5 Lineare Regressionsanalyse

Bemerkung: In diesem Kapitel legen wir bei allen Tests ein Signifikanzniveau von 5% zu Grunde.

5.1 Modellannahmen

Mit Hilfe der Regressionsanalyse kann der Einfluss einer oder mehrerer unabhängiger Variablen $x_1,...,x_k$ auf eine abhängige Variable Y untersucht werden. Hierbei liegt allgemein das folgende Modell zu Grunde:

$$Y = f(x_1,...,x_k) + E.$$

Falls es nur eine unabhängige Variable gibt, spricht man von der einfachen linearen Regression und sonst von der multiplen linearen Regression. E ist hierbei eine Zufallsvariable, welche wir als normalverteilt ansehen mit der Varianz σ^2 und dem Erwartungswert 0. Im Fall der linearen Regressionsanalyse ist f eine lineare Funktion:

$$f(x_1,...,x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k.$$

Die Aufgabe der Regressionsanalyse ist es nun, die unbekannten Parameter $\beta_0, \beta_1, ..., \beta_k$ zu schätzen, Tests bezüglich dieser Parameter durchzuführen, sowie die Güte des gewählten Modells zu beurteilen.

Die Schätzung der Parameter kann über die Methode der kleinsten Quadrate durchgeführt werden. Dabei werden die Parameter so geschätzt, dass die Abweichungsquadrate minimal werden:

Gesucht wir das Minimum von

$$Q(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}))^2 \ .$$

Dazu müssen zunächst mindestens n = k+1 Beobachtungen (x_{i1}, x_{i2}, …, x_{ik}, y_i) gegeben sein.

Im Fall der einfachen linearen Regression (d.h. für k = 1) liegen nur Wertepaar (x_{i1}, y_i) bzw. (x_i, y_i) vor und hier ergibt sich das Minimum mit folgenden Schätzern (wenn $n \geq 2$ und falls mindestens zwei x-Werte verschieden sind, d.h. $x_i \neq x_j$ für $i \neq j$) für die beiden Parameter:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

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

In Matrix-Vektor-Schreibweise wird das Ganze etwas übersichtlicher. Das Modell stellt sich hier wie folgt dar:

$$\vec{Y} = X\vec{\beta} + \vec{E}$$

X ist die Designmatrix mit k+1 Spalten und n (= Anzahl der Beobachtungen) Zeilen. In der ersten Spalte stehen nur Einsen. In der zweiten Spalte stehen die Beobachtungen der ersten unabhängigen Variablen, in der dritten die der zweiten unabhängigen Variablen, bis zur (k+1)-ten Spalte, wo die Beobachtungen der k-ten unabhängigen Variablen stehen. In unserem Modell gehen wir davon aus, dass X nicht stochastisch ist. $\vec{\beta}$ ist der unbekannte Parametervektor mit k+1 Komponenten und $\vec{E}$ der normalverteilte Zufallsvektor mit dem Erwartungswert $\vec{0}$ und der Varianz-Kovarianzmatrix σ^2 I (I ist die Einheitsmatrix mit n Zeilen und Spalten). Somit werden die einzelnen

Komponenten des Zufallsvektors $\vec{E}$ als paarweise unkorreliert vorausgesetzt.

Die zu minimierende Funktion Q hat dann die folgende Gestalt:

$$Q(\vec{\beta}) = \left(\vec{y} - X\vec{\beta}\right)^T \cdot \left(\vec{y} - X\vec{\beta}\right)$$

Der Schätzvektor ergibt sich dann (falls die Spalten von X linear unabhängig sind) durch:

$$\hat{\vec{\beta}} = \left(X^T X\right)^{-1} X^T \vec{y}$$

Wir führen nun zwei Beispiele vor, wobei wir am ersten Beispiel den Output allgemein beschreiben. Beiden Beispiele beziehen sich auf den unten zu sehenden Datensatz. Dabei verwenden wir v1 für die abhängige Variable in unserem Regressionsmodell. Im ersten Beispiel verwenden wir nur eine unabhängige Variable (v2) und im zweiten Beispiel zwei unabhängige Variablen (v2 und v3). Im Datensatz stehen unter v1 Körpergewichte in kg, unter v2 Körpergrößen in cm und unter v3 Altersangaben in Jahren.

Die Daten:

v1	**v2**	**v3**
54	170	20
67	170	21
60	167	22
63	177	23
75	182	23
63	167	24
56	164	25
65	170	25
61	169	26
80	176	27