

Multiple Regression

Eine multiple Regression verwendet man, wenn mehr als eine unabhängige Größe x vorhanden ist. Das einfache lineare Modell lautet:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots$$

Es wird vorausgesetzt, dass die Merkmale in der Grundgesamtheit normalverteilt sind und linear eingehen. Nicht lineare Größen lassen sich in den meisten Fällen durch Umformung oder durch Hinzunahme eines quadratischen Terms realisieren z.B.:

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_2 + \dots$$

Bei Vorliegen von Tabellenwerten heißt das, dass man die Spalte mit den Werten x kopiert, und quadriert in eine neue Spalte hinzufügt. Es kann auch eine Verknüpfung zweier Einflüsse erfolgen, die eine Wechselwirkung darstellt, z.B.:

$$y = b_0 + b_1 x_1 + b_2 x_1 x_2 + b_3 x_2 + \dots$$

Die entsprechenden Tabellenspalten für x und x' sind dann in einer neuen Spalte als Produkt einzufügen. Weitere Umrechnungen sind möglich, um auf das lineare Modell zu gelangen. Für den Umfang an Modelltermen sind eine entsprechende Anzahl von Datenreihen erforderlich. In Matrizenform schreibt man die Modellgleichung:

$$\hat{y} = b X$$

mit \hat{y} = Vektor der entsprechenden Modell-Ergebnisse für die jeweiligen Faktoreinstellungen, X = Matrix der aktuellen Einstellungen und b = Vektor der Koeffizienten

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{z1} \\ 1 & x_{12} & \dots & x_{z2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{zn} \end{bmatrix} \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_z \end{bmatrix}$$

Hinweise: 1. Spalte in X steht für konstanten Anteil

Der gesuchte Vektor b mit den Koeffizienten bestimmt über die Matrizen-Operation

$$b = (X^T X)^{-1} X^T y$$

Beispiel: Es liegt ein Modell mit einer Wechselwirkung vor:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$$

Die einzelnen Schritte der Gleichung

$b = (X^T X)^{-1} X^T y$ ergeben sich wie folgt:

Versuchsplan:	Ergebnisse Y		
V_1	-1	-1	3
V_2	1	-1	5
V_3	-1	1	7
V_4	1	1	11
V_5	0	0	6

$$X' = X^T X \quad \text{mit} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{z1} \\ 1 & x_{12} & \dots & x_{z2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{zn} \end{bmatrix} \quad z+1 \text{ Spalten und } n \text{ Zeilen}$$

Die jeweiligen Zellen berechnen sich nacheinander entsprechend mit:

$$x'_{j,i} = \sum_{k=1}^n x_{k,i}^{(T)} x_{j,k} \quad (\text{erster Index} = \text{Spalte}, \text{zweiter Index} = \text{Zeilen})$$

Die erste Spalte mit jeweils einer 1 steht für die Konstante b_0 . Die beiden weiteren für die Hauptfaktoren x_1 und x_2 und die letzte Spalte errechnet sich aus dem Produkt der Spalte 2 und 3 (Wechselwirkung von x_1 und x_2).

$$X = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & 0 \\ -1 & -1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 1 & 0 \end{bmatrix}$$

z.B. Zelle

$$j=1 \quad i=1$$

$$x'_{1,1} = (1) \cdot (1) + (1) \cdot (1) + (1) \cdot (1) + (1) \cdot (1) + (1) \cdot (1) = 5$$

$$j=2 \quad i=2$$

$$x'_{2,2} = (-1) \cdot (-1) + (1) \cdot (1) + (-1) \cdot (-1) + (1) \cdot (1) + (0) \cdot (0) = 4$$

gesamthaft ergibt sich:

$$X' = X^T X = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

und invertiert:

$$(X^T X)^{-1} = \begin{bmatrix} 1/5 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}$$

und über den Zwischenschritt:

$$X^T y = \begin{bmatrix} 32 \\ 6 \\ 10 \\ 2 \end{bmatrix}$$

erhält man das Ergebnis für die gesuchten Koeffizienten:

$$b = (X^T X)^{-1} X^T y = \begin{bmatrix} 6,4 \\ 1,5 \\ 2,5 \\ 0,5 \end{bmatrix}$$

Die Gleichung vom Anfang lautet also:

$$y = 6,4 + 1,5 x_1 + 2,5 x_2 + 0,5 x_1 x_2$$

Kategoriale Faktoren

Der Aufbau der Matrix für kategoriale Faktoren ist im Kapitel D-Optimale Versuchspläne dargestellt. Hat der kategoriale Faktor die Merkmale (Einstellungen) A, B, C usw., so besitzt das Regressionsmodell keinen Term für A. A stellt die Grundstellung dar, von der aus variiert wird. Erst ab dem zweitem Merkmal B werden Terme erzeugt. Der nicht benötigte Koeffizient für A ergibt sich aus der Summe der negierten anderen Koeffizienten. Aufgrund der entsprechenden Spalten für einen kategorialen Faktor nehmen die zur Verfügung stehenden Freiheitsgrade schnell ab, besonders wenn Wechselwirkungen aller Einstellungen ermittelt werden müssen.