Korrelationskoeffizient

Wie linear ist der Zusammenhang zwischen x und y

Für eine Punktewolke $P_1, P_2, \ldots, P_n$ gibt der Korrelationskoeffizient r an, wie linear die x- und y-Werte von einander abhängig sind.

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

Wie kommt es zu so einer Definition?

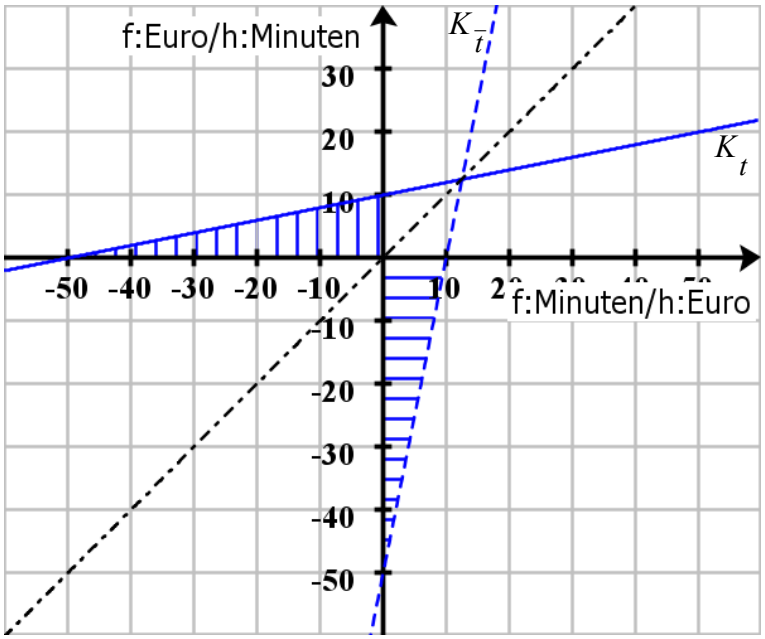
„Echter“ linearer Zusammenhang

Ein Mobilfunktarif setzt sich aus einer Grundgebühr und einem Minutenpreis zusammen.

Minuten	0	5	15
Euro	10	11	13

Der Preis ist linear von der Dauer in Minuten abhängig: t: Preis$=0{,}2\cdot$Minuten$+10$.

Aus diesem Grund lässt sich aus dem bezahlten Preis die gesamt Dauer berechnen: $\bar{t}$: Minuten$=5\cdot$Preis-50 (Umkehrfunktion).



Aus dem Schaubild lässt sich ableiten, dass $m_{\bar{t}}=\frac{1}{m_t} \Rightarrow m_t\cdot m_{\bar{t}}=1$

Was lernen wir daraus?

Besteht zwischen $x_1, x_2, \ldots, x_n$ und $y_1, y_2, \ldots, y_n$ ein linearer Zusammenhang, so ist

$y_i=m_x x_i+b_x$ und $x_i=m_y y_i+b_y$ und es ist $m_x\cdot m_y=1$

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.
2016 Henrik Horstmann

Was hat das mit Regressionsgeraden zu tun?

Beispiel: $P_1(2|1);\ P_2(5|4);\ P_3(11|4)\ \Rightarrow\ x_1=2,\ x_2=5,\ x_3=11\ \wedge\ y_1=1, y_2=4,\ y_3=4$

Die berechnete Gleichung der Regressionsgerade lautet $R_x:\ y=\frac{2}{7}x+\frac{9}{7}$

Wenn ein lineare Zusammenhang zwischen den x- und y- Werten besteht, muss die Regressionsgerade R_y mit der y-Werte in x-Werte berechnet werden eine Umkehrfunktion von R_x sein.

Bestimme R_y:

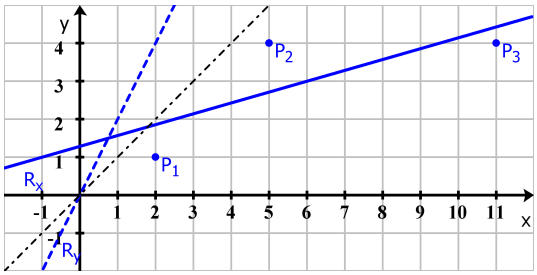
$\bar{x}=6 \qquad \bar{y}=3$

$$y_d=\sum_{i=1}^{3}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)=(2-6)(1-3)+(5-6)(4-3)+(11-6)(4-3)=12$$

$$x_d=\sum_{i=1}^{3}\left(y_i-\bar{y}\right)^2=(1-3)^2+(4-3)^2+(4-3)^2=6$$

$$m_R=\frac{12}{6}=2$$

$$R_Y:\ y=2(x-3)+6=2x$$



Es ist $m_x\cdot m_y=\frac{2}{7}\cdot 2=\frac{4}{7}\approx 0{,}57$

Damit ist der Zusammenhang nicht „echt“ linear.

Zurück zum Korrelationskoeffizienten

m_x und m_y berechnen sich mit:

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

damit ist

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

und es ist

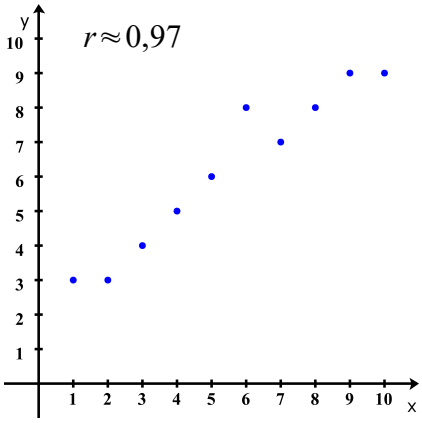
$$\sqrt{m_x\cdot m_y}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\cdot\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}=r$$

Es kann gezeigt werden, dass $-1\le r\le 1$.

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.
2016 Henrik Horstmann

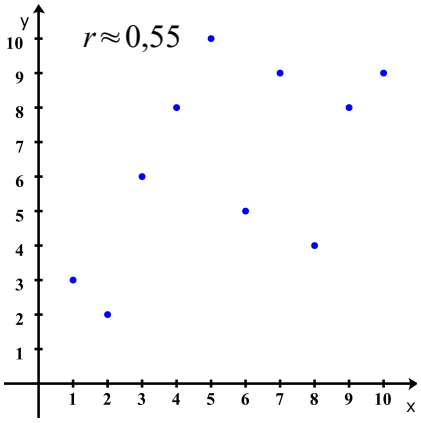
Güteklassen

Für $|r|=1$ besteht ein „echter“ linearer Zusammenhang.



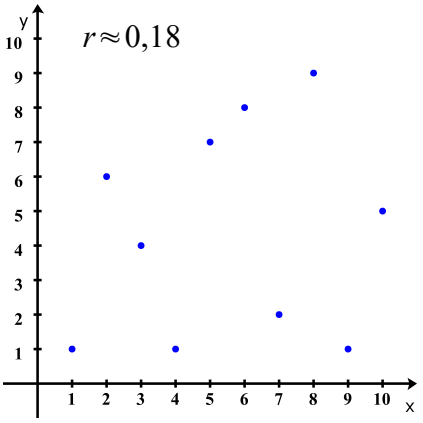
$0{,}8 \leq |r|$

Hohe Korrelation (Punkte liegen sehr nahe an der Regressionsgeraden)



$0{,}3 \leq |r| < 0{,}8$

Schwache Korrelation (wage ist ein linearer Zusammenhang erkennbar)



$|r| < 0{,}3$

Keine Korrelation (kein linearer Zusammenhang erkennbar)

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.
2016 Henrik Horstmann