

Hintergrundmaterial zur Vorlesung Statistik 2

Kap. 15.2: Herleitung des Zerlegungssatzes der Gesamtstreuung bei linearer Regression

Die Gleichung für die Zerlegung bei linearer Regression lautet bekanntlich:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \quad (1)$$

Betrachten wir zunächst die Quadratsumme der Residualabweichung $\sum_{i=1}^n (y_i - \hat{y}_i)^2$:
Einsetzen des Ergebnisses für die Ausgleichsgerade,

$$\hat{y} = a + bx = \bar{y} + b(x - \bar{x}),$$

ergibt zunächst:

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)^2 &= \sum_{i=1}^n (y_i - \bar{y} - b(x_i - \bar{x}))^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2b \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

Setzt man nun auch für den Ausgleichskoeffizienten b das Regressionsergebnis $b = s_{xy}/s_x^2$ ein, ergibt sich für die Quadratsumme der Residualabweichung:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - n \frac{s_{xy}^2}{s_x^2}.$$

Für die Quadratsumme der erklärten Abweichung (zweiter Summand der rechten Seite in Gl. (1)) ergibt sich analog:

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b^2 \sum_{i=1}^n (x_i - \bar{x})^2 = n \frac{s_{xy}^2}{s_x^2}.$$

Damit heben sich die Kovarianzterme $\propto s_{xy}^2$ weg und man erhält

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

was zu beweisen war.

Zu 15.3 Herleitung des Zusammenhangs Rangkorrelation-Bestimmtheitsmaß im linearen Fall

Das Bestimmtheitsmaß der Regression $\hat{y}(x)$ eines Streudiagramms $\{(x_i, y_i), i = 1, \dots, n\}$ lautet bekanntlich

$$B = 1 - U = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Mit dem für lineare Regressionen gültigen Zerlegungssatz (1) erhält man zunächst

$$B = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (2)$$

Damit nun die Korrelationsfunktion ins Spiel kommt, muss man die Regressionsfunktion irgendwie durch r_{xy} ausdrücken:

$$\hat{y}(x) = a + bx = \bar{y} + b(x - \bar{x}) = \bar{y} + \frac{r_{xy}s_y}{s_x}(x - \bar{x}). \quad (3)$$

Dabei hat man

$$b = \frac{s_{xy}}{s_x^2} = \left(\frac{s_{xy}}{s_x s_y} \right) \frac{s_y}{s_x} = r_{xy} \frac{s_y}{s_x}$$

ausgenutzt.

Damit wird der Zähler von (2):

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \frac{s_y^2 r_{xy}^2 (x_i - \bar{x})^2}{s_x^2} = \frac{s_y^2 r_{xy}^2}{s_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 = n s_y^2 r_{xy}^2.$$

Der Nenner von (2) ist $\sum_{i=1}^n (y_i - \bar{y})^2 = n s_y^2$ und damit letztendlich

$$B = \frac{n s_y^2 r_{xy}^2}{n s_y^2} = r_{xy}^2,$$

was zu zeigen war.

Kap. 15.3(c): Herleitung des Rangkorrelationskoeffizienten nach Spearman aus dem Maßkorrelationskoeffizient nach Pearson

Ausgangspunkt: Maßkorrelationskoeffizient für einen Scatter-Plot der Werte x_i und y_i , $i = 1, \dots, n$, von kardinalskalierten Größen X und Y :

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Erzeugen einer ordinalskalierten Rangfolgenliste der x_i und y_i :

Sei $R_i \in \{1, 2, \dots, n\}$, $i = 1, 2, \dots, n$ die Rangfolge der Werte x_i und $S_i \in \{1, 2, \dots, n\}$, $i = 1, 2, \dots, n$ die Rangfolge der Werte y_i . Alle Werte müssen voneinander verschieden sein, so dass die Reihenfolge genau von 1 bis n geht! (bzw. allgemeiner von $(k+1)$ bis $(k+n)$ mit k einer beliebigen ganzen Zahl). Zum Beispiel gilt $R_j = 1$ für den Index j , für den der zugehörige Wert x_j am kleinsten ist und $R_k = n$ für den Index k , für den x_k am größten ist. Falls zwei oder mehr der x_i oder y_i gleich sind, wählt man die Rangfolge willkürlich.

Anwendung der Maßkorrelation auf die Rangfolgenlisten:

$$r_{xy} \Rightarrow r_{RS} = \frac{s_{RS}}{s_R s_S} \quad (4)$$

mit

$$s_R^2 = \frac{1}{n} \sum_{i=1}^n (R_i - \bar{R})^2, \quad (5)$$

$$s_S^2 = \frac{1}{n} \sum_{i=1}^n (S_i - \bar{S})^2, \quad (6)$$

$$s_{RS} = \frac{1}{n} \sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S}). \quad (7)$$

Ausnutzen der speziellen Struktur der R_i und S_i :

$$\bar{R} = \bar{S} = \frac{n+1}{2}, \quad (8)$$

$$\sum_{i=1}^n (R_i - \bar{R})^2 = \sum_{i=1}^n (S_i - \bar{S})^2 = \sum_{i=1}^n (i^2 - 2\bar{R}i + \bar{R}^2) = \sum_{i=1}^n i^2 - n \left(\frac{n+1}{2} \right)^2. \quad (9)$$

Mit

$$\sum_{i=1}^n i^2 = \frac{1}{6}(n + 3n^2 + 2n^3)$$

ergibt sich

$$s_R^2 = s_S^2 = \frac{(n^2 - 1)}{12}. \quad (10)$$

Wir benötigen nun noch die Summe $\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})$:

$$\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S}) = \sum_{i=1}^n (R_i S_i - R_i \bar{S} - \bar{R} S_i + \bar{R} \bar{S}) + \frac{1}{2} \sum_{i=1}^n (R_i - S_i)^2 - \frac{1}{2} \sum_{i=1}^n (R_i + S_i)^2 \quad (11)$$

Die letzten beiden Summanden wurden *pro Forma* hinzuaddiert; sie ergeben natürlich zusammen Null. Fasst man die beiden ersten Summanden zusammen und beachtet Gln. (8) und $\sum_{i=1}^n R_i = \sum_{i=1}^n S_i$, erhält man

$$\begin{aligned} \sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S}) &= \sum_{i=1}^n (-2R_i\bar{R} + \bar{R}^2 + R_i^2) - \frac{1}{2} \sum_{i=1}^n (R_i - S_i)^2 \\ &= ns_R^2 - \frac{1}{2} \sum_{i=1}^n (R_i - S_i)^2. \end{aligned}$$

Setzt man dies in Gl. (4) ein und beachtet man Gl. (7), $s_{RS} = s_R^2$ und Gl. (10), erhält man den Rangkorrelationskoeffizienten nach Spearman:

$$r_{RS} = \frac{n \frac{n^2-1}{12} - \frac{1}{2} \sum_{i=1}^n (R_i - S_i)^2}{n \frac{n^2-1}{12}} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n (R_i - S_i)^2. \quad (12)$$

Summa summarum: Der Rangkorrelationskoeffizienten nach Spearman ist der Maßkorrelationskoeffizient nach Pearson, angewandt auf die natürlichzahligen Rangfolgelisten R_i und S_i von ordinal- oder kardinalskalierten Werten x_i und y_i .