmals.

4.3 Erweiterung auf Kombination von diskreten und stetigen Merkmalen (Kovarianzanalyse)

Beispiel für Design-Matrix X für $K = 3$ Gruppen mit je $n_k = 2$ Beobachtungen pro Gruppe und stetigem Merkmal x:

$$X = \begin{pmatrix} 1 & 0 & 0 & x_1 \\ 1 & 0 & 0 & x_2 \\ 0 & 1 & 0 & x_3 \\ 0 & 1 & 0 & x_4 \\ 0 & 0 & 1 & x_5 \\ 0 & 0 & 1 & x_6 \end{pmatrix} \beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_4 \end{pmatrix}$$



Interpretation: In den drei Gruppen drei parallele Geraden mit Achsenabschitt α_i und Steigung β_4

4.3.1 Erweiterung auf Geraden mit versch. Steigung

Modell:

$$Y_{kl} = \alpha_k + \beta_k X_{kl} + \varepsilon_{kl} \tag{4.8}$$

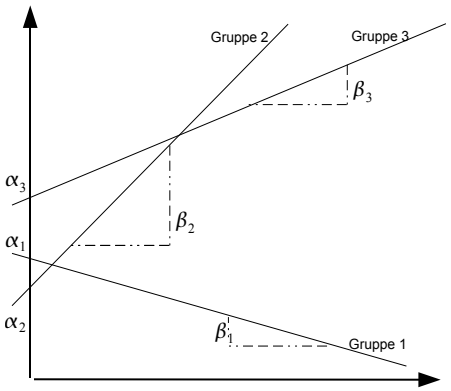
Matrixdarstellung (3 Gruppen 2 Beobachtungen pro Gruppe)

$$X = \begin{pmatrix} 1 & 0 & 0 & x_1 & 0 & 0 \\ 1 & 0 & 0 & x_2 & 0 & 0 \\ 0 & 1 & 0 & 0 & x_3 & 0 \\ 0 & 1 & 0 & 0 & x_4 & 0 \\ 0 & 0 & 1 & 0 & 0 & x_5 \\ 0 & 0 & 1 & 0 & 0 & x_6 \end{pmatrix} \beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Interaktion bedeutet, dass die Steigungen verschieden sind.

▷ Vorsicht bei der Interpretation von Modellen mit Interaktionen. Interaktion bedeutet **nicht**, dass die Variablen selber in Korrelation stehen, sondern nur, dass ihr Einfluss auf die Y-Variable voneinander abhängig ist.

Test auf Interaktion: $H_0 : \beta_1 = \beta_2 = \beta_3$



4.3.2 Darstellung mit Referenzkodierung

Modell:

$$\begin{aligned} Y_{kl} &= \alpha_3 + \alpha_k + \beta_3\, X_{kl} + \beta_k\, X_{kl} + \varepsilon_{kl} (k = 1, 2) \\ Y_{kl} &= \alpha_3 + \alpha_3 X_{kl} + \varepsilon_{kl} (k = 3) \end{aligned}$$

Matrixdarstellung (3 Gruppen 2 Beobachtungen pro Gruppe)

$$X = \begin{pmatrix} 1 & 1 & 0 & x_1 & x_1 & 0 \\ 1 & 1 & 0 & x_2 & x_2 & 0 \\ 1 & 0 & 1 & x_3 & 0 & x_3 \\ 1 & 0 & 1 & x_4 & 0 & x_4 \\ 1 & 0 & 0 & x_5 & 0 & 0 \\ 1 & 0 & 0 & x_6 & 0 & 0 \end{pmatrix} \beta = \begin{pmatrix} \alpha_3 \\ \alpha_1 \\ \alpha_2 \\ \beta_3 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Interaktion bedeutet Steigungen verschieden.
Test auf Interaktion: $H_0 : \beta_1 = \beta_2 = 0$
Interpretation von β_1, β_2 als Unterschied der Steigung zur Referenzkategorie.

5 Behandlung von metrischen Einflussgrößen

▷ Ziel: Flexible metrische Regression: Man möchte so nah wie möglich an die Einfachheit der nicht-parametrischen Regression ($y = f(x) + \epsilon$, wobei f eine glatte Funktion ist) herankommen.

5.1 Einfach linear

$$y = \beta_0 + \beta_1 x + \varepsilon$$

5.2 Transformiert

$$y = \beta_0 + \beta_1 T(x) + \varepsilon$$

Beachte: Andere Interpretation von β_1 z.B.:

Logarithmisch:	$T(x)$	$=$	$\ln(x)$
Logarithmisch mit Nullpunkt-Erhaltung:	$T(x)$	$=$	$\ln(1+x)$
Exponentiell mit bekanntem c:	$T(x)$	$=$	x^c

▷ logarithmische Abhängigkeit von y : $y = \beta_0 + \beta_1 ln(x) + \epsilon$
▷ Interpretation von β_1: Erhöhung von $ln(x)$ um 1 bewirkt eine Erhöhung von y um β_1 (also multiplikativer Faktor)

5.3 Als Polynom

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots \beta_k x^k$$

▷ $X = \begin{pmatrix} 1 & x_1 & {x_1}^2 & {x_1}^3 \\ 1 & x_2 & {x_2}^2 & {x_2}^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & {x_n}^2 & {x_n}^3 \end{pmatrix}$
ergibt ein Polynom dritten Grades, das durch Taylor approximierbar ist und endlich viele Nullstellen hat.

Problem: Bestimmung von k

▷ Vorsicht: Die Interpretation der Schätzwerte für β ändert sich komplett.

5.4 Stückweise konstante Funktion

$$y = \begin{cases} \beta_0 & \text{für} \quad x \leq x_0 \\ \beta_1 & \text{für} \quad x_0 < x < x_1 \\ \vdots & \\ \beta_h & \text{für} \quad x > x_{h-1} \end{cases}$$

Dies entspricht der Kategorisierung der x-Variablen.

▷ Hierbei zwingt man den Daten stückweise ein konstantes Modell auf
▷ Problem: Wie groß sind die Kategorien? Welche Abschnitte können zusammengefasst werden?

5.5 Stückweise linear

$$y = \beta_0 + \beta_1 x + \beta_2 (x - g_1)_+ + \beta_3 (x - g_2)_+ + \ldots \beta_h (x - g_k)_+$$

mit bekannten Knoten g_k und $t_+ = max(t, 0)$.

▷ Hierbei zwingt man den Daten stückweise ein lineares Modell auf
▷ Angenommen eine Funktion hat Hochpunkte in g_1 und g_2 (Knoten):

$$y = \begin{cases} a_0 + b_0 x & \text{für} x \leq g_1 \\ a_1 + b_1 x & \text{für} g_1 \leq x \leq g_2 \\ a_2 + b_2 x & \text{für} x \geq g_2 \end{cases}$$

und Stetigkeit, d.h. $a_0 + b_0 g_1 = a_1 + b_1 g_1$ und $a_2 + b_2 g_2 = a_1 + b_1 g_2$

5.6 Regressionssplines

▷ Erweiterung: Nicht nur stetige Funktion, sondern stetig diffbare Funktion

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - g_1)_+^3 + \beta_5 (x - g_2)_+^3$$

Polynom 3. Grades 2 x stetig differenzierbar da x^3 2 x stetig differenzierbar in 0.

▷ Knoten g_i sind dabei bekannt.
▷ Splines sind ein sehr flexibles Werkzeug, wo man auch noch weitere Größen mit dazu nehmen kann. Sie bilden eine wichtige Grundlage der nichtparametrischen Regression.

5.7 Trigonometrische Polynome zur Modellierung von periodischen Termen (Saisonfigur)

Beispiel:

$$y = \beta_0 + \beta_1 \sin(\frac{2\pi}{T} \cdot x) + \beta_2 \cos(\frac{2\pi}{T} \cdot x) + \beta_3 \sin(\frac{2\pi}{T} \cdot 2x) + \beta_4 \cos(\frac{2\pi}{T} \cdot 2x)$$

T: Periodenlänge, x: Zeit
Alternative: Saison- Dummy (Indikator) Variablen
Beachte:

$$A_1 \cdot cos(x) + A_2 \cdot sin(x) = A_3 \cdot sin(x + \phi)$$

Beispiel: Trendmodell für die Populationsgröße von Füchsen in Baden-Württemberg

Gegeben: Sogenannte Jagdstrecken Y = Anzahl der geschossenen Füchse als Indikator für die Populationsgröße

Modelle:

$$\begin{aligned} ln(Y) &= \beta_0 + \beta_1 t + \beta_2 * t^2 \\ ln(Y) &= \beta_0 + \beta_1 t + \beta_2 * t^2 + \beta_3 t^3 + \beta_4 (t - 70)_+^3 + \beta_5 (t - 85)_+^3 \\ &etc. \end{aligned}$$

Versuchen Sie eine Modellierung von ln (Hase) !!

6 Probleme bei der Regression und Diagnose

Gegeben sei das multiple Regressionsmodell (2.1) und (2.5):

$$Y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0, \sigma^2 I)$$

Es geht darum herauszufinden, ob das Modell zur Analyse der jeweiligen Daten geeignet ist. Es werden also die Modellannahmen ueberprueft. Da sich diese auf die Störterme beziehen, werden die typischerweise die Residuen betrachtet. Beachte, dass sich die Annahmen nicht auf die Randverteilung von Y beziehen.

6.1 Verschiedene Typen von Residuen

▷ Residuen sind die Schätzungen für die Abweichungen vom Regressionsmodell, wobei Störterme die Abweichungen von dem wahren Modell sind.
▷ Residuen und Störterme sind nicht gleichzusetzen:

Beachte: Die Residuen ergeben sich aus den (unbekannten) Störtermen ε durch

$$\hat{\varepsilon} = Q\varepsilon = (I - X(X'X)^{-1}X')\varepsilon = (I - P)\varepsilon$$

Daher gilt:

$$Var(\hat{\varepsilon}_i) = q_{ii}\sigma^2$$

▷ Hier gibt es wegen der Idempotenz von Q einen linearen Zusammenhang: $Var(\hat{\epsilon}_i) = QIQ' = Q^2\sigma^2 = Q\sigma^2$

Beispiel:

$$X := \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \quad \Rightarrow \quad Q := \begin{bmatrix} 2/5 & -2/5 & -1/5 & 0 & 1/5 \\ -2/5 & \frac{7}{10} & -1/5 & -1/10 & 0 \\ -1/5 & -1/5 & 4/5 & -1/5 & -1/5 \\ 0 & -1/10 & -1/5 & \frac{7}{10} & -2/5 \\ 1/5 & 0 & -1/5 & -2/5 & 2/5 \end{bmatrix}$$

▷ Die Regressionsgerade wird durch Ausreißer und Punkte mit starken Einfluss leicht verschoben.

6.1.1 Standardisierte Residuen:

$$r_i := \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \quad i = 1, \ldots, n \tag{6.1}$$

▷ Die Residuen ($Var(\hat{\varepsilon}_i) = q_{ii}\sigma^2$) werden mit $\sigma\sqrt{q_{ii}}$ standardisiert: $r_i := \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{q_{ii}}}$.
▷ Wenn man Residuen betrachtet, sollte man grundsätzlich die standardisierten Residuen benutzen (z.B. bei Plots), um die Vergleichbarkeit zu gewährleisten.

6.1.2 Studentisierte Residuen:

Problem: Bei der Schätzung von σ geht das Residuum mit ein. Dies kann insbesondere bei kleinen Stichproben ein Problem sein. Daher definiert man:

$$r_i^* := \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{q_{ii}}} \quad i = 1, \ldots, n \tag{6.2}$$

$\hat{\sigma}_{(i)}$:= Schätzung von σ **ohne** die Beobachtung i.

▷ Studentisierte Residuen basieren auf der Schätzung der Regression ohne die vermutliche Ausreißer-Beobachtung
▷ Sehr aufwendig, da das Modell n-mal geschätzt wird.
▷ Vergleich mit den standardisierten Residuen: Wenn $r_i^* < r_i \Rightarrow\ x_i$ könnte Ausreißer sein.

6.1.3 Rekursive Residuen:

Bei Zeitreihen verwendet man häufig:

$$\omega_i := \frac{y_i - x_i' \hat{\beta}_{[i-1]}}{\sqrt{1 - x_i'(X_{[i-1]}' X_{[i-1]})^{-1} x_i}} \quad i = p' + 1, \ldots, n \tag{6.3}$$

$\hat{\beta}_{[i-1]}$: Schätzung von β aus den ersten $i-1$ Beobachtungen

$X_{[i-1]}$: X-Matrix der ersten $i-1$ Beobachtungen

▷ Rekursive Residuen entsprechen der Vorstellung der sequentiellen Quadratsummen.

6.1.4 Kreuzvalidierungs - Residuen:

▷ Wie ist die Prognose für die einzelnen Werte?

$$e_{(i)} := y_i - x_i' \hat{\beta}_{(i)} \tag{6.4}$$

$\hat{\beta}_{(i)}$: Schätzung von β ohne Beobachtung i

PRESS: (Predicted Residual Sum of Squares)

$$PRESS := \sum_{i=1}^{n} e_{(i)}^2$$

Es gilt:

$$e_{(i)} = e_i / (1 - h_{ii})$$

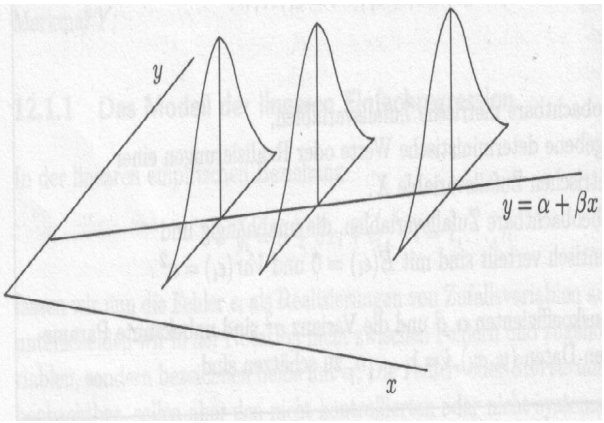
▷ Die Regressionsgerade wird unabhängig von der i-ten Beobachtung geschätzt. Diese Beobachtung wird dadurch prognostiziert. ϵ_i ist dann die Abweichung zwischen wahren Wert und der Prognose.
▷ $1 - h_{ii}$ sind die Diagonalelemente der Q-Matrix
▷ $Var(\epsilon_{(i)})$ ist die Abweichung der i-ten Beobachtung und es gilt: $Var(\epsilon_{(i)}) > \sigma^2$, da der Prognosefehler zum Tragen kommt.
▷ Kreuzvalidierungs-Residuen entsprechen etwa der Vorstellung der partiellen Quadratsummen.

6.2 Diagnose und Therapie von Problemen bei Regression

6.2.1 Die Störterme ϵ_i sind nicht normalverteilt

▷ Betrachtung von $\epsilon_i \sim N(0, \sigma^2)$

Ursachen:
Die Y-Variable stellt eine Zählgröße, eine Überlebenszeit, oder einen Anteil dar. Y ist nicht-negativ etc.



Folgen:
Der KQ-Schätzer $\hat{\beta}$ ist immer noch erwartungstreuer Schätzer mit kleinster Varianz (▷ also BLUE).
Der F-Test ist i. a. robust (▷ d.h. F-Test ist gültig, auch wenn die Modellannahmen nicht exakt erfüllt sind).
Problematisch sind insbesondere bei kleinen Stichprobenumfängen die Konfidenzintervalle der Parameter.
Außerdem sind die Prognoseintervalle nicht mehr gültig, da hierbei die NV-Annahme (▷ in Form der t-Verteilung) besonders eingeht.

Diagnose:
Betrachtung der Schiefe und Kurtosis der Verteilung der Residuen.
Betrachtung von Normal-Plots der standardisierten Residuen
QQ-PLots: ▷ Es sollten etwa 5% der Werte außerhalb von $[-2, 2]$ liegen
▷ 50% sollten links bzw. rechts von Null liegen
▷ Insgesamt sollte der Plot eine Gerade durch den Nullpunkt zeigen (bei Verwendung der standardisierten Residuen eine Winkelhalbierende)

Therapie:
Transformationen der Y-Variablen (▷ z.B. logarithmische Transformation)
Verwendung von generalisierten linearen Modellen.

6.2.2 Heterogene Varianzen (Heteroskedastizität)

Die Varianz der Störterme ε_i ist von i abhängig.

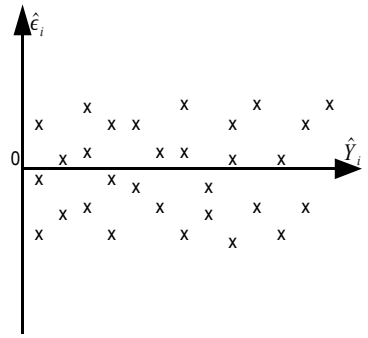
▷ Betrachtung von $\epsilon_i \overset{iid.}{\sim} N(0, \sigma^2)$

Ursachen:
Multplikative Fehlerstruktur, d. h. σ_i ist abhängig von der Größe von Y_i.
Y Zähldaten, Anteile.
Gruppierte Daten führen zu verschieden Residualvarianzen innerhalb der Gruppen.

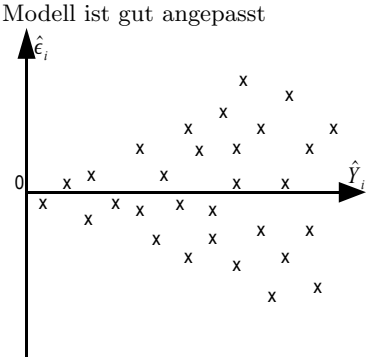
Folgen:
Schätzer für β ist erwartungstreu, aber er hat nicht mehr die kleinste Varianz.
Konfidenzintervalle und Tests für β nicht mehr korrekt

Diagnose:
Residualplot der standardisierten $\hat{\varepsilon}_i$ auf $\hat{Y}_i$ (▷ hier sollte kein Zusammenhang erkennbar sein)
Plot von x_i gegen $\hat{Y}_i$ (typisch für einen linearen Zusammenhang ist eine „Trompetenform“)
Berechnung der Residualvarianzen in den einzelnen Gruppen (bei gruppierten Daten)

Therapie:
Transformation der Y-Variable (▷ z.B. logarithmische Transformation bei Trompetenform des Residualplots)
Gewichtete KQ-Schätzung (siehe Kapitel 8).



Modell ist gut angepasst



Hinweis auf Heteroskedastizität

6.2.3 Korrelation zwischen den Störtermen

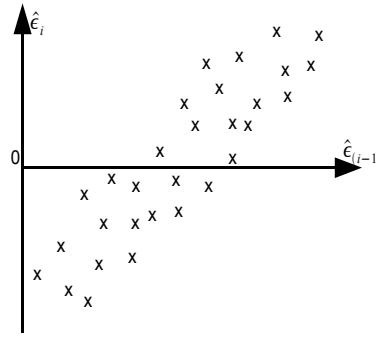
Es gilt $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$ für einige $i \neq j$.

Ursachen:
Zeitreihenstruktur oder räumliche Struktur der Daten führen zu positiver Korrelation von aufeinander folgenden (bzw. nahen) Beobachtungen (z.B. Tiefbohrprojekt).
Residuen bei gruppierten Beobachtungen, bei denen die Gruppenzugehörigkeit nicht zusätzlich modelliert wird, sind häufig positiv korreliert.

Folgen:
Schätzer von β erwartungstreu aber nicht mit geringster Varianz (BLUE).
Bias bei der Varianzschätzung führt zu fehlerhaften Konfidenzintervallen und zu Problemen bei den F-Tests.

Diagnose:
Analyse der Zeitreihenstruktur der Residuen, z.B. mit Durbin–Watson–Test (siehe unten)
Plots der Residuen gegen die Zeit (▷ Gibt es bezüglich des zeitlichen Verlaufs noch eine weitere Struktur (Autokorrelation))
Plots von $\hat{\varepsilon}_i$ gegen $\hat{\varepsilon}_{i-1}$ (▷ entspricht der Korrelation zwischen $\hat{\varepsilon}_i$ und $\hat{\varepsilon}_{i-1}$).

Therapie:
Verwendung von Zeitreihenmethoden
Einbeziehung von (Zeit-)Trend und Saison (▷ z.B. durch periodische Funktionen (z.B. sin) oder zusätzlicher Variable $Jahr$ oder $Jahr^2$)
Gewichtete KQ-Methode



Durbin-Watson-Test

▷ Frage: Gibt es eine zeitliche Struktur in den Daten?

Um zu Testen, ob die Störterme ε_i und ε_{i-1} korreliert sind benutzt man folgende Testgröße:

$$d := \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2} \approx 2(1-\hat{\rho})$$

$\hat{\rho}$: Korrelation zwischen $\hat{\varepsilon}_i$ und $\hat{\varepsilon}_{i-1}$.

Beweis:

$$\begin{aligned} d &:= \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2} = \frac{\sum\limits_{i=1}^{n} \hat{\varepsilon_i}^2 + \sum\limits_{i=1}^{n} \hat{\varepsilon}_{i-1}^2 - 2\sum\limits_{i=1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2} \\ &= \frac{\sum\limits_{i=1}^{n} \hat{\varepsilon_i}^2}{\sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2} + \frac{\sum\limits_{i=1}^{n} \hat{\varepsilon}_{i-1}^2}{\sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2} - 2\frac{\sum\limits_{i=1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum\limits_{i=1}^{n} \hat{\varepsilon_i}^2} \approx 2 - 2\hat{\rho} = 2(1-\hat{\rho}) \end{aligned}$$

weil:

$$\hat{\rho} = \frac{1}{n} \frac{\sum\limits_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\varepsilon})(\hat{\varepsilon}_{i-1} - \bar{\varepsilon})}{\sqrt{Var(\hat{\varepsilon}_i)}\sqrt{Var(\hat{\varepsilon}_{i-1})}} = \frac{\sum\limits_{i=1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum\limits_{i=1}^{n} \hat{\varepsilon_i}^2}$$

□

Lehne H_0: $\rho = 0$ ab, falls $d > d_1$ oder $d < d_2$. (d_1, d_2 sind von p und n abhängige und fest tabellierte Werte).

Kleine Werte von d: positives ρ
Große Werte von d: negatives ρ
$d \approx 2 \longrightarrow$ keine Autokorrelation

▷ Korrelation zwischen $\hat{\epsilon}_i$ und $\hat{\epsilon}_{i-1}$ entspricht dem Plot von $\hat{\epsilon}_i$ gegen $\hat{\epsilon}_{i-1}$.
▷ Autokorrelation: Man geht von einer geordneten Folge von Zufallsvariablen aus. Wenn zwischen den Gleidern der Folge eine Beziehung/Korrelation besteht, spricht man von Autokorrelation.

6.2.4 Ausreißer und Punkte mit starkem Einfluss

Einflussreiche Beobachtungen (high leverage points) sind in den X-Werten weit vom Zentrum der Daten entfernt. Sie können die Regressionsgerade durch ihre große Hebelwirkung leicht verschieben.

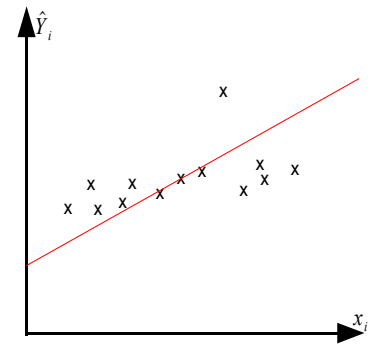
Ausreißer haben dem Betrag nach sehr große Störterme.

Ursachen: Falsche Erhebung, Beobachtung gehört nicht zur Grundgesamtheit, Besonderheiten bei einzelner Untersuchungseinheit

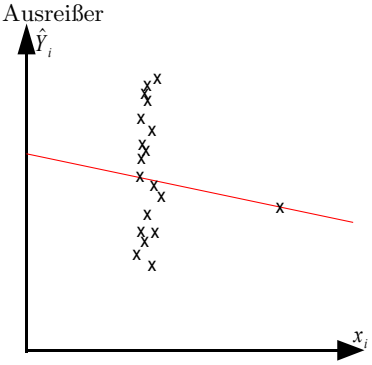
Folgen:
Einflussreiche Beobachtungen wirken stark auf die Schätzung von β. Ausreißer können zu erheblicher Verzerrung der Schätzung von β führen. Dies gilt besonders für Ausreißer, die gleichzeitig high leverage points sind.

Diagnose:
Analyse der Diagonalelemente der Hat-Matrix P zum Auffinden von high leverage points, verschiedene Residuenplots zur Ausreißeranalyse, Influence-Statistiken.

Therapie:
Fehlerhafte Daten weglassen, aber nur wenn entweder gute fachliche Gründe vorliegen (Messfehler, Punkt gehört nicht zur Grundgesamtheit) oder das Weglassen des Punktes verändert die Aussage nicht (Test auf Signifikanz).
Robuste Regression
Gewichtete Regression (siehe Kapitel 8).



Ausreißer



High Leverage Point

Wichtige Einflussmaße

(1) **Leverage:**
Das i-te Diagonalelement der Hat-Matrix P

$$h_{ii} := x_i'(X'X)^{-1}x_i \tag{6.5}$$

heißt **Leverage** der Beobachtung x_i.

Es gilt: $\frac{1}{n} \leq h_{ii} \leq 1$

Normalwert: $h_{ii} = \frac{p'}{n}$
großer Wert: $h_{ii} > \frac{2p'}{n}$
▷ Der Leverage misst, wo die Punkte bezüglich des Zentrums der Daten (x-Werte) liegen: Wie stark schwanken meine Elemente auf der Regressionsgerade?

(2) **Cook's Distanz:**

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{\hat{\sigma}^2 p'} \tag{6.6}$$

$\hat{\beta}_{(i)}$: Schätzung von β ohne Beobachtung i. Es gilt:

$$D_i = \frac{r_i^2}{p'} \cdot \frac{h_{ii}}{1 - h_{ii}} \tag{6.7}$$

$$D_i = \frac{\left(\hat{Y}_{(i)} - \hat{Y}\right)' \left(\hat{Y}_{(i)} - \hat{Y}\right)}{p'\hat{\sigma}^2} \tag{6.8}$$

▷ Wegen $(\hat{\beta}_{(i)} - \hat{\beta})'X = \hat{\beta}_{(i)}X - \hat{\beta}X = (\hat{Y}_{(i)} - \hat{Y})$
▷ In Cook's Distanz geht das Residuum r_i^2 und der Leverage ein.
▷ Normierung durch $(X'X)$, da die Streuung von β ist: $Var(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$
▷ Das Maß D_i ist der mit $\hat{\sigma}^2$ standardisierte Abstand zwischen $\hat{\beta}$ und $\hat{\beta}_{(i)}$ bzw. $\hat{y}$ und $\hat{y}_{(i)}$

(3) **DFFITS:**

$$\text{DFFITS}_i := \frac{\hat{Y}_{(i)i} - \hat{Y}_i}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

DFFITS_i misst den Einfluss der i-Beobachtung auf die Schätzung $\hat{Y}_i$.
Es gilt:

$$D_i = \frac{\hat{\sigma}^2_{(i)}}{p'\hat{\sigma}^2}\text{DFFITS}_i^2.$$

(4) **DFBETAS:**

$$\text{DFBETAS}_{ki} := \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\hat{\sigma}_{(i)}\sqrt{c_{kk}}}$$

misst den Einfluss der Beobachtung i auf einzelne Parameterschätzungen $\hat{\beta}_k$.

(5) **Varianzverhältnis:**

$$\text{COVRATIO}_i = \frac{\det\left(\hat{\sigma}^2_{(i)}\left(X'_{(i)}X_{(i)}\right)^{-1}\right)}{\det\left(\hat{\sigma}^2(X'X)^{-1}\right)}$$

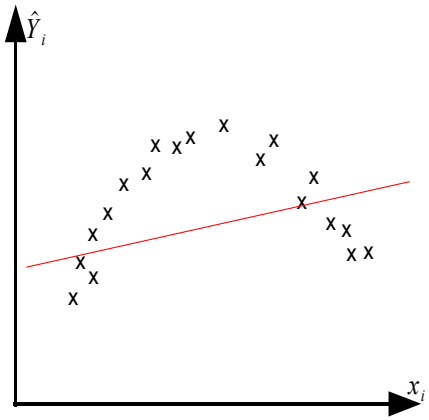
misst die Veränderung der Varianz von $\hat{\beta}$ durch Weglassen der Beobachtung i.

6.2.5 Regressionsgleichung ist nicht korrekt

Die Gleichung $y = X\beta + \varepsilon$ ist fehlerhaft.

Ursachen:
Variable wurden weggelassen oder überflüssigerweise in das Modell einbezogen.
Der Zusammenhang ist nicht linear.
Interaktionen werden nicht in das Modell einbezogen

▷ z.B.: Es wird ein lineares Modell verwendet, obwohl der Zusammenhang quadratisch ist (siehe 1.1.4).



Folgen:
Systematische Fehler bei der Schätzung der Modellparameter und bei der Prognose, aber Modellschätzung liefert häufig brauchbare Näherung

Diagnose:
Im 2-dimensionalen Fall: Scatterplot mit Regressionsgerade.
Im mehrdimensionalen Fall:
Residuenplots $\hat{\varepsilon}_i$ gegen $\hat{y}_i$.
F-Tests auf Einfluss von weiteren Variablen, Interaktionen, Polynomterme höherer Ordnung etc.

Therapie:
Modellerweiterung, Transformationen der Einflussgrößen, Variablenselektionsverfahren.

6.2.6 Partial Leverage Plot

y^* auf x_k^* mit
$y^* := Q_{(k)}y$
$x_k^* := Q_{(k)}x_k$

$Q_{(k)}$: Q-Matrix der Einflussgrößen **ohne** Variable k

$\Longrightarrow$ Darstellung des Zusammenhangs zwischen y und der Einflussgröße x_k unter Berücksichtigung der übrigen Einflussgrößen.

▷ Idee: Der mehrdimensionale Scatterplot ist wegen der Interaktionen der Variablen nicht besonders aussagekräftig.
▷ z.B.: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ mit $x_2 \in \{0, 1\}$.
▷ Hier ist die Bereinigung um den Einfluss der Gruppenvariable (Abziehen der Gruppenmittelwerte; siehe 3.5.) und damit die Verschiebung der Gearde in den Ursprung sinnvoll.

6.2.7 Kollinearität

Die Spalten von X sind (annähernd) linear abhängig.

Ursachen
Hohe Korrelation zwischen den Einflussgrößen, Ungünstiges Versuchs-Design, Codierung von diskreten Variablen

Folgen:
Ungenaue Schätzung von β, häufig sogar falsches Vorzeichen und damit ist auch $\hat{\sigma}_{\hat{\beta}}$ sehr groß
Aber: Konfidenzintervalle korrekt und damit entsprechend groß.

Diagnose:
Analyse der Matrix $(X'X)$ und der Korrelationsmatrix der metrischen Einflussgrößen
Regression der abhängigen Variablen $x_1 = \alpha_0 + \alpha_1 x_2$: R^2 zu groß = gute Anpassung $\Rightarrow$ starker Zusammenhang.

Therapie:
Zusammenfassen bzw. Weglassen von Einflussgrößen (Bei starkem Zusammenhang der Variablen).
Verwendung von anderen Schätzmethoden, z.B.: Ridge-Regression

▷ Extreme Multikollinearität bedeutet geometrisch, dass die zu x_k und x_l gehörenden Datenvektoren auf der gleichen Geraden liegen, also einen Raum der Dimension 1 bilden. Man kann also einen Vektor o.B.d.A. als eine Linearkombination des anderen darstellen.
$\Rightarrow$ Regressionskoeffizient β ist nicht mehr identifizierbar (nicht eindeutig).
▷ Korrelation ist ein hinreichendes Kriterium für Kollinearität. D.h. starke Korrelation kann ein Kollinearitätsproblem andeuten, schwache Korrelation lässt aber nicht den Schluss zu, es existiere keine Kollinearität.

▷ **Beispiel:**

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ mit X_1: Einkommen nach Steuern, X_2: Einkommen vor Steuern.

Hängt X_1 von X_2 ab? Oder anders herum? Wo kommt der Einfluss her?
Angenommen wird eine einfaches System: $x_1 = 0,8x_2 \;\Rightarrow\; X$ hat Rangdefizit
Es folgt:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + 1,25\beta_2 x_1 + \epsilon \\ &= \beta_0 + (\beta_1 + 1,25\beta_2) x_2 + \epsilon \\ &= \beta_0 + (\beta_1 + 1,25\beta) 0,8 x_2 + \epsilon \end{aligned}$$

Problem: Modell ist nicht eindeutig identifizierbar, wenn die Variablen vollständig linear abhängig sind.

▷ Kollinearität ist keine Verletzung der Vorraussetzungen.
▷ Wenn man aber die Möglichkeit hat einen kontrollierten Versuch zu machen, dann kümmert man sich um ein orthogonales Design (der x-Werte), d.h. $x_i \perp x_j \;\forall i \neq j \;\Rightarrow$ alle x sind unabhängig voneinander.

Kollinearitätsdiagnostik

▷ Betrachtung von Zusammenhängen von mehr als zwei Variablen wie z.B. $x_1 \sim \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2$

(1) **Konditionszahl:**

$$K(X) \quad := \quad \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{6.9}$$

$\lambda_{min}, \lambda_{max}$: minimaler, bzw. maximaler Eigenwert von X'X

▷ Kollinearität $\Leftrightarrow$ mindestens ein Eigenwert λ von $X'X$ ist Null (oder nahe bei Null), weil $\exists\ e \neq 0 : (X'X)e = \lambda e = 0e \ \Rightarrow$ Matrix hat nicht vollen Rang $\Rightarrow$ Kollinearität.
▷ Skalierung durch $\frac{\lambda_{max}}{\lambda_{min}}$

(2) **Varianz Inflationsfaktor:**

▷ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
▷ Möglicherweise gibt es ein Kollinearitätsproblem mit x_1: $x_1 \sim \gamma_0 + \gamma_1 x_2 + \gamma_2 x_3 + \gamma_3 x_4$. Eine Regression ergibt R_1^2.

$$\text{VIF}_j \quad := \quad \frac{1}{1 - R_j^2} \tag{6.10}$$

R_j^2: Bestimmheitsmaß der Regression von x_j auf die übrigen x.

▷ $R_j^2 \to 1 \Rightarrow VIF_J \to \infty$ ($R_j^2 = 0 \Rightarrow VIF_j = 1, R_j^2 = 0,99 \Rightarrow VIF_j = 10$)
▷ Man kann allerdings keine Aussage über den Grad der Kollinearität machen.

Es gilt für die Varianz von β_j:

$$\sigma_{\hat{\beta}_j} = \frac{\sigma^2}{(x_j - \bar{x})'(x_j - \bar{x})} \text{VIF}_j \tag{6.11}$$

▷ $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1} \Rightarrow \ Var(\beta_j) = \sigma^2 c_{jj}$, wobei c_{jj} das Diagonalelement ist.
▷ Beachte, dass man VIF auch für binäre Einflussgrößen (Indikatorvariablen) sinnvoll verwendet werden kann.

Alternative Schätzfunktionen: Shrinkage-Schätzung

Problem: Wenn $|\hat{\beta}_i|$ zu groß ist (durch große Werte), explodieren die Abweichungen.

Idee: Minimiere

$$\sum_{i=1}^{n} (y_i - \hat{y})^2 + \lambda(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p) = SSE + \text{Strafterm} \tag{6.12}$$

mit λ als Strafparameter (Shrinkage-Parameter) um $\hat{\beta}_i$ klein zu halten.

- Ridge-Regression $\sum_{i=1}^{n} (y_i - \hat{y})^2 + \lambda \sum_{i=1}^{n} \hat{\beta}_i^2$
- Lasso-Regression $\sum_{i=1}^{n} (y_i - \hat{y})^2 + \lambda \sum_{i=1}^{n} |\hat{\beta}_i|$
- Elastic Net $\sum_{i=1}^{n} (y_i - \hat{y})^2 + \lambda_1 \sum_{i=1}^{n} |\hat{\beta}_i| + \lambda_2 \sum_{i=1}^{n} \hat{\beta}_i^2$

▷ λ wird durch Kreuzvalidierung geschätzt (oder willkürlich gewählt).
▷ Die X-Variable sollte bei dieser Schätzung normiert werden.

6.2.8 Fehler in X-Variablen

Die Einflussgrößen sind fehlerhaft gemessen bzw. erhoben.

Ursachen
Messfehler im engeren Sinne (Messgerät) und im weiteren Sinne (z.B. falsche Beantwortung von Fragen)

Folgen
Meist systematische betragsmäßige Unterschätzung der zu den fehlerhaft gemessenen Größen gehörigen β_i. Geringere Power der entsprechenden F-Tests.

Diagnose
Mehrfach-Messungen der entsprechenden Größen

Therapie
Verwendung von Korrektur-Verfahren $\longrightarrow$ Theorie der Fehler-in-den-Variablen-Modelle

7 Modellwahl

7.1 Zielsetzung der Modellierung

▷ Ausgangssituation: Es sind viele mögliche Einflussgrößen vorhanden. Nun möchte man durch Regression folgende Ziele erreichen:

(a) gute Beschreibung des Verhaltens der Zielgröße
▷ Exploration und Deskription, Erklären und Verstehen des Verhaltens der Zielgröße.

(b) Vorhersage zukünftiger Werte der Zielgröße und Schätzung des Mittels der Zielgröße
▷ Prognose

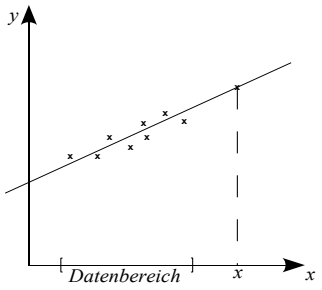
a) b) Modellgenauigkeit wichtig
→ viele Variablen → Kausale Beziehung nicht nötig "Variable enthält Information"

(c) Extrapolation auf Bereiche außerhalb der X-Daten

▷ besonders kritisch, ohne Kausalität nicht möglich.

Wenn ein Wert außerhalb des Datenbereichs der x-Werte liegt (Exploration – keine Prognose), braucht man eine gute inhaltlich Begründung, dass der Zusammenhang auch noch außerhalb des Datenbereichs gilt.

Bsp.: Körpergröße ∼ Alter (10-17) kann nicht auf 30 Jahre übertragen werden.



(d) Schätzung von Parametern
▷ Gehört eher zu (a)
Bias ergibt sich durch Weglassen von Variablen, Erhöhung der Varianz von Schätzern durch überflüssige Variablen
Beachte: Interpretation der Regressionskoeffizienten "bei Festhalten der anderen Variablen".
→ Einschränkung durch viele Kovariablen.

(e) Kontrolle eines Prozesses durch Variation des Inputs
▷ Output soll y sein. Wie muss dann x entsprechend verändert werden?
→ Kausalität nötig

(f) Entwicklung realistischer Modelle für einen Prozess
▷ Kausalzusammenhänge sollen festgestellt werden.
→ Kausalität nötig
Realistische Beschreibung → Sparsames klares Modell

7.2 Allgemeines

"Tradeoff" zwischen Modell-Genauigkeit (R^2) und Einfachheit.
Je mehr Variablen ⇒ R^2 steigt entspricht steigender Komplexität.

Kein Verfahren kann (zunächst bei den Zielen (c,d,e,f)) das Fachwissen ersetzen
→ Verfahren eher explorativ.

Bei der Prognose können bei größeren Datenmengen Variablenselektionsverfahren sehr effizient sein.

Folgende Punkte sollte man zusätzlich im Auge behalten:

1. Diskrete Variablen und Interaktionseffekte bereiten zusätzliche Probleme: Es gibt "Regeln" z.B.,

a) Interaktionen nicht ohne Haupteffekte ins Modell
b) Effekte von kategoriellen Variablen nur als Ganzes ins Modell.

Eine andere Möglichkeit ist das Verwenden von Indikator-Variablen → Variablen, die nicht im Modell sind, sind gemeinsame Referenzkategorie

2. Multikollinearität kann ein erhebliches Problem sein. → Korrelation der Kandidaten - Einflussgrößen analysieren.
3. Transformation und quadratische Terme liefern weitere Möglichkeiten

▷ **Ad-hoc-Kriterium:**
Wahl zwischen zwei Modellen über die Prüfung von $H_0 : \beta_j = 0$. Da eine „nestet“ -Situation vorliegt, wendet man den F-Test an (Auswahl von k Regressoren):

$$F = \frac{(SSE_k - SSE_{p'})/(p' - k)}{SSE_{p'}/(n - p')} = \frac{n - p'}{p' - k} \frac{\hat{\epsilon}_k - \hat{\epsilon}_{p'}}{\hat{\epsilon}_{p'}}$$

H_0 ablehnen, falls $F > F_{1-\alpha}(p' - k, n - p')$, d.h. volles Modell bevorzugen.

7.3 Maße für die Modellgüte

Gegeben sei das lineare Modell $Y = X\beta + \varepsilon$

a) **Bestimmtheitsmaß:**

$$R^2 = \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{7.1}$$

▷ R^2 ungeeignet, da R^2 mit der Anzahl der hinzugenommenen Variablen wächst. Deshalb verwendet man lieber das mit der Anzahl der Parameter adjustierte $R^2_{adj.}$

b) **Adjustiertes Bestimmtheitsmaß:**

$$R^2_{adj} := 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{\hat{\sigma}^2}{\text{SST}/(n-1)} \tag{7.2}$$

c) **Akaikes Informationskriterium AIC:**

$$AIC = n \ln(\text{SSE}) + 2p' - n \ln(n) \tag{7.3}$$

▷ Das kleinste AIC „gewinnt“ (smaller is better)
▷ $AIC \propto$ normierter SSE + 2 · Anzahl der Parameter
▷ Anzahl der Parameter als Strafterm

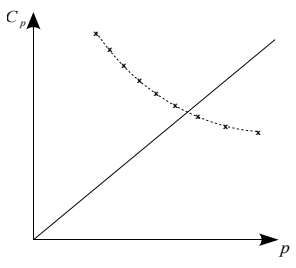
d) **Schwarz'sches Bayes-Kriterium SBC (=BIC):**

$$\text{SBC} = n \ln(\text{SSE}) + \ln(n)p' - n \ln(n) \tag{7.4}$$

▷ Das kleinste BIC „gewinnt“ (smaller is better)
▷ BIC hat generell die Tendenz weniger Variablen zuzulassen als AIC, d.h. BIC ist strenger, weil hier der Strafterm $ln(n)p'$

e) **Mallows C_p:**

$$C_p = \frac{\text{SSE}}{\hat{\sigma}^2_G} + 2p' - n \tag{7.5}$$



$\hat{\sigma}_G$: Schätzung aus vollem Modell

▷ C_p sollte etwa gleich p' sein. Wenn also das erste Mal $C_p < p' = p + 1$ (=Anzahl der Einflussgrößen), dann ergibt sich das beste Modell.

7.4 Variablenselektionsverfahren

Gegeben ist eine Zielgröße Y und mehrere mögliche Einflussgrößen $x_k, k = 1, \ldots, K$. Gesucht ist ein möglichst gutes Modell

$$Y = \beta_0 + \sum_{j=1}^{L} \beta_j x_{k_l},$$

wobei die $x_{k_l}, l = 1, \ldots, L$ ausgewählt werden sollen.

▷ Variablenselektionsverfahren sind allgemeine Verfahren zur Modellanpassung (auch für nicht-lineare Regressionen oder nicht-Regressionen).
▷ Hier wird der Trade-Off der Modellgenauigkeit gegen die Modellkomplexität abgewogen (Trade-Off).

1. **Auswahl nach einem Kriterium**
   Wähle aus allen möglichen 2^k Modellen das Modell mit optimalem Kriterium C aus. C ist in der Regel ein Kriterium aus 7.3. (R^2, R^2_{adj}, AIC, BIC oder/und C_p).

2. **Vorwärtsselektion**
   a) Wähle im Anfangsschritt das Modell $Y = \beta_0$.
   b) Im ersten Schritt wird die Variable in das Modell aufgenommen, die zu dem höchsten R^2 führt.
   c) In den weiteren Schritten wird jeweils eine Einflussgröße in das Modell zusätzlich aufgenommen. Es wird jeweils die Variable, die zu dem höchsten R^2 des resultierenden Modells führt, aufgenommen.
   d) **Stoppregel:** Die Prozedur wird beendet, falls ein bestimmtes Zielkriterium erfüllt ist, z.B. p-Wert des zu der aufgenommenen Variablen gehörigen F-Tests überschreitet einen bestimmten Wert p_0.

   ▷ Die Signifikanz der einzelnen Variablen ist nicht. Durch Hinzunahme einer neuen variable, kann die Erklärungskraft einer anderen abgemildert werden.

3. **Rückwärts-Selektion**
   a) Wähle im Anfangsschritt das volle Modell $Y = \sum_{k=0}^{K} \beta_k x_k$
   b) In den weiteren Schritten wird jeweils eine Einflussgröße aus dem Modell genommen. Es wird jeweils die Einflussgröße, die zu dem höchsten R^2 des resultierenden Modells führt, ausgeschlossen.
   c) **Stoppregel:** Nach bestimmtem Zielkriterium, z.B. p-Wert der F-Tests, die zu den ausgeschlossenen Variablen gehören, unterschreitet einen bestimmten Wert p_0.

   ▷ Hier ist die Signifikanz der einzelnen Variablen gesichert, da sie bei jedem Schritt überprüft wird. Die Signifikanz, die aus Kombination einzelner Variablen entstehen kann, kann hier hingegen nicht berücksichtigt werden.

4. **Schrittweise Selektion:**
   Kombination aus Vorwärts- und Rückwärtsselektion. Er wird eine Vorwärtsselektion und nach jedem Schritt eine Rückwärtsselektion mit geeignetem Stoppkriterium durchgeführt.

   ▷ Hier wird nach jedem Vorwärtsschritt die Signifikanz untersucht und gegebenenfalls eine Variable wieder entfernt. Also ist bei dieser Methode die Signifikanz gesichert.

Beispiel:

$$
\begin{aligned}
v_1 & & \\
v_1\ v_2 & & \\
v_1\ v_2\ v_3 & \Rightarrow & v_2\ v_3 \\
v_2\ v_3\ v_4 & & \\
v_2\ v_3\ v_4\ v_5 & \Rightarrow & v_2\ v_4\ v_5
\end{aligned}
$$

7.5 Beispiel: Tiefbohrprojekt

Modellwahl mit **Modellgütemaße**:

R^2 wird mit steigender Komplexität besser ⇒ nicht besonders gut geeignet.

R^2_{adj} steigt bis 13 Einflussgrößen und sinkt dann wieder ⇒ Modell mit 13 Einflussgrößen.

AIC hat Minimum bei 7 ⇒ Modell mit 7 Einflussgrößen.

BIC hat Minimum bei 7, wobei der Unterschied ($6 \rightarrow 7$) nicht so stark ist wie beim AIC ⇒ Modell mit 7 Einflussgrößen.

C_p sollte etwa einen Wert haben, der ähnlich zu der Anzahl der Parameter ⇒ Modell mit 6 Einflussgrößen.

8 Das allgemeine lineare Modell

▷ Mittels Gewichtung kann man bei bekannten Strukturen der Störterme das gemischte Modell auf das einfache lineare Modell zurückführen. Man setzt keine einheitliche Varianz, aber Unkorreliertheit der Störterme vorraus.

8.1 Der gewichtete KQ-Schätzer

▷ Hier existiert keine einheitliche Varianz der Störterme, aber auch keine Korrelation.

Das lineare Modell mit heteroskedastischen Störgrößen ist gegeben durch:

$$\begin{aligned} Y &= X\beta + \varepsilon && (8.1)\\ \varepsilon &\sim N(0, \sigma^2 V) && (8.2)\\ V &= \text{diag}(v_1, v_2, \ldots, v_n) && (8.3) \end{aligned}$$

V bekannte Matrix zur Beschreibung der Varianzstruktur

Gewichtsmatrix: $W = V^{-1/2} = \text{diag}\ (v_1^{-1/2}, v_2^{-1/2} \ldots, v_n^{-1/2})$

Der gewichtete KQ-Schätzer hat die Form:

$$\begin{aligned} \hat{\beta}_W &:= (X'V^{-1}X)^{-1}X'V^{-1}Y && (8.4)\\ \hat{\sigma}^2 &= (\hat{\varepsilon}'V^{-1}\hat{\varepsilon})/(n-p') && (8.5) \end{aligned}$$

$\hat{\beta}_W$ ist ML-Schätzer und minimiert die **gewichtete Residuenquadratsumme**

$$(Y - X\beta)'V^{-1}(Y - X\beta). \qquad (8.6)$$

8.1.1 Herleitung durch Transformation

Das Modell (8.1)-(8.3) lässt sich in ein gewöhnliches lineares Modell transformieren:

$$\begin{aligned} Y^* &:= WY && (8.7)\\ X^* &:= WX && (8.8)\\ \varepsilon^* &:= W\varepsilon && (8.9) \end{aligned}$$

Dann gilt:

$$\begin{aligned} Y^* &= X^*\beta + \varepsilon^* && (8.10)\\ \varepsilon^* &\sim N(0, \sigma^2 I) && (8.11) \end{aligned}$$

$\hat{\beta}_W$ ist KQ-Schätzer im transformierten Modell

▷ **Idee:** Standardisieren/Normieren von $\epsilon \sim N(0, \sigma^2 V)$:
$\epsilon^* = \frac{\epsilon}{\sqrt{V}}$, wobei $V = TT' \Rightarrow \sqrt{V} = T \Rightarrow \sqrt{V}^{-1} = T^{-1} = W$
$\Rightarrow$ Gewichtung der KQ-Methode, sodass $\epsilon^* \sim N(0, \sigma^2 I)$ gilt.
▷ Ohne Gewichtung ist die Schätzung zwar erwartungstreu, aber nicht optimal, d.h. die KQ-Schätzung hat nicht die kleinste Varianz.

8.1.2 Bemerkung

Beim gewichteten KQ-Schätzer wird angenommen, dass V mit v_i bekannt ist. Das ist in der Praxis meist nicht der Fall.

Abhilfe:
1. Schritt:

Schätzung von v_i durch Berechung der KQ-Schätzung und Betrachtung der $\hat{\epsilon}_i^2$.
Dabei ist ein Parameter für eine Beobachtung nicht sinnvoll. Besser ist ein Model für die Varianzen v_i.

Beispiel: Gruppierte Daten (3 Gruppen)

Angenommen wird, dass v_i in den Gruppen gleich sind. Dann gilt:

$$\hat{\sigma}_1^2 = \frac{\sum_{i\in\sigma_1}\hat{\epsilon}_i^2}{|\sigma_1| - p'}, \hat{\sigma}_2^2 = \frac{\sum_{i\in\sigma_2}\hat{\epsilon}_i^2}{|\sigma_2| - p'}$$

$$\Rightarrow V = \begin{pmatrix} \begin{pmatrix} \sigma_1^2 & & \\ & \cdots & \\ & & \sigma_1^2 \end{pmatrix} & & \\ & \begin{pmatrix} \sigma_2^2 & & \\ & \cdots & \\ & & \sigma_2^2 \end{pmatrix} & \\ & & \begin{pmatrix} \sigma_3^2 & & \\ & \cdots & \\ & & \sigma_3^2 \end{pmatrix} \end{pmatrix}$$

2. Schritt:
Gewichtete KQ-Methode mit $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ und $\hat{\sigma}_3^2$.
Hier gelten die Eigenschaften des Modells dann aber nur approximativ.

8.2 Verallgemeinerte KQ-Methode

▷ Verwendung bei korrelierten Störtermen.
Das lineare Modell mit allgemeiner Varianzstruktur ist gegeben durch:

$$Y = X\beta + \varepsilon \tag{8.12}$$
$$\varepsilon \sim N(0, \sigma^2 V) \tag{8.13}$$

$V \in R^{n\times n}$: beliebige bekannte Kovarianzmatrix mit vollem Rang

▷ V zeigt die Korrelation zwischen den Störtermen an. Die Korrelation zwischen den Störtermen resultiert, dass die Information über die Störterme geringer ist.

Dann gibt es eine invertierbare Matrix T mit

$$TT' = V, \;\; W = T^{-1} \textbf{ Gewichtsmatrix.}$$

Der **verallgemeinerte KQ-Schätzer** ist gegeben durch:

$$\hat{\beta}_W := (X'V^{-1}X)^{-1}X'V^{-1}Y \tag{8.14}$$
$$\hat{\sigma}^2 = (\hat{\varepsilon}'V^{-1}\hat{\varepsilon})/(n - p') \tag{8.15}$$

Das Modell (8.12)-(8.13) lässt sich wie oben in ein gewöhnliches lineares Modell transformieren:

$$Y^* := WY \tag{8.16}$$
$$X^* := WX \tag{8.17}$$
$$\varepsilon^* := W\varepsilon \tag{8.18}$$

Dann gilt:

$$Y^* = X^*\beta + \varepsilon^* \tag{8.19}$$
$$\varepsilon^* \sim N(0, \sigma^2 I) \tag{8.20}$$

▷ W transoformiert das Modell Y zu Y^*.
▷ $Var(\epsilon^*) = \sigma^2 WVW' = \sigma^2 WTT'W' = \sigma^2 I$ (wegen $TT' = V$, $W = T^{-1}$).

▷ **Beispiel:**

$$V = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \ddots & & \vdots \\ \vdots & & \ddots & & \\ \rho^{n-1} & & \dots & & 1 \end{pmatrix} \Rightarrow W = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 \\ -\rho & 1 & 0 & \dots \\ 0 & -\rho & 1 & \vdots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix} \text{ mit } |\rho| < 1.$$

$Y^* = WY$ ist nicht eindeutig lösbar:

$$\begin{aligned} y_1^* &= \sqrt{1-\rho^2}y_1 \\ y_2^* &= y_2 - \rho y_1 \\ y_3^* &= y_3 - \rho y_2 \end{aligned} \qquad \begin{aligned} y_{11} &= \sqrt{1-\rho^2}x_{11} \\ y_{12} &= x_{12} - \rho x_{11} \\ &\vdots \end{aligned}$$

Mittels Gewichtung kann bei bekannten Strukturen die Störterme auf das einfache lineare Modell zurückführen.

8.2.1 Eigenschaften des verallgemeinerten KQ-Schätzers

Gegeben sei das Modell (8.12) bis (8.13). Dann gilt:

$$E(\hat{\beta}_W) = \beta \tag{8.21}$$
$$V(\hat{\beta}_W) = \sigma^2(X'V^{-1}X)^{-1} \tag{8.22}$$

Alle Testverfahren und Quadratsummenzerlegungen lassen sich im Modell
$Y^* = X^*\beta + \varepsilon^*$ betrachten und damit auf den Fall homogener Varianzen zurückführen.

8.3 Allgemeines Gauss-Markov-Theorem

Sei das Modell

$$\begin{aligned} Y &= X\beta + \varepsilon, \quad \text{rg } X = p' \\ E(\varepsilon) &= 0 \\ V(\varepsilon) &= \sigma^2 V \end{aligned}$$

gegeben.

Dann ist $\hat{\beta}_W$ unter den erwartungstreuen linearen Schätzern derjenige mit der kleinsten Varianz: $\hat{\beta}_W$ ist BLUE-Schätzer (**b**est **l**inear **u**nbiased **e**stimator).

8.4 Beispiele für Varianzstrukturen

- AR(1) (allgemeine Zeitreihenstruktur): $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ mit v_t *iid.*
  ρ entspricht der Veränderung von Zeitpunkt $(t-1)$ zum Zeitpunkt t , wobei v_t als weiterer Einfluss eingerechnet wird. Die Veränderungen setzen sich dann systematisch fort (von $(t-2)$ zu t: ρ^2; von $(t-3)$ zu t: ρ^3 usw.) $\Rightarrow$ V.

- Longitudinale Daten (Mehrdimensionale Zeitreihen) Blockdiagonale Struktur
- Symmetrische Struktur (gemischte Modelle)

8.4.1 Weitere Schätzstrategien

- REML
- Robuste Varianzschätzung mit "Working correlation"

8.4.2 Beispiel: Tiefbohrung

Varianzstruktur: Abhängig vom Abstand

$$\hat{\beta} = \begin{pmatrix} -1.8 \\ 2.25 \\ 3.22 \\ -0.023 \end{pmatrix}$$

Alle Schätzungen sind signifikant.

Hinweis auf Autokorrelation wird durch den Durbin-Watson-Test untersucht. Dieser ergibt 0.82, obwohl sie bei 2 liegen sollte. Wegen $0.82 < 2$ kann eine positive Autokorrelation festgestellt werden.

8.4.3 Beispiel: Wildzeitreihen

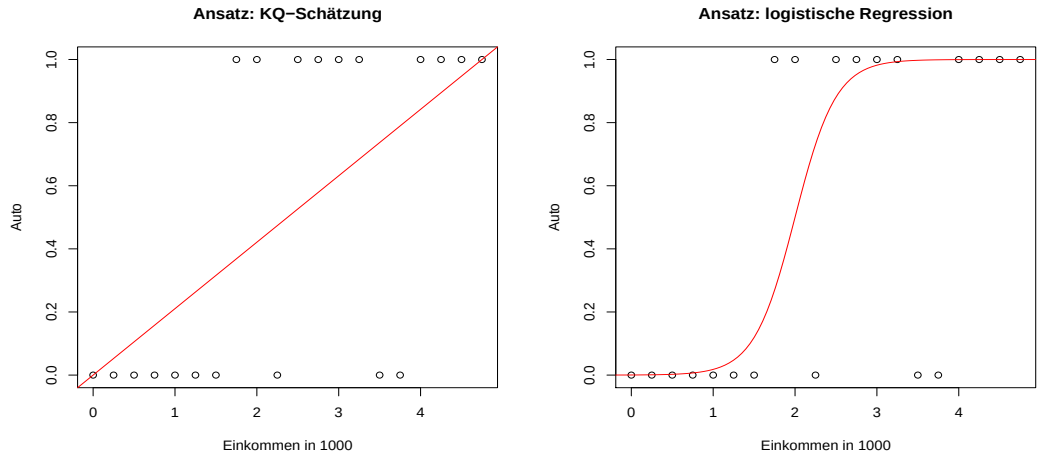
Varianzstruktur: Unabhängigkeit der Einzelzeitreihen, aber AR(1)- Struktur für jede einzelne Zeitreihe

9 Das logistische Regressionsmodell

▷ **Ziel:** Untersuchung des Einflusses von x auf y mit y als binäre Zielgröße.

9.1 Beispiel: Einkommen ∼ Besitz von Auto

Dieses Beispiel wurde anhand der Mitschriften aus der Vorlesung eingefuegt.
X ... Einkommen (stetig)
Y ... Besitz eines Auto (nein=0/ja=1)



9.1.1 Ansatz: KQ-Schätzung

Sei $y = \beta_0 + \beta_1 * \text{Einkommen} + \epsilon$.
Hier wird angenommen, dass $E(\epsilon) = 0$, so dass: $E(y) = \beta_0 + \beta_1 * \text{Einkommen}$ (siehe oben).

Problem:

1. Das Regressionsmodell ergibt auch Werte, die außerhalb des Intervalls $(0, 1)$ liegen, z.B. Bei hohem Einkommen kann die Zielvariable auch angeben, ob jemand zwei Autos hat, was in diesem Fall uninteressant ist und nicht beachtet werden soll.
2. Y ist binomialverteilt ($Y \sim Bin(1, E(Y))$). Es ergibt sich also keine konstante Varianz: $Var(Y) = E(Y)(1 - E(Y))$, was zum Problem der Heteroskedastizität (Vgl. 6.2.2.) führt.

9.1.2 Ansatz: lineares Wahrscheinlichkeitsmodell

Sei $P(Y = 1) = \beta_0 + \beta_1 \cdot \text{Einkommen}$.

Probleme ergeben die Wahrscheinlichkeiten, die weit weg von 1 liegen.

9.1.3 Ansatz: logistisches Regressionsmodell

Hier geht die Regressionsgerade nicht über $(0, 1)$ hinaus (Vgl. Grafik 2).

Sei also $P(Y = 1|\text{Einkommen}) = G(\beta_0 + \beta_1 * Einkommen)$, wobei $G : [-\infty, \infty] \longrightarrow (0, 1)$ und monoton steigend sein sollte, sodass G invertierbar ist.

Man wählt für G die Verteilungsfunktion der logistischen Funktion, die die Voraussetzungen erfüllt:

$$G(t) = [1 + exp(-t)]^{-1}$$

9.2 Definition des logistischen Regressionsmodells

$$\begin{aligned}
\pi_i = P(Y_i = 1|x_i) &= G(x_i'\beta) && (9.1)\\
\ln\frac{\pi_i}{1-\pi_i} &= x_i'\beta \quad \triangleright\ = G^{-1}(\pi_i) && (9.2)\\
Y_i, i = 1,\ldots,n &\quad \text{unabhängig (bei gegebenem festen } X) && (9.3)\\
G(t) &= (1+\exp(-t))^{-1} && (9.4)\\
&= \frac{\exp(t)}{1+\exp(t)}
\end{aligned}$$

Y_i: binäre Zielgröße
x_i: Vektor der Einflussgrößen
X: Design-Matrix der Einflussgrößen mit vollem Rang

Bezeichnungen

$\ln\frac{\pi_i}{1-\pi_i}$	:	"Logarithmierte Chance" Log -odds
$x_i'\beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$	:	linearer Prädiktor
G	:	Response-Funktion (Inverse Link-Funktion)

▷ Der lineare Prädikator summiert die Einflussgrößen auf:
z.B. $P(Y = 1|\text{Einkommen}) = G(\beta_0 + \beta_1 * \text{Einkommen} + \beta_2 * (\text{Zahl der Kinder}) + \beta_3 * \text{Umweltbewusstsein})$.

9.3 Interpretation

Die Wahl von G (Verteilungsfunktion der logistischen Verteilung) als Responsefunktion ermöglicht folgende Interpretation:

Z.B. im einfachen Modell: $y = \beta_0 + \beta_1 x + \epsilon$

$$\begin{aligned}
P(Y = 1|x_0) &= G(\beta_0 + \beta_1 * x_0)\\
P(Y = 1|x_0 + 1) &= G[\beta_0 + \beta_1 * (x_0 + 1)]\\
\frac{P(Y = 1|x_0 + 1)/(1 - P(Y = 1|x_0 + 1))}{P(Y = 1|x_0)/(1 - P(Y = 1|x_0))} &= exp(\beta_1) \quad \triangleright \text{ Odds Ratio von } \pi_i\\
\ln\frac{P(Y = 1|x_0 + 1)}{[1 - P(Y = 1|x_0 + 1)]} - \ln\frac{P(Y = 1|x_0)}{1 - P(Y = 1|x_0)} &= \beta_1 \quad \triangleright \text{ logarithmierter Odds Ratio von } \pi_i
\end{aligned}$$

Das logistische Regressionsmodell nimmt einen linearen Zusammenhang zwischen den "Log odds" von Y und den Einflussgrößen X an.

- Wenn x_k um einen Einheit steigt, so ändert sich die logarithmierte Chance von Y um β_k.
- Wenn x_k um einen Einheit steigt, so ändert sich die Chance von Y um den Faktor $\exp(\beta_k)$.
- Das Odds Ratio (Chancenverhältnis) zwischen Y bei x_k und Y bei $x_k + 1$ ist $\exp(\beta_k)$.

W'keit	0.01	0.05	0.1	0.3	0.4	0.5	0.6	0.7	0.9	0.95	0.99
Odds	1/99	1/19	1/9	3/7	2/3	1	1.5	7/3	9	19	99
Log odds	-4.6	-2.9	-2.2	-0.85	-0.41	0	0.41	0.85	2.2	2.9	4.6

▷ Je größer β, desto steiler die Kurve.
▷ Symmetrie, aber nicht Linearität (die Abstände verändern sich).

9.4 Bemerkungen

Die Varianten des multiplen linearen Regressionsmodells lassen sich direkt auf das logistische Modell übertragen:

- Behandlung von Guppenvergleichen (ANOVA) mit Hilfe von Indikatorvariablen
- Behandlung von diskreten Einflussgrößen: verschiedene Codierungen, Interaktionen, etc.

- Behandlung von stetigen Einflussgrößen (Polynome, Splines etc.)

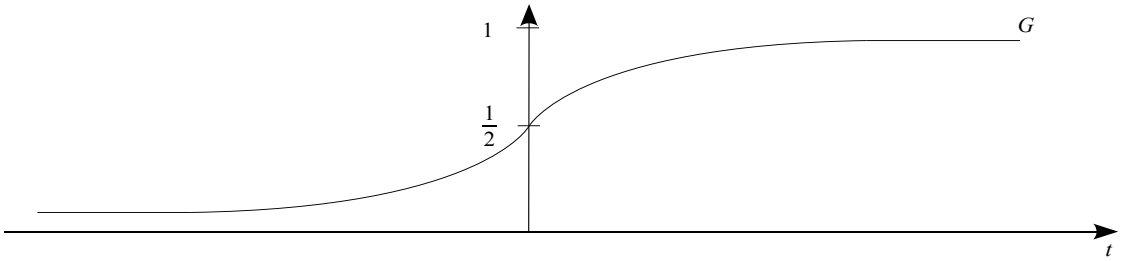
Beachte: Beim logistischen Modell "fehlt" der Varianz-Parameter σ, da $Var(Y) = E(Y) * [1 - E(Y)]$. Weiter ist keine Verteilungsannahme nötig, da Y immer Bernoulli-verteilt ist.

▷ Also sind die Methoden übertragbar.
▷ Y hat Bernoulli-Verteilung mit fester Varianz $Var(Y) = E(Y)(1 - E(Y))$.
▷ Da der Varianzparameter nicht benötigt wird, muss ein Parameter weniger geschätzt werden (Veränderung der Freiheitsgrade).
▷ ABER: Es ist immer zu hinterfragen, ob das Modell wirklich passt.

9.4.1 Herleitung der logistischen Funktion – Wieso wählt man gerade die logistische Verteilungsfunktion für G?

1. Ansatz:

$$\begin{array}{lcll} \frac{dN(t)}{dt} & = & \alpha N(t) & \text{gilt für } N(t) = c\ exp(\alpha t) \\ \frac{dF(t)}{dt} & = & \alpha(r - N(t)) & \text{gilt für } F(t) = \frac{N(t)}{r} \\ \frac{dF(t)}{dt} & = & \beta\ F(t)(1 - F(t)) & \text{gilt für } F(t) = G(\alpha + \beta t) \text{ wegen } G'(t) = G(t)(1 - G(t)) \end{array}$$



2. Ansatz: Grenzwert-/Schwellenwertkonzept

Y wird durch latente Variable Z gesteuert. Man legt fest: $Y = 1 \Leftrightarrow\ Z \geq z_0$ (Z: Nutzen)
In der Population hat der Schwellenwert eine bestimmt Verteilung: $z_0 \sim F_{\alpha,\beta}$ (z_0 hat also Verteilung F mit den Parameter α, β).

Was gilt jetzt für $P(Y_i = 1|z_i)$? $\quad P(Y_i = 1|z_i) = P(z_i \geq\ z_0) = F_{\alpha,\beta}(z_i)$

Sei nun $F_{\alpha,\beta}$ die logistische Verteilung, woraus sich das Logit-Modell ergibt.
Falls $F_{\alpha,\beta} \sim\ N(\alpha, \beta)$, dann folgt das Probit-Modell $\Rightarrow\ \Phi(\frac{z_i - \alpha}{\beta}) = \Phi(\tilde{\alpha} + \tilde{\beta} z_i)$

3. Ansatz: Alternatives Grenzwertkonzept

Es wird ein fester Grenzwert angenommen (z.B. Null). Des Weiteren gilt: $Y_i = 1 \Leftrightarrow\ Z_i > 0$ und $Z_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$ sei der Nutzen.
Wenn nun ϵ_i logistisch verteilt ist, erhält man das logistische Modell.

4. Ansatz: Gruppenkonzept

Angenommen werden zwei Gruppen: $Y = 1 \Rightarrow\ X \sim N(\mu_1, \sigma^2)$ und $Y = 0 \Rightarrow\ X \sim N(\mu_0, \sigma^2)$

$$\begin{array}{lcl} P(Y = 1|x) & = & \frac{P(Y = 1)\, P(x|Y = 1)}{P(x)} = \frac{P(Y = 1)\, \frac{1}{\sigma}\, \varphi(\frac{x - \mu_1}{\sigma})}{P(Y = 1)\, \frac{1}{\sigma}\, \varphi(\frac{x - \mu_1}{\sigma}) + P(Y = 0)\, \frac{1}{\sigma}\, \varphi(\frac{x - \mu_0}{\sigma})} \\ & = & \frac{c_1\, exp(\frac{1}{2}\, (\frac{x - \mu_1}{\sigma})^2)}{c_1\, exp(\frac{1}{2}\, (\frac{x - \mu_1}{\sigma})^2) + c_2\, exp(\frac{1}{2}\, (\frac{x - \mu_0}{\sigma})^2)} = \frac{1}{1 + \frac{c_2}{c_1}\, exp(\frac{1}{2\sigma^2}\, ((x - \mu_1)^2 - (x - \mu_0)^2))} \\ & = & \frac{1}{1 + exp(\alpha + \beta x)} \end{array}$$

Hier gilt $\beta = 0 \Leftrightarrow \ \mu_1 = \mu_0$.

▷ Die Varianzen der Gruppen müssen gleich sein, da sich sonst das x nicht weg kürzt. Für ungleiche Varianzen erhält man einen Ausdruck der Form: $P(Y = 1|x) = G(\alpha + \beta x + \gamma x^2)$.

9.5 Logistische Regression als Klassifikationsproblem

- Prognose in der Logistischen Regression entspricht Klassifikationsproblem mit 2 Gruppen
  ▷ Man nutzt die Information der x-Werte, um die y-Werte durch logistische Regression der entsprechenden Gruppe zuzuordnen.
- Analogien zu Verfahren der Diskriminanzanalyse
- Diskriminanzregeln aus logistischer Regression möglich

9.6 Beispiele

- Y: Kreditwürdigkeit, X: Personenmerkmale (Schufa-Projekt)
- Y: Auftreten einer Krankheit innerhalb einer bestimmten Zeit, X: Exposition, Geschlecht, Alter etc.
- Y: Auffinden der korrekten Blüte, X: Zeit (Trend), Art (Fledermaus)
- Y: Präferenz für eine Partei, X: Persönlichkeitsmerkmale
- Y: Bestehen eines Tests, X: Lehrmethode, Geschlecht etc.

▷ Y ist hier immer binär (nein, ja) $\hat{=}\ (0, 1)$

9.6.1 logistische Regression einer 4-Felder-Tafel

Man kann eine 4-Felder-Tafel direkt in eine logistische Regression übersetzen. Die Bernoulliverteilung ist in eine Binomialverteilung zusammenfassbar. Die Vorraussetzungen der Unabhängigkeit (durch die Modellannahmen) und der gleichen Wahrscheinlichkeiten π (da x_i alle gleich) sind erfüllt.

	Krank	Gesund	
Placebo	22	15	37
Medikament	10	20	30

$\Rightarrow$

y		x_1	
0	1	0	
⋮	1	⋮	} 22*mal*
0	⋮	0	
1	⋮	0	
⋮	⋮	⋮	} 15*mal*
1	1	0	

9.7 ML-Schätzung im logistischen Regressionsmodell

Sei das Modell (9.1)–(9.4) gegeben.

$$\hat{\beta}_{ML} := \arg\max L(\beta) = \arg\max \ln L(\beta) \tag{9.5}$$

$$L(\beta) = \prod_{i=1}^{n} G(x_i'\beta)^{Y_i}(1 - G(x_i'\beta))^{1-Y_i} \tag{9.6}$$

$$\ln L(\beta) = \sum_{i=1}^{n} Y_i \ln(G(x_i'\beta)) + (1 - Y_i) \ln(1 - G(x_i'\beta)) \tag{9.7}$$

Beweis:

X und $Y = (y_1, \ldots, y_n)'$ sind die Beobachtungen mit

$$\begin{aligned} P(Y_i = 1) &= G(x_i'\beta) \\ P(Y_i = 1) &= G(x_i'\beta)(1 - G(x_{i+1}'\beta)) \\ &\Rightarrow L(\beta) \end{aligned}$$

Ableiten nach β und Null setzen liefert unter Benutzung von $G' = G(1-G)$ die Score-Gleichungen für $\hat{\beta}_{ML}$:

$$\begin{aligned} \frac{d\, lnL(\beta)}{d\ \beta} &= \sum_{i=1}^{n} Y_i \frac{1}{G(x_i'\beta)} G(x_i'\beta)\,(1 - G(x_{i+1}'\beta))\, x_i + (1 - Y_i) \frac{1}{1 - G(x_i'\beta)} \left(-G(x_i'\beta)\,(1 - G(x_i'\beta))\, x_i \right. \\ &= \sum_{i=1}^{n} (Y_i\,(1 - G(x_{i+1}'\beta)) + (1 - Y_i)(-G(x_{i+1}'\beta)))\, x_i \\ &= \sum_{i=1}^{n} (Y_i - G(x_{i+1}'\beta))\, x_i \\ \Rightarrow \qquad & s(\hat{\beta}_{ML}) := \sum_{i=1}^{n} (Y_i - G(x_i'\hat{\beta}_{ML}))x_i = 0. \end{aligned}$$

□

9.7.1 Eigenschaften des ML-Schätzers

Die allgemeine Theorie der Maximum Likelihood - Schätzung liefert:
Für $n \to \infty$ gilt unter Regularitätsbedingungen:

$$\begin{aligned} \hat{\beta}_{ML} &\to N(\beta, F^{-1}(\beta)) \qquad \text{asymptotische Normalverteilung} && (9.8) \\ F(\beta) &= X'D(\beta)X && (9.9) \\ D(\beta) &= \text{diag}\{(G(x_i'\beta)(1 - G(x_i'\beta)))\} && (9.10) \\ \hat{\beta}' F(\beta)\hat{\beta} &\to \chi^2(p') \qquad \text{asymptotisch } \chi^2 \text{ verteilt} && (9.11) \end{aligned}$$

- Die asymptotische Varianzmatrix ergibt sich als Inverse der Fischer-Information (Ableitung der Score-Funktion)
- Die asymptotische Varianzmatrix entspricht auch der Varianzmatrix aus der gewichteten (heteroskedastischen) Regression, das $Var(Y) = D(\beta) = \text{diag}(G(x_i'\beta)(1 - G(x_i'\beta)))$
- Die numerische Berechnung des ML- Schätzers erfolgt nach der Methode der "iterierten gewichteten kleinsten Quadrate" (IWLS Iteratively Weighted Least Squares)

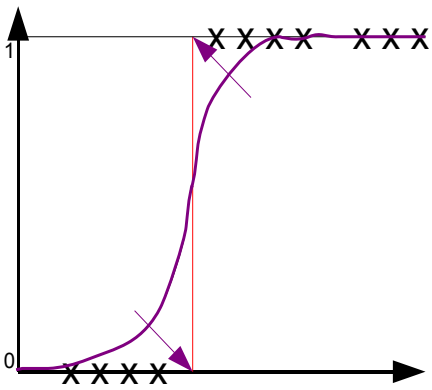
▷ Hier gilt das Gauß-Markow-Theorem nicht, weil nur asymptotische Erwartungstreue vorliegt.

9.7.2 Existenz und Eindeutigkeit des ML-Schätzers im logistischen Modell

Eindeutigkeit: Da die Likelihood-Funktion konkav ist, ist die Lösung der Score-Gleichung immer eindeutig

Existenz: Der ML- Schätzer existiert $\Leftrightarrow$ Die Werte 0 und 1 sind nicht linear trennbar,
d.h. es existiert kein α mit $Y = 1$ für $x'\alpha > 0$ und $Y = 0$ für $x'\alpha < 0$
Im Fall der Nicht- Existenz geht mindestens eine Komponente gegen ∞.
Im einfachen Modell bedeutet die Bedingung, dass Y=1 für $x > c$ und Y=0 für $x < c$.

▷ Sobald ein Wert außerhalb des Datenschwerpunktes für das entsprechende Y liegt, existiert der ML-Schätzer.
▷ Wenn der Schätzer existiert, ist er auch eindeutig.
▷ Wenn die Daten trennbar sind, divergiert der ML-Schätzer, d.h. $\hat{\beta} \to \infty$. Das ist bei der Anwendung erkennbar, wenn der ML-Schätzer und die Varianz ungewöhnlich hohe Werte haben (hohe Zahl von Fisher-Scoring Iterationen). In solchen Fällen ist eine Gruppierung sinnvoll.
▷ Für endliche Stichproben ist der ML-Schätzer nicht erwartungstreu, nur asymptotisch erwartungstreu.



9.8 Inferenz im logistischen Regressionsmodell

Beachte: Alle Aussagen gelten - im Gegensatz zum linearen Regressionsmodell – nur asymptotisch, d.h. für hinreichend große Stichprobenumfänge!

▷ $P(\beta \in [\text{untere Grenze, obere Grenze}]) = 1 - \alpha$
$P(\exp(\beta) \in [\exp(\text{untere Grenze}), \exp(\text{obere Grenze})]) = 1 - \alpha$
$\Rightarrow$ Tests und KIs lassen sich übertragen

9.8.1 Wald-Test für die lineare Hypothese

Sei das logistische Regressionsmodell (9.1)–(9.4) gegeben.

$H_0 : A\beta = c$ mit rg(A) =a.

Analog zum linearen Modell wird folgende quadratische Form betrachtet:

$$W = (A\hat{\beta} - c)' \underbrace{(AF^{-1}(\hat{\beta})A')^{-1}}_{=\text{ Varianz}}(A\hat{\beta} - c)$$

W heißt Wald-Statistik. Aus der asymptotischen Normalität folgt unmittelbar:

$$W \overset{as.}{\sim} \chi^2(a)$$

Mit dieser Statistik lässt sich die allgemeine lineare Hypothese testen.

Wald-Konfidenzintervalle

Wir benutzen die asymptotische Normalität (▷ wir haben hier keine t-Verteilunng wie im allgemeinen Modell, da dort die Varianz geschätzt werden muss)und erhalten folgende Konfidenzintervalle für β zum Niveau α :

$$\hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} z_{1-\alpha/2}$$
$$\hat{\sigma}_{\hat{\beta}_k} = \sqrt{c_{kk}} \text{ (k-tes Diagonalelement der Matrix } F^{-1}(\hat{\beta}))$$

Für die Odds-ratios $\exp(\beta_k)$ ergibt sich das transformierte Konfidenzintervall zum Niveau α :

$$exp\left[\hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} z_{1-\alpha/2}\right]$$

Da kein Varianzparameter zu schätzen ist, kommt die t-Verteilung hier nicht vor.

9.8.2 Likelihood-Quotienten -Test für die lineare Hypothese

Sei das logistische Regressionsmodell (9.1)–(9.4) gegeben.

$H_0 : A\beta = c$ mit $rg(A) = a$.

Wir definieren folgende Teststatistik:

$$\begin{aligned} LQ &= -2\left\{\ln L(\hat{\tilde{\beta}}) - \ln L(\hat{\beta})\right\} \\ \hat{\tilde{\beta}} &: \text{ML-Schätzer unter } H_0 \end{aligned}$$

Aus der allgemeine Theorie von Likelihood-Quotienten-Tests folgt:
Es gilt unter H_0:

$$LQ \overset{as}{\sim} \chi^2(a)$$

Beachte: Der LQ-Test ist mit dem Wald-Test für endliche Stichproben nicht äquivalent. Äquivalenz gilt nur asymptotisch.

▷ Nullhypothese H_0 ablehnen, wenn $LQ > \chi^2(a)$

Likelihood-Quotienten Konfidenzintervalle

Mit Hilfe des LQ-Tests lassen sich auch Konfidenzintervalle zum Niveau α konstruieren:

$$KI := \{\tilde{\beta}_k | H_0 : \beta_k = \tilde{\beta}_k \text{wird mit LQ-Test zum Niveau } \alpha \text{ nicht abgelehnt}\}$$

▷ Überdeckungswahrscheinlichkeit für große Stichprobenumfänge $1-\alpha$, sonst approximativ $1-\alpha$.
▷ KIs sind nicht symmetrisch.

9.8.3 Score-Test für die lineare Hypothese

Sei das logistische Regressionsmodell (9.1)–(9.4) gegeben.

$H_0 : A\beta = c$ mit rg(A) =a.

Wir definieren folgende Teststatistik:

$$SC = s(\hat{\tilde{\beta}})' F^{-1}(\hat{\tilde{\beta}})\, s(\hat{\tilde{\beta}})$$

$$\hat{\tilde{\beta}} : \text{ML-Schätzer unter } H_0$$

Es gilt unter H_0:

$$SC \overset{as}{\sim} \chi^2(a)$$

▷ Score-Gleichung (siehe (9.8.)): $s(\hat{\beta}_{ML}) := \sum_{i=1}^{n}(Y_i - G(x_i'\hat{\beta}_{ML}))x_i = 0$
▷ H_0 ablehnen, wenn $SC > \chi^2(a)$
▷ Vorteil: $\hat{\beta}_{ML}$ muss man nicht ausrechnen (ist u.U. nicht analytisch berechenbar).

9.8.4 Zusammenfassung: Tests für die lineare Hypothese

Sei das logistische Regressionsmodell (9.1)–(9.4) gegeben. $H_0 : A\beta = c$ mit rg(A) =a.

$$W = (A\hat{\beta} - c)'(AF^{-1}(\hat{\beta})A')^{-1}(A\hat{\beta} - c) \tag{9.12}$$

$$LQ = -2\left\{\ln L(\hat{\tilde{\beta}}) - \ln L(\hat{\beta})\right\} \tag{9.13}$$

$$SC = s(\hat{\tilde{\beta}})' F^{-1}(\hat{\tilde{\beta}})\, s(\hat{\tilde{\beta}}) \tag{9.14}$$

$$\hat{\tilde{\beta}} : \text{ML-Schätzer unter} H_0$$

W: Wald-Statistik
LQ: Likelihood-Quotienten-Statistik Es gilt unter H_0:
SC: Score-Statistik

$$W \overset{as}{\sim} \chi^2(a) \tag{9.15}$$
$$LQ \overset{as}{\sim} \chi^2(a) \tag{9.16}$$
$$SC \overset{as}{\sim} \chi^2(a) \tag{9.17}$$

Außerdem sind alle 3 Tests asymptotisch äquivalent. Sie unterscheiden sich aber für kleine Stichproben.

▷ Küchenhoffs Empfehlung: Anwendung der LQ-Statistik, wobei sich in der Praxis kaum ein Unterschied ergibt.

9.8.5 Devianz im logistischen Modell

▷ Je flacher die Regressionsgerade ist, umso kleiner ist R^2, woraus sich schlussfolgern lässt, dass R^2 als Maß für die Modellgüte im logistischen Modell nicht besonders gut geeignet ist
Analog zur ANOVA- Tafel betrachtet man im logistischen Modell die Log-Likelihood als Maß für die Modellgüte: Dabei wird definiert:

Modell mit Konstante (SST):	$P(Y_i = 1) = G(\beta_0)$
Modell (SSE):	$P(Y_i = 1) = G(x_i'\beta)$
"volles Modell"	$P(Y_i = 1) = p_i$

Von diesen Modellen wird jeweils der Wert von -2 log(L) verglichen.

9.9 Das logistische Modell für gruppierte Daten

Wir betrachten das logistische Regressionsmodell mit **gruppierten** Daten: Jeweils n_j; Datenpunkte werden zu einer Gruppe zusammengefasst. Dabei sind in einer Gruppe die Kovariablen identisch. Sei $\hat{\pi}_j := G(x_j'\hat{\beta})$.
Y_j : Anzahl der Erfolge in Gruppe j. Das Modell ist dann:

$$Y_j|x_j \sim B(n_j, G(x_j'\beta)), \quad j = 1 \dots g. \tag{9.18}$$

$$D = -2\sum_{j=1}^{g}(\ln L(\hat{\beta}) - \ln L(y_j)) \tag{9.19}$$

heißt **Devianz**. Es gilt:

$$D = 2\sum_{j=1}^{g} y_j \ln \frac{y_j/n_j}{G(x_j'\hat{\beta})} + (n_j - y_j)\ln \frac{(n_j - y_j)/n_j}{(1 - G(x_j'\hat{\beta}))} \tag{9.20}$$

▷ $\ln L(\hat{\beta})$ ist der Likelihoodbeitrag der j-ten Gruppe für das ideale Modell
▷ $\ln L(y_j)$ist der Likelihoodbeitrag der j-ten Gruppe für das geschätzte Modell
▷ Devianz entspricht der LQ-Statistik des idealen Modells zum geschätzten Modell.

9.9.1 Anpassungstests

a) **Pearson-Statistik**

$$\chi_P^2 = \sum_{j=1}^{g} n_j \frac{(y_j/n_j - G(x'\hat{\beta}))^2}{\hat{\pi}_j(1 - \hat{\pi}_j)}$$

▷ ist abgeleitet aus dem generalisierten linearen Modell.

b) **Devianz (siehe (9.19))** Verteilungsapproximation ($n_i/n \to \lambda_i$)

$$\chi_p^2, D \overset{(a)}{\sim} \chi^2(g - p')$$

▷ Hier ist g die Anzahl der Gruppen/Parameter im perfekten Modell und p' Anzahl der geschätzten Parameter.

c) Bei kleinen Gruppenumfängen oder im Fall $n_i = 1$: **Hossmer-Lemeshow-Test**
Bilde ca. g= 10 Gruppen nach der Größe des linearen Prädiktors $x'\hat{\beta}$ und bilde Anpassungsstatistik wie unter a). Die Testverteilung ist eine $\chi^2(g-2)$-Verteilung.

▷ (a), (b) basieren auf der Binomialverteilung. Sie sind also nur auf große Stichprobenumfänge anwendbar, da sonst die Approximation noch nicht greift.
▷ zu (c): Die Gruppen erhält man, indem man sich den linearen Prädiktor anschaut und dementsprechend eine Aufteilung nach Quantilen oder Ähnlichem wählt.

9.9.2 Residuen im logistischen Regressionsmodell

Wir betrachten wie oben das logistische Regressionsmodell mit gruppierten Daten. Sei $\hat{\pi}_j := G(x_j'\hat{\beta})$.

a) **Devianzresiduen**

$$d_j = \underbrace{\text{sign}(y_j - n_j\hat{\pi}_j)}_{\text{Richtung der Devianz-Res.}} \underbrace{\sqrt{y_j \ln \frac{y_j/n_j}{\hat{\pi}_j} + (n_j - y_j)\ln\frac{(n_j-y_j)/n_j}{(1-\hat{\pi}_j)}}}_{\text{Devianzresiduen (vgl. (9.19))}} \tag{9.21}$$

▷ d_j = Richtung * „$\sqrt{D_j}$“(siehe 9.21)

b) **Pearson-Residuen**

$$r_j = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)}} \tag{9.22}$$

c) **Standardisierung der Residuen**

$$H := D^{\frac{1}{2}}X(X'DX)^{-1}X'D^{\frac{1}{2}} \tag{9.23}$$
$$D = \text{diag}(n_j\hat{\pi}_j(1-\hat{\pi}_j)) \tag{9.24}$$
$$d_j^* := d_j/\sqrt{1-h_{jj}} \tag{9.25}$$
$$r_j^* := r_j/\sqrt{1-h_{jj}} \tag{9.26}$$

d) **Likelihood-Residuen**

$$lr_j := \text{sign}(y_j - n_jG(x_j'\beta))\sqrt{2(\ln L(\tilde{\beta},\hat{\gamma}_j) - \ln L(\hat{\beta}))} \tag{9.27}$$

$L(\tilde{\beta},\hat{\gamma}_j)$: Likelihood des Modells mit dem der zusätzlichen Indikatorvariablen für die Beobachtung j mit zugehörigem Parameter γ_j.

▷ H ist hier eine Verallgemeinerung der Hat-Matrix P. Dient also zur Standardisierung.

9.10 Maße für die Modellanpassung

▷ R^2 ist als Maß für die Modellgüte im logistischen Modell nicht besonders gut geeignet. Deshalb verwendet man andere Maße.

Wir betrachten das logistische Regressionsmodell mit $\hat{\pi}_j := G(x_j'\hat{\beta})$.

a) **Likelihood-Quotienten-Index**

$$R^2_{LQ} = 1 - \frac{log(L(\hat{\beta}))}{log(L_0)} \tag{9.28}$$

▷ L_0: Modell mit Konstanten
▷ $L(\hat{\beta})$: volle Modell
▷ ${R_{LQ}}^2 \in [0, 1]$

b) **Vorhersagefehler** Sei $\hat{Y}_i := 1$, falls $\hat{\pi}_i \geq p_0$ und $\hat{Y}_i := 0$ falls $\hat{\pi}_i < p_0$ (z. B. p_0 =0.5)
Nun analysiert man die Vierfeldertafel, die durch die binären Größen Y_i und $\hat{Y}_i$ gegeben ist.

▷ Nur bei starkem Effekt und kleiner Tafel hilfreich (wie z.B. Rauchen $\sim$ chronische Bronchitis).

c) **Zusammenhang von Y_i und $\hat{\pi}_i$** Für beobachtete Paare sei
N : Anzahl von Paaren mit unterschiedlichen Response, d.h. $y_{i_1} \neq y_{i_2}$,
N_c: Anzahl konkordanten Paare, d.h. mit $\text{sign}(y_{i_1} - y_{i_2}) = \text{sign}(\hat{\pi}_{i_1} - \hat{\pi}_{i_2})$
N_d: Anzahl der Paare, die diskordant sind.

Kendalls τ

$$\tau_a = \frac{N_c - N_d}{n(n-1)/2} \tag{9.29}$$

▷ ist auch Korrelationsmaß zwischen ordinalen Daten

Goodman & Kruskals γ-Koeffizient

$$\gamma = \frac{N_c - N_d}{N_c + N_d} \tag{9.30}$$

Somers D

$$D = \frac{N_c - N_d}{N} \tag{9.31}$$

▷ Goodman & Kruskals γ-Koeffizient und Somers D dienen zum Vergleich von Modellen

9.11 ROC-Kurven-Analyse

Allgemein
$Y = 1 \longrightarrow$ Ausfall (krank)
$Y = 0 \longrightarrow$ kein Ausfall (gesund)

In der medizinischen Literatur ist das Testergebnis m:

$$\hat{Y}_i = 1 \Leftrightarrow m_i \geq c \tag{9.32}$$

In der Literatur zum Kreditrisiko ist der Score s:

$$\hat{Y}_i = 1 \Leftrightarrow s_i \leq c \tag{9.33}$$

c ist dabei ein Grenzwert. Beide Ansätze sind offensichtlich äquivalent: Betrachte dazu $m_i = -s_i$

9.11.1 Sensitivität und Spezifität

Richtig Positiv = Sensitivität:

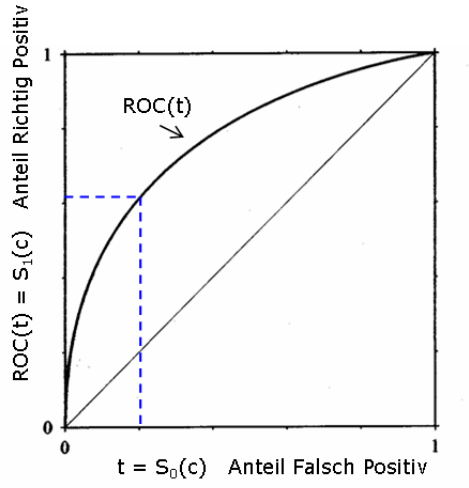
$$(m) \quad P(\hat{Y} = 1|Y = 1) = P(m \geq c|Y = 1) = S_1(c) \tag{9.34}$$
$$(k) \quad P(\hat{Y} = 1|Y = 1) = P(s \leq c|Y = 1) = F_1(c) \tag{9.35}$$

$S_1(c)$ stellt die Survivorfunktion dar, $F_1(c)$ die Verteilungsfunktion.

Falsch Positiv = 1- Spezifität:

$$(m)P(\hat{Y} = 1|Y = 0) = P(m \geq c|Y = 0) = S_0(c) \tag{9.36}$$
$$(k)P(\hat{Y} = 1|Y = 0) = P(s \leq c|Y = 0) = F_0(c) \tag{9.37}$$

Die ROC-Kurve besteht aus den Punkten $(S_0(c), S_1(c))$ bzw. $(F_0(c), F_1(c))$.



9.11.2 Zusammenhang von logistischer Regression und ROC-Kurve

$$\begin{aligned} m_i &= G(x_i'\beta) \\ m_i &= x_i'\beta \end{aligned}$$

Beachte die Invarianz der ROC Kurve bzgl. monotoner Funktionen! Die Verteilung von F_0, F_1 bzw. S_0, S_1 kann aus den Daten geschätzt werden
$\Rightarrow$ ROC-Kurve

Alternative: Schätze F_0, F_1 bzw. S_0, S_1 aus Validierungsdaten.

9.11.3 Maße zur Bewertung der Kurve

AUC: Fläche unter der Kurve

$$AUC = \int_{t=0}^{1} ROC(t)dt = P(m_1 \leq m_0) \tag{9.38}$$

m_1 ist dabei aus der Verteilung $m|Y = 1$
m_0 ist dabei aus der Verteilung $m|Y = 0$

Daher ist das empirische AUC:

$$\widehat{AUC} = \frac{N_c}{N} \tag{9.39}$$

Dabei bezeichnet N_c die Anzahl der konkordanten Paare und N die Anzahl der Paare mit unterschiedlichem Y.
Mit N_d gleich der Anzahl der diskordanten Paare ist der GINI dann:

GINI- Koeffizient: Normierte Fläche zwischen Winkelhalbiernder und ROC- Kurve

$$GINI = 2 \cdot (AUC - \frac{1}{2}) = 2 \cdot AUC - 1 \tag{9.40}$$

$$\widehat{GINI} = \frac{N_c - N_d}{N} \tag{9.41}$$

Der empirische GINI entspricht dem Somers D.

9.11.4 Die logistische Regression für Fall-Kontroll-Studien

Sei in der Grundgesamtheit folgende Beziehung gegeben:

$$P_0(Y = 1|X = x) = G(\alpha + \beta x) \tag{9.42}$$

mit $G(t) = (1 + \exp(-t))^{-1}$

X: Exposition
Y: Erkrankung

Es wird nun aus der Grundgesamtheit gezogen:
n_1 Fälle $(Y = 1)$ und n_2 Kontrollen $(Y = 0)$

Gesucht ist

$$P_S(Y = 1|X = x) \tag{9.43}$$

Mit P_S werden die Wahrscheinlichkeiten (Dichten) in der Stichprobe bezeichnet, mit P_0 die in der Grundgesamtheit.

Berechnung von P_S:

$$P_S(Y = 1|X = x) = \frac{P_S(Y = 1, X = x)}{P_S(X = x)} =$$
$$\frac{P_S(Y = 1)P_S(X = x|Y = 1)}{P_S(Y = 1)P_S(X = x|Y = 1) + P_S(Y = 0)P_S(X = x|Y = 0)} =$$
$$\frac{c_1 P_0(X = x|Y = 1)}{c_1 P_0(X = x|Y = 1) + c_2 P_0(X = x|Y = 0)}$$

Die letzte Identität gilt wegen

$$P_0(X = x|Y = 1) = P_S(X = x|Y = 1) \tag{9.44}$$

und mit $c_1 = \frac{n_1}{n_1 + n_2}$ und $c_2 = \frac{n_2}{n_1 + n_2}$

$$c_1 P_0(X = x|Y = 1) = \frac{c_1}{P_0(Y = 1)} P_0(X = x) P(Y = 1|X = x)$$
$$c_2 P_0(X = x|Y = 0) = \frac{c_2}{P_0(Y = 0)} P_0(X = x) P(Y = 0|X = x)$$

Mit $\frac{c_1}{P_0(Y = 1)} = d_1$ und $\frac{c_2}{P_0(Y = 0)} = d_2$ folgt:

$$\begin{aligned} P_S(Y = 1|X = x) &= \frac{d_1 P(Y = 1|X = x)}{d_1 P(Y = 1|X = x) + d_2(1 - P(Y = 1|X = x))} \\ &= \frac{1}{1 + \frac{d_2}{d_1}\frac{1 - G(\alpha + \beta x)}{G(\alpha + \beta x)}} \\ &= \frac{1}{1 + \exp(-\alpha - \ln(\frac{d_1}{d_2}) - \beta x)} \end{aligned}$$

Die letzte Gleichung folgt aus $\frac{1 - G(t)}{G(t)} = \frac{1}{G(t)} - 1 = \exp(-t)$.

Insgesamt gilt:

$$P_S(Y = 1|X = x) = G(\alpha + \ln\left(\frac{d_1}{d_2}\right) + \beta x) \tag{9.45}$$

Die Auswertung bezüglich β kann also durch eine logistische Regression erfolgen. Der Parameter α in der Grundgesamtheit kann dabei nicht geschätzt werden.

10 Das gemischte lineare Regressionsmodell ("Linear mixed Model")

10.1 Das Modell mit einem einfachen zufälligen Effekt

auch: Varianzkomponenten-Modell, **Random-Intercept Modell**

Wir betrachten gruppierte Daten mit Gruppenindex i:

$$\begin{aligned} Y_{ij} &= x'_{ij}\beta + \gamma_i + \varepsilon_{ij} \quad i = 1, \ldots, g; \;\; j = 1 \ldots n_i && (10.1)\\ \varepsilon &\sim N(0, \sigma^2 I) && (10.2)\\ \gamma_i &\sim N(0, \sigma^2_\gamma) && (10.3)\\ & \quad \gamma_i \text{ und } \varepsilon \text{ unabhängig} \end{aligned}$$

$i \ldots$ Gruppe, $ij \ldots$ Einzeilbeobachtung in der Gruppe i
γ_i: Random Intercept (Zufällige Effekte), zufällige Bereinigung um den Gruppeneffekt. Das entspricht dem Abstand zwischen Gesamtmittel und Gruppenmittel.

▷ β ist für alle Beobachtungen gleich ⇒ jede Gruppe hat gleiche Steigung.
▷ Random-Intercept: Es variiert der Intercept für jede Gruppe

▷ **Varianzkomponenten-Modell**:
Zerlegung der „Gesamt"-Varianz in die Varianz **zwischen** den Klassen ($\rightarrow$ γ_i mit σ^2)) und die Varianz **innerhalb** einer Klasse ($\rightarrow$ ϵ_{ij} mit ${\sigma_\gamma}^2$)

10.1.1 Das marginale Modell

▷ Zurückführen des Modells mit einfachen zufälligen Effekten auf das allgemeine lineare Regressionsmodell:

▷ 2 Stufen:
(1) Gruppen: $\gamma_i \sim N(0, {\sigma_\gamma}^2)$
(2) Einzelbeobachtungen: $y_{ij}|\gamma_i \sim N(x'_{ij}\beta + \gamma_i, \sigma^2)$
⇒ marginales Modell mit $Y_{ij} \sim N(x'_{ij}\beta, \sigma^2 + {\sigma_\gamma}^2)$

Das obige Modell kann kann umgeformt werden zu dem **marginalen Modell**

$$\begin{aligned} Y_{ij} &= x'_{ij}\beta + \delta_{ij} = x'_{ij}\beta + \gamma_i + \varepsilon_{ij} && (10.4)\\ \delta_{ij} &= \varepsilon_{ij} + \gamma_i && (10.5)\\ Var(\delta_{ij}) &= \sigma^2 + \sigma^2_\gamma && (10.6)\\ cov(\delta_{i_1 j_1}, \delta_{i_1 j_2}) &= \sigma^2_\gamma && (10.7)\\ cov(\delta_{i_1 j_1}, \delta_{i_2 j_2}) &= 0 \text{ für } i_1 \neq i_2 && (10.8) \end{aligned}$$

Darstellung als allgemeines lineares Modell:

$$\begin{aligned} Y &= X\beta + \delta\\ \delta &\sim N(0, \sigma^2 I + diag[\sigma^2_\gamma e_i e'_i])\\ e_i &:= \text{1-Vektor der Länge } n_i \end{aligned}$$

10.2 Das Modell mit allgemeinen zufälligen Effekten

▷ Random-Intercept UND Random-Slope

Zweistufiges Modell mit individuellem linearem Trend, dessen Steigung individuell geschätzt wird

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}\\ \beta_{0i} &= \beta_0 + \gamma_{0i}\\ \beta_{1i} &= \beta_1 I_m(i) + \beta_2 * I_B(i) + \gamma_{1i} \end{aligned}$$

Einsetzen ergibt:

$$Y_{ij} = \beta_0 + \beta_1 I_m(i) * t_{ij} + \beta_2 * I_B(i) * t_{ij} + \gamma_{0i} + \gamma_{1i} t_{ij}$$

Annahme : γ_{0i} und zufällige Effekte,
Alle $\boldsymbol{\gamma_i}$ sind unabhängig, und $\boldsymbol{\gamma_i} \sim N(\mathbf{0}, G), i = 1, \ldots, g$,
β_1 und β_2: eigentlich interessierende Populationseffekte

▷ γ_{0i} Random-Intercept
▷ γ_{1i} Random-Slope

10.2.1 Ein hierarchisches Modell für longitudinale Daten Stufe 1

Sei $\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ der Vektor der wiederholten Messungen für das i-te Subjekt zu den Zeiten $t_{ij}, j = 1, \ldots, n_i$, für $i = 1, \ldots, g$.

$$\boldsymbol{Y_i} = Z_i \boldsymbol{\beta_i} + \boldsymbol{\varepsilon_i} \tag{10.9}$$

- Z_i eine $(n_i \times q)$-Matrix bekannter Kovariablen, die modellieren, wie sich die Zielgröße für das i-te Subjekt über die Zeit verhält
- $\boldsymbol{\beta_i}$ ein q-dimensionaler Vektor unbekannter subjektspezifischer Regressionskoeffizienten
- $\boldsymbol{\varepsilon_i}$ ein n_i-dimensionaler Vektor mit Residuen für das i-te Individuum
- **Annahme:**
  Alle $\boldsymbol{\varepsilon_i}$ sind unabhängig und $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i), i = 1, \ldots, N$, Σ_i unbekannte Kovarianzmatrix.
  (Meist Zusatzannahme: Σ_i hängt von i nur über n_i ab.)

10.2.2 Ein hierarchisches Modell für longitudinale Daten Stufe 2

Ein lineares Modell für die subjektspezifischen Regressionskoeffizienten $\boldsymbol{\beta_i}$:

$$\boldsymbol{\beta_i} = K_i \boldsymbol{\beta} + \boldsymbol{\gamma_i} \tag{10.10}$$

- K_i eine $(q \times p)$-Matrix bekannter Kovariablen
- $\boldsymbol{\beta}$ ein p-dimensionaler Vektor unbekannter Regressionsparameter
- **Annahme:** Alle $\boldsymbol{\gamma_i}$ sind unabhängig, und $\boldsymbol{\gamma_i} \sim N(\mathbf{0}, G), i = 1, \ldots, g$,
  G unbekannte Kovarianzmatrix.

10.2.3 Das lineare gemischte Modell für longitudinale Daten

Substitution von (10.10) in (10.9) ergibt

$$\boldsymbol{Y_i} = X_i \boldsymbol{\beta} + Z_i \boldsymbol{\gamma_i} + \boldsymbol{\varepsilon_i} \tag{10.11}$$

mit $X_i = Z_i K_i$, das **lineare gemischte Modell** mit **fixed effects** (festen Effekten) $\boldsymbol{\beta}$ und **random effects** (Zufallseffekten) $\boldsymbol{\gamma_i}$.

▷ $Y_i = Z_i K_i \beta + Z_i \gamma_i + \varepsilon_i = Z_i (K_i \beta + \gamma_i) + \varepsilon_i$

Annahme:

$$\left.\begin{array}{l} \boldsymbol{\gamma_i} \sim N(\mathbf{0}, G), \boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i) \\ \boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_g}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_g} \text{ unabhängig.} \end{array}\right\} \Rightarrow \begin{array}{l} \boldsymbol{Y_i} \sim N(X_i \boldsymbol{\beta}, Z_i G Z_i' + \Sigma_i) \\ \text{(marginales Modell)} \end{array} \tag{10.12}$$

10.3 Das lineare gemischte Modell (LMM) in allgemeiner Darstellung

$$Y = X\beta + Z\gamma + \varepsilon \tag{10.13}$$

$$\begin{pmatrix} \gamma \\ \varepsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{pmatrix}\right) \tag{10.14}$$

X und Z	:	feste bekannte Designmatrizen
β	:	Vektor der festen Effekte
γ	:	Vektor der zufälligen Effekte
R	:	Kovarianzmatrix der Störterme, also $Cov(\varepsilon_i, \varepsilon_j)$
G	:	Kovarianzmatrix der zufälligen Effekte, also $Cov(\gamma_i, \gamma_j)$

Bemerkungen:

Das obige Modell ist sehr flexibel und enthält als Spezialfälle das gemischte Modell für Longitudinaldaten und das Varianzkomponenten-Modell

Im **Modell für Longitudinaldaten** gilt:

$$R = diag(\Sigma_i)$$

Im **Varianzkomponentenmodell** gilt:

$$R = \sigma^2 I(\text{ Dimension: Anzahl Beobachtungen} \sum_{i=1}^{g} n_i)$$

$$G = \sigma_\gamma^2 I(\text{ Dimension: g })$$

10.3.1 Marginales und bedingtes (konditionales) Modell

Marginales Modell:

$$Y = X\beta + \delta \tag{10.15}$$

$$\delta = Z\gamma + \varepsilon \tag{10.16}$$

$$\delta \sim N(0, R + ZGZ') \tag{10.17}$$

Bedingtes Modell:

$$Y|\gamma \sim N(X\beta + Z\gamma, \boldsymbol{R}) \tag{10.18}$$

$$\gamma \sim N(0, \boldsymbol{G}) \tag{10.19}$$

10.4 Inferenz im gemischten linearen Modell

Die Inferenz erfolgt zunächst mit Hilfe des marginalen Modells: Sei $\boldsymbol{\vartheta}$ ein Vektor aller Parameter, die in G und R vorkommen. $\boldsymbol{\vartheta}$ und β können nach der Maximum-Likelihood Methode geschätzt werden: Als Log-Likelihood ergibt sich (von additiven Konstanten abgesehen):

$$l(\beta, \boldsymbol{\vartheta}) = -\frac{1}{2}\left(\ln|V(\boldsymbol{\vartheta})| + (\boldsymbol{Y} - X\beta)'V^{-1}(\boldsymbol{\vartheta})(Y - X\beta)\right) \tag{10.20}$$

wobei $V = ZG(\boldsymbol{\vartheta})Z' + R(\boldsymbol{\vartheta})$. Ist $\boldsymbol{\vartheta}$ bekannt, so ist der ML-Schätzung von $\boldsymbol{\beta}$ bedingt auf $\boldsymbol{\vartheta}$ (gewichteter KQ-Schätzer:)

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) = \left(X'V(\boldsymbol{\vartheta})^{-1}X\right)^{-1}X'V^{-1}(\boldsymbol{\vartheta})Y. \tag{10.21}$$

Einsetzen liefert die Profil-Log-Likelihood:

$$l(\boldsymbol{\vartheta}) = -\frac{1}{2}\left(\ln|V(\boldsymbol{\vartheta})| + (\boldsymbol{Y} - X\beta(\boldsymbol{\vartheta}))'V^{-1}(\boldsymbol{\vartheta})(Y - X\beta(\boldsymbol{\vartheta}))\right) \tag{10.22}$$

10.4.1 ML und REML-Schätzer

Maximieren von (10.22) bezüglich ϑ liefert ML-Schätzer. Da dieser nicht erwartungstreu ist, verwendet man häufig den sogenannten restringierteten ML-Schätzer: Dieser maximiert

$$L_R(\vartheta) = l(\boldsymbol{\vartheta}) - \frac{1}{2} ln|X'V(\vartheta)^{-1}X| \tag{10.23}$$

Im einfachen linearen Modell entspricht der REML-Schätzer dem erwartungstreuen Schätzer von σ^2.

10.4.2 Inferenz bezüglich von β im linearen gemischten Modell II

Unter dem marginalen Modell (10.11) und bedingt auf $\boldsymbol{\vartheta}$ folgt $\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta})$ einer multivariaten Normalverteilung mit Erwartungswert $\boldsymbol{\beta}$ und Kovarianzmatrix

$$Var(\hat{\boldsymbol{\beta}}) \quad = \quad (X'V^{-1}X)^{-1} \tag{10.24}$$

Da V unbekannt ist, wird es durch den (RE)ML-Schätzer $V(\hat{\vartheta})$ ersetzt.
Zur Konstruktion von Konfidenzintervallen und entsprechenden Tests nimmt man an, dass β asymptotisch normalverteilt ist. Für spezielle Modelle ist dies bewiesen, aber eine allgemeingültige asymptotische Normalverteilungsaussage ist nicht nachgewiesen.
Da die Varianzmatrix V nur geschätzt wird, werden in der Praxis deshalb häufig approximative t-Tests und entsprechende Konfidenzintervalle benutzt, die die Verteilung von $(\hat{\beta}_j - \beta_j)/\hat{se}(\hat{\beta}_j)$ durch eine t-Verteilung approximieren und die zugehörigen Freiheitsgrade geeignet schätzen.
▷ Dabei ist $\hat{se}(\hat{\beta}_j) = \sqrt{\widehat{Var(\hat{\beta}_j)}}$ der Standarderror des Schätzers

10.4.3 Schätzung der zufälligen Effekte

In manchen Fällen ist die Schätzung der Zufälligen Effekte von Interesse. Dazu betrachten wir das gesamte Modell in folgender Form:

$$\begin{pmatrix} \gamma \\ Y \end{pmatrix} \quad \sim \quad N\left(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{ZG} \\ \boldsymbol{GZ'} & \boldsymbol{V} \end{pmatrix}\right) \tag{10.25}$$

Nun erhält man den bedingten Erwartungswert von γ bei gegebenem Y nach den allgemeinen Regeln für die multivariate NV:

$$E(\gamma|Y) = \boldsymbol{GZ'V}^{-1}(Y - X\beta) \tag{10.26}$$

Ersetzen durch die Schätzer ergibt:

$$\hat{\gamma} = \hat{\boldsymbol{G}}\boldsymbol{Z'}\hat{\boldsymbol{V}}^{-1}(Y - X\hat{\beta}) \tag{10.27}$$

10.5 Praktisches Umsetzen von gemischten Modellen mit SAS

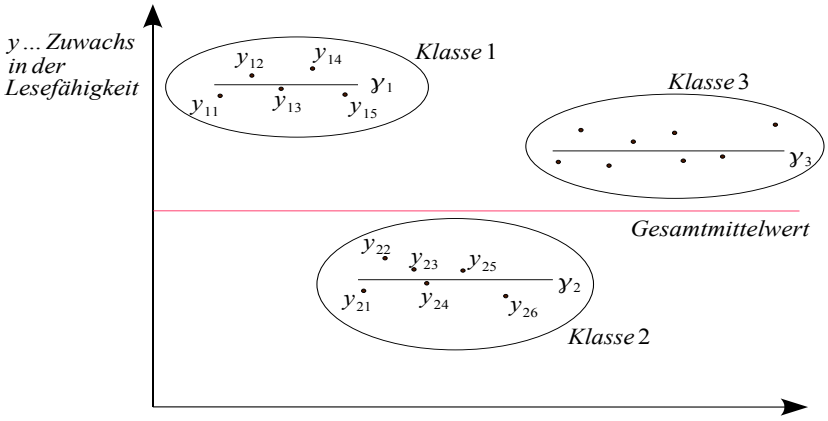
- Proc MIXED
- Random-Statement legt den zufälligen Effekt fest, auch Intercept zugelassen, d.h. die Matrix Z. Type legt Kovarianzstruktur von G. Subject legt Einheiten fest
- Repeated-Statement legt Varianzstruktur R fest Subject-Statement legt Blockdiagonale Struktur fest
- Model-Statement legt Modell mit den festen Effekte fest; X-Matrix
- Viele Varianten und Optionen möglich

10.6 Beispiele

10.6.1 Beispiel: Studie zur Leseföderung

- Zielgröße: Verbesserung der Lesefähigkeit
- Einflussgrößen: spezielle Förderung
- Störgröße : Ausgangsniveau
- Problem : Versuch wurde klassenweise durchgeführt

Voraussetzung der Unabhängigkeit der Störterme nicht erfüllt (Cluster Daten), weil von der Ähnlichkeit der Schüler einer Klasse ausgegangen werden kann.
Abhilfe: Einführung eines Klasseneffekts
Problem: Zu viele Parameter
Abhilfe: Klasseneffekt wird als zufälliger Effekt eingeführt.



▷ i - Klasse, j - Schüler in Klasse i
▷ γ_i Random-Intercept

Das marginale Modell

▷ Zurückführen des Modells mit einfachen zufälligen Effekten auf das allgemeine lineare Regressionsmodell:
▷ 2 Stufen:
(1) Klassen: $\gamma_i \sim N(0, {\sigma_\gamma}^2)$
(2) Schüler in den Klassen: $y_{ij}|\gamma_i \sim N(x'_{ij}\beta + \gamma_i, \sigma^2)$
$\Rightarrow$ marginales Modell mit $Y_{ij} \sim N(x'_{ij}\beta, \sigma^2 + {\sigma_\gamma}^2)$, d.h. $Y_{ij} = x'_{ij}\beta + \gamma_i + \varepsilon_{ij}$

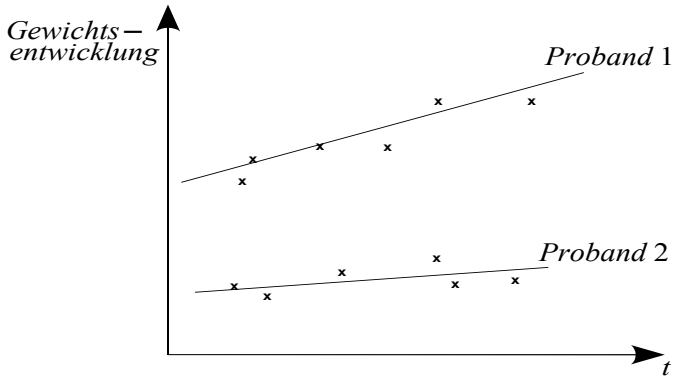
▷ Also ergibt sich, wenn man 3 Schüler in Klasse 1 und 2 Schüler in Klasse 2 annimmt:

$$V = Var(Y_{ij}) = \left(\begin{array}{ccc|cc} \sigma^2+\sigma^2_{\gamma_1} & \sigma^2_{\gamma_1} & \sigma^2_{\gamma_1} & & \\ \sigma^2_{\gamma_1} & \sigma^2+\sigma^2_{\gamma_1} & \sigma^2_{\gamma_1} & \multicolumn{2}{c}{0} \\ \sigma^2_{\gamma_1} & \sigma^2_{\gamma_1} & \sigma^2+\sigma^2_{\gamma_1} & & \\ \hline & & & \sigma^2+\sigma^2_{\gamma_2} & \sigma^2_{\gamma_2} \\ \multicolumn{3}{c|}{0} & \sigma^2_{\gamma_2} & \sigma^2+\sigma^2_{\gamma_2} \end{array}\right)$$

$$\begin{aligned} \triangleright\ Cov(Y_{ij}, Y_{nk}) &= Cov(\gamma_i + \epsilon_{ij}, \gamma_n + \epsilon_{nk}) \\ &= Cov(\gamma_i, \gamma_n) = \begin{cases} \sigma^2_\gamma & , i = n \text{ (also die Schüler in der gleichen Klasse)} \\ 0 & , i \neq n \end{cases} \end{aligned}$$

d.h. Random-Intercept unabhängig

10.6.2 Beispiel: Gewichtsentwicklung



- Zielgröße: Gewichtsentwicklung (in nur unregelmäßigen Abständen erhoben)
- Einflussgrößen: Geschlecht I_M, Art der Intervention I_B

Zweistufiges Modell mit individuellem linearem Trend, dessen Steigung von Geschlecht (Indikator I_m) und Art der Intervention (IndikatorI_B) abhängt

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij} \\ \beta_{0i} &= \beta_0 + \gamma_{0i} \\ \beta_{1i} &= \beta_1 I_M(i) + \beta_2 * I_B(i) + \gamma_{1i} \end{aligned}$$

Hier werden die Messungen für den jeweiligen Probanden gruppiert. (i „Proband“ , j „Messung“)

$\Rightarrow$ Für jeden Probanden wird ein eigenes Intercept und eine eigene Steigung geschätzt.

$\Rightarrow$ Dann erst überlegt man sich, ob die individuellen Schätzungen in Beziehung zu anderen Einflussgrößen (z.B. Geschlecht) stehen, d.h. man macht eine zweite Regression.

$\Rightarrow$ Y_{ij} = normale Regression (Einfluss von Geschlecht usw.) + individuelle Einflussgrößen.

$\Rightarrow$ Ein mögliches Ergebnis könnte sein, dass Probanden mit hohem Startgewicht, auch mehr zunehmen.

Individuelle Regression (Vgl. 10.2.1):

Regression für die einzelnen Probanden, d.h. **innerhalb der einzelnen Gruppen** der Daten.

$$Z_i = \begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in} \end{pmatrix} \beta_i = (\beta_{0i}, \beta_{1i})$$

Untersuchung auf die Zusammenhänge im kompletten Modell (Vgl. 10.2.2):

Regression zwischen den Gruppen, wobei die Gruppen hier die Zusammenfassung der Messdaten für den jeweiligen Probanden sind.

$$\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & I_m(i) & I_b(i) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \gamma_{0i} \\ \gamma_{1i} \end{pmatrix}$$

$$\begin{aligned} \beta_i &= \kappa * \beta + \gamma_i \\ \Rightarrow \quad G &= \begin{pmatrix} {\sigma_{\gamma_0}}^2 & {\sigma_{\gamma_0\gamma_1}}^2 \\ {\sigma_{\gamma_0\gamma_1}}^2 & {\sigma_{\gamma_1}}^2 \end{pmatrix} \end{aligned}$$

$I_M(i)$ Indikator für den Einfluss des Geschlechts
$I_B(i)$ Indikator für den Einfluss der Intervention

γ_i sind die individuellen Einflüsse, die ich nicht erklären kann.

Stufe 1 und 2 durch Substitution zu einem Modell zusammenfassen (Vgl. 10.2.3):

$$\begin{aligned} Y_i &= x_i'\beta + Z_i\gamma_i + \varepsilon_i \\ E(Y_i) &= x_i'\beta + 0 + 0 = x_i'\beta \\ Cov(Y_i) &= 0 + Z_i \underbrace{Cov(\gamma_i)}_{G} Z_i' + \Sigma_i \qquad \text{da } \gamma_i, \varepsilon_i \text{ unabhangig} \\ &\Rightarrow \quad Y_I \sim N(x_i\beta, Z_iGZ_i'\Sigma_i) \end{aligned}$$

11 Messfehler: Modelle und Effekte

Die Unterlagen sind Teil des Kurses "Measurement error in epidemiological studies Short course at KU Leuven vom 17/18.11.2003"

Literatur:

- Carroll R. J. , D. Ruppert, L. Stefanski and C. Crainiceanu: Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition Crcpress.
- Carroll, R.J. Measurement error in epidemiologic studies. In: Encyclopedia of Biostatistics, ed. by Armitage, P.and Colton, T., 2491- 2519. Wiley, Chichester.
- Kuha, J., C. Skinner and J. Palmgren. Misclassification error. In: Encyclopedia of Biostatistics, ed. by Armitage, P.and Colton, T., 2615- 2621. Wiley, Chichester.

Ursachen von Messfehlern

- Es wird davon ausgegangen, dass Daten nicht genau erhoben werden können ⇒ Messfehler des Messgerätes
- Was ist eigentlich die richtige Variable? (z.B. Blutdruck/Fettverzehr)
- Die Angaben der Patienten muss nicht richtig sein
- Ausreißer-Werte (Bsp.: Fettverzehr – Sonntagsbraten)
- Problematisch wird zusätzlich, dass die x-Werte für die Prognose ebenfalls ungenau erhoben werden, woraus resultiert, dass der Fehler für die Prognose steigt.

Beispiele

- Studie zur chronischen Bronchitis (Bronchitis ∼ Staubbelastung, Rauchen)
- Job-Ausgesetztsein-Matrix (Exposition ∼ Lebenslauf, Arbeitsstelle)
- Studie zum Fettverzehr (über Tagebuchführung)

11.1 Modelle für Messfehler

- Systematisch vs. Zufall
- Klassisch vs. Berkson
- Additiv vs multiplikativ
- Homoskedastisch vs Heteroskedastisch
- Differentiell vs Nicht-Differentiell

11.1.1 Klassischer additiver zufälliger Messfehler

X_i : Wahrer Wert
W_i : Messung von X, d.h. mit Messfehler

$$\begin{aligned} W_i &= X_i + U_i \quad (U_i, X_i) \text{ unabh.} \\ E(U_i) &= 0 \\ V(U_i) &= \sigma_U^2 \\ U_i &\sim N(0, \sigma_U^2) \end{aligned}$$

▷ additiver Fehler: $W_i = X_i \pm \text{Messfehler} = X_i + U_i \qquad (X_i, U_i)\text{unabh.})$

Das Modell ist passend für

- Messfehler des Messinstruments
- Eine Messung wird für den Mittelwert verwendet
- Messfehler ist bedingt durch den Arzt

11.1.2 Additiver Berkson-Fehler

$$\begin{aligned} X_i &= W_i + U_i \quad (U_i, W_i) \text{ unabh.} \\ E(U_i) &= 0 \\ V(U_i) &= \sigma_U^2 \\ U_i &\sim N(0, \sigma_U^2) \end{aligned}$$

▷ wahrer Wert = gemessener Wert + Störterm
▷ Unterschied zum klassischen Modell: X_i, W_i sind unabhängig (gegenüber X_i, U_i unabhängig)

Das Modell ist passend für

- durchschnittliche Exposition der Region W anstatt der individuellen Exposition X.
- Arbeitsplatzmessung
- Dosis in einem kontrollierten Experiment

Beachte, dass in einem Berkson-Fall:

$$\begin{aligned} E(X|W) &= W \\ Var(X) &= Var(W) + Var(U) \end{aligned}$$

▷ Var(wahrer Wert) $> Var$(gemessener Wert), wegen $Var(X) = Var(W) + Var(U)$.

11.1.3 Multiplikativer Messfehler

$$\begin{aligned} W_i &= X_i * U_i \quad (U_i, X_i) \text{ unabh.} \\ \text{Klassisch} & \\ X_i &= W_i * U_i \quad (U_i, W_i) \text{ unabh.} \\ \text{Berkson} & \\ E(U_i) &= 1 \\ U_i &\sim \textit{Lognormal} \end{aligned}$$

- Additiv auf einer logarithmischen Skala
- Wird benutzt für Expositionen durch Chemikalien oder Strahlung

▷ Prozentfehler, d.h. Messung wird ungenauer, je größer der Wert ist.

11.1.4 Messfehler in der Zielgröße

Einfache lineare Regression

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon \\ Y^* &= Y + U \text{ additiver Messfehler} \\ \longrightarrow Y^* &= \beta_0 + \beta_1 X + \varepsilon + U \end{aligned}$$

Neuer Ausgleichsfehler: $\varepsilon + U$

Annahme : U und X unabhängig, U und ϵ unabhängig
$\longrightarrow$ Größere Varianz von ε
$\longrightarrow$ Inferenz weiterhin richtig
Fehler bei dem Ausgleich- oder Messfehler sind nicht anders zu behandeln.

▷ $Y^* = Y + U$ additiver Messfehler. Es ergibt sicht ein Modell mit einem zusätzlichen Fehler.

11.1.5 Messfehler in den Einflussgrößen/Kovariablen

Wir betrachten nun genauer die Messfehler in den Kovariablen in Regressionsmodellen.

Haupt-Modell:

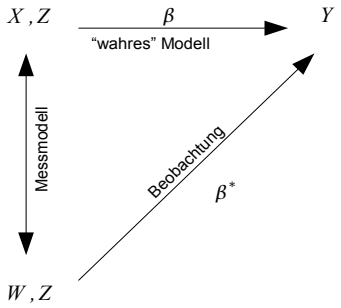
$$E(Y) = f(\beta, X, Z)$$

Wir sind an β_1 interessiert, d.h. den Zusammenhang zwischen Y und den Kovariablen X
Z ist eine weitere Kovariable, gemessen ohne Fehler

Fehler-Modell:

$$X \longleftrightarrow W$$

$$E(Y) = f^*(W, Z, \beta^*)$$



Naive Schätzung:
Beobachtetes Modell = Hauptmodell
aber in vielen Fällen: $f^* \neq f,\ \beta^* \neq \beta$

▷ Das naive Modell kümmert sich nicht um Messfehler, wie z.B. im Modell mit Messfehlern in der Zielgröße.

11.1.6 Differential and non differential measurement error

Annahme eines differential Messfehlers, der die Zielvariable beeinflusst:

$$[Y|X, W] = [Y|X]$$

Für Y gibt es keine weitere Information in U oder X, wenn X bekannt ist.
Dann kann das Fehler- und Haupt-Modell aufgesplittet werden.

$$[Y, W, X] = [Y|X][W|X][X]$$

▷ $f(Y|X)\ldots$ Wahrscheinlichkeit einen Herzinfarkt zu bekommen, wenn ein bestimmter (bekannter) Blutdruck vorliegt.

Aus der substanziellen Sicht:

- Messprozess und Y sind unabhängig
- Blutdruck an einem bestimmten Tag ist irrelevant für das Herzinfarktrisiko, wenn ein Langzeit-Mittel bekannt ist.
- Durchschnitts-exposition ist irrelevant, wenn die individuelle Exposition bekannt ist.
- **Aber** Menschen können ihr Ernährungsverhalten anders betrachten, wenn sie bereits einen Herzinfarkt hatten.

11.2 Einfache lineare Regression

Wir nehmen ein non differential additiven normalen Messfehler an

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon && (10.1)\\ W &= X + U,\ \ (U, X, \varepsilon)\ \text{ unabh.} && (10.2)\\ U &\sim N(0, \sigma_u^2) && (10.3)\\ \varepsilon &\sim N(0, \sigma_\varepsilon^2) && (10.4) \end{aligned}$$

11.2.1 SAS-Simulation für ein lineares Messfehler-Modell

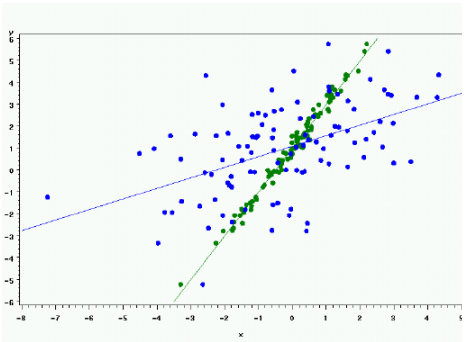
```
/* Simulation von X ~ N(0,1) */
data sim ;
do i=1 to 100; x=rannor(137);
output;
end;
run;

/* Simulation von Y = 1+2*x + epsilon*/

/* Simulation eines Stellvertreters mit additiven Messfehler */
data sim ;
set sim; su=2; /* Messfehler std*/
y= 1+2*x+0.3*rannor(123); w= x+su*rannor (167) ;
run;

/* Plot Zusätze grün für die wahren und blau für die Stellvertreter */
symbol1 c = green V = dot; symbol2 c = blue V = dot;
proc gplot data = sim;
symbol i=none; plot y*x y*w /overlay; /* Scatterplot */
symbol1 i=r; /* Regressionsgeraden */ symbol2 i=r;
plot y*x y*w/ overlay;
run;
```

Results



Effekt eines additiven Messfehlers auf lineare Regession

▷ Angenommen wird hier ein Messfehler, der nicht in Abhängigkeit mit der Zielgröße steht (extremes Beispiel)
▷ Messfehler $\rightarrow \infty \quad \Rightarrow \quad \beta \rightarrow 0$, d.h. Steigung der Regressionsgerade $\rightarrow 0$.

11.2.2 Das beobachtete Modell in der linearen Regression

$E(Y|W) = \beta_0 + \beta_1 E(X|W)$

Angenommen $X \sim N(\mu_x, \sigma_x^2)$, das beobachtete Modell ist:

$$\begin{aligned} E(Y|W) &= \beta_0^* + \beta_1^* W \\ \beta_1^* &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta_1 \\ \beta_0^* &= \beta_0 + \left(1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right)\beta_1 \mu_x \\ Y - \beta_0^* - \beta_1^* W &\sim N\left(0, \sigma_\varepsilon^2 + \frac{\beta_1^2 \sigma_u^2 \sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right) \end{aligned}$$

- Das beobachtete Modell ist immer noch eine lineare Regression !
- Abschwächung von β_1 durch den Faktor $\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$
  "Reliability ratio"
- Verlust von Präzision (höherer Fehler-Term)

$$\triangleright \text{ Wenn } \quad \begin{pmatrix} X \\ W \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} {\sigma_X}^2 & {\sigma_X}^2 \\ {\sigma_X}^2 & {\sigma_X}^2 + {\sigma_U}^2 \end{pmatrix}\right]$$

$$\left[Cov(X,W) = Cov(X, X+U) = Cov(X,X) + Cov(X,U) = {\sigma_X}^2 + 0 \quad \text{, da X,U unabhängig}\right]$$

$$\Rightarrow \quad E(X|W) \text{ ist explizit berechenbar:}$$
$$E(X|W) = \frac{{\sigma_X}^2}{{\sigma_X}^2 + {\sigma_U}^2} W + \left(\mu_X - \frac{{\sigma_X}^2}{{\sigma_X}^2 + {\sigma_U}^2}\mu_X\right)$$

11.2.3 Identifikation

▷ Ich habe gemischte Daten (W und Y). Kann ich meine Daten schätzen?

$$\begin{aligned} (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\varepsilon^2) &\longrightarrow [Y, W] \\ &\longrightarrow \mu_y, \mu_w, \sigma_y^2, \sigma_w^2, \sigma_{wy} \end{aligned}$$

$(\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\varepsilon^2)$ und $(\beta_0^*, \beta_1^*, \mu_x, \sigma_x^2 + \sigma_u^2, 0, \sigma_\varepsilon)$ ergibt die gleichen Verteilungen von (Y,W)

▷ Ich habe 5 Parameter (rechts) zur Verfügung, aber 6 Einflussgrößen (links) – Wenn ich also keine Informationen zu den Messfehlern bekomme, kann ich nichts machen.

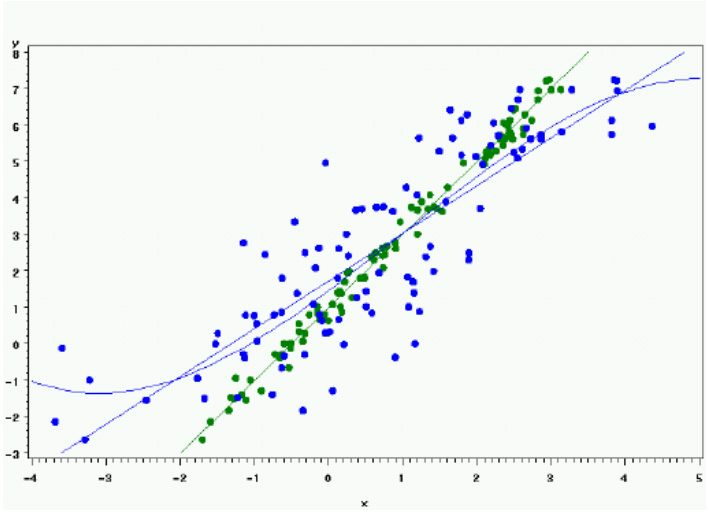
$\Longrightarrow$ Die Modell-Parameter sind nicht identifizierbar

Wir brauchen zusätzliche Informationen (können z.B. aus anderen Studien herangezogen werden), z.B.

- σ_u ist bekannt oder kann geschätzt werden
- $\sigma_u/\sigma_\varepsilon$ ist bekannnt (orthogonale Regression mit exakten Zusammenhang $Y = \beta_0 + \beta_1 X$)

Das Modell mit einer anderen Verteilung von X ist identifizierbar mit höhere Momente.

Beachte, dass das beobachtete Modell von der Verteilung von X abhängt. Es ist keine lineare Regression, wenn X nicht normal ist:



Effekt eines Messfehlers auf die lineare Regression, wenn X gemischt normal ist

11.2.4 Naive KQ-Schätzung

Steigung:

$$\hat{\beta}_{1n} = \frac{S_{yw}}{S_w^2} \overset{n\to\infty}{\longrightarrow} \frac{\sigma_{yw}}{\sigma_w^2} \overset{(*)}{=} \frac{\sigma_{yx}}{\sigma_x^2+\sigma_u^2} \overset{(**)}{=} \beta_1 * \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$$

Intercept:

$$\hat{\beta}_{0n} = \bar{Y} - \beta_{1n}\bar{W} \overset{n\to\infty}{\longrightarrow} = \mu_y - \beta_1 * \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} * \mu_w = \beta_0 + \beta_1 * \left(1 - \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\right) * \mu_x$$

Residuen-Varianz:

$$MSE = S_{Y-\beta_0 n-\beta_1 nW} \overset{n\to\infty}{\longrightarrow} = \sigma_\varepsilon^2 + \frac{\beta_1^2\sigma_u^2\sigma_x^2}{\sigma_x^2+\sigma_u^2}$$

▷ $(*)\, Cov(Y,W) = Cov(Y, X+U) = Cov(Y,X) + Cov(Y,U) = Cov(Y,X)$, da Y,U unabh.
▷ $(**)\, Cov(Y,X) = Cov(\beta_0 + \beta_1 X + \varepsilon, X) = \beta_1 Cov(X,X) = \beta_1\sigma_X^2$
▷ standardmäßige Übertragung auf die multiple Regression.

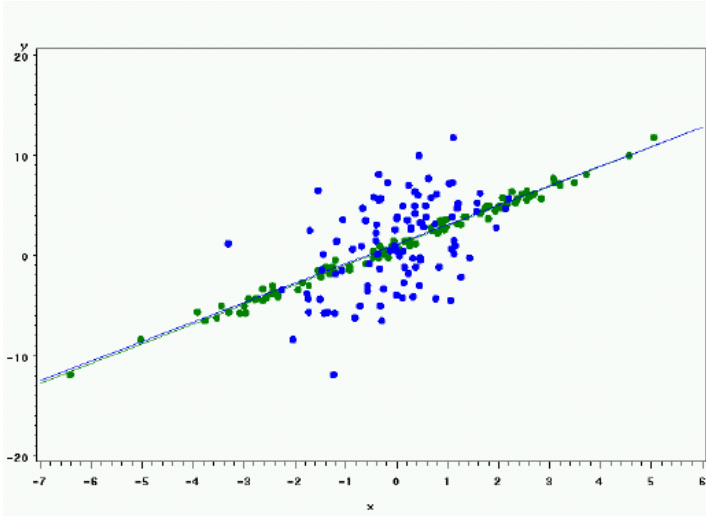
11.2.5 Korrektur von Abschwächung

▷ Um Abschwächung zu korrigieren wird „hochmultiplizieren“ notwendig. Dies geschieht mittels dem bekannten σ_u^2, wordurch die Varianz „aufgebläht“ wird: $V(\hat{\beta}_1) > V(\beta_{1n})$

$$\begin{aligned}
\hat{\beta}_1 &= \hat{\beta}_{1n}\frac{\sigma_x^2+\sigma_u^2}{\sigma_x^2} \\
\hat{\beta}_1 &= \hat{\beta}_{1n}\frac{S_w^2}{S_w^2-\sigma_u^2} \\
\hat{\beta}_0 &= \hat{\beta}_{0n} - \hat{\beta}_1\left(\frac{S_w^2-\sigma_u^2}{S_w^2}\right)\bar{W}
\end{aligned}$$

11.2.6 Berkson-Fehler in einfacher linearer Regression

$$\begin{aligned}
Y &= \beta_0 + \beta_1 X + \varepsilon \\
X &= W + U, \;\; U, (W,Y) \text{ indep.}, \; E(U) = 0
\end{aligned}$$



Effekt von Berkson-Fehler in der linearen Regression

▷ Die Punktewolke verändert sich, aber die Steigung der Regressionsgerade bleibt gleich.

▷ **Herleitung:**

$$E(Y|W) = \beta_0 + \beta_1 E(X|W) = \beta_0 + \beta_1 E(U+W|W) = \beta_0 + \beta_1 W + \beta 1 E(U) = \beta_0 + \beta_1 W$$

Der Unterschied ergibt sich durch die Varianz: $Var(Y|W) = \sigma_\varepsilon^2 + \beta_1\sigma_U^2$

11.2.7 Beobachtete Modell

$$\begin{aligned} E(Y|W) &= \beta_0 + \beta_1 W \\ V(Y|W) &= \sigma_\varepsilon^2 + \beta_1^2 * \sigma_u^2 \end{aligned}$$

- Regressionmodell mit gleichen β
- Messfehler vernachlässigbar
- Verlust von Präzision

11.2.8 Binäre Regression

Logistisch mit additiven non differential Messfehler

$$\begin{aligned} P(Y = 1|X) &= G(\beta_0 + \beta_1 X) \\ G(t) &= (1 + \exp(-t))^{-1} \\ W &= X + U \end{aligned}$$

Beobachtete Modell:

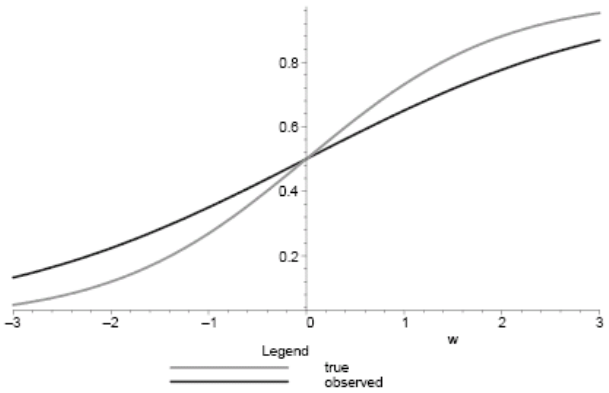
$$P(Y = 1|W) = \int P(Y = 1|X, W) f_{X|W} dx \tag{10.5}$$

$$= \int P(Y = 1|X) f_{X|W} dx \tag{10.6}$$

Wenn wir einen additiven Messfehler brauchen und X und U sind normal, dann ist $X|W$ auch normal

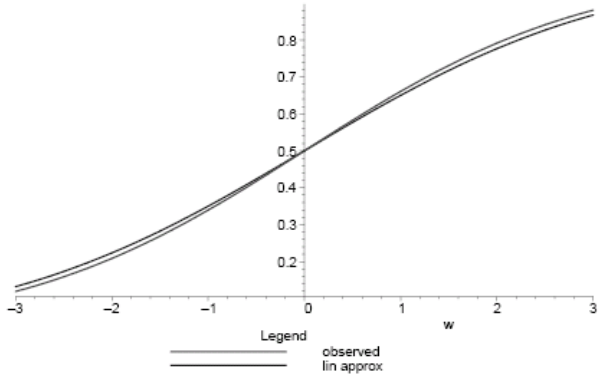
$$P(Y = 1|W) = \int G(\beta_0 + \beta_1 X) f_{X|W} dx$$

11.2.9 Einfach Logistisch



Effekt eines additiven Messfehler in logistischer Regression

11.2.10 Linear Approximation



Beobachter und $G(\beta_0^* + \beta_1^* X)$in logistischer Regression

11.3 Methoden

- Regression Kalibrierung
- Simulation und Extrapolationsverfahren (SIMEX)

11.3.1 Regression Kalibierung

Diese einfache Methode wird weit verbreitet eingesetzt. Sie ist durch verschiedene Autoren angeregt worden: Rosner et al. (1989) Carroll and Stefanski(1990)

1. Finde ein Modell für $E(X|W, Z)$ durch Bereinigung der Daten oder Reproduktion
2. Ersetze das nicht beobachtete X durch die Schätzung $E(X|W, Z)$ im Haupt-Modell
3. Korrigiere die Varianzschätzung durch Bootstrap-Verfahren oder asymtotische Methoden

- Gute Methode in vielen praktischen Situationen
- Kalibrierte Daten können eingebunden werden
- Probleme in höheren nicht-linearen Modellen
- Berkson Fall: $E(X|W) = W \longrightarrow$
  einfache Schätzung = Regression Kalibrierung
- Klassisch : Lineare Regression von X auf W

$$E(X|W) \quad = \quad \frac{\sigma_x^2}{\sigma_w^2} * W + \mu_X * (1 - \frac{\sigma_x^2}{\sigma_w^2})$$

  Korrektur für die Schwächung in linearen Modellen

- Klassisch mit der Annahme von gemischten Modellen:

$$\begin{aligned} E(X|W) = P(I = 1|W) * E(X|W, I = 1) \\ +P(I = 2|W) * E(X|W, I = 2) \end{aligned}$$

11.3.2 SIMEX: Grundidee

Der Effekt eines Messfehlers auf einfache Schätzung wird durch ein Simulations-Experiment analysiert:

Lineare Regression:
Für die Messfehlervarianz σ_u^2, in einer einfachen linearen Regression von Y auf Wwird

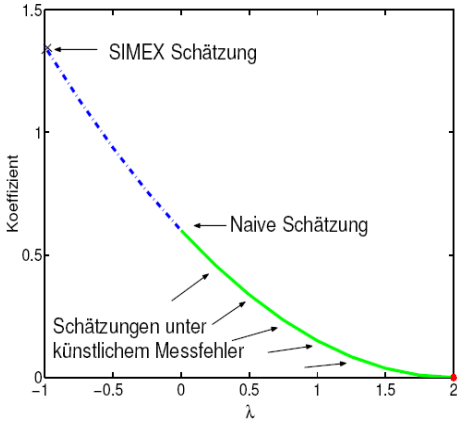
$$p\lim \widehat{\beta}_{\xi*,m} = \frac{\sigma_\xi^2}{\sigma_\xi^2 + (1+\lambda_m)\cdot\sigma_\delta^2} \cdot \beta_\xi$$

geschätzt.
Im Allgemienen haben wir eine Funktion $\beta_n(\sigma_u^2)$.
Offensichtlich $\beta_n(0) = \beta$ (kein Messfehler)

11.3.3 Der SIMEX Algorithmus

Wir nehmen den additiven Messfehler an mit bekannter oder geschätzter Varianz σ_u^2

1. Berechne den einfachen Schätzer $\hat{\beta}_n := \bar{\beta}(0)$
2. **Simulation:**
   Für ein festen Grid von λ,z. B. $\lambda = 0.5, 1, 1.5, 2$
   Für $b = 1, \dots B$
   Simuliere neue fehlerbehaftete Regressoren
   $W_{ib} = W_i + \sqrt{\lambda} * U_{i,b}, U_{i,b} \sim N(0, \sigma_u^2)$
   Berechne die (einfache) Schätzung von $\hat{\beta}_{b,\lambda}$ basierend auf $[Y_i, W_{i,b}]$
   Berechne den durchschnitt über b $\bar{\beta}(\lambda)$
3. **Extrapolation:**
   Passe die Regression $\hat{\beta}(\lambda) = g(\lambda, \gamma)$ an
   $\hat{\beta}_{simex} := g(-1, \hat{\gamma})$



11.3.4 Extrapolation Funktionen

$$\begin{aligned} \text{Linear} &: g(\lambda) = \gamma_0 + \gamma_1\lambda \\ \text{Quadratisch} &: g(\lambda) = \gamma_0 + \gamma_1\lambda + \gamma_2 * \lambda^2 \\ \text{Nicht-Linear} &: g(\lambda) = \gamma_1 + \frac{\gamma_2}{\gamma_3 * \lambda} \end{aligned}$$

- Nicht-lineare Regression ist motiviert durch lineare Regression
- Quadratisch ist für viele Beispiele gut geeignet

11.4 Zusammenfassung

1. Für das Messfehler-Modell grundlegend:
   In vielen Fällen hohe Differenz zwischen Berson-Effekt und Klassischen Messfehler!
2. Additiver klassischer non differential Messfehler führt zu einer Abschwächung
3. Man kann viele Methoden benutzen, z.B. Likelihood
4. Regression Kalibrierung funktioniert in vielen Fällen

12 Bayesianische Inferenz im linearen Modell

12.1 Ansatz

Parameter des Modells: Zufallsgrößen mit (unbekannten) Verteilungen.
Vor der Erhebung: a priori-Verteilung $p(\theta)$
Nach der Erhebung: a posteriori-Verteilung $p(\theta|D)$
Der Satz von Bayes liefert das Werkzeug zur Berechnung von $p(\theta|D)$

$$\begin{array}{rcccc} p(\theta|D) = \frac{P(D|\theta)\cdot p(\theta)}{\int p(D|\theta)\cdot p(\theta)d\theta} & \propto & p(D|\theta) & \cdot & p(\theta) \\ & & \uparrow & & \uparrow \\ \text{Posteriori} & \propto & \text{Likelihood} & & \text{Priori} \end{array}$$

12.2 Gammaverteilung

Eine ZV X ist *gammaverteilt*, wenn für seine Dichtefunktion gilt:

$$f(x) \propto x^{a-1}\exp(-bx)$$

für $a > 0, b > 0$. Schreibweise: $X \sim G(a,\, b)$.
Es gilt:

$$\begin{array}{rcl} E(X) & = & \frac{a}{b} \\ \text{Var } (X) & = & \frac{a}{b^2} \\ \text{Modus}(X) & = & \frac{a-1}{b} \text{ für } a > 1 \end{array}$$

Speziell: $a = \frac{n}{2} \qquad b = \frac{1}{2}$
$\Longrightarrow \quad X \sim \chi^2(n,\, 0)$ X ist (zentral) χ^2-verteilt mit n Freiheitsgraden.

12.3 Inverse Gammaverteilung

Sei $X \sim G(a,\, b)$, dann ist $Y = X^{-1}$ *invers gammaverteilt* mit Parametern a und b.
Schreibweise: $Y \sim$ IG $(a,\, b)$.
Es gilt:

$$\begin{array}{rcl} E(Y) & = & \frac{b}{a-1} \text{ für } a > 1 \\ \text{Var } (Y) & = & \frac{b^2}{(a-1)^2(a-2)} \text{ für } a > 2 \\ \text{Modus } (Y) & = & \frac{b}{a+1} \end{array}$$

$$f(y) \propto y^{-a-1}\exp\left(-\frac{b}{y}\right).$$

12.4 Multivariate t-Verteilung

Ein p-dimensionaler Zufallsvektor X heißt *multivariat t-verteilt*, falls für eine Dichtefunktion gilt:

$$f(x) \propto \left[1 + \frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{\nu}\right]^{-\frac{\nu+p}{2}}.$$

Schreibweise:

$$X \sim t(\nu, \mu, \Sigma).$$

Es gilt:

$$\begin{array}{rcl} E(X) & = & \mu \qquad \text{für } \nu > 1 \\ V(X) & = & \frac{\nu}{\nu-2}\Sigma \text{ für } \nu > 2. \end{array}$$

Jeder Subvektor von X ist wieder multivariat t verteilt mit ν Freiheitsgraden und den entsprechenden Subvektoren/Submatrix aus μ und Σ.

12.5 Normal-Gamma-Verteilung

Eine Zufallsgröße $\theta = (\beta, \tau)$ mit
β: $p' \times 1$-dim. Zufallsvektor
τ: Skalar besitzt eine *Normal-Gamma-Verteilung*
Schreibweise: $NG(\beta_0, \Sigma_0, a, b)$, wenn

$$f(\theta) = f(\beta, \tau) = f_1(\beta \,|\, \tau) \cdot f_2(\tau)$$

mit

$$\beta \,|\, \tau \sim N(\beta_0, \tau^{-1}\Sigma_0) \quad \text{und} \quad \tau \sim G(a, b).$$

Es gilt: Ist $(\beta, \tau) \sim$ NG $(\beta_0, \Sigma_0, a, b)$, so ist die marginale Verteilung von β multivariat t-verteilt:

$$\beta \sim t\left(2a, \beta_0, \frac{b}{a}\Sigma_0\right).$$

12.6 Inferenz bei bekannter Kovarianzmatrix Σ

$$\begin{array}{l} y|\beta \sim N(X\beta, \Sigma) \quad \Sigma \quad \text{bekannt} \\ \beta \sim N(\beta_0, \Sigma_0) \end{array}$$

Posteriori-Verteilung von β ist Normalverteilung mit

$$\begin{array}{rcl} E(\beta|Y) & = & (X'\Sigma^{-1}X + \Sigma_0^{-1})^{-1}(X'\Sigma^{-1}Y + \Sigma_0^{-1}\beta_0) \\ V(\beta|Y) & = & (X'\Sigma^{-1}X + \Sigma_0^{-1})^{-1} \end{array}$$

12.7 Andere Darstellung und Spezialfälle

$$P = \Sigma^{-1}, P_0 = \Sigma^{-1} \quad \text{“Präzision”}$$

$$\beta|Y \quad \text{ist NV mit} \quad E(\beta|Y) = (X'PX + P_0)^{-1}(X'PX + P_0\beta_0)$$

$$\text{und Präzision} \quad X'PX + P_0$$

Sonderfälle:
Homoskedastischer i.i.d. Fall: $\Sigma = \sigma^2 I$

$$\beta|Y \sim N\left[(\frac{1}{\sigma^2}X'X + P_0)^{-1}(\frac{1}{\sigma^2}X'Y + P_0\beta_0), \sigma^2(X'X + P_0^{-1})^{-1}\right]$$

“Völlige Unwissenheit” $P_0 \to 0$

$$\beta|Y \sim N((X'X)^{-1}X'Y, \sigma^2(X'X)^{-1})$$

12.8 Inferenz bei unbekannter Präzision τ

Sei $Y \,|\, \beta, \tau \sim N(X\beta, \tau^{-1}I)$ und $(\beta, \tau) \sim$ NG $(\beta_0, \Sigma_0, a, b)$ mit
Y: $n \times 1$-dim. Zufallsvektor
β: $p' \times 1$-dim. Zufallsvektor.
Dann ist $\beta, \tau \,|\, Y \sim$ NG$(\beta^*, \Sigma^*, a^*, b^*)$ mit

$$\begin{array}{rcl} \beta^* & = & (X'X + \Sigma_0^{-1})^{-1}(X'Y + \Sigma_0^{-1}\beta_0) \\ \Sigma^* & = & (X'X + \Sigma_0^{-1})^{-1} \\ a^* & = & a + \frac{n}{2} \\ b^* & = & b + \frac{1}{2}[Y'Y - B^*] \\ B^* & = & (X'Y - \Sigma_0^{-1}\beta_0)'(X'X + \Sigma_0^{-1})^{-1}(X'Y + \Sigma_0^{-1}\beta_0). \end{array}$$

12.9 Inferenz mit “Jeffrey's prior”

Sei $Y \,|\, \beta, \tau \sim N(X\beta, \tau^{-1}I)$ und $f(\beta, \tau) \propto \tau^{-1}$ (“Jeffrey's prior”).
Dann ist $\beta, \tau \,|\, Y \sim$ NG$(\beta^*, \Sigma^*, a^*, b^*)$ mit

$$\begin{array}{rcl} \beta^* & = & (X'X)^{-1}X'Y \\ \Sigma^* & = & (X'X)^{-1} \\ a^* & = & \frac{n-p'}{2} \\ b^* & = & \frac{1}{2}\left[Y'Y - Y'X(X'X)^{-1}X'Y\right]. \end{array}$$