## Einfaches lineares Regressionsmodell

Das einfache lineare Regressionsmodell hat die Form

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

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

## Kleinste Quadrate (KQ) Schätzer

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

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

und nennen $(\hat\beta_0, \hat\beta_1)$ den KQ-Schätzer von $(\beta_0, \beta_1)$ und $\hat\varepsilon_i := Y_i - (\hat\beta_0 + \hat\beta_1 x_i)$ die Residuen.

## Existenz und Berechnung vom KQ Schätzer

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

$\bar{y} \sum (x_i - \bar{x}) = \bar{y} \cdot 0 = 0$

$$\hat\beta_1 = \frac{S_{xY}}{S_x^2} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2} \overset{!}{=} \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} y_i \sim \mathcal{N}(\dots)$$

$$\hat\beta_0 = \overline{Y} - \hat\beta_1 \overline{x}.$$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$ : $\hat\beta_1 = \frac{\beta_1 \sum (x_i - \bar{x})^2 + \sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2 \mid S_x^2 \cdot n} = \beta_1 + \frac{1}{S_x^2 n} \sum (x_i - \bar{x})\varepsilon_i \sim \mathcal{N}(\dots)$

Durch differenzieren von der Gleichung (1) erhält man $(\hat\beta_0, \hat\beta_1)$ als Lösung der Normalengleichungen

$$\sum_{i=1}^n \hat\varepsilon_i = 0$$

$$\sum_{i=1}^n \hat\varepsilon_i x_i = 0 \qquad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \perp \begin{pmatrix} \hat\varepsilon_1 \\ \hat\varepsilon_2 \\ \vdots \\ \hat\varepsilon_n \end{pmatrix}$$

## 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 $\beta_1$ Einheiten.
- Es gilt $\beta_0 = \mathbb{E}(Y | X = 0)$.
- Der Parameter $\sigma$ die erwartete Abweichung der

## 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 $\sigma^2$

Gegeben dem einfachen linearen Modell mit $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, gilt $\hat\sigma^2 = \frac{RSS}{n-2}$

$$\hat\sigma^2 := \frac{1}{n-2} \sum_{i=1}^n \boxed{\hat\varepsilon_i^2} \left(y_i - (\hat\beta_0 + \hat\beta_1 x_i)\right)^2$$

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

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

Der KQ-Schätzer $(\hat\beta_0, \hat\beta_1)$ und der Schätzer $\hat\sigma^2$ sind stoch.unabhängig.

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

Gegeben dem einfachen linearen Modell mit $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, gilt für $\hat\beta_1$ und $\hat\beta_0$ $\hat\beta_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$

$$\frac{\hat\beta_1 - \beta_1}{\hat\sigma_{\hat\beta_1}} \sim \underset{\text{FG (2 verbraucht für Schätz. von } \hat\beta_0, \hat\beta_1)}{t_{n-2}} \text{ mit } \hat\sigma_{\hat\beta_1} := \sqrt{\frac{\hat\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \mid nS_x^2$$

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

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

$$[\hat\beta_1 - \hat\sigma_{\hat\beta_1} \underbrace{t_{1-\alpha/2}}_{\text{Quantil}}(n-2); \hat\beta_1 + \hat\sigma_{\hat\beta_1} t_{1-\alpha/2}(n-2)]$$

$$[\hat\beta_0 - \hat\sigma_{\hat\beta_0} t_{1-\alpha/2}(n-2); \hat\beta_0 + \hat\sigma_{\hat\beta_0} t_{1-\alpha/2}(n-2)]$$

Varianz von Prognosefehler $= Var\left((\hat\beta_0 + \hat\beta_1 x_{n+1}) - \underbrace{(\beta_0 + \beta_1 x_{n+1})}_{\text{const.}}\right) + Var(\varepsilon_{n+1})$ ($\varepsilon_{n+1}$ unabh. von $(\hat\beta_0, \hat\beta_1)$)

$= \underbrace{Var(\hat\beta_0 + \hat\beta_1 x_{n+1})}_{Var(\hat y_{n+1})} + \underbrace{\sigma^2}_{Var(\varepsilon_{n+1})}$

mit $Var(\hat y_{n+1}) = Var(\hat\beta_0 + \hat\beta_1 x_{n+1}) = Var(\hat\beta_0) + x_{n+1}^2 Var(\hat\beta_1) + 2 x_{n+1} \underset{-\bar{x}\sigma^2 / S_x^2}{cov(\hat\beta_0, \hat\beta_1)}$

u. $\mathbb{V}(\hat\beta_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)$, $\mathbb{V}(\hat\beta_1) = \frac{\sigma^2}{nS_x^2}$

## 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 $Y$ (weil $r_{xy}$ invariant)
- $R$ 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: $\hat y_{n+1} - y_{n+1} = [\hat\beta_0 + \hat\beta_1 x_{n+1} - \beta_0 - \beta_1 x_{n+1}] - \varepsilon_{n+1}$

## 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

$= \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{n+1}$ $= \overbrace{E(\hat{\beta}_0)}^{\beta_0} - \beta_0 + x_{n+1} \overbrace{E(\hat{\beta}_1)}^{\beta_1} - x_{n+1}\beta_1 - \overbrace{E(\varepsilon_{n+1})}^{0}$

$$\mathbb{E}(\hat{Y}_{n+1} - Y_{n+1}) = 0$$ $E(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} - \beta_0 - \beta_1 x_{n+1} - \varepsilon_{n+1}) = 0$

$= \mathbb{V}(\hat{Y}_{n+1}) + \mathbb{V}(\varepsilon_{n+1})$

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

$\varepsilon_{n+1}$ unabh. von Schätzung $(\hat{\beta}_0, \hat{\beta}_1)$ $\mathbb{V}(\hat{Y}_{n+1}) = \sigma^2\left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n s_X^2}\right)$, $\mathbb{V}(\varepsilon_{n+1}) = \sigma^2$

## 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

eigtl. $\hat{\sigma}_{pred, n+1} = \sqrt{Var(\hat{Y}_{n+1} - Y_{n+1})}$

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

unbekanntes $\sigma^2$ schätzen

## R-Code

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

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

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

## Interpretation von transformierten Modellen

- Log-Log-Modell: $\log(10)$: $\log(10) x_i \uparrow$ um eine Einheit $\Rightarrow Y_i$ steigt im EW um $e^{\beta_1}$ $[10^{\beta_1}]$

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

  $E(Y|ax) = e^{\beta_0} \cdot (ax)^{\beta_1} = e^{\beta_0} \cdot a^{\beta_1} x^{\beta_1} = a^{\beta_1} \cdot e^{\beta_0} \cdot x^{\beta_1} = a^{\beta_1} \cdot E(Y|X)$

  Wenn $x_i$ um den Faktor $a$ steigt, dann steigt $Y_i$ im Erwartungswert um den Faktor $\underline{a^{\beta_1}} = e^{\beta_1 log(a)}$. $e^{\beta_1 \log(x_i)} = e^{\log(x_i^{\beta_1})} = x_i^{\beta_1}$ genauso bei log10()!

  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 $\beta_1$% Einheiten.

- Log-Linear-Modell:

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

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

## 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 (also $Var(X) \uparrow$ doppelt gut)

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.

**Beachte:** Grundsätzlich gelten alle Aussagen nur unter der zentralen Modellannahme des linearen Zusammenhangs von $E(Y)$ und $X$!