

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

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

und nennen $(\hat{\beta}_0, \hat{\beta}_1)$ den **KQ-Schätzer von (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 $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$. 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 β_1 Einheiten.
- Es gilt $\beta_0 = E(Y | X = 0)$.
- 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: $E(\hat{\beta}_0, \hat{\beta}_1) = (\beta_0, \beta_1)$.
- $V(\hat{\beta}_1) = \frac{\sigma^2}{n S_x^2}$ 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 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

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

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 $\hat{\beta}_0$

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

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

Damit können wir Konfidenzintervalle zum Niveau $1 - \alpha$ für 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 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

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

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

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)
```

Kapitel 2 - Das multiple lineare Regressionsmodell

Überschrift