Lineare Regression und der Korrelationskoeffizient

Arno Fehringer, Juni 2018

Quellen:

Athen, Hermann : Wahrscheinlichkeitsrechnung und Statistik. Verlag Schroedel Schöningh, 2. Aufl. 1968

Heynkes, Roland :
http://www.heynkes.de/anna/index.html
http://www.heynkes.de/anna/Statistik2.pdf [11.06.2018]

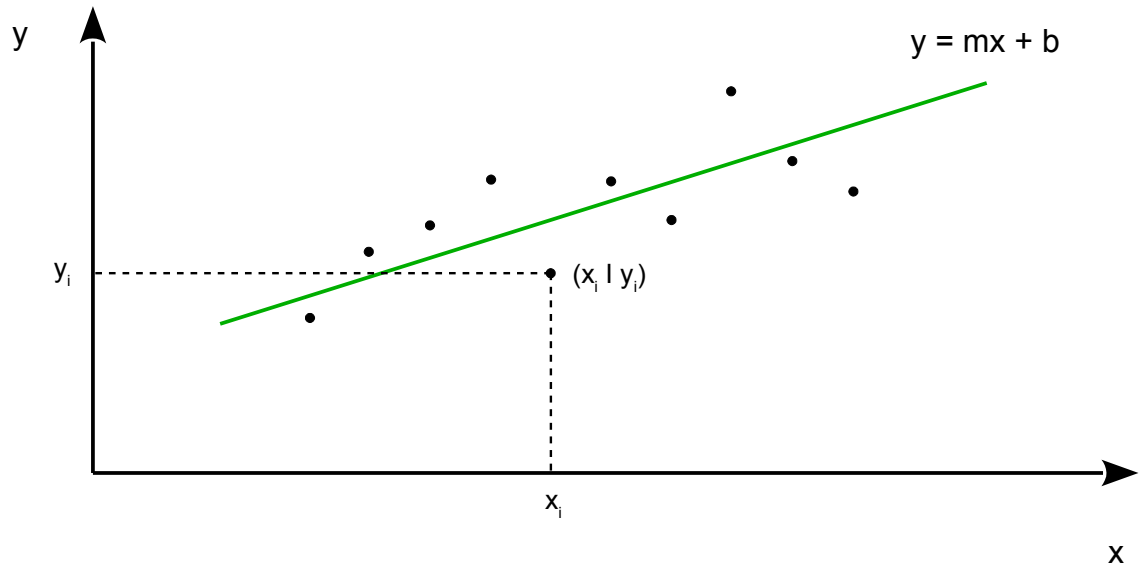
Ringel, C. M. (Fakultät für Mathematik, Universität Bielefeld) :
https://www.math.uni-bielefeld.de/~sek/funktion/
https://www.math.uni-bielefeld.de/~sek/funktion/leit03.pdf
https://www.math.uni-bielefeld.de/~sek/funktion/leit03-2.pdf [11.06.2018]

Madincea, Arne (Herder-Oberschule in Berlin-Charlottenburg) :
http://www.madincea.privat.t-online.de/aufg11-P.htm
https://www.herder-oberschule.de/madincea/aufg0011/induktio.pdf [11.06.2018]

Gegeben seien zwei Zufallsvariable x und y , denen man einen funktionalen Zusammenhang unterstellt.

Die Variable x könnte zum Beispiel die Größe und die Variable y das Körpergewicht eines Erwachsenen darstellen.

Aus Erfahrung weiß man, dass zu einer bestimmten Größe nicht ein genau bestimmtes Gewicht gehört, sondern dass es innerhalb eines gewissen Bereiches streut.
Eine entsprechende Erhebung vom Umfang n würde in der graphischen Darstellung eine „Punktwolke“ liefern .



Die Aufgabe besteht darin, eine möglichst gut passende Gerade $y = mx + b$, die sogenannte **Regressionsgerade bezüglich der Variablen** y , zu finden .

Als Bedingung soll die Summe der Abweichungsquadrate $q(m,b)$ der Messwerte y_i vom theoretischen Wert y minimal sein :

$$q(m,b) = \sum_{i=1}^{n} (y_i - y)^2 = \sum_{i=1}^{n} (y_i - mx_i - b)^2 \quad \text{minimal}$$

Eine notwendige Bedingung ist, dass die Parteiellen Ableitungen gleich Null sind :

$$\frac{\partial}{\partial m} q(m,b) = 0 \qquad \frac{\partial}{\partial b} q(m,b) = 0$$

$$\sum_{i=1}^{n} 2(y_i - mx_i - b)(-x_i) = 0 \qquad \sum_{i=1}^{n} 2(y_i - mx_i - b)(-1_i) = 0$$

$$-2\left(\sum_{i=1}^{n} x_i y_i - m\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i\right) = 0 \qquad -2\left(\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i - nb\right) = 0$$

Jetzt formuliert man (vorübergehend) folgenden Mittelwerte

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \quad , \quad \overline{x^2} := \frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad , \quad \overline{xy} := \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

mit den entsprechenden Umformungen

$$\sum_{i=1}^{n} x_i = n\bar{x} \quad , \quad \sum_{i=1}^{n} y_i = n\bar{y} \quad , \quad \sum_{i=1}^{n} x_i^2 = n\overline{x^2} \quad , \quad \sum_{i=1}^{n} x_i y_i = n\overline{xy} \quad ,$$

so dass folgt :

$$-2\left(\sum_{i=1}^{n} x_i y_i - m\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i\right) = 0 \qquad -2\left(\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i - nb\right) = 0$$

$$-2\left(n\overline{xy} - mn\overline{x^2} - bn\bar{x}\right) = 0 \qquad -2\left(n\bar{y} - mn\bar{x} - nb\right) = 0$$

$$-2n\left(\overline{xy} - m\overline{x^2} - b\bar{x}\right) = 0 \qquad -2n\left(\bar{y} - m\bar{x} - b\right) = 0$$

$$\overline{xy} - m\overline{x^2} - b\bar{x} = 0 \qquad \bar{y} - m\bar{x} - b = 0$$

$$b = \bar{y} - m\bar{x}$$

$$\overline{xy} - m\overline{x^2} - (\bar{y} - m\bar{x})\bar{x} = 0$$

$$\overline{xy} - m\overline{x^2} - \bar{x}\bar{y} + m\bar{x}^2 = 0$$

$$\overline{xy} - \bar{x}\bar{y} = m\overline{x^2} - m\bar{x}^2$$

$$\overline{xy} - \bar{x}\bar{y} = m\left(\overline{x^2} - \bar{x}^2\right)$$

$$\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = m$$

Jetzt hat man die gesuchten Parameter zur erforderlichen Geraden

$$\boxed{m = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}} \qquad \boxed{b = \bar{y} - \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}\,\bar{x}}$$

Der Nenner von m , also $\overline{x^2} - \bar{x}^2$, ist gerade die Varianz v_x der Variablen x , denn :

$$v_x = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

$$v_x = \frac{1}{n}\sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

$$v_x = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \overline{x}^2\right)$$
$$v_x = \frac{1}{n}\left(n\overline{x^2} - 2\overline{x}n\overline{x} + n\overline{x}^2\right)$$
$$v_x = \frac{1}{n}\left(n\overline{x^2} - n\overline{x}^2\right)$$
$$v_x = \overline{x^2} - \overline{x}^2$$

Analog kann man zeigen, dass der Zähler $\overline{xy} - \overline{x}\,\overline{y}$ von m gleich dem Ausdruck $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$ ist, denn :

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{x}\,\overline{y}\right)$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + n\overline{x}\,\overline{y}\right)$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \frac{1}{n}\left(n\overline{xy} - \overline{y}n\overline{x} - \overline{x}n\overline{y} + n\overline{x}\,\overline{y}\right)$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \frac{1}{n}\left(n\overline{xy} - n\overline{x}\,\overline{y}\right)$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \overline{xy} - \overline{x}\,\overline{y}$$

Die Parameter für die **Regressionsgerade bezüglich der Variablen** y kann man also auch wie folgt schreiben :

$$m = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

$$b = \overline{y} - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\,\overline{x}$$

Man könnte jetzt auch für den umgekehrten funktionalen Zusammenhang $x = m'y + b'$ die **Regressionsgerade bezüglich der Variablen** x die entsprechenden Gleichungen aufstellen :

$$m' = \frac{\overline{yx} - \overline{y}\,\overline{x}}{\overline{y^2} - \overline{y}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

$$b' = \overline{x} - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}\,\overline{y}$$

<u>Bemerkung</u> :

Beide Regressionsgeraden gehen durch den Punkt $\left(\overline{x} \mid \overline{y}\right)$!

Falls nun alle Punkte der Punktwolke auf einer Geraden lägen, falls also der lineare Zusammenhang zwischen den Variablen x und y vollständig gegeben wäre, stimmten die beiden Regressionsgeraden überein, und es gälte

$$m = \frac{1}{m'}$$

$$m\,m' = 1$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1$$

$$\frac{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1$$

$$\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)^2 = 1$$

Der Wert 1 wird genau dann erreicht, wenn der Ausdruck

$$r := \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \pm 1 \quad \text{ist}$$

Definition :

Der Ausdruck $r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ heißt **Korrelationskoeffizient** .

Die beiden Wurzelterme im Nenner sind die **Standardabweichungen** der Variablen x und y ,

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad , \quad \sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad .$$

Schreibt man für den Zähler $\sigma_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, so erhält man folgende Schreibweise für den Korrelationskoeffizienten :

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Jetzt kann man mit der **Cauchy-Schwarzschen Ungleichung (CSU)** zeigen, dass der Korrelationskoeffizient nur Werte zwischen -1 und +1 annehmen kann.

(CSU) Für alle $\vec{x}\ ,\ \vec{y}\ \in\ \mathbb{R}^n$ gilt : $\boxed{\left|\sum_{i=1}^{n} x_i y_i\right| \le \sqrt{\sum_{i=1}^{n} x_i^2}\cdot\sqrt{\sum_{i=1}^{n} y_i^2}}$

$$|r| = \left|\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right|$$

$$|r| = \left|\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right|$$

$$|r| = \frac{\left|\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right|}{\left|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right|\cdot\left|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right|}$$

$$|r| \le \frac{\left|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right|\cdot\left|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right|}{\left|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right|\cdot\left|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right|}$$

$$|r| \le 1$$

Für Werte $|r| < 1$ ist der funktionale Zusammenhang schwächer als für $|r| = 1$. Die Werte streuen, anstatt auf einer Geraden zu liegen.

Übereinkunft :

$|r| > 0{,}8$: starke Korrelation

$|r| < 0{,}5$: schwache Korrelation

Für $r = 0$ gilt $mm' = r^2 = 0$, also $m = 0$, $m' = 0$

<u>Bemerkung</u> :

Aus einer starken Korrelation kann man nicht auf einen funktionalen Zusammenhang schließen !

Paradebeispiel: Die Zunahme der Störche und die Steigung der Geburtenzahl sind nicht funktional gegeben .

Die **Cauchy-Schwarzsche Ungleichung (CSU) :**

(CSU) Für alle $\vec{x}\ ,\ \vec{y}\ \in\ \mathbb{R}^n$ gilt :

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \le \left(\sum_{i=1}^{n} x_i^2\right)\cdot\left(\sum_{i=1}^{n} y_i^2\right)$$

$$\Leftrightarrow \quad \left|\sum_{i=1}^{n} x_i y_i\right| \le \sqrt{\sum_{i=1}^{n} x_i^2}\cdot\sqrt{\sum_{i=1}^{n} y_i^2}$$

Beweis :

n=1: $\left(\sum_{i=1}^{1} x_i y_i\right)^2 \le \left(\sum_{i=1}^{1} x_i^2\right)\cdot\left(\sum_{i=1}^{1} y_i^2\right)$

$$\left(x_1 y_1\right)^2 \le x_1^2\cdot y_1^2$$

$$-x_1^2\cdot y_1^2 \le -\left(x_1 y_1\right)^2 \le \left(x_1 y_1\right)^2 \le x_1^2\cdot y_1^2$$

$$\left|x_1 y_1\right| \le \sqrt{x_1^2}\cdot\sqrt{y_1^2} \qquad \text{[wahr]}$$

IV Sei die Aussage für $k\ \in\ \mathbb{N}$ wahr: $\left(\sum_{i=1}^{k} x_i y_i\right)^2 \le \left(\sum_{i=1}^{k} x_i^2\right)\cdot\left(\sum_{i=1}^{k} y_i^2\right)$

$$\Leftrightarrow \quad \left|\sum_{i=1}^{k} x_i y_i\right| \le \sqrt{\sum_{i=1}^{k} x_i^2}\cdot\sqrt{\sum_{i=1}^{k} y_i^2}$$

IS Man muss zeigen, dass dann die Aussage auch für $k+1\ \in\ \mathbb{N}$ wahr ist :

$$\left(\sum_{i=1}^{k+1} x_i y_i\right)^2 = \left(\sum_{i=1}^{k} x_i y_i + x_{k+1} y_{k+1}\right)^2$$

$$\left(\sum_{i=1}^{k+1} x_i y_i\right)^2 = \left(\sum_{i=1}^{k} x_i y_i\right)^2 + 2\left(\sum_{i=1}^{k} x_i y_i\right)\left(x_{k+1} y_{k+1}\right) + x_{k+1}^2 y_{k+1}^2$$

$$\left(\sum_{i=1}^{k+1} x_i y_i\right)^2 \le \left(\sum_{i=1}^{k} x_i^2\right)\left(\sum_{i=1}^{k} y_i^2\right) + 2\left(\sum_{i=1}^{k} x_i y_i\right)\left(x_{k+1} y_{k+1}\right) + x_{k+1}^2 y_{k+1}^2$$

Wegen $\sum_{i=1}^{k}\left(x_{k+1} y_i - y_{k+1} x_i\right)^2 \ge 0$

$$\sum_{i=1}^{k}\left(x_{k+1}^2 y_i^2 - 2x_{k+1} y_i y_{k+1} x_i + y_{k+1}^2 x_i^2\right) \ge 0$$

$$\sum_{i=1}^{k} x_{k+1}^2 y_i^2 - \sum_{i=1}^{k} 2x_{k+1} y_i y_{k+1} x_i + \sum_{i=1}^{k} y_{k+1}^2 x_i^2 \ge 0$$

$$x_{k+1}^2\sum_{i=1}^{k} y_i^2 - 2x_{k+1} y_{k+1}\sum_{i=1}^{k} y_i x_i + y_{k+1}^2\sum_{i=1}^{k} x_i^2 \ge 0$$

$$x_{k+1}^2\sum_{i=1}^{k} y_i^2 + y_{k+1}^2\sum_{i=1}^{k} x_i^2 \ge 2x_{k+1} y_{k+1}\sum_{i=1}^{k} y_i x_i$$

$$2x_{k+1} y_{k+1}\sum_{i=1}^{k} y_i x_i \le x_{k+1}^2\sum_{i=1}^{k} y_i^2 + y_{k+1}^2\sum_{i=1}^{k} x_i^2$$

folgt $$\left(\sum_{i=1}^{k+1} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{k} x_i^2 \right)\left(\sum_{i=1}^{k} y_i^2 \right) + x_{k+1}^2 \sum_{i=1}^{k} y_i^2 + y_{k+1}^2 \sum_{i=1}^{k} x_i^2 + x_{k+1}^2 y_{k+1}^2$$

$$\left(\sum_{i=1}^{k+1} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{k} x_i^2 \right)\left(\sum_{i=1}^{k} y_i^2 + y_{k+1}^2 \right) + x_{k+1}^2 \left(\sum_{i=1}^{k} y_i^2 + y_{k+1}^2 \right)$$

$$\left(\sum_{i=1}^{k+1} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{k} x_i^2 + x_{k+1}^2 \right)\left(\sum_{i=1}^{k} y_i^2 + y_{k+1}^2 \right)$$

$$\left(\sum_{i=1}^{k+1} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{k+1} x_i^2 \right)\left(\sum_{i=1}^{k+1} y_i^2 \right)$$

$\Leftrightarrow$ $$\left| \sum_{i=1}^{k+1} x_i y_i \right| \leq \sqrt{\sum_{i=1}^{k+1} x_i^2} \cdot \sqrt{\sum_{i=1}^{k+1} y_i^2}$$