

# Kapitel 1 - Das einfache lineare Regressionsmodell

## Einfaches lineares Regressionsmodell

Das **einfache lineare Regressionsmodell** hat die Form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

für ein festes numerisches  $x_i$  und  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ . Beachte, dass per Definition gilt  $Y_i | x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$

## Kleinste Quadrate (KQ) Schätzer

Wir schätzen die Parameter  $(\beta_0, \beta_1)$  durch

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2 \quad (1)$$

und nennen  $(\hat{\beta}_0, \hat{\beta}_1)$  den **KQ-Schätzer von  $(\beta_0, \beta_1)$**  und  $\hat{\varepsilon}_i := Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$  die **Residuen**.

## Existenz und Berechnung vom KQ Schätzer

Der KQ-Schätzer existiert und ist eindeutig, falls  $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$ . Dieser lässt sich berechnen als

$$\hat{\beta}_1 = \frac{S_{xY}}{S_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Durch differenzieren von der Gleichung (1) erhält man  $(\hat{\beta}_0, \hat{\beta}_1)$  als Lösung der **Normalengleichungen**

$$\sum_{i=1}^n \hat{\varepsilon}_i = 0$$
$$\sum_{i=1}^n \hat{\varepsilon}_i x_i = 0$$

## Interpretation der Modellparameter

Für  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ,  $i = 1, \dots, n$  mit  $E(Y_i | x_i) = \beta_0 + \beta_1 x_i$  gilt,

- wenn  $x$  um eine **Einheit** steigt, dann steigt  $Y$  im **Erwartungswert** um  $\beta_1$  Einheiten.
- Es gilt  $\beta_0 = E(Y | X = 0)$ .
- Der Parameter  $\sigma$  die erwartete Abweichung der  $Y_i$ -Werte von der Regressionsgerade an.

## Eigenschaften des KQ-Schätzers

Gegeben dem einfachen linearen Modell, gilt für den KQ-Schätzer  $(\hat{\beta}_0, \hat{\beta}_1)$

- Erwartungstreue:  $E(\hat{\beta}_0, \hat{\beta}_1) = (\beta_0, \beta_1)$ .
- $V(\hat{\beta}_1) = \frac{\sigma^2}{n S_x^2}$  und  $V(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{n S_x^2} \right)$ .
- $(\hat{\beta}_0, \hat{\beta}_1)$  ist der maximum-likelihood Schätzer.

## Schätzer für $\sigma^2$

Gegeben dem einfachen linearen Modell mit  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ , gilt

$$\hat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

ist ein erwartungstreuer Schätzer von  $\sigma^2$  und

$$\frac{n-2}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-2}^2.$$

Der KQ-Schätzer  $(\hat{\beta}_0, \hat{\beta}_1)$  und der Schätzer  $\hat{\sigma}^2$  sind stoch.unabhängig.

## Konfidenzintervalle für $\beta_0$ und $\beta_1$

Gegeben dem einfachen linearen Modell mit  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ , gilt für  $\hat{\beta}_1$  und  $\hat{\beta}_0$

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

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

Damit können wir Konfidenzintervalle zum Niveau  $1 - \alpha$  für  $\beta_1$  und  $\beta_0$  erzeugen:

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

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

## Quadratsummenzerlegung

Gegeben sei ein einfaches lineares Modell mit  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$  und  $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$ . Dann gilt

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

SST(otal): Gesamtstreuung von Y  
SSE(rror): Streuung der Residuen  
SSM(odel): Streuung, die das Modell erklärt

## Bestimmtheitsmaß

Unter Verwendung der obigen Notation definieren wir das **Bestimmtheitsmaß** als

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

Es gilt

$$R^2 = r_{xY}^2 = \frac{S_{xY}}{S_x S_Y},$$

wobei  $r_{xY}$  der Bravais-Pearson Korrel.koeffizient ist.

## Interpretation von $R^2$

- $R^2$  beschreibt den Anteil der Varianz von  $Y$ , die durch  $x$  erklärt wird.
- $R$  ist invariant gegenüber linearen linearen Transformationen von  $x$  und  $Y$ .
- $R$  ist symmetrisch bzgl.  $x$  und  $Y$ .
- !  $R^2$  hängt auch von der Streuung von  $x$  in der Stichprobe ab.

## Prognosewert

Gegeben sei ein einfaches lineares Modell mit  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$  und  $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$ ,  $i = 1, \dots, n$ . Sei nun eine weitere Beobachtung  $x_{n+1}$  mit zugehörigem  $Y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$  gegeben. Der **Prognosewert von  $Y_{n+1}$**  ist definiert als  $\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$

### Prognosefehler

Gegeben sei ein einfaches lineares Modell, sowie eine weitere Beobachtung  $x_{n+1}$  mit zugehörigem  $Y_{n+1}$  sowie der Prognosewert  $\hat{Y}_{n+1}$ . Dann gilt

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

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

### Prognoseintervall

Gegeben sei ein einfaches lineares Modell, sowie eine weitere Beobachtung  $x_{n+1}$  mit zugehörigem  $Y_{n+1}$  sowie der Prognosewert  $\hat{Y}_{n+1}$ . Dann können wir für  $Y_{n+1}$  ein Konfidenzintervall zum Niveau  $1 - \alpha$  konstruieren:

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

mit

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

### R-Code

```
# simuliere aus einfachem lin. Modell
beta0 <- 3
beta1 <- 1
sigma <- 2
x <- seq(from = 0, to = 10, by = 0.5)
e <- rnorm(length(x), sd = sigma)
y <- beta0 + beta1 * x + e
dat <- data.frame(x, y)

# Lineares Modell erzeugen
reg = lm(y ~ x, data = dat)
summary(reg)

# Konfidenzintervalle
confint(reg, level = 0.95)
```

### Interpretation von transformierten Modellen

- Log-Log-Modell:

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$$

Wenn  $x_i$  um den Faktor  $a$  steigt, dann steigt  $Y_i$  im Erwartungswert um den Faktor  $a^{\beta_1} = e^{\beta_1 \log(a)}$ .

Alternativ: Wenn  $x_i$  um 1% steigt, dann steigt  $Y_i$  im Erwartungswert um  $(e^{\beta_1 \log(1.01)} - 1)\%$ .

- Linear-Log-Modell:

$$Y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$$

Wenn  $x_i$  um  $p\%$  steigt, dann steigt  $Y_i$  im Erwartungswert um  $\beta_1 \cdot \log(1 + p)\%$ .

Alternativ: Wenn  $x_i$  um 1% steigt, dann steigt  $Y_i$  im Erwartungswert um approximativ  $\beta_1$  Einheiten.

- Log-Linear-Modell:

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Wenn  $x_i$  um eine Einheit steigt, dann steigt  $Y_i$  im Erwartungswert um den Faktor  $e^{\beta_1}$ .

### Vorlesung

$R^2$  ist abhängig von  $X$ . Das heißt über mehrere Studien hinweg, die das gleiche messen, ist  $R^2$  nur vergleichbar, wenn auch  $X$  vergleichbar ist. Je sicherer wir mit unserem Schätzer sein wollen, desto höher sollten wir die Varianz von  $X$  einstellen. Gegeben, dass der Zusammenhang tatsächlich linear ist, würde eine höhere Varianz von  $X$  zu einer geringeren Varianz von  $\hat{\beta}_1$  führen.

Im multiplen Reg.modell ist es KEINE Annahme, dass  $x_i, x_j$  unabhängig voneinander sind. Es wäre nur praktisch für die Interpretation der Effekte. Das „magische“ am multiplen Reg.modell ist, dass ich für verschiedene Größen kontrollieren/korrigieren kann.

Erwartungstreue gilt auch bei Abhängigkeit und normalverteilt ist nicht nötig. Varianzformel benötigt Unabhängigkeit.

### Annahmen des linearen Regressionsmodells

Gegeben sei das einfache (oder multiple) lineare Regressionsmodell mit

$$Y = X\beta + \varepsilon \quad (1)$$

$$E(\varepsilon) = 0 \quad (2)$$

$$V(\varepsilon_i) = \sigma^2, \text{ für alle } i = 1, \dots, n \quad (3)$$

$$\varepsilon_i \text{ sind paarweise unabhängig voneinander} \quad (4)$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ für alle } i = 1, \dots, n \quad (5)$$

Die folgende Tabelle gibt an, welche Annahmen für die jeweiligen Schätzer, Eigenschaften, Größen etc. benötigt werden.

	(1)	(2)	(3)	(4)	(5)
KQ-Schätzer	✓				
ML-Schätzer	✓	✓	✓	✓	✓
$E(\hat{\beta}) = \beta$	✓	✓			
$\hat{\beta} \sim \mathcal{N}(\beta, V(\hat{\beta}))$	✓	✓	✓	✓	(✓)
Konfidenzintervalle	✓	✓	✓	✓	(✓)
Prädiktionsintervalle	✓	✓	✓	✓	✓

(✓) bedeutet, dass die Annahme nicht benötigt wird, wenn der Stichprobenumfang  $n$  groß genug ist.

# Kapitel 2 - Das multiple lineare Regressionsmodell

## Multiple lineares Regressionsmodell

Das **multiple lineare Regressionsmodell** hat die Form

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \dots + \beta_p x_{ip}}_{\mathbf{x}_i^\top = (1, x_{i1}, \dots, x_{ip})} + \varepsilon_i; i = 1, \dots, n$$

oder in Matrix-Vektor Notation:  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  mit

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Wir nehmen dabei an, dass  $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$  eine feste Design-Matrix mit vollem Rang ist und dass  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ . Wir definieren  $p' := p + 1$ .

## Kleinste Quadrate (KQ) Schätzer

Wir schätzen den Parameter(vektor)  $\boldsymbol{\beta}$  durch

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{p'}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (6)$$

und nennen  $\hat{\boldsymbol{\beta}}$  den **KQ-Schätzer von  $\boldsymbol{\beta}$**  und  $\hat{\varepsilon}_i := Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$  die **Residuen**.

## Existenz und Berechnung vom KQ Schätzer

Der KQ-Schätzer existiert und ist eindeutig, falls  $\mathbf{X}^\top \mathbf{X}$  invertierbar ist. Dieser lässt sich berechnen als

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

Durch differenzieren von der Gleichung (2) erhält man  $\hat{\boldsymbol{\beta}}$  als Lösung der **Normalengleichung**

$$\mathbf{X}^\top \hat{\boldsymbol{\varepsilon}} = 0$$

## Gauss-Markov-Theorem

Sei das Modell  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  gegeben mit  $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$  und  $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ . Dann ist der KQ-Schätzer  $\hat{\boldsymbol{\beta}}$  der beste lineare erwartungstreue Schätzer (best linear unbiased estimator, BLUE) von  $\boldsymbol{\beta}$ .

Das heißt, dass für jeden anderen linearen erwartungstreuen Schätzer  $\tilde{\boldsymbol{\beta}}$  von  $\boldsymbol{\beta}$  gilt  $\mathbb{V}(\hat{\boldsymbol{\beta}}) \leq \mathbb{V}(\tilde{\boldsymbol{\beta}})$ .

## Interpretation der Modellparameter

- **ceteris paribus**: alle anderen x-Variablen bleiben konstant.
- (Theoretische) Interpretation: Steigt  $x_k$  um eine Einheit, so steigt  $Y$  (ceteris paribus) im Erwartungswert um  $\beta_k$  Einheiten.
- (Empirische) Interpretation: Steigt  $x_k$  um eine Einheit, so steigt  $Y$  (ceteris paribus) im Durchschnitt um  $\hat{\beta}_k$  Einheiten.
- !  $\beta_k$  charakterisiert den Einfluss von  $x_k$  unter Berücksichtigung der übrigen Variablen (Confounder-Korrektur). Das heißt, dass in einem einfachen linearen Regressionsmodell mit  $Y_i = \beta_0 + \beta'_k x_{ik} + \varepsilon_i$  wäre im Allgemeinen  $\beta'_k \neq \beta_k$ .

## Eigenschaften des KQ-Schätzers

Gegeben dem multiplen linearen Modell, gilt für den KQ-Schätzer  $\hat{\boldsymbol{\beta}}$

- Erwartungstreue:  $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ .  
! Gilt auch ohne die Annahme  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ , solange  $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ .  
! Gilt auch ohne die Annahme  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ , solange  $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$
- $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$

## Hat-Matrix und Residualmatrix

Gegeben dem multiplen linearen Modell mit  $\text{rang}(\mathbf{X}) = p'$  gilt

$$\hat{\mathbf{Y}} := \mathbf{X} \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}}_{\hat{\boldsymbol{\beta}}}$$

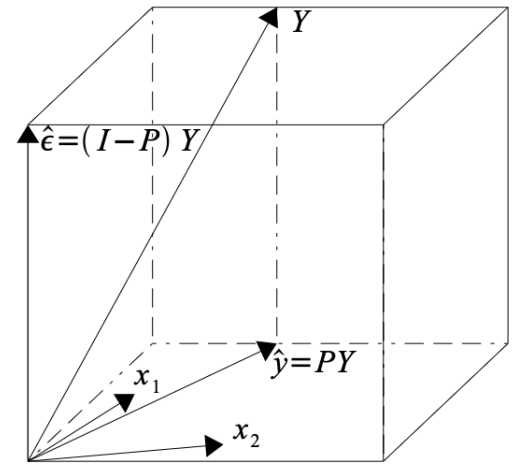
$$\mathbf{P} := \underbrace{\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{n \times n}$$

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

$$\mathbf{Q} := \mathbf{I} - \mathbf{P}$$

$\mathbf{P}$  heißt **Hat-Matrix** und  $\mathbf{Q}$  heißt **Residualmatrix**.

## Geometrische Interpretation



Die KQ-Schätzung ist eine orthogonale Projektion von  $\mathbf{Y}$  auf den von den  $\mathbf{x}$ -Vektoren aufgespannten Unterraum.

## Eigenschaften von $\mathbf{P}$ und $\mathbf{Q}$

Die Hat-Matrix  $\mathbf{P}$  und die Residualmatrix  $\mathbf{Q}$  sind Projektionsmatrizen und zueinander orthogonal:

$$\mathbf{P}^\top = \mathbf{P} \text{ und } \mathbf{P}^2 = \mathbf{P}$$

$$\mathbf{Q}^\top = \mathbf{Q} \text{ und } \mathbf{Q}^2 = \mathbf{Q}$$

$$\mathbf{P}\mathbf{Q} = \mathbf{Q}\mathbf{P} = \mathbf{0}.$$

Daraus folgt

$$\mathbb{V}(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{P}$$

$$\mathbb{V}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2 \mathbf{Q}, \text{ da } \hat{\boldsymbol{\varepsilon}} = \mathbf{Q}\boldsymbol{\varepsilon}$$

## Schätzer für $\sigma^2$

Gegeben dem multiplen linearen Modell, gilt

$$\hat{\sigma}^2 := \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n - p'} = \frac{1}{n - p'} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

ist ein erwartungstreu Schätzer von  $\sigma^2$ .

! Gilt auch ohne die Annahme  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ , solange  $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$  und  $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$

# Kapitel 3 - Quadratsummenzerlegung und statistische Inferenz im multiplen linearen Regressionsmodell

## Quadratsummenzerlegung

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$ . Dann gilt

$$\underbrace{(\mathbf{Y} - \bar{\mathbf{Y}})^\top (\mathbf{Y} - \bar{\mathbf{Y}})}_{SST} = \underbrace{(\mathbf{Y} - \hat{\mathbf{Y}})^\top (\mathbf{Y} - \hat{\mathbf{Y}})}_{SSE} + \underbrace{(\hat{\mathbf{Y}} - \bar{\mathbf{Y}})^\top (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})}_{SSM}.$$

$SST(\text{otal})$ :	Gesamt-Quadratsumme (korrigiert)
$SSE(\text{rror})$ :	Fehler-Quadratsumme
$SSM(\text{odel})$ :	Modell-Quadratsumme

## Quadratsummenzerlegung ohne $\beta_0$

Gegeben sei das multiple lineare Regressionsmodell mit, aber ohne Absolutglied  $\beta_0$ . Dann gilt

$$\underbrace{\mathbf{Y}^\top \mathbf{Y}}_{SST^*} = \underbrace{(\mathbf{Y} - \hat{\mathbf{Y}})^\top (\mathbf{Y} - \hat{\mathbf{Y}})}_{SSE} + \underbrace{\hat{\mathbf{Y}}^\top \hat{\mathbf{Y}}}_{SSM^*}.$$

$SST^*$ :	Gesamt-Quadratsumme (nicht korrigiert)
$SSE$ :	Fehler-Quadratsumme (wie zuvor)
$SSM^*$ :	Modell-Quadratsumme (nicht korrigiert)

## Erwartungswerte der Quadratsummen

Gegeben sei das multiple lineare Regressionsmodell mit den üblichen Annahmen. Wir definieren

$$\mathbf{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ und } \mathbf{P}_e = \mathbf{e}(\mathbf{e}^\top \mathbf{e})^{-1} \mathbf{e}^\top \text{ und } \mathbf{Q}_e = \mathbf{I} - \mathbf{P}_e.$$

Dann gilt

$$\begin{aligned} \mathbf{P}_e \mathbf{Y} &= \bar{\mathbf{Y}} \quad \text{und} \quad \mathbf{Q}_e \mathbf{Y} = \mathbf{Y} - \bar{\mathbf{Y}} \\ \mathbb{E}(SST^*) &= \sigma^2 n + \beta^\top \mathbf{X}^\top \mathbf{X} \beta \\ \mathbb{E}(SST) &= \sigma^2 (n - 1) + \beta^\top (\mathbf{Q}_e \mathbf{X})^\top (\mathbf{Q}_e \mathbf{X}) \beta \\ \mathbb{E}(SSE) &= \sigma^2 (n - p') \\ \mathbb{E}(SSM^*) &= \sigma^2 p' + \beta^\top \mathbf{X}^\top \mathbf{X} \beta \\ \mathbb{E}(SSM) &= \sigma^2 (p' - 1) + \beta^\top (\mathbf{Q}_e \mathbf{X})^\top (\mathbf{Q}_e \mathbf{X}) \beta \end{aligned}$$

Wir können diese Eigenschaften zur Konstruktion von Tests verwenden. Es gilt nämlich unter anderem

$$\beta = 0 \implies \mathbb{E}(SST^*) = \sigma^2 n$$

$$\beta_1 = \dots = \beta_p = 0 \implies \mathbb{E}(SSM) = \sigma^2 (p' - 1)$$

## Mittlere Quadratsummen

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

$$\begin{aligned} \text{MSE} &= \frac{SSE}{n - p'} = \hat{\sigma}^2 \\ \text{MSM} &= \frac{SSM}{p' - 1} \\ \text{MST} &= \frac{SST}{n - 1} \\ \text{MSM}^* &= \frac{SSM^*}{p'} \\ \text{MST}^* &= \frac{SST^*}{n} \end{aligned}$$

## Multivariate Normalverteilung

Eine Zufallsvariable  $\mathbf{Z} \in \mathbb{R}^n$  heißt **multivariat normalverteilt** mit Erwartungswert  $\boldsymbol{\mu} \in \mathbb{R}^n$  und positiv definiter Kovarianzmatrix  $\Sigma \in \mathbb{R}^{n \times n}$ , wenn ihre Dichte gegeben ist durch

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right).$$

Wir schreiben  $\mathbf{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ .

## Eigenschaften von $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$

Sei  $\mathbf{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$  und  $\mathbf{A} \in \mathbb{R}^{m \times n}$  mit  $\text{rang}(\mathbf{A}) = m$ . Dann gilt

- $\mathbb{E}(\mathbf{Z}) = \boldsymbol{\mu}$
- $\mathbb{V}(\mathbf{Z}) = \Sigma$
- $\mathbf{AZ} \sim \mathcal{N}_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\Sigma\mathbf{A}^\top)$
- Es existiert eine orthogonale Matrix  $\mathbf{T} \in \mathbb{R}^{n \times n}$  mit  $\mathbf{T}\Sigma\mathbf{T}^\top = \text{diag}(\lambda_1, \dots, \lambda_n)$ , so dass

$$\mathbf{TZ} \sim \mathcal{N}_n(\mathbf{T}\boldsymbol{\mu}, \text{diag}(\lambda_1, \dots, \lambda_n))$$

## Chi-Quadrat Verteilung

Sei  $\mathbf{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{I})$ , so heißt  $\mathbf{W} = \mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^n Z_i^2$  (nicht-zentral) **Chi-Quadrat-verteilt** und wir schreiben

$$\mathbf{W} \sim \chi^2(n, \delta).$$

Wir nennen  $n$  die **Zahl der Freiheitsgrade** und  $\delta = \boldsymbol{\mu}^\top \boldsymbol{\mu}$  den **Nicht-Zentralitätsparameter**. Es gilt

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

## t-Verteilung

Seien  $Z \sim \mathcal{N}(\delta, 1)$  und  $W \sim \chi^2(n, 0)$  unabhängig. Dann heißt  $T = \frac{Z}{\sqrt{W/n}}$  (nicht-zentral) **t-verteilt** mit  $n$  **Freiheitsgraden** und **Nicht-Zentralitätsparameter**  $\delta$  und wir schreiben

$$T \sim t(n, \delta).$$

Es gilt

$$\mathbb{E}(T) = \delta \sqrt{\frac{n}{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \text{ für } n > 1$$

## F-Verteilung

Sei  $W_1 \sim \chi^2(n_1, \delta)$  und  $W_2 \sim \chi^2(n_2, 0)$  unabhängig. Dann heißt  $X = \frac{W_1/n_1}{W_2/n_2}$  (nicht-zentral) **F-verteilt** mit  $n_1$  und  $n_2$  **Freiheitsgraden** und **Nicht-Zentralitätsparameter**  $\delta$  und wir schreiben

$$X \sim F(n_1, n_2, \delta).$$

Es gilt

$$\mathbb{E}(X) = \frac{n_2 + \frac{n_2 \delta}{n_1}}{n_2 - 2} \text{ für } n_2 > 2$$

### Satz von Cochran

Sei

- $\mathbf{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ ,
- $\mathbf{A} \in \mathbb{R}^{n \times n}$  mit  $\text{rang}(\mathbf{A}) = r$  und  $\mathbf{A}^2 = \mathbf{A}$ ,
- $\mathbf{B} \in \mathbb{R}^{n \times n}$  mit  $\mathbf{B}^2 = \mathbf{B}$  und  $\mathbf{B}^\top = \mathbf{B}$ ,
- $\mathbf{C} \in \mathbb{R}^{m \times n}$ .

dann gilt

$$\mathbf{Z}^\top \mathbf{A} \mathbf{Z} \sim \chi^2(r, \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu})$$

$$\mathbf{C} \mathbf{A} = \mathbf{0} \implies \mathbf{C} \mathbf{Z} \text{ und } \mathbf{Z}^\top \mathbf{A} \mathbf{Z} \text{ sind unabhängig.}$$

$$\mathbf{A} \mathbf{B} = \mathbf{0} \implies \mathbf{Z}^\top \mathbf{A} \mathbf{Z} \text{ und } \mathbf{Z}^\top \mathbf{B} \mathbf{Z} \text{ sind unabhängig.}$$

### Verteilung des KQ-Schätzers

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$  und den üblichen Annahmen über  $\varepsilon$ . Dann gilt für den KQ-Schätzer  $\hat{\beta}$ :

$$\hat{\beta} \sim \mathcal{N}_{p'}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$

$$\hat{\Sigma}_{\hat{\beta}} := \hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}$$

$$(n - p') \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - p')$$

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

$$\hat{\sigma}_{\hat{\beta}_k}^2 := (\hat{\Sigma}_{\hat{\beta}})_{kk}$$

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}_{\hat{\beta}_k}^2}} \sim t(n - p', 0)$$

Um exakte Tests durchzuführen, ist die Normalverteilungsannahme für  $\varepsilon$  notwendig. Jedoch gelten einige Eigenschaften auch approximativ ohne diese Annahme. Nehmen wir stattdessen an, dass gilt

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^\top \mathbf{X} = \mathbf{V}, \quad \mathbf{V} \text{ positive definit.}$$

Dann gilt weiterhin, dass  $\hat{\beta}$  und  $\hat{\sigma}^2$  consistent sind und

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \sigma^2 \mathbf{V}^{-1}).$$

Daraus folgt die für die praxis essenzielle Eigenschaft

$$\hat{\beta} \stackrel{a}{\sim} \mathcal{N}_p(\beta, \hat{\sigma}^2 \frac{(\mathbf{X}^\top \mathbf{X})^{-1}}{n}) \text{ für großes } n.$$

### Overall-Test

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$ . Dann gilt für die mittleren Quadratsummen

$$F_0 = \frac{MSM}{MSE} \sim F(p' - 1, n - p', \frac{\beta^\top (\mathbf{Q}_e \mathbf{X})^\top (\mathbf{Q}_e \mathbf{X}) \beta}{\sigma^2})$$

Wir können damit den **Overall-Test** durchführen, um die Hypothese

$$H_0^O : \beta_1 = \dots = \beta_{p'} = 0$$

zu testen. Wir lehnen  $H_0^O$  ab, wenn  $F_0 > F_{1-\alpha}(p' - 1, n - p')$ .

### Allgemeine lineare Hypothese

Es sollen Hypothesen der Form  $\mathbf{H}_0 : \mathbf{A}\beta = \mathbf{c}$  getestet werden, wobei  $\mathbf{A} \in \mathbb{R}^{a \times p'}$  mit  $\text{rang}(\mathbf{A}) = a$  und  $\mathbf{c} \in \mathbb{R}^a$ .

Wir definieren

$$SSH := (\mathbf{A}\hat{\beta} - \mathbf{c})^\top (\mathbf{A}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{A}^\top)^{-1} (\mathbf{A}\hat{\beta} - \mathbf{c})$$

$$MSH := \frac{SSH}{a}$$

$$\delta_{SSH} := (\mathbf{A}\beta - \mathbf{c})^\top (\mathbf{A}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{A}^\top)^{-1} (\mathbf{A}\beta - \mathbf{c}).$$

SSH beschreibt die Quadratsumme der Abweichung von der Hypothese  $\mathbf{A}\beta = \mathbf{c}$ .

Es gilt

$$\frac{SSH}{\sigma^2} \sim \chi^2(a, \frac{\delta_{SSH}}{\sigma^2}),$$

$$\frac{MSH}{MSE} \sim F(a, n - p', \frac{\delta_{SSH}}{\sigma^2}).$$

Damit können wir nun die Hypothese  $H_0 : \mathbf{A}\beta = \mathbf{c}$  testen. Wir lehnen  $H_0$  ab, wenn

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

Dieses Vorgehen können wir als Wald-Test identifizieren und in diesem Fall entspricht dieser einem Likelihood-Quotienten-Test, ist also optimal.

Der Test vergleicht intuitiv den SSE des Modells mit dem SSE des Modells unter  $H_0 : \mathbf{A}\beta = \mathbf{c}$ .

Für  $n \rightarrow \infty$  gilt  $\frac{MSH}{MSE} \rightarrow \frac{SSH}{\sigma^2}$ , d.h. im Allgemeinen ist die F-Verteilung asymptotisch Chi-Quadrat-verteilt.

### Partielle Quadratsummen

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$ . Die zu der Hypothese  $\mathbf{H}_0 : \beta_k = 0$  gehörende Quadratsumme bzgl. des Gesamtmodells heißt **partielle Quadratsumme** und wird definiert als

$$R(\beta_k | \beta_1, \dots, \beta_{k-1}, \beta_{k+1}, \dots, \beta_{p'}) = SSE(M_{-k}) - SSE$$

wobei  $M_{-k}$  das Modell mit  $\beta_k = 0$  ist.

### Sequentielle Quadratsummen

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$ . Wir definieren das Modell  $M_k$  als das Modell

$$M_k : \mathbf{Y} = \beta_0 + \beta_1 \mathbf{x} + \dots + \beta_k \mathbf{x}_k + \varepsilon$$

Die zu dem Modell  $M_k$  gehörende Quadratsumme heißt **sequentielle Quadratsumme** und wird definiert als

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

für  $k = 1, \dots, p'$ .

Es gilt

$$SST = \sum_{k=1}^{p'} R(\beta_k | \beta_0, \dots, \beta_{k-1}) + SSE.$$

### R-Code

```
# Teste lineare Hypothese der Form
# H_0: A*beta = c
A <- matrix(c(...))
c <- c(...)
car::linearHypothesis(model, A, c)

# Wenn c != 0, dann benutzen wir im
# model einen offset().
```

### Konfidenzellipsoid

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$ . Das **Konfidenzellipsoid** für  $\beta$  zum Niveau  $1 - \alpha$  ist gegeben als

$$\left\{ \beta \in \mathbb{R}^{p'} \mid \frac{1}{p'} (\beta - \hat{\beta})^\top \hat{\Sigma}_{\hat{\beta}}^{-1} (\beta - \hat{\beta}) \leq F_{1-\alpha}(p', n - p') \right\}$$

### Prognosewert

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$ . Sei nun eine weitere Beobachtung  $\mathbf{x}_{n+1}$  mit zugehörigem unbekannten  $Y_{n+1}$  gegeben. Der **Prognosewert von  $Y_{n+1}$**  ist gegeben als  $\hat{Y}_{n+1} = \mathbf{x}_{n+1}^\top \hat{\boldsymbol{\beta}}$

### Prognosefehler und Prognoseintervall

Gegeben sei das multiple lineare Regressionsmodell mit  $\text{rang}(\mathbf{X}) = p'$ . Sei nun eine weitere Beobachtung  $\mathbf{x}_{n+1}$  mit zugehörigem unbekannten  $Y_{n+1}$  gegeben. Sei  $\hat{Y}_{n+1}$  der Prognosewert. Dann gilt

$$\mathbb{E}(\hat{Y}_{n+1} - Y_{n+1}) = 0$$

$$\mathbb{V}(\hat{Y}_{n+1} - Y_{n+1}) = \sigma^2 \left[ 1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1} \right]$$

Wir konstruieren das **Prognoseintervall** für  $Y_{n+1}$  zum Niveau  $1 - \alpha$  als

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

mit

$$\hat{\sigma}_{\hat{Y}_{n+1}} = \hat{\sigma}^2 \left[ 1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1} \right].$$

Wir konstruieren das **Prognoseintervall** für  $\mathbb{E}(Y_{n+1}) = \mu_{n+1}$  zum Niveau  $1 - \alpha$  als

$$[\hat{Y}_{n+1} - \hat{\sigma}_{\hat{\mu}_{n+1}}^2 t_{1-\alpha/2}(n-p'); \hat{Y}_{n+1} + \hat{\sigma}_{\hat{\mu}_{n+1}}^2 t_{1-\alpha/2}(n-p')]$$

mit

$$\hat{\sigma}_{\hat{\mu}_{n+1}}^2 = \hat{\sigma}^2 \left[ \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1} \right].$$



# Kapitel 4 - Diskrete Einflußgrößen

## Kodierung

Sei  $C$  eine nominale Variable mit  $K$  Ausprägungen.

### Dummy/Referenz-Kodierung:

Wir definieren  $K$  neue Variablen  $Z_1, \dots, Z_K$  als

$$Z_k(C) = \begin{cases} 1, & \text{falls } C = k \\ 0, & \text{sonst} \end{cases}$$

$Z_1, \dots, Z_K$  sind abhängig, da  $Z_K = 1 - \sum_{k=1}^{K-1} Z_k$

**Effekt-Kodierung:** Wir definieren  $K - 1$  neue Variablen  $Z_1^e, \dots, Z_{K-1}^e$  als

$$Z_k^e(C) = \begin{cases} 1, & \text{falls } C = k \\ -1, & \text{falls } C = K \\ 0, & \text{sonst} \end{cases}$$

Note:  $Z_k(C) = \begin{pmatrix} Z_k(C_1) \\ \vdots \\ Z_k(C_n) \end{pmatrix}$  und  $Z_k^e(C) = \begin{pmatrix} Z_k^e(C_1) \\ \vdots \\ Z_k^e(C_n) \end{pmatrix}$

## Setup einfache Varianzanalyse

Im folgenden betrachten wir die einfache Varianzanalyse mit nur einer diskreten Einflußgröße  $C = \begin{pmatrix} C_1 \\ \vdots \\ C_n \end{pmatrix}$  mit  $K$  Ausprägungen. Sei  $n_k$  dabei die Anzahl der Beobachtungen mit  $C_i = k$ .

## Mittelwertsmodell

Das **Mittelwertsmodell** ist gegeben durch

$$Y_{kl} = \mu_k + \epsilon_{kl} \quad l = 1, \dots, n_k \quad k = 1, \dots, K$$

oder in Matrix-Vektor Notation:

$$\mathbf{Y} = (Z_1(C) \cdots Z_K(C)) \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} + \epsilon$$

Bei dem Mittelwertsmodell gibt es keinen Intercept und die  $\mu_k$  sind die Mittelwerte der  $k$ -ten Gruppe. Der Effekt der  $k$ -ten Gruppe ist also  $\mu_k$ .

## Mittelwertsmodell Beispiel

Für  $K = 3$  Ausprägungen und  $n_k = 2$  für alle  $k = 1, 2, 3$  erhalten wir als Mittelwertsmodell:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

## Modell mit Effekt-Kodierung

Das **Modell mit Effekt-Kodierung** ist gegeben durch

$$Y_{kl} = \mu + \tau_k + \epsilon_{kl}; \quad \tau_K = - \sum_{k=1}^{K-1} \tau_k$$

für  $l = 1, \dots, n_k \quad k = 1, \dots, K$  oder in Matrix-Vektor Notation:

$$\mathbf{Y} = (e \ Z_1^e(C) \cdots Z_{K-1}^e(C)) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K-1} \end{pmatrix} + \epsilon$$

Bei dem Modell mit Effekt-Kodierung gibt es einen Intercept  $\mu$  und die  $\tau_k$  sind die Abweichungen der  $k$ -ten Gruppe vom Gesamtmittelwert bzw. vom Intercept  $\mu$ . Der Effekt der  $k$ -ten Gruppe ist also  $\mu + \tau_k$ .

## Modell mit Effekt-Kodierung Beispiel

Für  $K = 3$  Ausprägungen und  $n_k = 2$  für alle  $k = 1, 2, 3$  erhalten wir als Modell mit Effekt-Kodierung:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

## Modell mit Referenz-Kodierung

Das **Modell mit Referenz-Kodierung** ist gegeben durch

$$Y_{kl} = \mu_K + \tau_k + \epsilon_{kl}; \quad \tau_K = 0$$

für  $l = 1, \dots, n_k \quad k = 1, \dots, K$  oder in Matrix-Vektor Notation:

$$\mathbf{Y} = (e \ Z_1(C) \cdots Z_{K-1}(C)) \begin{pmatrix} \mu_K \\ \tau_1 \\ \vdots \\ \tau_{K-1} \end{pmatrix} + \epsilon$$

Beim Modell mit Referenz-Kodierung gibt es einen Intercept  $\mu_K$  der den Mittelwert der  $K$ -ten Gruppe angibt und die  $\tau_k$  sind die Abweichungen der  $k$ -ten Gruppe vom Mittelwert der  $K$ -ten Referenz-Gruppe. Der Effekt der  $k$ -ten Gruppe ist also  $\mu_K + \tau_k$  für  $k = 1, \dots, K - 1$  und  $\mu_K$  für  $k = K$ .

## Modell mit Referenz-Kodierung Beispiel

Für  $K = 3$  Ausprägungen und  $n_k = 2$  für alle  $k = 1, 2, 3$  erhalten wir als Modell mit Referenz-Kodierung:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_3 \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

## Bemerkungen-Kodierung

Alle Modellvarianten führen zur gleichen Modellanpassung ( $R^2$ ). Die Parameter haben aber unterschiedliche Interpretationen. Parameter und deren Schätzer sind aber ineinander umrechenbar.

Wir können folgende Nullhypothese für den Effekt von  $C$  testen:

Mittelwertsmodell	$H_0 : \mu_1 = \dots = \mu_K$
Effekt-Kodierung	$H_0 : \tau_1 = \dots = \tau_{K-1} = 0$
Referenz-Kodierung	$H_0 : \tau_1 = \dots = \tau_{K-1} = 0$

## Setup zweifaktorielle Varianzanalyse

Im folgenden betrachten wir zwei diskrete Einflußgrößen  $\mathbf{C} = \begin{pmatrix} C_1 \\ \vdots \\ C_n \end{pmatrix}$  und  $\mathbf{D} = \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix}$  mit  $K_C$  bzw.  $K_D$  Ausprägungen. Sei  $n_{k,l}$  dabei die Anzahl der Beobachtungen mit  $C_i = k$  und  $D_j = l$ .

! Hier ist die Mittelwertsdarstellung bzw. das Mittelwertsmodell nicht möglich, da dieser davon abhängig ist, welche Variable zuerst kodiert wird.

## Modell mit Effekt-Kodierung (mehrfaktoriell)

Das **Modell mit Effekt-Kodierung** ist gegeben durch

$$\mathbf{Y} = (\mathbf{e} \quad \mathbf{Z}_1^e(\mathbf{C}) \cdots \mathbf{Z}_{K_C-1}^e(\mathbf{C}) \quad \mathbf{Z}_1^e(\mathbf{D}) \cdots \mathbf{Z}_{K_D-1}^e(\mathbf{D})) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K_C-1} \\ \gamma_1 \\ \vdots \\ \gamma_{K_D-1} \end{pmatrix} + \epsilon$$

mit  $\tau_{K_C} = -\sum_{k=1}^{K_C-1} \tau_k$  und  $\gamma_{K_D} = -\sum_{k=1}^{K_D-1} \gamma_k$ .

Bei dem Modell mit Effekt-Kodierung gibt es einen Intercept  $\mu$  und die  $\tau_k$  und  $\gamma_l$  sind die Abweichungen der Gruppe mit  $C = k$  bzw.  $D = l$  vom Gesamtmittelwert bzw. vom Intercept  $\mu$ .

## Modell mit Referenz-Kodierung (mehrfakt.)

Das **Modell mit Referenz-Kodierung** ist gegeben durch

$$\mathbf{Y} = (\mathbf{e} \quad \mathbf{Z}_1(\mathbf{C}) \cdots \mathbf{Z}_{K_C-1}(\mathbf{C}) \quad \mathbf{Z}_1(\mathbf{D}) \cdots \mathbf{Z}_{K_D-1}(\mathbf{D})) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K_C-1} \\ \gamma_1 \\ \vdots \\ \gamma_{K_D-1} \end{pmatrix} + \epsilon$$

mit  $\tau_{K_C} = 0$  und  $\gamma_{K_D} = 0$ .

Bei dem Modell mit Referenz-Kodierung gibt es einen Intercept  $\mu$  der den Mittelwert der Gruppe mit  $C = K_C$  und  $D = K_D$  angibt und die  $\tau_k$  und  $\gamma_l$  sind die Abweichungen der Gruppe mit  $C = k$  bzw.  $D = l$  vom Mittelwert der Gruppe mit  $C = K_C$  und  $D = K_D$ .

## Kodierung Vergleich (mehrfaktoriell) Beispiel

Sei  $K_C = 2$  und  $K_D = 3$  mit  $n_{k,l} = 2$  für alle  $k = 1, 2, 3$  und  $l = 1, 2$ .

Dann erhalten wir als Designmatrix für das Modell mit

**Effekt-Kodierung:**

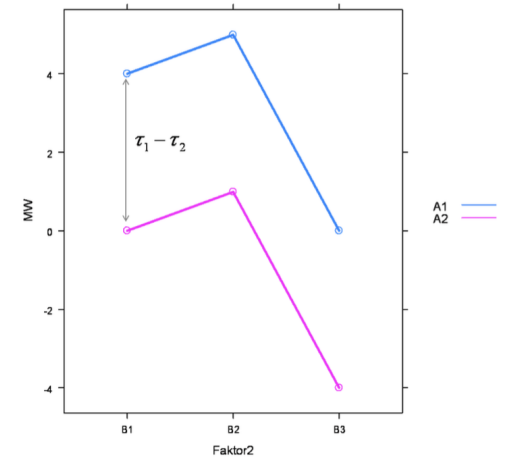
$$\mathbf{X} = (\mathbf{e} \quad \mathbf{Z}_1^e(\mathbf{C}) \quad \mathbf{Z}_1^e(\mathbf{D}) \quad \mathbf{Z}_2^e(\mathbf{D})) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

**Referenz-Kodierung:**

$$\mathbf{X} = (\mathbf{e} \quad \mathbf{Z}_1(\mathbf{C}) \quad \mathbf{Z}_1(\mathbf{D}) \quad \mathbf{Z}_2(\mathbf{D})) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

## Visualisierung Beispiel

Wir können die Effekte visualisieren, indem wir die Mittelwerte der Gruppen betrachten:



Note: „Faktor 2“ ist hier  $D$  und „A1“ und „A2“ sind hier  $C = 1$  und  $C = 2$ . Auf der y-Achse ist der Mittelwert der Gruppe dargestellt.

In beiden Fällen werden folgende Gleichungen erfüllt:

$$\begin{aligned} \hat{\mu} + \hat{\tau}_1 + \hat{\gamma}_1 &= 4 & \hat{\mu} + \hat{\tau}_2 + \hat{\gamma}_1 &= 0 \\ \hat{\mu} + \hat{\tau}_1 + \hat{\gamma}_2 &= 5 & \hat{\mu} + \hat{\tau}_2 + \hat{\gamma}_2 &= 1 \\ \hat{\mu} + \hat{\tau}_1 + \hat{\gamma}_3 &= 0 & \hat{\mu} + \hat{\tau}_2 + \hat{\gamma}_3 &= -4 \end{aligned}$$

**Effekt-Kodierung:**

$$\begin{aligned} \hat{\mu} &= 1 & \hat{\gamma}_1 &= 1 \\ \hat{\tau}_1 &= 2 & \hat{\gamma}_2 &= 2 \\ \hat{\tau}_2 &= -2 & \hat{\gamma}_3 &= -3 \end{aligned}$$

! Der Verlauf für  $C = 1$  und  $C = K_C = 2$  ist parallel mit Abstand  $\hat{\tau}_1 - \hat{\tau}_2$ .

**Referenz-Kodierung:**

$$\begin{aligned} \hat{\mu} &= -4 & \hat{\gamma}_1 &= 4 \\ \hat{\tau}_1 &= 4 & \hat{\gamma}_2 &= 5 \\ \hat{\tau}_2 &= 0 & \hat{\gamma}_3 &= 0 \end{aligned}$$

! Der Verlauf für  $C = 1$  und  $C = K_C = 2$  ist parallel mit Abstand  $\hat{\tau}_1$ . Man beachte auch, dass sich die Schätzer direkt in dem Plot ablesen lassen.



## Interaktion

In den obigen Modellen haben wir die Interaktion zwischen den Einflußgrößen  $C$  und  $D$  nicht berücksichtigt. Interaktion bedeutet, dass der Effekt von  $C$  von  $D$  abhängt und umgekehrt.

Beispiele:

- Die Wirkung des Medikaments ist bei Männern anders als bei Frauen.
- Die Wirkung des Medikaments ist bei jungen Menschen anders als bei alten Menschen.
- Die Wirkung des Lesetrainings ist bei guten Schülern geringer als bei schwachen Schülern.

! Der Begriff Interaktion ist in anderen Fachbereichen auch bekannt als Moderation (Psychologie), Synergieeffekte (Wirtschaft).

Wir können die Interaktion zwischen  $C$  und  $D$  berücksichtigen, indem wir die Effekte von  $C$  und  $D$  nicht additiv, sondern multiplikativ betrachten.

## Interaktionsmodell (Effekt-Kodierung)

In dem **Modell mit Effekt-Kodierung und Interaktion** ist die Designmatrix gegeben durch die Spalten der Designmatrix aus dem Modell ohne Interaktion  $\mathbf{X}^e$  und zusätzlich die Spalten der Interaktionsterme:

$$\mathbf{Z}^e = (z_1^e(C)z_1^e(D) \cdots z_1^e(C)z_{K_D-1}^e(D) \cdots z_{K_C-1}^e(C)z_{K_D-1}^e(D))$$

Die Designmatrix ist also gegeben durch  $(\mathbf{X}^e \quad \mathbf{Z}^e)$

Die Parameter sind gegeben durch  $\mu \quad \tau_k \quad \gamma_l \quad (\tau\gamma)_{k,l}$  mit  $k = 1, \dots, K_C - 1 \quad l = 1, \dots, K_D - 1$ .

Die Modellgleichung ist gegeben durch

$$Y_{k,l} = \mu + \tau_k + \gamma_l + (\tau\gamma)_{k,l} + \epsilon_{k,l}$$

mit den Nebenbedingungen

$$\sum_{k=1}^{K_C-1} \tau_k = 0 \quad \text{und} \quad \sum_{l=1}^{K_D-1} \gamma_l = 0$$

sowie

$$\forall l : \sum_{k=1}^{K_C} (\tau\gamma)_{k,l} = 0 \quad \text{und} \quad \forall k : \sum_{l=1}^{K_D} (\tau\gamma)_{k,l} = 0$$

## Interpretation Interaktion (Effekt-Kodierung)

- $\mu$  ist der Gesamtmittelwert.
- $\tau_k$  ist der Unterschied zwischen dem Mittelwert der Gruppe mit  $C = k$  und dem Gesamtmittelwert.
- $\gamma_l$  ist der Unterschied zwischen dem Mittelwert der Gruppe mit  $D = l$  und dem Gesamtmittelwert.
- $(\tau\gamma)_{k,l}$  ist der Effekt der Interaktion auf die Basiseffekte  $\tau_k$  und  $\gamma_l$  durch die Ausprägungen  $C = k$  und  $D = l$ .

## Interaktionsmodell (Referenz-Kodierung)

In dem **Modell mit Referenz-Kodierung und Interaktion** ist die Designmatrix gegeben durch die Spalten der Designmatrix aus dem Modell ohne Interaktion  $\mathbf{X}$  und zusätzlich die Spalten der Interaktionsterme:

$$\mathbf{Z} = (z_1(C)z_1(D) \cdots z_1(C)z_{K_D-1}(D) \cdots z_{K_C-1}(C)z_{K_D-1}(D))$$

Die Designmatrix ist also gegeben durch  $(\mathbf{X} \quad \mathbf{Z})$ .

Die Parameter sind gegeben durch  $\mu \quad \tau_k \quad \gamma_l \quad (\tau\gamma)_{k,l}$  mit  $k = 1, \dots, K_C - 1 \quad l = 1, \dots, K_D - 1$ .

Die Modellgleichung ist gegeben durch

$$Y_{k,l} = \mu + \tau_k + \gamma_l + (\tau\gamma)_{k,l} + \epsilon_{k,l}$$

mit den Nebenbedingungen

$$\tau_{K_C} = 0 \quad \text{und} \quad \gamma_{K_D} = 0$$

sowie

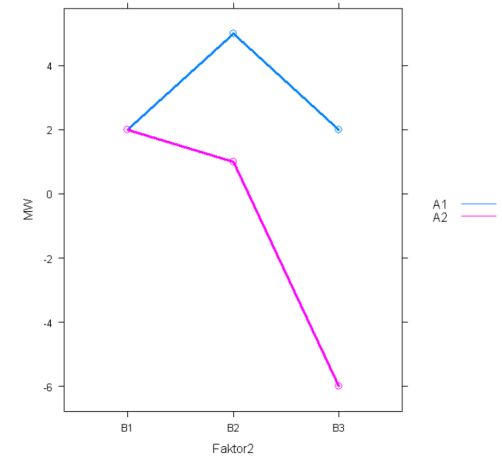
$$\forall l : (\tau\gamma)_{K_C,l} = 0 \quad \text{und} \quad \forall k : (\tau\gamma)_{k,K_D} = 0$$

## Interpret. Interaktion (Referenz-Kodierung)

- $\tau_k$  ist der Unterschied zwischen dem Mittelwert der Gruppe mit  $C = k$  zu der Referenzgruppe mit  $C = K_C$ .
- $\gamma_l$  ist der Unterschied zwischen dem Mittelwert der Gruppe mit  $D = l$  zu der Referenzgruppe mit  $D = K_D$ .
- $(\tau\gamma)_{k,l}$  ist der Effekt der Interaktion auf die Basiseffekte  $\tau_k$  und  $\gamma_l$  zur jeweiligen Referenzgruppe durch die Ausprägungen  $C = k$  und  $D = l$ .

## Visualisierung Beispiel

Wir können die Effekte visualisieren, indem wir die Mittelwerte der Gruppen betrachten:



In beiden Fällen werden folgende Gleichungen erfüllt:

$$\hat{\mu} + \hat{\tau}_1 + \hat{\gamma}_1 + (\widehat{\tau\gamma})_{1,1} = 2$$

$$\hat{\mu} + \hat{\tau}_1 + \hat{\gamma}_2 + (\widehat{\tau\gamma})_{1,2} = 5$$

$$\hat{\mu} + \hat{\tau}_1 + \hat{\gamma}_3 + (\widehat{\tau\gamma})_{1,3} = 2$$

$$\hat{\mu} + \hat{\tau}_2 + \hat{\gamma}_1 + (\widehat{\tau\gamma})_{2,1} = 2$$

$$\hat{\mu} + \hat{\tau}_2 + \hat{\gamma}_2 + (\widehat{\tau\gamma})_{2,2} = 1$$

$$\hat{\mu} + \hat{\tau}_2 + \hat{\gamma}_3 + (\widehat{\tau\gamma})_{2,3} = -6$$

### Effekt-Kodierung:

$$\hat{\mu} = 1 \quad \hat{\gamma}_1 = 1 \quad (\widehat{\tau\gamma})_{1,1} = -2 \quad (\widehat{\tau\gamma})_{2,1} = 2$$

$$\hat{\tau}_1 = 2 \quad \hat{\gamma}_2 = 2 \quad (\widehat{\tau\gamma})_{1,2} = 0 \quad (\widehat{\tau\gamma})_{2,2} = 0$$

$$\hat{\tau}_2 = -2 \quad \hat{\gamma}_3 = -3 \quad (\widehat{\tau\gamma})_{1,3} = 2 \quad (\widehat{\tau\gamma})_{2,3} = -2$$

### Referenz-Kodierung:

$$\hat{\mu} = -6 \quad \hat{\gamma}_1 = 8 \quad (\widehat{\tau\gamma})_{1,1} = -2 \quad (\widehat{\tau\gamma})_{2,1} = 0$$

$$\hat{\tau}_1 = 8 \quad \hat{\gamma}_2 = 7 \quad (\widehat{\tau\gamma})_{1,2} = 1 \quad (\widehat{\tau\gamma})_{2,2} = 0$$

$$\hat{\tau}_2 = 0 \quad \hat{\gamma}_3 = 0 \quad (\widehat{\tau\gamma})_{1,3} = 0 \quad (\widehat{\tau\gamma})_{2,3} = 0$$

### Kodierung Vergleich (mehrfaktoriell) Beispiel

Sei  $K_C = 2$  und  $K_D = 3$  mit  $n_{k,l} = 2$  für alle  $k = 1, 2, 3$  und  $l = 1, 2$ .

Dann erhalten wir als Designmatrix für das Modell mit Interaktionstermen

#### Effekt-Kodierung:

$$(X^e \quad Z^e) = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}$$

#### Referenz-Kodierung:

$$(X \quad Z) = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

### Kombination mit stetigen Merkmalen

Sei  $E$  eine metrische Variable und  $C$  eine diskrete Variable mit  $K$  Ausprägungen. Dann ergeben sich die Modelle von oben, wenn wir bei den Modellen ohne Interaktion in der Designmatrix eine Spalte mit den Werten von  $E$  hinzufügen und bei den Modellen mit Interaktion in der Designmatrix die Spalten hinzufügen, die sich ergeben, wenn man die Spalten der Designmatrix ohne  $E$  mit den Werten von  $E$  punktweise multipliziert.

### Interaktion: zwei kategorische Variablen

Gegeben sei das Regressionsmodell

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 I(z_1 = B) + \beta_4 I(z_1 = C) + \beta_5 I(z_2 = 1) + \beta_6 I(z_1 = B, z_2 = 1) + \beta_7 I(z_1 = C, z_2 = 1) + \varepsilon$$

Die geschätzten Koeffizienten sind

	Estimate
(Intercept)	2.30320
x1	2.01652
x2	3.00275
z1B	-2.70533
z1C	-1.27543
z21	2.27501
z1B:z21	-0.96322
z1C:z21	-0.79880

Die Interaktionskoeffizienten  $z_1 B:z_2 1$  und  $z_1 C:z_2 1$  haben nun folgenden Einfluß auf die Interpretation:

Wenn  $z_1 = B$  und  $z_2 = 1$ , dann steigt  $Y$  im Durchschnitt (im Vergleich zur Referenzgruppe  $z_1 = A$  und  $z_2 = 0$ ) (cet.par.) um  $-2.71 + 2.28 - 0.96 = -1.39$  Einheiten.

Wenn  $z_1 = C$  und  $z_2 = 1$ , dann steigt  $Y$  im Durchschnitt (im Vergleich zur Referenzgruppe  $z_1 = A$  und  $z_2 = 0$ ) (cet.par.) um (cet.par.) um  $-1.28 + 2.28 - 0.80 = 0.2$  Einheiten.

### Interaktion: kategorisch und numerisch

Gegeben sei das Regressionsmodell

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 I(z_1 = B) + \beta_4 I(z_1 = C) + \beta_5 I(z_2 = 1) + \beta_6 x_1 I(z_1 = B) + \beta_7 x_1 I(z_1 = C) + \varepsilon$$

Die geschätzten Koeffizienten sind

	Estimate
(Intercept)	0.38990
x1	4.00464
x2	2.99193
z1B	-0.19605
z1C	0.02191
z21	1.65244
x1:z1B	-2.98711
x1:z1C	-1.49294

Der Interaktionskoeffizienten  $x_1:z_1 B$  und  $x_1:z_1 C$  haben nun folgenden Einfluß auf die Interpretation:

Wenn  $x_1$  um eine Einheit steigt und  $z_1 = A$ , dann steigt  $Y$  im Durchschnitt (cet.par.) um 4.0 Einheiten. Wenn  $x_1$  um eine Einheit steigt und  $z_1 = B$ , dann steigt  $Y$  im Durchschnitt (cet.par.) um  $4.0 - 2.98 = 1.02$  Einheiten.

Wenn  $x_1$  um eine Einheit steigt und  $z_1 = C$ , dann steigt  $Y$  im Durchschnitt (cet.par.) um  $4.0 - 1.49 = 2.51$  Einheiten.

### Interaktion: zwei numerische Variablen

Gegeben sei das Regressionsmodell

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 I(z_1 = B) + \beta_4 I(z_1 = C) + \beta_5 I(z_2 = 1) + \beta_6 x_1 x_2 + \varepsilon$$

Die geschätzten Koeffizienten sind

	Estimate
(Intercept)	2.247787
x1	2.011725
x2	3.023465
z1B	-2.894061
z1C	-1.536237
z21	1.738088
x1:x2	0.499418

Der Interaktionskoeffizient  $x_1:x_2$  hat nun folgenden Einfluß auf die Interpretation:

Wenn  $x_1$  um eine Einheit steigt, dann steigt  $Y$  im Durchschnitt (cet.par.) um  $2.012 + x_2 \cdot 0.5$  Einheiten. Wenn  $x_2$  um eine Einheit steigt, dann steigt  $Y$  im Durchschnitt (cet.par.) um  $3.02 + x_1 \cdot 0.5$  Einheiten.

## ANOVA und ähnliche Begriffe

- **ANOVA:**

Analysis of Variance bzw. Varianzanalyse testet, ob sich der Mittelwert (einer metrischen Zielvariable) in verschiedenen Gruppen (diskrete Einflussvariable) unterscheidet.

- **ANCOVA:**

Analysis of Covariance bzw. Kovarianzanalyse ist identisch zur ANOVA, es wird jedoch noch für weitere Kovariablen kontrolliert. Sprich man testet also auf Mittelwertsgleichheit zwischen den Ausprägungen (Gruppen) der diskreten Kovariable und kontrolliert zusätzlich dabei auf weitere (metrische oder diskrete) Kovariablen, falls diese vermutete Störfaktoren sind oder eventuell einen Effekt haben, der nicht von Interesse ist.

- **t-Test:**

Der klassische t-Test für Mittelwertsvergleich (mit Varianzhomogenität) ist ein Spezialfall der ANOVA. Die diskrete Einflussgröße hat hier nur zwei Ausprägungen, wohingegen bei der ANOVA mehrere Ausprägungen der diskreten Kovariable möglich sind.

- **Einfaktorielle Varianzanalyse:**

Die einfaktorielle Varianzanalyse ist eine ANOVA mit nur einer interessierenden diskreten Einflussgröße (welche mehrere Ausprägungen haben kann).

- **Mehrfaktorielle Varianzanalyse:**

Bei der mehrfaktoriellen Varianzanalyse ist man an mehreren diskreten Einflussgrößen (welche je mehrere Ausprägungen haben können) interessiert. Üblicherweise wird hier auch eine Interaktion zwischen den Größen mit aufgenommen.

Alle Methoden testen, ob diskrete Einflussgrößen (Gruppen) einen signifikanten Effekt auf die metrische Zielgröße haben (es also Unterschiede zwischen den Gruppen gibt). Dabei hat der t-Test eine Einflussgröße mit zwei Ausprägungen, die einfaktorielle Varianzanalyse eine Einflussgröße mit mehreren Ausprägungen, die mehrfaktorielle Varianzanalyse mehrere Einflussgrößen mehreren Ausprägungen.

# Kapitel 5 - Metrische Einflußgrößen

## Interaktion metrischer Variablen

Seien  $X_1, X_2$  zwei metrische Variablen mit den Ausprägungen  $x_{1i}, x_{2i}$  für  $i = 1, \dots, n$ . Die Modellgleichung für das Modell mit Interaktion lautet:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i \\ &= \beta_0 + \beta_2 x_{2i} + (\beta_1 + \beta_3 x_{2i}) x_{1i} + \epsilon_i \\ &= \beta_0 + \beta_1 x_{1i} + (\beta_2 + \beta_3 x_{1i}) x_{2i} + \epsilon_i \end{aligned}$$

Interpretation der Modellparameter:

Die Parameter  $\beta_1, \beta_2$  geben die Steigung bei  $x_1 = x_2 = 0$  an. I.d.R. nicht sinnvoll interpretierbar.

## Allgemeiner Ansatz mit Basisfunktionen

Sei  $X$  eine metrische Variable mit den Ausprägungen  $x_i$  für  $i = 1, \dots, n$ .

Allgemeiner Ansatz für Modelle mit Basisfunktionen:  
Seien  $B_1, B_2, \dots, B_k$  Basisfunktionen. Dann ist die allgemeine Modellgleichung:

$$Y_i = \beta_0 + \beta_1 B_1(x_i) + \beta_2 B_2(x_i) + \dots + \beta_k B_k(x_i) + \epsilon_i$$

## Modelle mit metrischen Variablen

Sei  $X$  eine metrische Variable mit den Ausprägungen  $x_i$  für  $i = 1, \dots, n$ . Typische Modelle sind:

- **Einfaches lineares Modell:**

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- **Transformiertes lineares Modell:**

$$Y_i = \beta_0 + \beta_1 T(x_i) + \epsilon_i, \quad \text{z.B. } T(x) = \log(x)$$

- **Polynomielles Modell:**

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d + \epsilon_i$$

Problem: Bestimmung von  $d$ .

Mögliche Lösung: Test auf Signifikanz der Koeffizienten  $\beta_2, \dots, \beta_d$  mittels sequentieller Quadratsummenzerlegung.

- **Stückweise konstantes Modell:**

$$Y_i = \beta_0 I_{[x_i \leq g_1]} + \beta_1 I_{[g_1 < x_i \leq g_2]} + \dots + \beta_{k-1} I_{[g_{k-1} < x_i \leq g_k]} + \beta_p I_{[x_i > g_p]} + \epsilon_i$$

Dies entspricht der Kategorisierung der  $x$ -Variablen.

- **Stückweise lineares Modell:**

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \max\{x_i - g_1, 0\} + \dots + \beta_p \max\{x_i - g_h, 0\} + \epsilon_i$$

mit bekannten Bruchpunkten (Knoten)  $g_j$ .

- **Regressionssplines:**

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k + \beta_{k+1} \max\{x_i - g_1, 0\}^3 + \dots + \beta_{k+h} \max\{x_i - g_h, 0\}^3 + \epsilon_i$$

Ein Polynom 3. Grades ist 2-mal stetig differenzierbar.

# Kapitel 6 - Modelldiagnose

## Arten von Residuen

Gegeben sei das multiple lineare Regressionsmodell mit den üblichen Annahmen über  $\varepsilon$  und  $\mathbf{X}$ . Aus Kapitel 2 wissen wir, dass  $\mathbb{V}(\hat{\varepsilon}) = \sigma^2 \mathbf{Q}$  und somit  $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2 q_{ii}$ , wobei  $q_{ii}$  das  $i$ -te Diagonalelement der Matrix  $\mathbf{Q}$  ist.

Wir definieren **standardisierte Residuen** als

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2 q_{ii}}}$$

Das Problem an den standardisierten Residuen ist, dass bei der Schätzung von  $\sigma^2$  die Residuen mit einbezogen werden. Dies führt zu einer Verzerrung der Residuen. Wir definieren daher zusätzlich **studentisierte Residuen** als

$$\begin{aligned} r_i^* &= \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}_{(i)}^2 q_{ii}}} \\ &= r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}} \sim t_{n-p-1} \end{aligned}$$

wobei  $\hat{\sigma}_{(i)}^2$  die Schätzung von  $\sigma^2$  ist, die ohne die  $i$ -te Beobachtung berechnet wurde. Man kann zeigen, dass die studentisierten Residuen  $t$ -verteilt sind.

Wir können

## Durbin-Watson-Test

Wir definieren die **Durbin-Watson-Teststatistik** als

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

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

Die Verteilung von  $d$  unter  $H_0 : \rho = 0$  ist schwierig allgemein herzuleiten. Es gilt heuristisch, dass wir  $H_0$  verwerfen, wenn  $d \approx 2$

Für genaue Tests können wir in R aus dem package `lmtest` die Funktion `dwtest` nutzen.

## Mögliche Probleme

Gegeben sei das multiple lineare Regressionsmodell mit den üblichen Annahmen über  $\varepsilon$  und  $\mathbf{X}$ . Folgende Probleme können typischerweise auftreten:

- **$\varepsilon_i$  nicht normalverteilt:**  
Die Fehlerterme sind nicht normalverteilt.
- **Heteroskedastizität:**  
Die Varianz der Fehlerterme ist nicht konstant bzw. von  $i$  abhängig.
- **Autokorrelation:**  
Die Fehlerterme sind korreliert.
- **Multikollinearität:**  
Die Kovariablen sind (annähernd) linear abhängig.
- **Ausreißer und Leverage Points:**  
Einzelne Beobachtungen haben einen starken Einfluss auf die Schätzungen.
- **Overfitting oder Underfitting:**  
Das Modell ist zu komplex oder zu einfach bzw. die Modellgleichung ist fehlerhaft.

## $\varepsilon_i$ nicht normalverteilt

- **Ursachen:** Die abhängige Variable  $Y$  kann bedingt auf  $\mathbf{x}$  nicht normalverteilt sein. Das ist z.B. der Fall, wenn  $Y$  eine Zählgröße, eine Überlebenszeit, ein Anteil, nicht-negativ oder eine binäre Variable ist.
- **Folgen:** KQ-Schätzer bleibt unbiased und F-Statistik ist i.A. robust. Aber Konfidenz-/Prognoseintervalle sind nicht mehr korrekt.
- **Diagnose:** Schiefe, Kurtosis, Normal-Plot der Residuen.
- **Therapie:** Transformation der abhängigen Variable  $Y$ . GLMs.

## Heteroskedastizität

- **Ursachen:** Die abhängige Variable  $Y$  stellt z.B. eine Zählgröße oder Anteil dar. Gruppierte Daten führen zu verschiedenen Residualvarianzen innerhalb der Gruppen. Multiplikative Fehlerstruktur, d.h.  $\sigma_i^2$  ist abhängig von der Größe von  $Y_i$ .
- **Folgen:** KQ-Schätzer bleibt unbiased, aber ist nicht mehr most efficient. Tests und Konfidenz-/Prognoseintervalle sind nicht mehr korrekt.
- **Diagnose:** Residuals vs. Fitted Plot. Berechnung der Residualvarianzen in den einzelnen Gruppen (bei gruppierten Daten).
- **Therapie:** Transformation der abhängigen Variable  $Y$ . Gewichtete KQ-Schätzung.

## Autokorrelation

- **Ursachen:** Zeitreihenstruktur oder räumliche Struktur der Daten führen zu positiver Korrelation von aufeinander folgenden (bzw. nahen) Beobachtungen. Residuen bei gruppierten Beobachtungen, bei denen die Gruppenzugehörigkeit nicht zusätzlich modelliert wird, sind häufig positiv korreliert.
- **Folgen:** KQ-Schätzer bleibt unbiased, aber ist nicht mehr most efficient. Tests und Konfidenz-/Prognoseintervalle sind nicht mehr korrekt.
- **Diagnose:** Analyse der Zeitreihenstruktur der Residuen, z.B. mit Durbin-Watson-Test; Plots der Residuen gegen die Zeit; Plots von  $\hat{\varepsilon}_i$  gegen  $\hat{\varepsilon}_{i-1}$ . ACP und PACP.
- **Therapie:** Verwendung von Zeitreihenmethoden; Einbeziehung von Trend und Saison; Gewichtete KQ-Methode.

## Multikollinearität

- **Ursachen:** Hohe Korrelation zwischen den Einflussgrößen; Ungünstiges Versuchs-Design; Codierung von diskreten Variablen.
- **Folgen:** Ungenauer KQ-Schätzer, häufig sogar mit falschem Vorzeichen. Aber Konfidenzintervalle sind korrekt (jedoch entsprechend sehr breit).
- **Diagnose:** Analyse der Matrix  $\mathbf{X}^\top \mathbf{X}$  und der Korrelationsmatrix der metrischen Einflussgrößen.

– Konditionszahl von  $\mathbf{X}$ :

$$\kappa(\mathbf{X}) = \sqrt{\frac{\lambda_{\max}(\mathbf{X}^\top \mathbf{X})}{\lambda_{\min}(\mathbf{X}^\top \mathbf{X})}}$$

$\kappa(\mathbf{X}) \gg 1$  deutet auf Multikollinearität hin.

– Varianz Inflationsfaktor:

$$V(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2},$$

wobei  $R_j^2$  das Bestimmtheitsmaß der Regression  $\mathbf{X}_j = \mathbf{X}_{-j}\alpha + \delta$  ist. Wir definieren den **Varianz Inflationsfaktor** als

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

Wenn  $VIF_j = 1$ , dann heißt das, dass  $\mathbf{X}_j$  orthogonal zu allen anderen Regressoren ist. Ein hohes  $VIF_j$  deutet auf Multikollinearität hin. Als Heuristik wird oft  $VIF_j > 10$  als kritisch angesehen.

- **Therapie:** Zusammenfassen bzw. Weglassen von Einflussgrößen; Verwendung von anderen Schätzmethoden, z.B.: Ridge-Regression.

## Ausreißer und Leverage Points

Wir unterscheiden zwischen Ausreißern und High Leverage Points (einflußreiche Beobachtungen). Ein **Ausreißer** ist eine Beobachtung, die in der abhängigen Variable  $Y$  stark von den anderen Beobachtungen abweicht (i.d.R. hoher Störterm). Ein **Leverage Point** ist eine Beobachtung, die in den unabhängigen Variablen  $\mathbf{X}$  stark von den anderen Beobachtungen abweicht.

- **Ursachen:** Falsche Erhebung; Beobachtung gehört nicht zur Grundgesamtheit; Besonderheiten bei einzelner Untersuchungseinheit.
- **Folgen:** High Leverage Points haben einen großen Einfluss auf  $\hat{\beta}$ . Ausreißer können zu erheblicher Verzerrung von  $\hat{\beta}$  führen.
- **Diagnose:** Analyse der Diagonalelemente der Hat-Matrix  $\mathbf{P}$  zum Auffinden von high leverage points; Verschiedene Residuenplots zur Ausreißeranalyse; Influence-Statistiken.
- **Therapie:** Fehlerhafte Daten weglassen (Sensitivitätsanalyse); Robuste Regression; Gewichete Regression.

Wir definieren **Leverage** als

$$\begin{aligned} h_{ii} &= \mathbf{P}_{ii} = \frac{V(\hat{Y}_i)}{\sigma^2} \\ &= \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i = \|\mathbf{x}_i\|_{(\mathbf{X}^\top \mathbf{X})^{-1}}^2 \end{aligned}$$

Optimalerweise gilt  $h_{ii} = \frac{p'}{n}$  und als Heuristik wird oft  $h_{ii} > \frac{2p'}{n}$  als kritisch angesehen.

! Der Leverage kann nach Transformation auch als quadratischer Mahalanobis-Abstand zum Mittelpunkt interpretiert werden.

Wir definieren **Cook's Distanz** als

$$\begin{aligned} D_i &:= \frac{(\hat{\beta}_{-i} - \hat{\beta})^\top (\mathbf{X}^\top \mathbf{X}) (\hat{\beta}_{-i} - \hat{\beta})}{\hat{\sigma}^2 p'} \\ &= \frac{(\hat{Y}_{-i} - \hat{Y})^\top (\hat{Y}_{-i} - \hat{Y})}{\hat{\sigma}^2 p'} \\ &= \frac{r_i^2}{p'} \cdot \frac{h_{ii}}{1 - h_{ii}} \end{aligned}$$

Als Heuristik wird oft verwendet, dass Beobachtungen mit  $D_i > 0.5$  auffällig sind und Beobachtungen mit  $D_i > 1$  auf jeden Fall untersucht werden sollten.

## Overfitting oder Underfitting

- **Ursachen:** Variablen wurden weggelassen oder überflüssigerweise in das Modell einbezogen; Der Zusammenhang ist nicht linear; Interaktionen werden nicht in das Modell einbezogen.
- **Folgen:** Systematische Fehler bei der Schätzung der Modellparameter und bei der Prognose; Dennoch: Modellschätzung liefert häufig brauchbare Näherung.
- **Diagnose:** Residuenplots  $\hat{\epsilon}$  gegen  $\hat{\mathbf{Y}}$ ; F-Tests auf Einfluss von weiteren Variablen; Interaktionen; Polynomterme höherer Ordnung, etc.
- **Therapie:** Modellerweiterung; Transformationen der Einflussgrößen; Variablenselektionsverfahren (z.B. LASSO).



# Kapitel 7 - Fortgeschrittenere lineare Modelle

## Das allgemeine lineare Regressionsmodell

Das **allgemeine lineare Regressionsmodell** hat die Form

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

wobei wir nun annehmen, dass  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{W}^{-1})$ , mit **bekannter** positiv definiter Matrix  $\mathbf{W}$ . Wir nennen  $\mathbf{W}$  die **Gewichtsmatrix**.

Im Falle von heteroskedastischen, aber unkorrelierten Fehlern ist  $\mathbf{W}$  eine Diagonalmatrix mit  $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$  und  $\mathbb{V}(\boldsymbol{\varepsilon}) = \text{diag}(\frac{\sigma^2}{w_1}, \dots, \frac{\sigma^2}{w_n})$ .

## Schätzer im allg. linearen Regressionsmodell

Gegeben sei das **allgemeine lineare Regressionsmodell**  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  mit  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{W}^{-1})$ .

Dann gilt:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y},$$

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$

$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$$

und  $\hat{\boldsymbol{\beta}}$  ist der **Beste lineare unverzernte Schätzer** (BLUE) für  $\boldsymbol{\beta}$ .

Der **Schätzer für  $\sigma^2$**  ist gegeben durch den REML-Schätzer

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{n - p'},$$

! Alle Schätzer setzen voraus, dass  $\mathbf{W}$  bekannt ist.

## Umformung

Wir können die symmetrische positiv definite Gewichtsmatrix  $\mathbf{W}$  zerlegen durch

$$\mathbf{W} = \mathbf{V} \mathbf{D} \mathbf{V}^\top,$$

wobei  $\mathbf{D}$  eine Diagonalmatrix ist. Dann definieren wir  $\mathbf{W}^{\frac{1}{2}} = \mathbf{V} \mathbf{D}^{\frac{1}{2}} \mathbf{V}^\top$ . Mit dieser Zerlegung können wir das allgemeine lineare Regressionsmodell in ein klassisches lineares Regressionsmodell umformen durch

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*, \quad \boldsymbol{\varepsilon}^* \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

wobei

$$\mathbf{Y}^* = \mathbf{W}^{\frac{1}{2}} \mathbf{Y}, \quad \mathbf{X}^* = \mathbf{W}^{\frac{1}{2}} \mathbf{X}, \quad \boldsymbol{\varepsilon}^* = \mathbf{W}^{\frac{1}{2}} \boldsymbol{\varepsilon}.$$

## AR(1) Zeitreihenmodell

Das **AR(1) Zeitreihenmodell** hat die Form

$$\varepsilon_t = \phi \varepsilon_{t-1} + \eta_t, \quad t = 2, \dots, n$$

wobei  $\eta_t \sim \mathcal{N}(0, \sigma^2)$ ,  $|\phi| < 1$  und  $\mathbb{V}(\varepsilon_1) = \frac{\sigma^2}{1 - \phi^2}$ .

Dann ist die Kovarianzmatrix  $\mathbf{W}^{-1}$  gegeben durch

$$\mathbf{W}^{-1} = \frac{1}{1 - \phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{n-1} \\ \phi & 1 & \phi & \dots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \dots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \dots & 1 \end{pmatrix}$$

Damit gilt also

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{W}^{-1}).$$

## Linear Mixed Modell

Wir betrachten die Beobachtungen  $(Y_{ij}, \mathbf{x}_{ij}^\top)$  mit  $i = 1, \dots, K$  und  $j = 1, \dots, n_i$ .

In der Regel benutzen wir dieses Setting in zwei Fällen:

1. Clustered Data:  $i$  beschreibt eine Gruppe und  $j$  ein Objekt innerhalb dieser Gruppe (z.B. Schüler  $j$  in einer Klasse  $i$ ).
2. Longitudinal Data:  $i$  beschreibt eine Untersuchungseinheit und  $j$  ein Beobachtung dieser Untersuchungseinheit zum Zeitpunkt  $t_{ij}$  (z.B. Messung von Patient  $i$  zum Zeitpunkt  $t_{ij}$ ).

Die Idee eines **Linear Mixed Modells** ist, dass wir ein Cluster-übergreifendes bzw. Untersuchungseinheiten-übergreifendes Modell haben, aber innerhalb der Cluster bzw. Untersuchungseinheiten können die Beobachtungen korreliert sein und zufällig vom übergreifenden Modell abweichen.

## Random Intercept Modell

Das **Random Intercept Modell** ist definiert durch

$$Y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \gamma_{0i} + \varepsilon_{ij}, \quad i = 1, \dots, K, \quad j = 1, \dots, n_i$$

mit  $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ , *i.i.d.* und  $\gamma_{0i} \sim \mathcal{N}(0, \sigma_{\gamma_0}^2)$ , *i.i.d.* und  $\sum_{i=1}^K n_i = n$  und der Annahme, dass die  $\varepsilon_{ij}$  unabhängig von den  $\gamma_{0i}$  sind.

Wir bezeichnen  $\beta_0$  als den fixed Population Intercept und  $\gamma_{0i}$  als die Cluster-spezifische bzw. Untersuchungseinheiten-spezifische zufällige Abweichung vom fixed Population Intercept. Zusammen bezeichnen wir  $\beta_0 + \gamma_{0i}$  als den **random Intercept** von Cluster/Untersuchungseinheit  $i$ . Die restlichen  $\beta_j$  sind die **fixed Effekte**.

Zwischen den Cluster/Untersuchungseinheiten sind die  $Y_{ij}$  unabhängig und es gilt das konditionale Modell für  $i = 1, \dots, K$

$$Y_{ij} | \gamma_{0i} \sim \mathcal{N}(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \gamma_{0i}, \sigma^2).$$

Innerhalb eines Clusters/Untersuchungseinheit sind die  $Y_{ij}$  korreliert. Wir bezeichnen diese Korrelation als **Intra-Class Correlation** (ICC). Es gilt

$$\text{Corr}(Y_{ij}, Y_{il}) = \frac{\sigma_{\gamma_0}^2}{\sigma_{\gamma_0}^2 + \sigma^2}, \quad j, l = 1, \dots, n_i, j \neq l$$

und

$$\mathbb{V}(\mathbf{Y}_i) = \sigma^2 \mathbf{I}_{n_i} + \sigma_{\gamma_0}^2 \mathbf{J}_{n_i}, \quad i = 1, \dots, K$$

wobei  $\mathbf{J}_{n_i}$  die  $n_i \times n_i$ -Matrix mit Einsen ist.

Daraus ergibt sich das **marginale Modell**

$$\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_i \boldsymbol{\beta}, \sigma^2 \mathbf{I}_{n_i} + \sigma_{\gamma_0}^2 \mathbf{J}_{n_i}).$$

! Bei dem Random Intercept Modell wird die Annahme gemacht, dass die Effekte von  $x$  auf  $Y$  innerhalb der Cluster/UEs und über diese hinweg gleich sind. Das kann zu falschen Schlüssen führen, wenn diese Annahme nicht zutrifft (z.B. Simpson's Paradoxon). Eine Möglichkeit das zu korrigieren, ist das erweiterte Random Intercept Modell.

$$Y_{ij} = \beta_0 + \beta_1(x_{ij} - \bar{x}_i) + \beta_2 \bar{x}_i + \gamma_{0i} + \varepsilon_{ij}$$

## Logistisches Regressionsmodell

Das **logistische Regressionsmodell** ist ein Modell für binäre abhängige Variablen, d.h.  $Y_i \sim \text{Ber}(\pi_i)$ . Das Ziel ist es, die Wahrscheinlichkeit

$$\pi_i = \mathbb{P}(Y_i = 1 \mid \mathbf{x}_i) = \mathbb{E}(Y_i \mid \mathbf{x}_i)$$

in Abhängigkeit von den Kovariablen zu modellieren. Wir definieren den **linearen Prädiktor** als

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Wir definieren die **Linkfunktion** als

$$g(\pi_i) = \eta_i$$

und ihre Umkehrfunktion nennen wir die **response function** oder **inverse Linkfunktion** und schreiben

$$\pi_i = g^{-1}(\eta_i) = h(\eta_i).$$

Im **logistischen Regressionsmodell (logit Modell)** sind die Link- und Responsefunktion definiert durch:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right), \quad h(\eta_i) = \frac{1}{1 + \exp(-\eta_i)}.$$

Das heißt, wir modellieren die log-odds der Wahrscheinlichkeit mittels einer multiplen linearen Regression.

## Interpretation Logit Modell

Gegeben sei das **logistische Regressionsmodell**

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Dann gilt:

- $\pi_i = \mathbb{P}(Y_i = 1 \mid \mathbf{x}_i) = h(\mathbf{x}_i^\top \boldsymbol{\beta})$
- Wenn sich  $x_k$  um eine Einheit erhöht, so ändert sich die log-odds der Wahrscheinlichkeit (ceteris paribus) im Erwartungswert um  $\beta_k$  Einheiten.
- Wenn sich  $x_k$  um eine Einheit erhöht, so ändern sich die Odds  $\frac{\pi_i}{1 - \pi_i}$  (ceteris paribus) im Erwartungswert um den Faktor  $\exp(\beta_k)$ .
- Wenn  $\beta_k > 0$ , so steigt die Wahrscheinlichkeit  $Y_i = 1$  mit steigendem  $x_k$  (und umgekehrt).

## Logit Modell in R

In R können wir ein logistisches Regressionsmodell mit der Funktion `glm()` schätzen. Die Syntax ist

```
glm(Y ~ X1 + X2 + ..., data = data,
     family = binomial(link = "logit"))
```

wobei  $Y$  die abhängige Variable und  $X1, X2, \dots$  die unabhängigen Variablen sind.

## MLE für Logit Modell

Die log-likelihood Funktion lautet

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i).$$

Mit der Linkfunktion und Responsefunktion können wir die Likelihood umschreiben als

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})).$$

Daraus ergibt sich die Scorefunktion

$$s(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \mathbf{x}_i (y_i - h(\mathbf{x}_i^\top \boldsymbol{\beta})).$$

Die (observed) Fisher Matrix ist

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbb{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top h(\mathbf{x}_i^\top \boldsymbol{\beta})(1 - h(\mathbf{x}_i^\top \boldsymbol{\beta})).$$

Daraus ergibt sich der **MLE Schätzer** für  $\boldsymbol{\beta}$  durch

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

und es gilt  $\mathbb{V}(\hat{\boldsymbol{\beta}}) \approx \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$ .