Korrelation

Ac 2017-2020

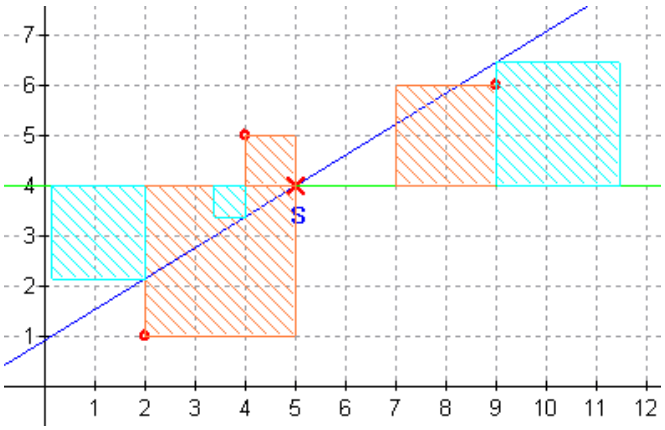
Korrelation ist ein Maß dafür, wie stark die Merkmalswerte (Daten) mit dem gewählten Regressionsmodell (z.B. Regressionsgerade oder Regressionspolynom) zusammenhängen.

Man definiert den **Betrag der Korrelation r** als Verhältnis (Anteil) der Standardabweichung, die durch das Regressionsmodell festgelegt ist zur Standardabweichung, die durch die Daten gegeben ist :

Kurzschreibweise: $$|r| = \frac{\sigma_{Modell}}{\sigma_{Daten}}$$

Was ist damit gemeint ?

Die folgende Grafik soll das erläutern (3 Punkte P1(2/1) , P2(4/5) , P3(9/6) seien gegeben):



Hier ist als Modell eine Gerade angenommen worden. Ihre Gleichung ist

$$y = \frac{8}{13}x + \frac{12}{13}$$

Für die beiden Standardabweichungen erhält man (Begründung siehe weiter unten):

$$\sigma_{Daten} = \sqrt{\frac{14}{3}} \qquad \sigma_{Modell} = \sqrt{\frac{128}{39}}$$

Das arithmetische Mittel der Daten heißt **Schwerpunkt S($\overline{x}$; $\overline{y}$)** , hier: S(5 ; 4) .

Es werden zunächst die roten Quadrate der vertikalen Abweichungen der Daten vom Schwerpunkt S betrachtet und ihre Flächeninhalte aufsummiert. $\sum_{i=1}^{n}(y_i - \overline{y})^2$

Teilt man diese Summe durch die Anzahl n der Daten (hier n = 3) und zieht dann die Wurzel, so ergibt sich die Standardabweichung $\sigma_{Daten} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}}$

Entsprechend geht man vor bei der Bestimmung von σ_{Modell} . Dort werden die blauen Quadrate der vertikalen Abweichungen des Modells „Gerade“ vom Schwerpunkt S betrachtet .

Für die Standardabweichung erhält man: $\sigma_{Modell} = \sqrt{\frac{\sum_{i=1}^{n}(f(x_i) - \overline{y})^2}{n}}$

Begründung der beiden Ergebnisse für die Standardabweichungen :

Für die roten Quadrate kann man die Summe der Flächeninhalte am Grafikraster ablesen:
Summe der Flächeninhalte = 14 Einheiten .

Für die blauen Quadrate rechnet man so:

Summe der Flächeninhalte = $(\frac{8}{13}\cdot 2+\frac{12}{13}-4)^2+(\frac{8}{13}\cdot 4+\frac{12}{13}-4)^2+(\frac{8}{13}\cdot 9+\frac{12}{13}-4)^2 = \frac{128}{13}\approx 9{,}85$

Teilt man σ_{Modell} durch σ_{Daten} , so erhält man $r = \sqrt{\frac{128\cdot 3}{14\cdot 13\cdot 3}} = \sqrt{\frac{64}{91}} \approx \underline{\underline{0{,}8386}}$

Dies ist der gesuchte Korrelationskoeffizient r für das obige Modell „Gerade“ .

Vereinfachung der Formel für den Korrelationskoeffizienten:

Bei der Division von σ_{Modell} durch σ_{Daten} kürzt sich n weg :

$$|r| = \frac{\sigma_{Modell}}{\sigma_{Daten}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(f(x_i)-\bar{y})^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}} = \frac{\sqrt{\sum_{i=1}^{n}(f(x_i)-\bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(f(x_i)-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$|r| = \sqrt{\frac{\sum_{i=1}^{n}(f(x_i)-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{Korrelationskoeffizient ;} \quad |r| \le 1$$

Der Korrelationskoeffizient liegt also zwischen -1 und 1 ; folgende Einteilung ist üblich:

|r| = 1	0,7 ≤ |r| < 1	0,3 ≤ |r| < 0,7	0 < |r| < 0,3	r = 0
volle Korrelation	starke Korrelation	mittlere Korrelation	schwache Korrelation	keine Korrelation

Eine starke Korrelation bedeutet nicht zwangsläufig, dass zwischen den betrachteten Merkmalen ein kausaler Zusammenhang besteht.

Ob ein solcher Zusammenhang besteht , muss immer auch sachlich beurteilt werden !

Anmerkung: Die Güte des angenommenen Regressionsmodells kann man auch dadurch beurteilen, dass man die Summe der Quadrate der Abweichungen der y_i zu den $f(x_i)$ berechnet, d.h. $\sum_{i=1}^{n}(y_i-f(x_i))^2$. Je kleiner die Summe, umso besser ist das Modell geeignet.

Korrelationskoeffizient für die Lineare Regression:

Ersetzt man in der obigen Formel f(x) durch m·x + b, so folgt:

$$|r| = \sqrt{\frac{\sum_{i=1}^{n}(mx_i + b - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \bar{y} = m\bar{x} + b \quad = \sqrt{\frac{\sum_{i=1}^{n}(mx_i + b - m\bar{x} - b)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^{n} m^2 \cdot (x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{m^2 \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Jetzt setzt man den aus der Formel für die Lineare Regression bekannten Term für m ein:

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

Hier ist schon die Unabhängigkeit von b,m zu sehen.

Weitere Umformungen:

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

Lässt man die Betragsstriche weg, so erhält man den vorzeichenbehafteten Wert von r :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$ Linearer Korrelationskoeffizient für $f(x) = m \cdot x + b$

Diese Formel lässt sich noch so umformen, dass die Rechenschritte vereinfacht werden:

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

$$\frac{\sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \sum_{i=1}^{n} y_i - \bar{y} \cdot \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x} \cdot \bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \bar{x}^2 - 2 \cdot \sum_{i=1}^{n} x_i \cdot \bar{x}) \cdot (\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \bar{y}^2 - 2 \cdot \sum_{i=1}^{n} y_i \cdot \bar{y})}} \quad \begin{matrix} \sum_{i=1}^{n} y_i = n \cdot \bar{y} \\ = \\ \sum_{i=1}^{n} \bar{x} \cdot \bar{y} = n \cdot \bar{x} \cdot \bar{y} \end{matrix}$$

$$\frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y} - n \cdot \bar{x} \cdot \bar{y} + n \cdot \bar{x} \cdot \bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 + n \cdot \bar{x}^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^{n} x_i) \cdot (\sum_{i=1}^{n} y_i^2 + n \cdot \bar{y}^2 - 2 \cdot \bar{y} \cdot \sum_{i=1}^{n} y_i)}} \quad \begin{matrix} \sum_{i=1}^{n} x_i = n \cdot \bar{x} \\ = \end{matrix}$$

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

Ergebnis:

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

Beispiel von oben: xi = {2, 4, 9} yi = {1, 5, 6}

$\bar{x} = 5$ $\bar{y} = 4$

$\sum xi \cdot yi = 2 \cdot 1 + 4 \cdot 5 + 9 \cdot 6 = 2 + 20 + 54 = 76$

$\sum xi^2 = 2^2 + 4^2 + 9^2 = 4 + 16 + 81 = 101$

$\sum yi^2 = 1^2 + 5^2 + 6^2 = 1 + 25 + 36 = 62$

$$r = \frac{76 - 3 \cdot 5 \cdot 4}{\sqrt{(101 - 3 \cdot 5^2) \cdot (62 - 3 \cdot 4^2)}} = \frac{16}{\sqrt{364}} \approx 0{,}8386278694$$

Anmerkung:

Sieht man den Korrelationskoeffizienten r in Bezug auf die beiden Regressionsgeraden

g1: $y = m \cdot x + b$ und g2: $y = m_2 \cdot x + b_2$,

so kann man folgende Beziehung erkennen:

$$r = \sqrt{\frac{m}{m_2}}$$

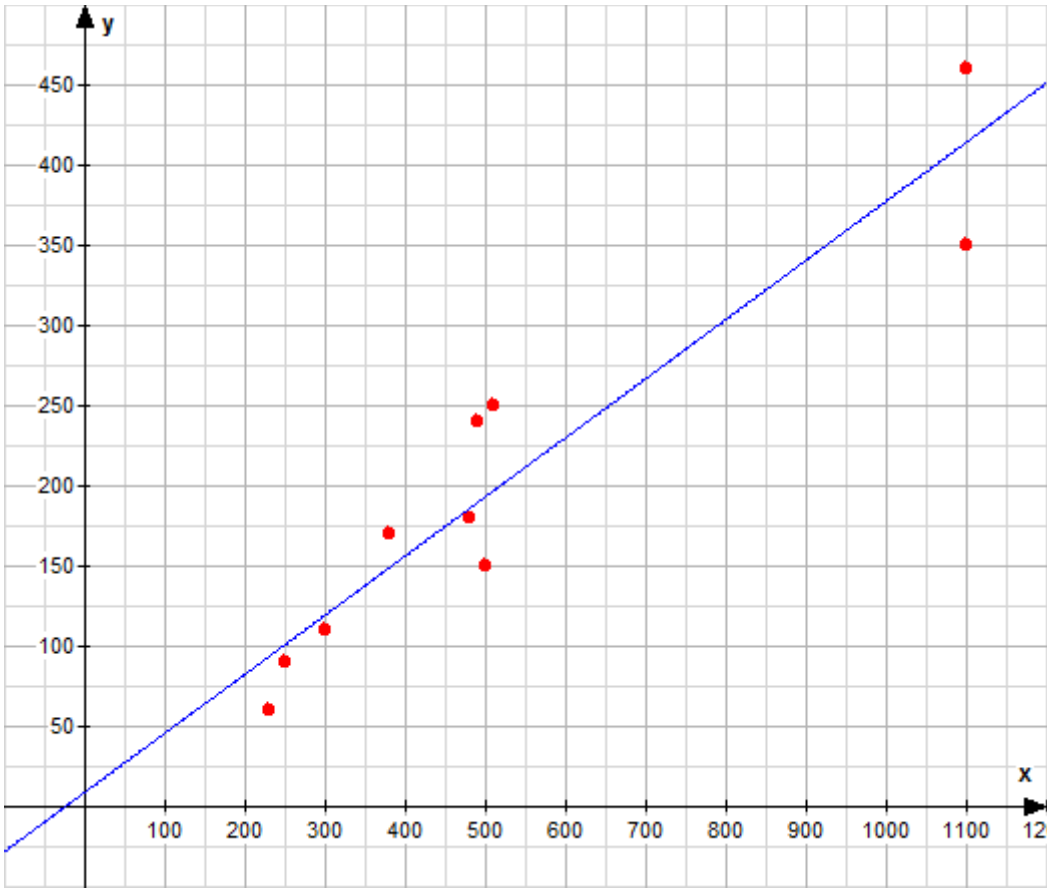
Aufgabe / Beispiel zur (Lin.) Korrelation:

R.Doll untersuchte 1955 als erster systematisch den möglichen Zusammenhang zwischen Zigarettenkonsum und Erkrankungen an Lungenkrebs. Er trug folgende Daten zusammen:

x = Zigarettenverbrauch pro Kopf 1930
y = Todesfälle an Lungenkrebs je Million 1950
z = Land, in dem die Untersuchung stattfand .

x	y	z
230	60	Island
250	90	Norwegen
300	110	Schweden
380	170	Dänemark
480	180	Australien
490	240	Niederlande
500	150	Kanada
510	250	Schweiz
1100	350	Finnland
1100	460	England

Lösung : y = 0,368653x + 9,139335 r = 0,942763



Es liegt eine starke Korrelation vor.
Dennoch ist es gewagt, von einem kausalen (ursächlichen) Zusammenhang zu sprechen.

Anmerkung:

Inzwischen ist dieser kausale Zusammenhang aufgrund weiterer Untersuchungen bewiesen.

Korrelationskoeffizient für den Linearen Spezialfall f(x) = m·x :

Die Gerade geht hier durch (0; 0), also verwenden wir die Steigungen m und m_2 der beiden Regressionsgeraden durch den Ursprung :

$$\text{Mit} \quad m = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \quad \text{und} \quad m_2 = \frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i y_i} \quad \text{folgt :}$$

$$r = \sqrt{\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \cdot \frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i y_i}} = \sqrt{\frac{(\sum_{i=1}^{n} x_i y_i)^2}{(\sum_{i=1}^{n} x_i^2) \cdot (\sum_{i=1}^{n} y_i^2)}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2) \cdot (\sum_{i=1}^{n} y_i^2)}}$$

$$r = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2) \cdot (\sum_{i=1}^{n} y_i^2)}} \qquad \text{Korrelationskoeffizient für } f(x) = m \cdot x$$

Beispiel von oben: xi = {2, 4, 9} yi = {1, 5, 6}

∑ xi·yi = 2 + 20 + 54 = 76
∑ xi² = 4 + 16 + 81 = 101
∑ yi² = 1 + 25 + 36 = 62

Also: $r = \frac{76}{\sqrt{101 \cdot 62}} \approx 0{,}9604108564...$