

Satz

Unter normalverteilten Fehlern ist  $\hat{\beta}_1$  normalverteilt mit

$$E(\hat{\beta}_1) = \beta_1, \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(n-1)s_{xx}}.$$

- Insbesondere ist  $\hat{\beta}_1$  erwartungstreu.
- Man kann auch zeigen, dass  $\hat{\beta}_0$  normalverteilt ist.
- Die unbekannte Varianz  $\sigma^2$  wird mithilfe der Residuen geschätzt.
- 

Residuen  $r_i$

$$r_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (i = 1, \dots, n).$$

- 
- [[Schätzer]] für  $\sigma^2$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n r_i^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

$$\text{Es gilt } E(\hat{\sigma}^2) = \sigma^2.$$

Verteilung der Schätzer

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)s_{xx}}\right) \iff \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{(n-1)s_{xx}}}} \sim N(0, 1).$$

Schätzen wir  $\sigma^2$  mittels  $\hat{\sigma}^2$ , folgt

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{(n-1)s_{xx}}}} \sim t_{n-2}.$$

- 
- Konfidenzintervall

$$CI_{1-\alpha}(\beta_1) = \hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{(n-1)s_{xx}}}.$$

## Konfidenz und Prädiktionsintervalle

- Konfidenzintervalle für die zu erwartende Response  $E(Y_h)$ ,
- Prädiktionsintervalle für die Response  $Y_h$ .

$$Y_h = \beta_0 + \beta_1 x_h + \epsilon_h.$$

$$\widehat{E(Y_h)} = \hat{\beta}_0 + \hat{\beta}_1 x_h.$$

- 
- Konfidenzintervall für zu erwartende Response

$$CI_{1-\alpha}(\mu_h) = \widehat{E(Y_h)} \pm t_{n-2, 1-\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{(n-1)s_{xx}}}.$$

Der unbekannte Erwartungswert von  $Y_h$  wird also mit Wahrscheinlichkeit  $1 - \alpha$  von  $CI_{1-\alpha}(\mu_h)$  überdeckt, das heißt

$$P(\mu_h \in CI_{1-\alpha}(\mu_h)) = 1 - \alpha.$$

- Prädiktionsintervalle

$$PI_{1-\alpha}(Y_h) = \widehat{E(Y_h)} \pm t_{n-2, 1-\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{(n-1)s_{xx}}}.$$

## Beispiele

- Festplatte Konfidenzintervall
  - siehe [[Lineare Regression]]

- Angenommen, wir haben Daten der Größe  $x_h = 250$  kB. Wie lange  
 \* benötigt die Festplatte im Mittel, um diese Daten auszulesen?  
 Vorherigen Berechnungen entnehmen wir

$$\widehat{E(Y_h)} = 15.41 + 0.34 \cdot 250 = 100.41,$$

- \* sowie  $\bar{x} = 550$  und  $(n-1)s_{xx}^2 = 825\,000$ . Für die geschätzte

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n r_i^2 = \frac{1}{8} \sum_{i=1}^{10} (Y_i - 15.41 - 0.34 x_i)^2 = 81.73.$$

- \* Wir sind an einem 95%igen Konfidenzintervall interessiert und lesen ab  $t_{n-2, 1-\alpha/2} = t_{8, 0.975} = 2.31$ . Daraus ergibt sich

$$\begin{aligned} \text{CI}_{0.95}(\mu_h) &= \widehat{E(Y_h)} \pm t_{n-2, 1-\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{(n-1)s_{xx}^2}} \\ &= 100.41 \pm 2.31 \cdot 9.04 \sqrt{\frac{1}{10} + \frac{(250 - 550)^2}{825\,000}} \\ &= [89.79, 108.85]. \end{aligned}$$

\*

- Festplatte Prädiktionsintervall

$$\begin{aligned} \text{PI}_{0.95}(Y_h) &= \widehat{E(Y_h)} \pm t_{n-2, 1-\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{(n-1)s_{xx}^2}} \\ &= 100.41 \pm 2.31 \cdot 9.04 \sqrt{1 + \frac{1}{10} + \frac{(250 - 550)^2}{825\,000}} \\ &= [76.40, 122.24]. \end{aligned}$$

- R

```
> x.h <- data.frame(length = 250)
> predict(lm.disk, x.h, interval = "confidence",
          level = 0.95)
      fit      lwr      upr
1 99.32018 89.78762 108.8527
> predict(lm.disk, x.h, interval = "predict",
          level = 0.95)
      fit      lwr      upr
1 99.32018 76.39717 122.2432
```