

5.2 Erstes Beispiel und Beschreibung des Outputs

5.2.1 Schätzung des Parametervektors und Varianzzerlegung

Wir wollen nun mit Hilfe der linearen Regressionsanalyse den Einfluss der Körpergröße (v2) auf das Gewicht (v1) untersuchen. In diesem Beispiel handelt es sich somit um eine einfache lineare Regression, der Output wird dann wie immer allgemein beschrieben.

Den folgenden Output erhalten Sie, wenn Sie unter Lineare Regression für x1 die Variable v2 und für y die Variable v1 auswählen und danach **→(Lineare) Regression →Lineare Regression** wählen.

Lineare Regression

Hier sind die eingegebenen Daten zu sehen:

Werte der unabhängigen Variablen	Werte der abhängigen Variable (y _i)
170	54
170	67
167	60
177	63
182	75
167	63
164	56
170	65
169	61
176	80

Die Regressionsfunktion: $y = b_1 \cdot x_1 + b_0$

Parameter	Schätzer	geschätzte Standardabweichung	Prüfgröße ('t-Value') für Test (H0: Parameter = 0; gegen H1: Parameter \neq 0)	p-Wert
b_0	-114.80118694363	61.824445611106	-1.8568898727497	0.1004
b_1	1.0467359050445	0.36095812446466	2.8998818258958	0.0199

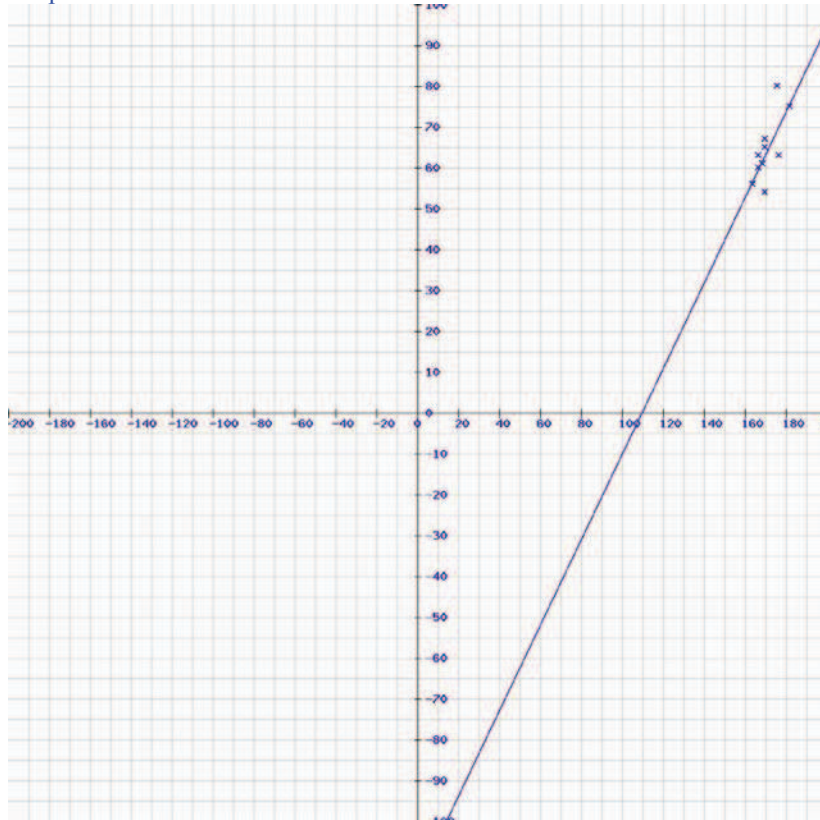
Stichprobenumfang n	10
geschätzte Fehlervarianz	35.126390949555
Bestimmtheitsmaß r^2	0.51247202013109

ANOVA-Tabelle

	Freiheitsgrade	Quadratsummen	mittlere Quadratsummen	Prüfgröße ('F-Value')	p-Wert
Modell (Model)	1	295.38887240356	295.38887240356	8.4093146041611	0.0199
Fehler (Error)	8	281.01112759644	35.126390949555		
Gesamt (Total)	9	576.4			

Obige Prüfgröße ist für den Test $H_0: b_i = 0$ mit $i \geq 1$; gegen H_1 : Es existiert mindestens ein $i \geq 1$ mit $b_i \neq 0$.

Graph



Es gilt im Beispiel:

$$\bar{y} = \begin{pmatrix} 54 \\ 67 \\ 60 \\ 63 \\ 75 \\ 63 \\ 56 \\ 65 \\ 61 \\ 80 \end{pmatrix} \quad \text{und} \quad X = \begin{pmatrix} 1 & 170 \\ 1 & 170 \\ 1 & 167 \\ 1 & 177 \\ 1 & 182 \\ 1 & 167 \\ 1 & 164 \\ 1 & 170 \\ 1 & 169 \\ 1 & 176 \end{pmatrix} .$$

Über die Methode der kleinsten Quadrate bestimmen wir den Schätzer $\hat{\beta}$ für den unbekannt Parametervektor β . Damit diese Schätzung möglich ist, wird die Designmatrix X als spaltenregulär vorausgesetzt, denn sonst existiert die Inverse von $X^T X$ nicht. Wäre dies nicht der Fall, wäre mindestens eine Spalte der Designmatrix von den anderen abhängig. Diese Spalte, beziehungsweise mindestens eine unabhängige Variable, könnte dann aus der Modellgleichung eliminiert werden. Im Beispiel gilt (siehe Output unter Schätzer):

$$\hat{\beta} = (X^T X)^{-1} X^T \bar{y} \approx \begin{pmatrix} -114,801 \\ 1,04674 \end{pmatrix}$$

Im nächsten Schritt betrachten wir die drei Quadratsummen zur Varianzzerlegung (ANOVA). Es gilt

$$SST = SSR + SSE,$$

mit

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i,$$

$$SSE = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}))^2 = Q(\hat{\beta}) \text{ und}$$

$$SSR = \sum_{i=1}^n (\bar{y} - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}))^2.$$

Wir bezeichnen hier die Quadratsummen mit großen Buchstaben, obwohl es sich um die Realisierungen und nicht um die Zufallsvariablen handelt. Dabei ist SSE (Sum of Squares due to Error) die Fehlerquadratsumme, SST (Sum of Squares Total) die Gesamtstreuung der Werte der abhängigen Variablen und SSR (Sum of Squares due to Regression) die Quadratsumme der Abweichungen

der Funktionswerte der geschätzten Regressionsfunktion vom arithmetischen Mittel der Werte der abhängigen Variablen. Mit dieser kann man prüfen, in wie weit die unabhängigen Variablen einen Einfluss auf die abhängige Variable haben.

Im Beispiel gilt:

$$SSE = 281,011\dots$$

$$SSR = 295,388\dots$$

$$SST = 576,4$$

Mit den Quadratsummen SSR und SST kann das Bestimmtheitsmaß r^2 berechnet werden. r^2 gibt uns den Anteil der Varianz an, die das gewählte Regressionsmodell im Verhältnis zur Gesamtvarianz erklärt. Das Bestimmtheitsmaß kann nur Werte zwischen 0 und 1 annehmen. Je größer das Bestimmtheitsmaß ist, umso besser ist die Anpassung des gewählten Regressionsmodells. Im Falle der einfachen linearen Regression ist das Bestimmtheitsmaß gleich dem Quadrat des empirischen Korrelationskoeffizienten zwischen der abhängigen und der unabhängigen Variable. Die Modellparameter sollten nur interpretiert werden, wenn das Bestimmtheitsmaß nicht zu klein ist, da sonst das gewählte Regressionsmodell nicht passend ist.

Es gilt:

$$r^2 = SSR/SST$$

Im Beispiel ist $r^2 = 0,51247\dots$