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: $\mathbb{E}(\hat{\beta}_0, \hat{\beta}_1) = (\beta_0, \beta_1)$.
- $\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n S_x^2}$ und $\mathbb{V}(\hat{\beta}_0) = \sigma^2 (\frac{1}{n} + \frac{\overline{x}^2}{n S_x^2})$.
- $(\hat{\beta}_0, \hat{\beta}_1)$ 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)
```

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 von einander 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.

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: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ mit

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

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

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

Kleinste Quadrate (KQ) Schätzer

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad (2)$$

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

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

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

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

Interpretation der Modellparameter

- Y_i hängt linear von $x_{i1}, \ldots, x_{in}$ ab.
- Steigt x_k um eine Einheit, so steigt Y (ceteris paribus) im Erwartungswert um β_k Einheiten, **wenn** alle anderen x-Variablen festgehalten werden.
- ! 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 $\hat{\boldsymbol{\beta}}$

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

Hat-Matrix und Residualmatrix

Gegeben dem multiplen linearen Modell mit $\mathrm{rang}(\boldsymbol{X}) = p + 1$ gilt

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

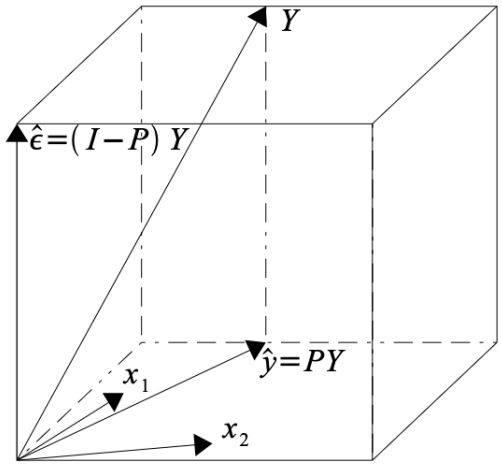
$$\boldsymbol{P} := \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{n \times n}$$

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

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

$\boldsymbol{P}$ 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:

$$\boldsymbol{P}^\top = \boldsymbol{P} \text{ und } \boldsymbol{P}^2 = \boldsymbol{P}$$
$$\boldsymbol{Q}^\top = \boldsymbol{Q} \text{ und } \boldsymbol{Q}^2 = \boldsymbol{Q}$$
$$\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{P} = \boldsymbol{0}.$$

Daraus folgt

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

Schätzer für σ^2

Gegeben dem multiplen linearen Modell, gilt

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

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

Überschrift