Einfache lineare Regression

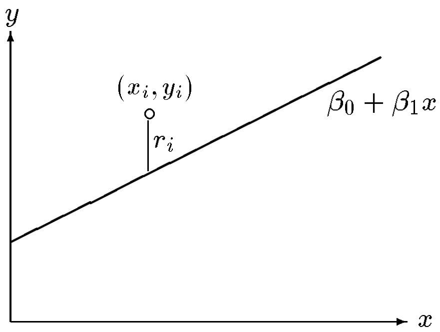
Das Modell

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

y_i ist die Zielvariable der i-ten Beobachtung.
x_i ist die erklärende Variable der i-ten Beobachtung. x_i ist eine feste, nicht zufällige Grösse.
β_0, β_1 sind unbekannte *Parameter*, die sog. Regressionskoeffizienten. Diese sollen mit Hilfe der vorhandenen Daten geschätzt werden.
ϵ_i ist der *zufällige Rest* oder *Fehler*, d. h. die zufällige Abweichung von y_i von der Geraden. Es wird vorausgesetzt, dass $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ ist und dass $Cov(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$.

Methode der Kleinsten Quadrate

Residuen: $r_i = y_i - (\beta_0 + \beta_1 x_i)$
Minimiere $Q(\beta_0, \beta_1) = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$



Normalgleichungen:

$$\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned}$$

Least Squares-Lösung:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regressionsgerade:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Eigenschaften der LS-Schätzer

a) erwartungstreu $E(\hat{\beta}_0) = \beta_0$ und $E(\hat{\beta}_1) = \beta_1$
b) **BLUE** Best Linear Unbiased Estimator

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right) \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Schätzung für σ^2 :

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$$

Tests und Vertrauensintervalle

Vor: ϵ_i normalverteilt und unabhängig.
Das Modell kann nun so geschrieben werden:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad y_i \text{ und } y_j \text{ unabhängig für } i \neq j$$

t-Test für $H_0 : \beta_1 = \beta$ gegen $H_A : \beta_1 \neq \beta$:

$$t^* = \frac{\hat{\beta}_1 - \beta}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta}{\sqrt{\hat{\sigma}^2 / \sum (x_i - \bar{x})^2}}$$

t^* hat eine t-Verteilung mit $n-2$ Freiheitsgraden. Die Grösse $se(\hat{\beta}_1)$ heisst *Standardfehler (standard error)* von $\hat{\beta}_1$. Verwerfe H_0, wenn $|t^*| > t_{97.5\%, n-2}$.

Ein 95%-Vertrauensintervall für β_1 ist:

$$\hat{\beta}_1 \pm t_{97.5\%, n-2} \cdot \sqrt{\hat{\sigma}^2 / \sum (x_i - \bar{x})^2}$$

Ein 95%-Vertrauensintervall für $\beta_0 + \beta_1 x_0$ ist

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{97.5\%, n-2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Ein 95%-*Prognoseintervall* für y_0 ist

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{97.5\%, n-2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Varianzanalyse

Zerlegung der Quadratsummen:

$$\begin{array}{ccc} \sum_{i=1}^{n} (y_i - \bar{y})^2 & = & \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ SST & = & SSR \quad + \quad SSE \\ \text{total sum of squares} & = & \text{regression sum of squares} + \text{error sum of squares} \end{array}$$

Mean squares of ... $= \dfrac{\text{sum of squares of ...}}{\text{Freiheitsgrade}}$
F-Test für $H_0 : \beta_1 = 0$ mit der Teststatistik:

$$F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

F^* hat unter H_0 eine F-Verteilung mit 1 und $n-2$ Freiheitsgraden. Verwerfe H_0, wenn $F^* > F_{95\%, 1, n-2}$

Anova-Tabelle:

Source of Variation	Sum of squares	Degrees of Freedom	Mean square	F^*
Regression	SSR	1	MSR	MSR/MSE
Residual	SSE	$n-2$	MSE	
Total	SST	$n-1$		

Bestimmtheitsmass R^2: Anteil an der Gesamtvariabilität, der „durch die Regression erklärt wird“:

$$R^2 = 1 - \frac{SSE}{SST}$$

Es gilt $R^2 = r^2$, wobei r die Korrelation zwischen x und y ist.