Einfache Lineare Regression

Wenn man einen linearen Zusammenhang zwischen zwei Zufallsvariablen X und Y vermutet, kann man ein Modell der Form

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

schreiben für Konstanten β_0, β_1 und eine Zufallsvariable ε mit Erwartungswert 0 und endlicher Varianz, die unabhängig von X ist.

Die Methode der kleinsten Quadrate (ordinary least squares) schätzt die Parameter β_0 und β_1 als die beiden Werte b_0 und b_1, welche die Summe der quadrierten Abweichungen

$$\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2 \tag{1}$$

minimieren, wobei $(x_1, y_1), \ldots, (x_n, y_n)$ n unabhängige Realisierungen des Paars (X, Y) sind. Wenn man (1) nach β_0 und β_1 ableitet, kriegt man die beiden Optimalitätsbedingungen

$$\sum_{i=1}^{n} b_0 + b_1 x_i - y_i = 0 \quad \text{und} \quad \sum_{i=1}^{n} (b_0 + b_1 x_i - y_i) x_i = 0,$$

welche man äquivalent schreiben kann als

$$b_0 + b_1 \bar{x} = \bar{y} \quad \text{und} \quad b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy}, \tag{2}$$

für

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \overline{x^2} := \frac{1}{n} \sum_{i=1}^{n} x_i^2 \quad \text{und} \quad \overline{xy} := \frac{1}{n} \sum_{i=1}^{n} x_i y_i.$$

Aus (2) kriegt man

$$b_0 \bar{x} + b_1 \bar{x}^2 = \bar{x}\bar{y} \quad \text{und} \quad b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy},$$

und deswegen

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} \quad \text{und} \quad b_0 = \bar{y} - \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} \bar{x},$$

wobei

$$\operatorname{var}(x) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{und} \quad \operatorname{cov}(x, y) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

die Stichprobenvarianz von $(x_1, \ldots, x_n)$ und Stichprobenkovarianz von $(x_1, y_1), \ldots, (x_n, y_n)$ sind.

Bemerkung 1 Die korrigierte Stichprobenvarianz (oder empirische Varianz)

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

ist eine erwartungstreue Schätzung von $\operatorname{Var}(X)$, und die korrigierte Stichprobenkovarianz (oder empirische Kovarianz)

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

ist eine erwartungstreue Schätzung von $\operatorname{Cov}(X, Y)$.

Aber für grosse n wird der Unterschied zwischen Stichprobenvarianz/Stichprobenkovarianz und korrigierter Stichprobenvarianz/Stichprobenkovarianz klein. Dann spielt es keine grosse Rolle, ob man $\operatorname{Var}(X)/\operatorname{Cov}(X, Y)$ mit Stichprobenvarianz/Stichprobenkovarianz oder korrigierter Stichprobenvarianz/Stichprobenkovarianz schätzt.