Dinge die man in zwei Dimensionen machen kann - Multiple lineare Regression

Norman Markgraf

2021-06-24

Wir wollen den Fall unterschen bei dem wir mit zwei statistischen Variabeln (X und Y) eine dritte Variable (Z) mittels einer multiplen linearen Regression modellieren.

Es seien die Datenpunkte $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ gegeben und wir wollen eine lineare Funktion $g(x, y)$ finden, so dass

$$z_i = g(x_i, y_i) + \epsilon_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot y_i + \epsilon_i$$

gilt und der Abweichungsterm ϵ_i möglichst klein ist.

Auf Grundlage unserer Datenpunkt wollen wir die Koeffizienten so schätzen, dass die Summe der quadratische Abweichungen minimal ist.

$$QS = QS(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} (z_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i - \hat{\beta}_2 \cdot y_i)^2$$

Das führt zu der folgenden, notwendigen Bedingen (für stationäre Punkte):

$$\nabla QS(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Im einzelnen heißt das:

$$\begin{aligned}\frac{\partial}{\partial\hat{\beta}_0}QS(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2) &= -2\cdot\sum_{i=1}^{n}\left(z_i-\hat{\beta}_0-\hat{\beta}_1\cdot x_i-\hat{\beta}_2\cdot y_i\right)\\ &= -2\cdot n\cdot\left(\bar{z}-\hat{\beta}_0-\hat{\beta}_1\cdot\bar{x}-\hat{\beta}_2\cdot\bar{y}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\hat{\beta}_1}QS(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2) &= -2\cdot\sum_{i=1}^{n}\left(z_i\cdot x_i-\hat{\beta}_0\cdot x_i-\hat{\beta}_1\cdot x_i\cdot x_i-\hat{\beta}_2\cdot y_i\cdot x_i\right)\\ &= -2\cdot\left(\sum_{i=1}^{n}z_i\cdot x_i-\hat{\beta}_0\cdot n\cdot\bar{x}-\hat{\beta}_1\cdot\sum_{i=1}^{n}x_i^2-\hat{\beta}_2\cdot\sum_{i=1}^{n}y_i\cdot x_i\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\hat{\beta}_2}QS(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2) &= -2\cdot\sum_{i=1}^{n}\left(z_i\cdot y_i-\hat{\beta}_0\cdot y_i-\hat{\beta}_1\cdot x_i\cdot y_i-\hat{\beta}_2\cdot y_i\cdot y_i\right)\\ &= -2\cdot\left(\sum_{i=1}^{n}z_i\cdot y_i-\hat{\beta}_0\cdot n\cdot\bar{y}-\hat{\beta}_1\cdot\sum_{i=1}^{n}x_i\cdot y_i-\hat{\beta}_2\cdot\sum_{i=1}^{n}y_i^2\right)\end{aligned}$$

Wir setzen die 1. Gleichung gleich Null und stellen nach $\hat{\beta}_0$ um:

$$\hat{\beta}_0=\bar{z}-\hat{\beta}_1\cdot\bar{x}-\hat{\beta}_2\cdot\bar{y}$$

Nun ersetzen wir $\hat{\beta}_0$ in den verbleibenden Gleichungen durch $z_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_2\cdot y_i$ und nutzen den Verschiebesatz:

$$\begin{aligned}\frac{\partial}{\partial\hat{\beta}_1}QS &= -2\cdot\left(\sum_{i=1}^{n}z_i\cdot x_i-(\bar{z}-\hat{\beta}_1\cdot\bar{x}-\hat{\beta}_2\cdot\bar{y})\cdot n\cdot\bar{x}-\hat{\beta}_1\cdot\sum_{i=1}^{n}x_i^2-\hat{\beta}_2\cdot\sum_{i=1}^{n}y_i\cdot x_i\right)\\ &= -2\cdot\left(\sum_{i=1}^{n}(z_i-\bar{z})(x_i-\bar{x})-\hat{\beta}_1\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2-\hat{\beta}_2\cdot\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\hat{\beta}_2}QS &= -2\cdot\left(\sum_{i=1}^{n}z_i\cdot y_i-(\bar{z}-\hat{\beta}_1\cdot\bar{x}-\hat{\beta}_2\cdot\bar{y})\cdot n\cdot\bar{y}-\hat{\beta}_1\cdot\sum_{i=1}^{n}x_i\cdot y_i-\hat{\beta}_2\cdot\sum_{i=1}^{n}y_i^2\right)\\ &= -2\cdot\left(\sum_{i=1}^{n}(z_i-\bar{z})(y_i-\bar{y})-\hat{\beta}_2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{\beta}_1\cdot\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\right)\end{aligned}$$

Wir setzen die beiden Gleichungen nun gleich Null und formen nach $\hat{\beta}_1$ und $\hat{\beta}_2$ um:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x}) - \hat{\beta}_2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y}) - \hat{\beta}_1 \cdot \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Durch Erweiterung von Zähler nun Nenner mit $\frac{1}{n-1}$ erhalten wir:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\frac{1}{n-1} \cdot \sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x}) - \hat{\beta}_2 \cdot \frac{1}{n-1} \cdot \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{s_{x,z} - \hat{\beta}_2 \cdot s_{x,y}}{s_x^2} = \frac{s_{x,z}}{s_x^2} - \hat{\beta}_2 \frac{s_{x,y}}{s_x^2}\end{aligned}$$

$$\begin{aligned}\hat{\beta}_2 &= \frac{\frac{1}{n-1} \cdot \sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y}) - \hat{\beta}_1 \cdot \frac{1}{n-1} \cdot \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{s_{y,z} - \hat{\beta}_1 \cdot s_{x,y}}{s_y^2} = \frac{s_{y,z}}{s_y^2} - \hat{\beta}_1 \frac{s_{x,y}}{s_y^2}\end{aligned}$$

wir setzen nun die erste in die zweite Gleichung ein und erhalten:

$$\begin{aligned}\hat{\beta}_2 &= \frac{s_{y,z}}{s_y^2} - \left(\frac{s_{x,z}}{s_x^2} - \hat{\beta}_2 \frac{s_{x,y}}{s_x^2}\right)\frac{s_{x,y}}{s_y^2} \\ &= \frac{s_{y,z}}{s_y^2} - \frac{s_{x,z}}{s_x^2}\frac{s_{x,y}}{s_y^2} + \hat{\beta}_2 \frac{s_{x,y}}{s_x^2}\frac{s_{x,y}}{s_y^2} \\ &= \frac{\frac{s_{y,z}}{s_y^2} - \frac{s_{x,z}}{s_x^2}\frac{s_{x,y}}{s_y^2}}{1 - \frac{s_{x,y}}{s_x^2}\frac{s_{x,y}}{s_y^2}} \\ &= \frac{\frac{s_{y,z} \cdot s_x^2 - s_{x,z}s_{x,y}}{s_x^2 \cdot s_y^2}}{\frac{s_x^2 s_y^2 - (s_{x,y})^2}{s_x^2 \cdot s_y^2}} = \frac{s_{y,z} \cdot s_x^2 - s_{x,z}s_{x,y}}{s_x^2 s_y^2 - (s_{x,y})^2}\end{aligned}$$

Und damit weiter:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{s_{x,z}}{s_x^2} - \hat{\beta}_2 \frac{s_{x,y}}{s_x^2} \\
&= \frac{s_{x,z}}{s_x^2} - \frac{s_{y,z} \cdot s_x^2 - s_{x,z} s_{x,y}}{s_x^2 s_y^2 - (s_{x,y})^2} \frac{s_{x,y}}{s_x^2} \\
&= \frac{s_{x,z}(s_x^2 s_y^2 - (s_{x,y})^2) - s_{y,z} s_{x,y} s_x^2 + s_{x,z} s_{x,y} s_{x,y}}{s_x^2 (s_x^2 s_y^2 - (s_{x,y})^2)} \\
&= \frac{s_{x,z} s_x^2 s_y^2 - s_{x,z}(s_{x,y})^2 - s_{y,z} s_{x,y} s_x^2 + s_{x,z}(s_{x,y})^2}{s_x^2 s_x^2 s_y^2 - s_x^2 (s_{x,y})^2}
\end{aligned}
$$

```
library(mosaic)

mtcars %>%
  select(mpg, hp, wt) -> dt

# Von R berechnete Koeffizienten:
coef(lm(mpg ~ hp + wt, data = dt))
#> (Intercept)          hp          wt
#> 37.22727012 -0.03177295 -3.87783074

mean_x = mean( ~ hp, data = dt)
mean_y = mean( ~ wt, data = dt)
mean_z = mean( ~ mpg, data = dt)

s_xy <- cov(hp ~ wt, data = dt)
s_xz <- cov(hp ~ mpg, data = dt)
s_yz <- cov(wt ~ mpg, data = dt)

var_x <- var(~ hp, data = dt)
var_y <- var(~ wt, data = dt)

b1 <- (s_xz*var_x*var_y - s_xz*(s_xy)**2 - s_yz*s_xy*var_x + s_xz*s_xy**2) / (var_x*var_
b2 <- (s_yz*var_x - s_xz*s_xy) / (var_x * var_y- s_xy*s_xy)
b0 <- mean_z - b1 * mean_x - b2 * mean_y

# Koeffizienten zur Ausgabe aufbereiten:
my_coef <- c(b0, b1, b2)
names(my_coef) <- c("(Intercept)", "hp", "wt")

# Von Hand berechnete Koeffizienten:
my_coef
#> (Intercept)          hp          wt
#> 37.22727012 -0.03177295 -3.87783074
```

Was passiert, wenn wir alle Datenpunkte studentisieren?

Wir rechnen um in:

$$x_i^{\text{stud}} = \frac{x_i - \bar{x}}{s_x}; \quad y_i^{\text{stud}} = \frac{y_i - \bar{y}}{s_y}; \quad z_i^{\text{stud}} = \frac{z_i - \bar{z}}{s_z}$$

Damit ist

$$\bar{x_i}^{\text{stud}} = 0; \quad \bar{y_i}^{\text{stud}} = 0; \quad \bar{z_i}^{\text{stud}} = 0$$

und

$$s_{x_i^{\text{stud}}} = 1; \quad s_{y_i^{\text{stud}}} = 1; \quad s_{z_i^{\text{stud}}} = 1$$

Zur Vereinfachung lassen wir die Kennzeichnung "stud" weg. Damit ist dann:

$$\hat{\beta}_0 = 0$$

$$\begin{aligned}
\hat{\beta}_1 &= \frac{s_{x,z} \cdot s_x^2 \cdot s_y^2 - s_{x,z} \cdot (s_{x,y})^2 - s_{y,z} \cdot s_{x,y} s_x^2 + s_{x,z} \cdot (s_{x,y})^2}{s_x^2 \cdot s_x^2 s_y^2 - s_x^2 \cdot (s_{x,y})^2} \\
&= \frac{s_{x,z} \cdot 1 \cdot 1 - s_{x,z} \cdot (s_{x,y})^2 - s_{y,z} \cdot s_{x,y} \cdot 1 + s_{x,z} \cdot (s_{x,y})^2}{1 \cdot 1 \cdot 1 - 1 \cdot (s_{x,y})^2} \\
&= \frac{s_{x,z} - s_{x,z} \cdot (s_{x,y})^2 - s_{y,z} \cdot s_{x,y} + s_{x,z} \cdot (s_{x,y})^2}{1 - (s_{x,y})^2} \\
&= \frac{s_{x,z} - s_{y,z} \cdot s_{x,y}}{1 - (s_{x,y})^2}
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_2 &= \frac{s_{y,z} \cdot s_x^2 - s_{x,z} \cdot s_{x,y}}{s_x^2 \cdot s_y^2 - (s_{x,y})^2} \\
&= \frac{s_{y,z} \cdot 1 - s_{x,z} \cdot s_{x,y}}{1 \cdot 1 - (s_{x,y})^2} \\
&= \frac{s_{y,z} - s_{x,z} \cdot s_{x,y}}{1 - (s_{x,y})^2}
\end{aligned}$$

Wir schauen uns ein paar Fälle genauer an:

1. Fall: X **und** Y **sind unabhängig.** Dann ist $s_{x,y} = 0$ und wir erhalten $\hat{\beta}_1 = s_{x,z} \in [-1; 1]$ und $\hat{\beta}_2 = s_{y,z} \in [-1; 1]$.
2. Fall: X **und** Y **sind abhängig.** Dann ist $|s_{x,y}| = 1$ und es gibt keine Lösung für $\hat{\beta}_1$ und $\hat{\beta}_2$.
3. Fall: $0 < |s_{x,y}| < 1$**.** ...