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 (β_0, β_1) durch

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

und nennen $(\hat{\beta}_0, \hat{\beta}_1)$ den KQ-Schätzer von (β_0, β_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 $\sum_{i=1}^{n} (x_i - \overline{x})^2 \neq 0$. Dieser lässt sich berechnen als

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

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

- wenn x um eine **Einheit** steigt, dann steigt Y **im Erwartungswert** um β_1 Einheiten.
- Es gilt $\beta_0 = \mathbb{E}(Y | X = 0)$.
- Der Parameter σ 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: $\mathbb{E}(\hat{\beta}_0, \hat{\beta}_1) = (\beta_0, \beta_1)$.
- $\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{nS_x^2}$ und $\mathbb{V}(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2})$.
- $(\hat{\beta}_0, \hat{\beta}_1)$ ist der maximum-likelihood Schätzer.

Schätzer für σ^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 σ^2 und

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

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

Konfidenzintervalle für β_0 und β_1

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

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

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

Damit können wir Konfidenzintervalle zum Niveau $1 - \alpha$ für β_1 und β_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 linearen Modell mit $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ und $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$. Dann gilt

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

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 linearen Modell mit $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ und $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n$. 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 linearen Modell, sowie eine weitere Beobachtung x_{n+1} mit zugehörigem Y_{n+1} sowie der Prognosewert $\hat{Y}_{n+1}$. Dann gilt

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

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

Prognoseintervall

Gegeben sei ein einfaches linearen 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

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

R-Code

```r
# 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 $100 \cdot (e^{\beta_1 log(1.01)} - 1)$%.

- 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 β_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^{β_1}.

Anmerkungen aus der 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 sichererer 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.

Annahmen des linearen Regressionsmodells

Gegeben sei das einfache (oder multiple) lineare Regressionsmodell mit

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

$$\mathbb{E}(\varepsilon) = \boldsymbol{0} \quad (2)$$

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

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

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

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	✓	✓	✓	✓	✓
$\mathbb{E}(\hat{\beta}) = \beta$	✓	✓			
$\hat{\beta} \sim \mathcal{N}(\beta, \mathbb{V}(\hat{\beta}))$	✓	✓	✓	✓	(✓)
Konfidenzintervalle	✓	✓	✓	✓	(✓)
Prädiktionsintervalle	✓	✓	✓	✓	✓

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

Kapitel 2 - Das multiple lineare Regressionsmodell

Multiples lineares Regressionsmodell

Das multiple lineare Regressionsmodell hat die Form

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

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

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

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

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

Kleinste Quadrate (KQ) Schätzer

Wir schätzen den Parameter(vektor) β durch

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p'}}{\arg\min} (\boldsymbol{Y} - \boldsymbol{X}\beta)^\top (\boldsymbol{Y} - \boldsymbol{X}\beta) \qquad (6)$$

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

Existenz und Berechnung vom KQ Schätzer

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

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

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

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

Gauss-Markov-Theorem

Sei das Modell $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ gegeben mit $\mathbb{E}(\varepsilon) = \boldsymbol{0}$ und $\mathrm{Cov}(\varepsilon) = \sigma^2 \boldsymbol{I}$. Dann ist der KQ-Schätzer $\hat{\beta}$ der beste lineare erwartungstreue Schätzer (best linear unbiased estimator, BLUE) von β.
Das heißt, dass für jeden anderen linearen erwartungstreuen Schätzer $\tilde{\beta}$ von β gilt $\mathbb{V}(\hat{\beta}) \leq \mathbb{V}(\tilde{\beta})$.

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 β_k Einheiten.
- (Empirische) Interpretation: Steigt x_k um eine Einheit, so steigt Y (ceteris paribus) im Durchschnitt um $\hat{\beta}_k$ Einheiten.
- ! β_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 $\hat{\beta}$

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

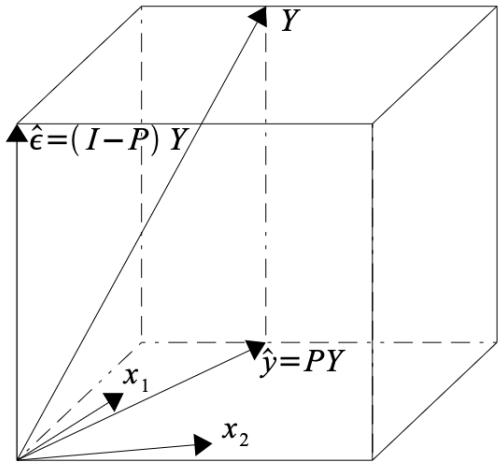
Hat-Matrix und Residualmatrix

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

$$\begin{aligned} \hat{\boldsymbol{Y}} &:= \boldsymbol{X} \underbrace{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}}_{\hat{\beta}} \\ \boldsymbol{P} &:= \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{n \times n} \\ \hat{\varepsilon} &= \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \\ \boldsymbol{Q} &:= \boldsymbol{I} - \boldsymbol{P} \end{aligned}$$

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

Geometrische Interpretation



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

Eigenschaften von P und Q

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

$$\begin{aligned} \boldsymbol{P}^\top &= \boldsymbol{P} \text{ und } \boldsymbol{P}^2 = \boldsymbol{P} \\ \boldsymbol{Q}^\top &= \boldsymbol{Q} \text{ und } \boldsymbol{Q}^2 = \boldsymbol{Q} \\ \boldsymbol{P}\boldsymbol{Q} &= \boldsymbol{Q}\boldsymbol{P} = \boldsymbol{0}. \end{aligned}$$

Daraus folgt

$$\begin{aligned} \mathbb{V}(\hat{\boldsymbol{Y}}) &= \sigma^2 \boldsymbol{P} \\ \mathbb{V}(\hat{\varepsilon}) &= \sigma^2 \boldsymbol{Q}, \text{ da } \hat{\varepsilon} = \boldsymbol{Q}\varepsilon \end{aligned}$$

Schätzer für σ^2

Gegeben dem multiplen linearen Modell, gilt

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

ist ein erwartungstreuer Schätzer von σ^2.
! Gilt auch ohne die Annahme $\varepsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, solange $\mathbb{E}(\varepsilon) = \boldsymbol{0}$ und $\mathrm{Cov}(\varepsilon) = \sigma^2 \boldsymbol{I}$

Kapitel 3 - Quadratsummenzerlegung und statistische Inferenz im multiplen linearen Regressionsmodell

Quadratsummenzerlegung

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

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

SST(otal):	Gesamt-Quadratsumme (korrigiert)
SSE(rror):	Fehler-Quadratsumme
SSM(odel):	Modell-Quadratsumme

Quadratsummenzerlegung ohne β_0

Gegeben sei das multiple lineare Regressionsmodell mit, aber ohne Absolutglied β_0. Dann gilt

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

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

$$\boldsymbol{e} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \text{ und } \boldsymbol{P_e} = \boldsymbol{e}(\boldsymbol{e}^\top\boldsymbol{e})^{-1}\boldsymbol{e}^\top \text{ und } \boldsymbol{Q_e} = \boldsymbol{I} - \boldsymbol{P_e}.$$

Dann gilt

$$\boldsymbol{P_e}\boldsymbol{Y} = \overline{\boldsymbol{Y}} \quad \textit{und} \quad \boldsymbol{Q_e}\boldsymbol{Y} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$$

$$\begin{aligned}
\mathbb{E}(SST^*) &= \sigma^2 n + \beta^\top \boldsymbol{X}^\top\boldsymbol{X}\beta \\
\mathbb{E}(SST) &= \sigma^2(n-1) + \beta^\top(\boldsymbol{Q_e}\boldsymbol{X})^\top(\boldsymbol{Q_e}\boldsymbol{X})\beta \\
\mathbb{E}(SSE) &= \sigma^2(n-p') \\
\mathbb{E}(SSM^*) &= \sigma^2 p' + \beta^\top \boldsymbol{X}^\top\boldsymbol{X}\beta \\
\mathbb{E}(SSM) &= \sigma^2(p'-1) + \beta^\top(\boldsymbol{Q_e}\boldsymbol{X})^\top(\boldsymbol{Q_e}\boldsymbol{X})\beta
\end{aligned}$$

Wir können diese Eigenschaften zur Konstruktion von Tests verwenden. Es gilt nämlich unter anderem
$\beta = 0 \implies \mathbb{E}(SST^*) = \sigma^2 n$
$\beta_1 = \cdots = \beta_p = 0 \implies \mathbb{E}(SSM) = \sigma^2(p'-1)$

Mittlere Quadratsummen

Wir definieren entsprechend der Zahl der Freiheitsgrade die mittleren Quadratsummen als

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

R^2

Gegeben sei das multiple lineare Regressionsmodell mit allen Annahmen. Dann definieren wir das Bestimmtheitsmaß R^2 als

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST} = r^2_{\boldsymbol{Y}\hat{\boldsymbol{Y}}} \in [0,1]$$

wobei $r_{\boldsymbol{Y}\hat{\boldsymbol{Y}}}$ der Korrelationskoeffizient zwischen $\boldsymbol{Y}$ und $\hat{\boldsymbol{Y}}$ ist.
Wir interpretieren R^2 als den Anteil der Varianz von $\boldsymbol{Y}$, die durch das Modell erklärt wird. Ein hohes R^2 deutet darauf hin, dass wir unser Modell gut nutzen können, um $\boldsymbol{Y}$ zu erklären. R^2 steigt mit steigender Anzahl an Kovariablen, auch wenn diese kaum/keine Erklärungskraft haben. Es ist daher nicht sinnvoll, R^2 zwischen Modellen zu vergleichen, die unterschiedlich viele Kovariablen haben. Hierfür nutzen wir das adjustierten Bestimmtheitsmaß

$$\begin{aligned}
R^2_{\text{adj}} &= \frac{MSM}{MST} = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/(n-1)}{SST/(n-p')} \\
&= 1 - \frac{n-1}{n-p'}(1-R^2)
\end{aligned}$$

Für $n \gg p'$ gilt $R^2_{\text{adj}} \approx R^2$.
! Bei einem Modell ohne Absolutglied ist R^2 nach obiger Definition nicht sinnvoll, da es dann auch negative Werte annehmen kann. Das kommt daher, dass die Zerlegung $SST = SSE + SSM$ nicht mehr gilt. Stattdessen ist es sinnvoll, R^2 als $\frac{SSM^*}{SST^*} = 1 - \frac{SSE}{SST^*}$ zu definieren.

Multivariate Normalverteilung

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

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

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

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

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

1. $\mathbb{E}(\boldsymbol{Z}) = \mu$
2. $\mathbb{V}(\boldsymbol{Z}) = \Sigma$
3. $\boldsymbol{A}\boldsymbol{Z} \sim \mathcal{N}_m(\boldsymbol{A}\mu, \boldsymbol{A}\Sigma\boldsymbol{A}^\top)$
4. Es existiert eine orthogonale Matrix $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ mit $\boldsymbol{T}\Sigma\boldsymbol{T}^\top = \text{diag}(\lambda_1, \ldots, \lambda_n)$, so dass

$$\boldsymbol{T}\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{T}\mu, \text{diag}(\lambda_1, \ldots, \lambda_n))$$

Chi-Quadrat Verteilung

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

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

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

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

t-Verteilung

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

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

dann gilt

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

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

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

Verteilung des KQ-Schätzers

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

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$$
$$\hat{\Sigma}_{\hat{\beta}} := \hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$
$$(n - p')\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - p')$$
$\hat{\sigma}^2$ und $\hat{\beta}$ sind unabhängig.
$$\hat{\sigma}^2_{\hat{\beta}_k} := (\hat{\Sigma}_{\hat{\beta}})_{kk}$$
$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_k}}} \sim t(n - p', 0)$$

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

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

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

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

Daraus folgt die für die praxis essenzielle Eigenschaft

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

Overall-Test

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

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

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

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

bzw. $R^2 = 0$ 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 $\boldsymbol{H}_0 : \boldsymbol{A}\beta = \boldsymbol{c}$ getestet werden, wobei $\boldsymbol{A} \in \mathbb{R}^{a\times p'}$ mit $\text{rang}(\boldsymbol{A}) = a$ und $\boldsymbol{c} \in \mathbb{R}^a$.
Wir definieren

$$SSH := (\boldsymbol{A}\hat{\beta} - \boldsymbol{c})^\top(\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{c})$$
$$MSH := \frac{SSH}{a}$$
$$\delta_{SSH} := (\boldsymbol{A}\beta - \boldsymbol{c})^\top(\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\beta - \boldsymbol{c}).$$

SSH beschreibt die Quadratsumme der Abweichung von der Hypothese $\boldsymbol{A}\beta = \boldsymbol{c}$.
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 $H_0 : \boldsymbol{A}\beta = \boldsymbol{c}$ 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 entsprich dieser einem Likelihood-Quotienten-Test, ist also optimal.

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

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

Partielle Quadratsummen

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

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

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

Sequentielle Quadratsummen

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

$$M_k : \boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{x} + \cdots + \beta_k \boldsymbol{x}_k + \epsilon$$

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

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

für $k = 1, \ldots, p'$.
Es gilt

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

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 $\text{rang}(\boldsymbol{X}) = p'$. Das Konfidenzellipsoid für β zum Niveau $1 - \alpha$ ist gegeben als

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

Prognosewert

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

Prognosefehler und Prognoseintervall

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

$$\mathbb{E}(\hat{Y}_{n+1} - Y_{n+1}) = 0$$
$$\mathbb{V}(\hat{Y}_{n+1} - Y_{n+1}) = \sigma^2 \left[1 + \boldsymbol{x}_{n+1}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{n+1} \right]$$

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

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

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

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

mit

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

Kapitel 4 - Diskrete Einflußgrößen

Kodierung

Sei C eine nominale Variable mit K Ausprägungen.

Dummy/Referenz-Kodierung:
Wir definieren K neue Variablen $Z_1, \ldots, Z_K$ als

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

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

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

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

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

Setup einfache Varianzanalyse

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

Mittelwertsmodell

Das Mittelwertsmodell ist gegeben durch

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

oder in Matrix-Vektor Notation:

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

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

Mittelwertsmodell Beispiel

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

$$\boldsymbol{Y} = \begin{pmatrix} Y_{1_1} \\ Y_{1_2} \\ Y_{2_1} \\ Y_{2_2} \\ Y_{3_1} \\ Y_{3_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1_1} \\ \varepsilon_{1_2} \\ \varepsilon_{2_1} \\ \varepsilon_{2_2} \\ \varepsilon_{3_1} \\ \varepsilon_{3_2} \end{pmatrix}$$

Modell mit Effekt-Kodierung

Das Modell mit Effekt-Kodierung ist gegeben durch

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

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

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

Bei dem Modell mit Effekt-Kodierung gibt es einen Intercept μ und die τ_k sind die Abweichungen der k-ten Gruppe vom Gesamtmittelwert bzw. vom Intercept μ. 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:

$$\boldsymbol{Y} = \begin{pmatrix} Y_{1_1} \\ Y_{1_2} \\ Y_{2_1} \\ Y_{2_2} \\ Y_{3_1} \\ Y_{3_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1_1} \\ \varepsilon_{1_2} \\ \varepsilon_{2_1} \\ \varepsilon_{2_2} \\ \varepsilon_{3_1} \\ \varepsilon_{3_2} \end{pmatrix}$$

Modell mit Referenz-Kodierung

Das Modell mit Referenz-Kodierung ist gegeben durch

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

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

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

Beim Modell mit Referenz-Kodierung gibt es einen Intercept μ_K der den Mittelwert der K-ten Gruppe angibt und die τ_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 $k = 1, \ldots, K-1$ und μ_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:

$$\boldsymbol{Y} = \begin{pmatrix} Y_{1_1} \\ Y_{1_2} \\ Y_{2_1} \\ Y_{2_2} \\ Y_{3_1} \\ Y_{3_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_3 \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1_1} \\ \varepsilon_{1_2} \\ \varepsilon_{2_1} \\ \varepsilon_{2_2} \\ \varepsilon_{3_1} \\ \varepsilon_{3_2} \end{pmatrix}$$

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

Setup zweifaktorielle Varianzanalyse

Im folgenden betrachten wir zwei diskrete Einflußgrößen $\boldsymbol{C} = \begin{pmatrix} C_1 \\ \vdots \\ C_n \end{pmatrix}$ und $\boldsymbol{D} = \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix}$ 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

$$\boldsymbol{Y} = (\boldsymbol{e} \;\; Z_1^e(\boldsymbol{C}) \cdots Z_{K_C-1}^e(\boldsymbol{C}) \;\; Z_1^e(\boldsymbol{D}) \cdots Z_{K_D-1}^e(\boldsymbol{D})) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K_C-1} \\ \gamma_1 \\ \vdots \\ \gamma_{K_D-1} \end{pmatrix} + \epsilon$$

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 μ und die τ_k und γ_l sind die Abweichungen der Gruppe mit $C = k$ bzw. $D = l$ vom Gesamtmittelwert bzw. vom Intercept μ.

Modell mit Referenz-Kodierung (mehrfakt.)

Das Modell mit Referenz-Kodierung ist gegeben durch

$$\boldsymbol{Y} = (\boldsymbol{e} \;\; Z_1(\boldsymbol{C}) \cdots Z_{K_C-1}(\boldsymbol{C}) \;\; Z_1(\boldsymbol{D}) \cdots Z_{K_D-1}(\boldsymbol{D})) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K_C-1} \\ \gamma_1 \\ \vdots \\ \gamma_{K_D-1} \end{pmatrix} + \epsilon$$

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

Bei dem Modell mit Referenz-Kodierung gibt es einen Intercept μ der den Mittelwert der Gruppe mit $C = K_C$ und $D = K_D$ angibt und die τ_k und γ_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:

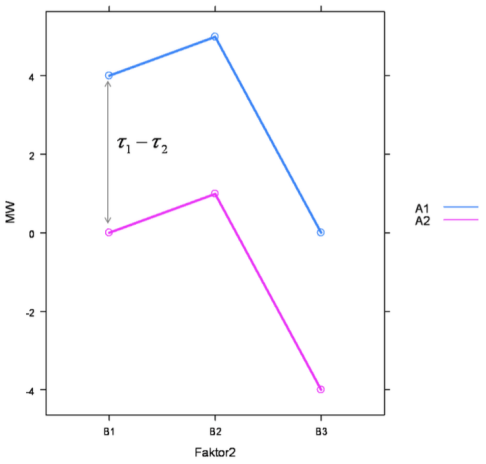
$$\boldsymbol{X} = (\boldsymbol{e} \; Z_1^e(\boldsymbol{C}) \; Z_1^e(\boldsymbol{D}) \; Z_2^e(\boldsymbol{D})) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

Referenz-Kodierung:

$$\boldsymbol{X} = (\boldsymbol{e} \; Z_1(\boldsymbol{C}) \; Z_1(\boldsymbol{D}) \; Z_2(\boldsymbol{D})) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

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:

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

Effekt-Kodierung:

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

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

Referenz-Kodierung:

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

! Der Verlauf für $C = 1$ und $C = K_C = 2$ ist parallel mit Abstand $\widehat{\tau}_1$. 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), Synergieefffekte (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 $\boldsymbol{X}^e$ und zusätzlich die Spalten der Interaktionsterme:

$$\boldsymbol{Z}^e = (Z_1^e(\boldsymbol{C})Z_1^e(\boldsymbol{D}) \cdots Z_1^e(\boldsymbol{C})Z_{K_D-1}^e(\boldsymbol{D}) \cdots Z_{K_C-1}^e(\boldsymbol{C})Z_{K_D-1}^e(\boldsymbol{D}))$$

Die Designmatrix ist also gegeben durch $\begin{pmatrix}\boldsymbol{X}^e & \boldsymbol{Z}^e\end{pmatrix}$

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

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)

- μ ist der Gesamtmittelwert.
- τ_k ist der Unterschied zwischen dem Mittelwert der Gruppe mit $C = k$ und dem Gesamtmittelwert.
- γ_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 τ_k und γ_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 $\boldsymbol{X}$ und zusätzlich die Spalten der Interaktionsterme:

$$\boldsymbol{Z} = (Z_1(\boldsymbol{C})Z_1(\boldsymbol{D}) \cdots Z_1(\boldsymbol{C})Z_{K_D-1}(\boldsymbol{D}) \cdots Z_{K_C-1}(\boldsymbol{C})Z_{K_D-1}(\boldsymbol{D}))$$

Die Designmatrix ist also gegeben durch $\begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{pmatrix}$.

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

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

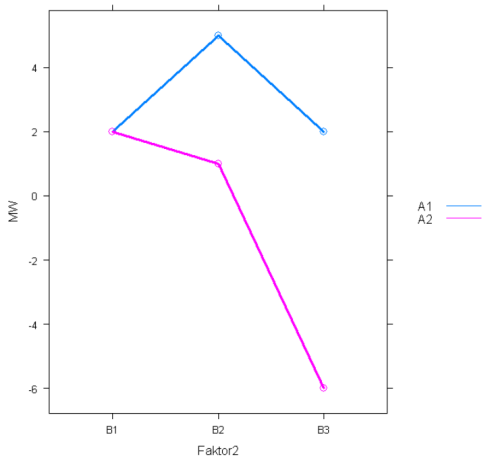
$$\forall l : (\tau\gamma)_{K_C,l} = 0 \quad \text{und} \quad \forall k : (\tau\gamma)_{k,K_D} = 0$$

Interpret. Interaktion (Referenz-Kodierung)

- τ_k ist der Unterschied zwischen dem Mittelwert der Gruppe mit $C = k$ zu der Referenzgruppe mit $C = K_C$.
- γ_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 τ_k und γ_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:

$$\begin{aligned}
\widehat{\mu} + \widehat{\tau}_1 + \widehat{\gamma}_1 + \widehat{(\tau\gamma)}_{1,1} &= 2 \\
\widehat{\mu} + \widehat{\tau}_1 + \widehat{\gamma}_2 + \widehat{(\tau\gamma)}_{1,2} &= 5 \\
\widehat{\mu} + \widehat{\tau}_1 + \widehat{\gamma}_3 + \widehat{(\tau\gamma)}_{1,3} &= 2 \\
\widehat{\mu} + \widehat{\tau}_2 + \widehat{\gamma}_1 + \widehat{(\tau\gamma)}_{2,1} &= 2 \\
\widehat{\mu} + \widehat{\tau}_2 + \widehat{\gamma}_2 + \widehat{(\tau\gamma)}_{2,2} &= 1 \\
\widehat{\mu} + \widehat{\tau}_2 + \widehat{\gamma}_3 + \widehat{(\tau\gamma)}_{2,3} &= -6
\end{aligned}$$

Effekt-Kodierung:

$$\begin{array}{llll}
\widehat{\mu} = 1 & \widehat{\gamma}_1 = 1 & \widehat{(\tau\gamma)}_{1,1} = -2 & \widehat{(\tau\gamma)}_{2,1} = 2 \\
\widehat{\tau}_1 = 2 & \widehat{\gamma}_2 = 2 & \widehat{(\tau\gamma)}_{1,2} = 0 & \widehat{(\tau\gamma)}_{2,2} = 0 \\
\widehat{\tau}_2 = -2 & \widehat{\gamma}_3 = -3 & \widehat{(\tau\gamma)}_{1,3} = 2 & \widehat{(\tau\gamma)}_{2,3} = -2
\end{array}$$

Referenz-Kodierung:

$$\begin{array}{llll}
\widehat{\mu} = -6 & \widehat{\gamma}_1 = 8 & \widehat{(\tau\gamma)}_{1,1} = -2 & \widehat{(\tau\gamma)}_{2,1} = 0 \\
\widehat{\tau}_1 = 8 & \widehat{\gamma}_2 = 7 & \widehat{(\tau\gamma)}_{1,2} = 1 & \widehat{(\tau\gamma)}_{2,2} = 0 \\
\widehat{\tau}_2 = 0 & \widehat{\gamma}_3 = 0 & \widehat{(\tau\gamma)}_{1,3} = 0 & \widehat{(\tau\gamma)}_{2,3} = 0
\end{array}$$

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:

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

Referenz-Kodierung:

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

Kombination mit stetigen Merkmalen

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

Interaktion: zwei kategorische Variablen

Gegeben sei das Regressionsmodell

$$\begin{aligned}\boldsymbol{Y} =& \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \\ & \beta_3 I(\boldsymbol{z}_1 = B) + \beta_4 I(\boldsymbol{z}_1 = C) + \beta_5 I(\boldsymbol{z}_2 = 1) + \\ & \beta_6 I(\boldsymbol{z}_1 = B, \boldsymbol{z}_2 = 1) + \beta_7 I(\boldsymbol{z}_1 = C, \boldsymbol{z}_2 = 1) + \varepsilon\end{aligned}$$

Die geschätzen 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 $z_1B{:}z_21$ und $z_1C{:}z_21$ 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

$$\begin{aligned}\boldsymbol{Y} =& \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \\ & \beta_3 I(\boldsymbol{z}_1 = B) + \beta_4 I(\boldsymbol{z}_1 = C) + \beta_5 I(\boldsymbol{z}_2 = 1) + \\ & \beta_6 \boldsymbol{x}_1 I(\boldsymbol{z}_1 = B) + \beta_7 \boldsymbol{x}_1 I(\boldsymbol{z}_1 = C) + \varepsilon\end{aligned}$$

Die geschätzen 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 $x_1{:}z_1B$ und $x_1{:}z_1C$ 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

$$\begin{aligned}\boldsymbol{Y} =& \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \\ & \beta_3 I(\boldsymbol{z}_1 = B) + \beta_4 I(\boldsymbol{z}_1 = C) + \beta_5 I(\boldsymbol{z}_2 = 1) + \\ & \beta_6 \boldsymbol{x}_1 \boldsymbol{x}_2 + \varepsilon\end{aligned}$$

Die geschätzen 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 einfaktorlielle 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.

ANOVA ein Modell

Gegeben sei ein Modell mit zwei kategorischen Einflussgrößen $z1$ und $z2$ und einer metrischen Einflussgröße $x1$ und einer Interaktion zwischen $z1$ und $z2$. Gegeben sei folgender Output einer ANOVA:

```
> anova(model)
Analysis of Variance Table

Response: y1
           Df Sum Sq Mean Sq  F value Pr(>F)
x1          1  35184   35184 136.2762 <2e-16 ***
z1          2    252     126   0.4875 0.6147
z2          1     13      13   0.0496 0.8239
z1:z2       2    289     144   0.5589 0.5724
Residuals 293  75647     258
```

- **df**: degrees of freedom geben bei den Variablen an, wie viele Parameter zu dieser Variable geschätzt wurden (ohne Intercept mitzuzählen). D.h. für $x1$ wurde 1 Parameter geschätzt, für $z1$ wurden 2 Parameter geschätzt und für $z2$ wurde 1 Parameter geschätzt und für die Interaktion wurden 2 Parameter geschätzt.
  Notiz: df für die Residuals ist $n - p - 1$. Kann man nutzen um n zu berechnen. Hier z.B.:

  $$\begin{aligned} n &= df_{x1} + df_{z1} + df_{z2} + df_{z1:z2} + df_{Residuals} + 1 \\ &= 1 + 2 + 1 + 2 + 293 + 1 = 300 \end{aligned}$$

- **Sum Sq**: Summe der Quadrate gibt an, wie viel Varianz durch die jeweilige Variable erklärt wird. Für die Variablen ist das jeweils der SSM und für die Residuen ist das der SSE. Es gilt:

  $$\begin{aligned} SST &= SSM + SSE \\ &= SSM_{x1} + SSM_{z1} + SSM_{z2} + SSM_{z1:z2} \\ &\quad + SSE_{Residuals} \end{aligned}$$

  und somit $R^2 = \frac{SSM_{x1}+SSM_{z1}+SSM_{z2}+SSM_{z1:z2}}{SST}$.
- **Mean Sq**: Analog zu Sum Sq, aber eben MSE und MSM. Berechnet als SSM/df und SSE/df. ! Werte sind hier gerundet.
- **F-Value**: Der F-Wert ist ein Teststatistik, die angibt, ob die erklärende Variable signifikant für das Modell ist. Berechnet sich als $F = \frac{MSM}{MSE}$. MSE ist für alle Variablen gleich, hier also $MSE_{Residuals} = 258$.

ANOVA mehrere Modelle

Gegeben seien drei Modelle mit kategorischen Einflussgrößen $z1$ und $z2$ und einer metrischen Einflussgröße $x1$. Wir vergleichen die drei Modelle mittels ANOVA:

```
> anova(model_ohne_z2, model_ohne_interaktion, model_full)
Analysis of Variance Table

Model 1: y1 ~ x1 + z1
Model 2: y1 ~ x1 + z1 + z2
Model 3: y1 ~ x1 + z1 * z2
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    296 75948
2    295 75935  1    12.809 0.0496 0.8239
3    293 75647  2   288.603 0.5589 0.5724
```

- **Res. Df**: degrees of freedom für die Residuals. Bei gleichem n kann man hier den Unterschied in der Anzahl der geschätzten Parameter sehen. Hier hat das dritte Modell die meisten Parameter geschätzt.
- **RSS**: Residual Sum of Squares gibt an, wie viel Varianz **nicht** durch das Modell erklärt wird. Je kleiner, desto besser.
- **Df**: degrees of freedom geben an, wie viel **mehr** oder **weniger** Parameter geschätzt wurden im Vergleich zum vorherigen Modell. Hier wurde bei Model 2 ein Parameter mehr geschätzt als bei Model 1 und bei Model 3 zwei Parameter mehr als bei Model 2.
- **Sum of Sq**: Wie viel Varianz durch das Modell **zusätzlich** erklärt wird im Vergleich zum vorherigen Modell. Einfach die Differenz der RSS-Werte.
- **F-value**: Gibt an, ob das Modell signifikant besser ist als das vorherige Modell. Berechnet sich als $F = \frac{\frac{\text{Sum of Sq}}{Df}}{(\frac{\text{RSS}}{\text{Res.Df}})}$.

Kapitel 5 - Metrische Einflußgrößen

Interaktion metrischer Variablen

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

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

Interpreation der Modellparameter:
Die Parameter β_1, β_2 geben die Steigung bei $x_1 = x_2 = 0$ an. I.d.R. nicht sinnvoll interpretierbar.

Modelle mit metrischen Variablen

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

- **Einfaches lineares Modell:**
  $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- **Transformiertes lineares Modell:**
  $Y_i = \beta_0 + \beta_1 T(x_i) + \epsilon_i, \quad$ z.B. $T(x) = \log(x)$
- **Polynomielles Modell:**
  $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \epsilon_i$
  Problem: Bestimmung von d.
  Mögliche Lösung: Test auf Signifikanz der Koeffizienten $\beta_2, \ldots, \beta_d$ mittels sequentieller Quadratsummenzerlegung.
- **Stückweise konstantes Modell:**
  $Y_i = \beta_0 I_{[x_i \leq g_1]} + \beta_1 I_{[g_1 < x_i \leq g_2]} + \cdots + \beta_{k-1} I_{[g_{k-1} < x_i \leq g_k]} + \beta_p I_{[x_i > g_p]} + \epsilon_i$
  Dies entspricht der Kategorisierung der x-Variablen.
- **Stückweise lineares Modell:**
  $Y_i = \beta_0 + \beta_1 x_i + \beta_2 \max\{x_i - g_1, 0\} + \cdots + \beta_p \max\{x_i - g_h, 0\} + \epsilon_i$
  mit bekannten Bruchpunkten (Knoten) g_j.
- **Regressionssplines:**
  $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \beta_{k+1} \max\{x_i - g_1, 0\}^3 + \cdots + \beta_{k+h} \max\{x_i - g_h, 0\}^3 + \epsilon_i$
  Ein Polynom 3.Grades ist 2-mal stetig differenzierbar.

Allgemeiner Ansatz mit Basisfunktionen

Sei X eine metrische Variable mit den Ausprägungen x_i für $i = 1, \ldots, n$.
Allgemeiner Ansatz für Modelle mit Basisfunktionen:
Seien $B_1, B_2, \ldots, B_k$ Basisfunktionen. Dann ist die allgemeine Modellgleichung:

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

Kapitel 6 - Modelldiagnose

Arten von Residuen

Gegeben sei das multiple lineare Regressionsmodell mit den üblichen Annahmen über ε und $\boldsymbol{X}$. Aus Kapitel 2 wissen wir, dass $\mathbb{V}(\hat{\varepsilon}) = \sigma^2\boldsymbol{Q}$ und somit $\mathbb{V}(\hat{\varepsilon}_i) = \sigma^2 q_{ii}$, wobei q_{ii} das i-te Diagonalelement der Matrix $\boldsymbol{Q}$ 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 σ^2 die Residuen mit einbezogen werden. Dies führt zu einer Verzerrung der Residuen. Wir definieren daher zusätzlich studentisierte Residuen als

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

wobei $\hat{\sigma}^2_{(i)}$ die Schätzung von σ^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 ε und $\boldsymbol{X}$. Folgende Probleme können typischerweise auftreten:

- ε_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.

ε_i nicht normalverteilt

- **Ursachen**: Die abhängige Variable Y kann bedingt auf $\boldsymbol{x}$ 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 verschieden Residualvarianzen innerhalb der Gruppen. Multiplikative Fehlerstruktur, d.h. σ_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 $\boldsymbol{X}^\top \boldsymbol{X}$ und der Korrelationsmatrix der metrischen Einflussgrößen.
  - Konditionszahl von $\boldsymbol{X}$:

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

    $\kappa(\boldsymbol{X}) \gg 1$ deutet auf Multikollinearität hin.
  - Varianz Inflationsfaktor:

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

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

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

    Wenn $VIF_j = 1$, dann heißt das, dass $\boldsymbol{X}_j$ 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 einer Einflußgröße $\mathbf{x}$, die stark von den anderen Beobachtungen in $\mathbf{x}$ abweicht.

- **Ursachen**: Falsche Erhebung; Beobachtung gehört nicht zur Grundgesamtheit; Besonderheiten bei einzelner UE.
- **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 $\boldsymbol{P}$ zum Auffinden von high leverage points; Verschiedene Residuenplots zur Ausreißeranalyse; Influence-Statistiken.
- **Therapie**: Fehlerhafte Daten weglassen (Sensitivitätsanalyse); Robuste Regression; Gewichtete Regression.

Wir definieren Leverage als

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

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

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

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 $\hat{\varepsilon}$ gegen $\hat{\boldsymbol{Y}}$; 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

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

wobei wir nun annehmen, dass $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{W}^{-1})$, mit **bekannter** positiv definiter Matrix $\boldsymbol{W}$. Wir nennen $\boldsymbol{W}$ die Gewichtsmatrix.
Im Falle von heteroskedastischen, aber unkorrelierten Fehlern ist $\boldsymbol{W}$ eine Diagonalmatrix mit $\boldsymbol{W} = \text{diag}(w_1, \cdots, w_n)$ und $\mathbb{V}(\varepsilon) = \text{diag}(\frac{\sigma^2}{w_1}, \cdots, \frac{\sigma^2}{w_n})$.

Schätzer im allg. linearen Regressionsmodell

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

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{Y},\\ \mathbb{E}(\hat{\beta}) &= \beta,\\ \mathbb{V}(\hat{\beta}) &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\end{aligned}$$

und $\hat{\beta}$ ist der Beste lineare unverzerrte Schätzer (BLUE) für β.
Der Schätzer für σ^2 ist gegeben durch den REML-Schätzer

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

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

Umformung

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

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

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

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

wobei

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

AR(1) Zeitreihenmodell

Das AR(1) Zeitreihenmodell hat die Form

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

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 $\boldsymbol{W}^{-1}$ gegeben durch

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

Damit gilt also

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

Linear Mixed Modell

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

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

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

Wir bezeichnen β_0 als den fixed Population Intercept und γ_{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 β_j sind die fixed Effekte.

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

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

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

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

und

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

wobei $\boldsymbol{J}_{n_i}$ die $n_i \times n_i$-Matrix mit Einsen ist.
Daraus ergibt sich das marginale Modell

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

! 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.

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

ICC Interpretation

Sei ein Random Intercept Modell gegeben durch

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

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

Dann gilt für die Intra-Class Correlation (ICC) folgende Interpretationen:

- **Varianzkomponenten**:
  Der ICC quantifiziert den Anteil der Gesamtvarianz, der auf die Gruppierungsstruktur in den Daten (die random effect variable) zurückzuführen ist.
- **Homogenität innerhalb von Gruppen**:
  Der ICC gibt an, wie stark die Beobachtungen innerhalb einer Gruppe korreliert sind. Bei Clustern bedeutet ein hoher ICC, dass die UEs innerhalb eines Clusters ähnliche Werte haben. Bei Longitudinaldaten bedeutet ein hoher ICC, dass die Beobachtungen innerhalb einer UE über die Zeit ähnliche Werte haben.

```
Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.131803 0.36305
 Residual             0.003219 0.05674
```

In diesem Beispiel ist die ICC gegeben durch

$$\begin{aligned} ICC &= \text{Corr}(Y_{ij}, Y_{il}) \\ &= \frac{\sigma^2_{\gamma_0}}{\sigma^2_{\gamma_0} + \sigma^2} \\ &= \frac{0.1318}{0.1318 + 0.003} = 0.977. \end{aligned}$$

Logistisches Regressionsmodell

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

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

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

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

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 defniert 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

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

Dann gilt:

- $\pi_i = \mathbb{P}(Y_i = 1 \mid \boldsymbol{x}_i) = h(\boldsymbol{x}_i^\top \beta)$
- Wenn sich x_k um eine Einheit erhöht, so ändert sich die log-odds der Wahrscheinlichkeit (ceteris paribus) um β_k Einheiten.
- Wenn sich x_k um eine Einheit erhöht, so ändern sich die Odds $\frac{\pi_i}{1-\pi_i}$ (ceteris paribus) 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, ...` die unabhängigen Variablen sind.

MLE für Logit Modell

Die log-likelihood Funktion lautet

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

Mit der Linkfunktion und Responsefunktion können wir die Likelihood umschreiben als

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

Daraus ergibt sich die Scorefunktion

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

Die (observed) Fisher Matrix ist

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

Daraus ergibt sich der MLE Schätzer für β durch

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

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