Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

Prof. Dr. Dirk Ostwald

(2) Korrelation

Grundlagen

Korrelation und Bestimmtheitsmaß

Korrelation und linear-affine Abhängigkeit

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie

Mehr Therapiestunden

$\Rightarrow$ Höhere Wirksamkeit?

Unabhängige Variable

- Anzahl Therapiestunden

Abhängige Variable

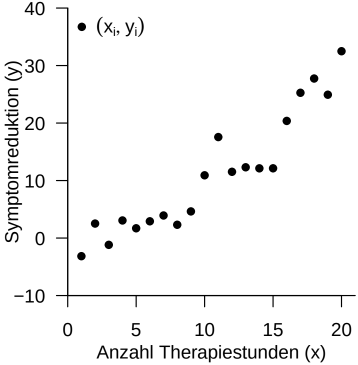
- Symptomreduktion

Beispieldatensatz

$i = 1, ..., 20$ Patient:innen, y_i Symptomreduktion bei Patient:in i, x_i Anzahl Therapiestunden von Patient:in i

y_i	x_i
-3.15	1
2.52	2
-1.18	3
3.06	4
1.70	5
2.91	6
3.92	7
2.31	8
4.63	9
10.91	10
17.56	11
11.52	12
12.31	13
12.12	14
12.13	15
20.37	16
25.26	17
27.75	18
24.93	19
32.49	20

Beispieldatensatz



Wie stark hängen Anzahl Therapiestunden und Symptomreduktion zusammen?

Definition (Korrelation)

Die *Korrelation* zweier Zufallsvariablen ξ und υ ist definiert als

$$\rho(\xi, \upsilon) := \frac{\mathbb{C}(\xi, \upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)} \tag{1}$$

wobei $\mathbb{C}(\xi, \upsilon)$ die Kovarianz von ξ und υ und $\mathbb{S}(\xi)$ und $\mathbb{S}(\upsilon)$ die Standardabweichungen von ξ und υ, respektive, bezeichnen.

Bemerkungen

- $\rho(\xi, \upsilon)$ wird auch *Korrelationskoeffizient* von ξ und υ genannt.
- Wir haben bereits gesehen, dass $-1 \le \rho(\xi, \upsilon) \le 1$ gilt.
- Wenn $\rho(\xi, \upsilon) = 0$ ist, werden ξ und υ *unkorreliert* genannt.
- Wir haben bereits gesehen, dass aus der Unabhängigkeit von ξ und υ, folgt dass $\rho(\xi, \upsilon) = 0$.
- Aus $\rho(\xi, \upsilon) = 0$ folgt aber wie bereits gesehen die Unabhängigkeit von ξ und υ im Allgemeinen nicht.

Definition (Stichprobenkorrelation)

$\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ sei ein Datensatz. Weiterhin seien:

- Die Stichprobenmittel der x_i und y_i definiert als

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \text{ und } \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{2}$$

- Die Stichprobenstandardabweichungen x_i und y_i definiert als

$$s_x := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ und } s_y := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3}$$

- Die Stichprobenkovarianz der $(x_1, y_1), ..., (x_n, y_n)$ definiert als

$$c_{xy} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{4}$$

Dann ist die *Stichprobenkorrelation* der $(x_1, y_1), ..., (x_n, y_n)$ definiert als

$$r_{xy} := \frac{c_{xy}}{s_x s_y} \tag{5}$$

und wird auch *Stichprobenkorrelationskoeffizient* genannt.

Beispiel

```r
# Laden des Beispieldatensatzes
fname = file.path(getwd(), "2_Daten", "2_Korrelation_Beispieldatensatz.csv")    # Dateipfad
D     = read.table(fname, sep = ",", header = TRUE)                              # Laden als Dataframe
x_i   = D$x_i                                                                    # x_i Werte
y_i   = D$y_i                                                                    # y_i Werte
n     = length(x_i)                                                              # n

# "Manuelle" Berechnung der Stichprobenkorrelation
x_bar = (1/n)*sum(x_i)                                                           # \bar{x}
y_bar = (1/n)*sum(y_i)                                                           # \bar{y}
s_x   = sqrt(1/(n-1)*sum((x_i - x_bar)^2))                                       # s_x
s_y   = sqrt(1/(n-1)*sum((y_i - y_bar)^2))                                       # s_y
c_xy  = 1/(n-1) * sum((x_i - x_bar) * (y_i - y_bar))                             # c_{xy}
r_xy  = c_xy/(s_x * s_y)                                                         # r_{xy}
print(r_xy)                                                                      # Ausgabe
```

```
> [1] 0.938
```

```r
# Automatische Berechnung mit cor()
r_xy  = cor(x_i,y_i)                                                             # r_{xy}
print(r_xy)                                                                      # Ausgabe
```

```
> [1] 0.938
```

$\Rightarrow$ Anzahl Therapiestunden und Symptomreduktion sind hochkorreliert.

Mechanik der Kovariationsterme

$(x_i - \bar{x}) < 0, (y_i - \bar{y}) > 0$	$(x_i - \bar{x}) > 0, (y_i - \bar{y}) > 0$
$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) < 0$	$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) > 0$
$(\bar{x}, \bar{y})$	
$(x_i - \bar{x}) < 0, (y_i - \bar{y}) < 0$	$(x_i - \bar{x}) > 0, (y_i - \bar{y}) < 0$
$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) > 0$	$\Rightarrow (x_i - \bar{x})(y_i - \bar{y}) < 0$

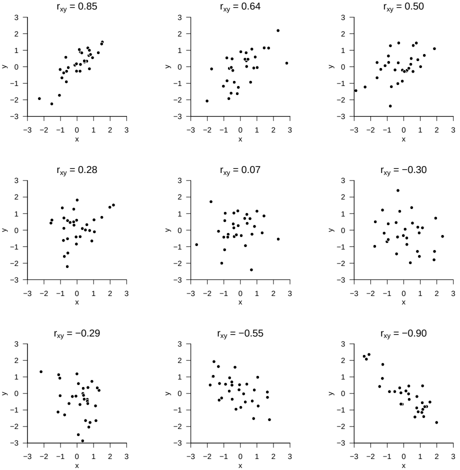
Häufige richtungsgleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ Positive Korrelation

Häufige richtungsungleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ Negative Korrelation

Keine häufigen richtungsgleichen oder -entgegengesetzten Abweichungen $\Rightarrow$ Keine Korrelation

Grundlagen

Beispiele



Theorem (Stichprobenkorrelation bei linear-affinen Transformationen)

Für einen Datensatz $\{(x_i, y_i)\}_{i=1,...n} \subset \mathbb{R}^2$ sei $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1,...n} \subset \mathbb{R}^2$ eine linear-affin transformierte Wertemenge mit

$$(\tilde{x}_i, \tilde{y}_i) = (a_x x_i + b_x, a_y y_i + b_y), a_x, a_y \neq 0. \tag{6}$$

Dann gilt

$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \tag{7}$$

Bemerkungen

- Der Betrag der Stichprobenkorrelation ändert sich bei linear-affiner Datentransformation nicht.
- Man sagt, dass die Stichprobenkorrelation im Gegensatz zur Stichprobenkovarianz *maßstabsunabhängig* ist.

Beweis

Es gilt

$$
\begin{aligned}
r_{\tilde{x}\tilde{y}} &:= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}\tilde{x}_i - \bar{\tilde{x}}\right)^2}\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}\tilde{y}_i - \bar{\tilde{y}}\right)^2}} \\
&= \frac{\sum_{i=1}^{n}(a_x x_i + b_x - (a_x\bar{x} + b_x))(a_y y_i + b_y - (a_y\bar{y} + b_y))}{\sqrt{\sum_{i=1}^{n}(a_x x_i + b_x - (a_x\bar{x} + b_x))^2}\sqrt{\sum_{i=1}^{n}(a_y y_i + b_y - (a_y\bar{y} + b_y))^2}} \\
&= \frac{a_x a_y \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{a_x^2\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{a_y^2\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{a_x a_y}{|a_x||a_y|}\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{a_x a_y}{|a_x||a_y|}\frac{c_{xy}}{s_x s_y} \\
&= \frac{a_x a_y}{|a_x||a_y|} r_{xy}.
\end{aligned}
\tag{8}
$$

Also folgt, durch Durchspielen aller möglichen Vorzeichenfälle, dass

$$
|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \tag{9}
$$

□

Grundlagen

Korrelation und Bestimmtheitsmaß

Korrelation und linear-affine Abhängigkeit

Selbstkontrollfragen

Überblick

Das sogenannte Bestimmtheitsmaß R^2 ist eine beliebte Statistik.

Numerisch ist R^2 das Quadrat des Stichprobenkorrelationskoeffizienten.

Ist die Stichprobenkorrelation $r_{xy} = 0.5$, dann ist $\mathrm{R}^2 = 0.25$, ist $r_{xy} = -0.5$, dann ist $\mathrm{R}^2 = 0.25$.

$\Rightarrow$ R^2 enthält also weniger Information über die Rohdaten als r_{xy}, da das Vorzeichen wegfällt.

$\Rightarrow$ *Perse* ist die Angabe von R^2 anstelle von r_{xy} im Kontext der Korrelation zweier Variablen wenig sinnvoll.

Ein tieferes Verständnis von R^2 erlaubt jedoch

(1) Einen Einstieg in das Konzept von Quadratsummenzerlegungen, einem wichtigen ALM Evaluationsprinzip.

(2) Einen Einstieg in das Verständnis der Zusammenhänge von Ausgleichsgerade und Stichprobenkorrelation.

(3) Einen ersten Einblick in die Tatsache, dass Korrelationen (nur) linear-affine Zusammenhänge quantifizieren.

Definition (Erklärte Werte und Residuen einer Ausgleichsgerade)

Gegeben seien ein Datensatz $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ und die zu diesem Datensatz gehörende Ausgleichsgerade

$$f_{\hat{\beta}} : \mathbb{R} \to \mathbb{R}, x \mapsto f_{\hat{\beta}}(x) := \hat{\beta}_0 + \hat{\beta}_1 x. \tag{10}$$

Dann werden für $i = 1, ..., n$

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{11}$$

die durch die Ausgleichsgerade *erklärten Werte* genannt und

$$\hat{\varepsilon}_i := y_i - \hat{y}_i \tag{12}$$

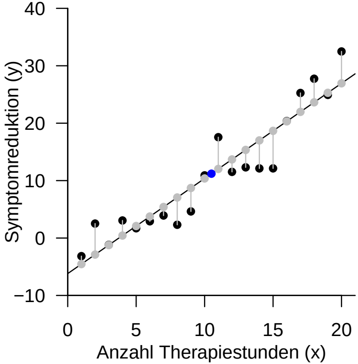
die *Residuen* der Ausgleichsgerade genannt.

Bemerkungen

- Die *erklärten Werte* sind die Datenvorhersage des Modells basierend auf den geschätzten Parameterwerten.
- Die *Residuen* sind die Differenzen zwischen der geschätzten Datenvorhersage und den beobachteten Datenwerten.

Korrelation und Bestimmtheitsmaß

Erklärte Werte und Residuen



$\bullet\ (x_i, y_i)$ $\quad$ $\bullet\ (\bar{x}, \bar{y})$ $\quad$ — $f_{\hat{\beta}}(x)$ $\quad$ $\bullet\ \hat{y}_i$ $\quad$ — $\hat{\varepsilon}_i$ $\quad$ $i = 1, ..., n$

Theorem (Quadratsummenzerlegung bei Ausgleichsgerade)

Für einen Datensatz $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ und seine zugehörige Ausgleichsgerade $f_{\hat{\beta}}$ seien für

$$\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \text{ und } \hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, \text{ für } i = 1, ..., n \tag{13}$$

das Stichprobenmittel der υ-Werte und die durch die Ausgleichsgerade erklärten Werte, respektive. Weiterhin seien

$$\mathsf{SQT} := \sum\nolimits_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{die } \textit{Total Sum of Squares}$$

$$\mathsf{SQE} := \sum\nolimits_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \text{die } \textit{Explained Sum of Squares}$$

$$\mathsf{SQR} := \sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \text{die } \textit{Residual Sum of Squares}$$

Dann gilt

$$\mathsf{SQT} = \mathsf{SQE} + \mathsf{SQR} \tag{14}$$

Korrelation und Bestimmtheitsmaß

Bemerkungen

- SQT repräsentiert die Gesamtstreung der y_i-Werte um ihren Mittelwert $\bar{y}$.
- SQE repräsentiert die Streuung der erklärten Werte $\hat{y}_i$ um ihren Mittelwert.

  $\Rightarrow$ Große Werte von SQE repräsentieren eine große absolute Steigung der y_i mit den x_i.

  $\Rightarrow$ Kleine Werte von SQE repräsentieren eine kleine absolute Steigung der y_i mit den x_i.

- SQE ist also ein Maß für die Stärke des linearen Zusammenhangs der x_i- und y_i-Werte.
- SQR ist die Summe der quadrierten Residuen, es gilt

$$\text{SQR} := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 := \sum_{i=1}^{n} \hat{\varepsilon}_i^2. \tag{15}$$

  $\Rightarrow$ Große Werte von SQR repräsentieren große Abweichungen der erklärten $\hat{y}_i$ von den beobachteten y_i-Werten.

  $\Rightarrow$ Kleine Werte von SQR repräsentieren geringe Abweichungen der erklärten $\hat{y}_i$ von den beobachteten y_i-Werten.

- SQR ist also ein Maß für die Güte der Beschreibung der Datenmenge durch die Ausgleichsgerade.

Korrelation und Bestimmtheitsmaß

Beweis

$$
\begin{aligned}
\mathsf{SQT} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^{n} \left((y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\right) \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \mathsf{SQE} + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \mathsf{SQR} \\
&= \mathsf{SQE} + \mathsf{SQR}
\end{aligned}
\tag{16}
$$

Beweis (fortgeführt)

Dabei ergibt sich die letzte Gleichung mit

$$\bar{\hat{y}} := \frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y} \quad (17)$$

und damit auch

$$\bar{\hat{y}} = \bar{y} \Leftrightarrow \frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}\sum_{i=1}^n y_i \Leftrightarrow \sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i \Leftrightarrow \bar{y}\sum_{i=1}^n \hat{y}_i = \bar{y}\sum_{i=1}^n y_i \quad (18)$$

sowie

$$\bar{\hat{y}} = \bar{y} \Leftrightarrow \frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}\sum_{i=1}^n y_i \Leftrightarrow \sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i \Leftrightarrow \sum_{i=1}^n y_i \hat{y}_i = \sum_{i=1}^n \hat{y}_i \hat{y}_i \quad (19)$$

aus

Beweis (fortgeführt)

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i\hat{y}_i - y_i\bar{y} - \hat{y}_i\hat{y}_i + \hat{y}_i\bar{y}) \\
&= \sum_{i=1}^{n} y_i\hat{y}_i - \sum_{i=1}^{n} y_i\bar{y} - \sum_{i=1}^{n}\hat{y}_i\hat{y}_i + \sum_{i=1}^{n}\hat{y}_i\bar{y} \\
&= \sum_{i=1}^{n} y_i\hat{y}_i - \sum_{i=1}^{n}\hat{y}_i\hat{y}_i + \bar{y}\sum_{i=1}^{n}\hat{y}_i - \bar{y}\sum_{i=1}^{n} y_i \\
&= 0 + 0 \\
&= 0
\end{aligned}
\tag{20}
$$

□

Definition (Bestimmtheitsmaß R^2)

Für einen Datensatz $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ und seine zugehörige Ausgleichsgerade $f_{\hat{\beta}}$ sowie die zugehörigen Explained Sum of Squares SQE und Total Sum of Squares SQT heißt

$$\mathrm{R}^2 := \frac{\mathrm{SQE}}{\mathrm{SQT}} \tag{21}$$

Bestimmtheitsmaß oder *Determinationskoeffizient*.

Theorem (Stichprobenkorrelation und Bestimmtheitsmaß)

Für einen Datensatz $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ sei R^2 das Bestimmtheitsmaß und r_{xy} sei die Stichprobenkorrelation. Dann gilt

$$\mathrm{R}^2 = r_{xy}^2. \tag{22}$$

Korrelation und Bestimmtheitsmaß

Bemerkungen

- Mit $-1 \leq r_{xy} \leq 1$ folgt aus dem Theorem direkt, dass $0 \leq \mathsf{R}^2 \leq 1$.
- Es gilt $\mathsf{R}^2 = 0$ genau dann, wenn $\mathsf{SQE} = 0$ ist

  $\Rightarrow$ Für $\mathsf{R}^2 = 0$ ist die erklärte Streuung der Daten durch die Ausgleichsgerade gleich null.

  $\Rightarrow \mathsf{R}^2 = 0$ beschreibt also den Fall einer denkbar schlechten Erklärung der Daten durch die Ausgleichsgerade.
- Es gilt $\mathsf{R}^2 = 1$ genau dann, wenn $\mathsf{SQE} = \mathsf{SQT}$ ist.

  $\Rightarrow$ Für $\mathsf{R}^2 = 1$ ist also die Gesamtstreuung gleich der durch die Ausgleichsgerade erklärten Streuung.

  $\Rightarrow$ $\mathsf{R}^2 = 1$ beschreibt also den Fall das sämtliche Datenvariabilität durch die Ausgleichsgerade erklärt wird.
- Man sagt, dass "R^2 die durch die Ausgleichsgerade erklärte Varianz an der Gesamtdatenvarianz" ist.

Korrelation und Bestimmheitsmaß

Beweis

Wir halten zunächst fest, dass mit

$$
\bar{\hat{y}} := \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 + \hat{\beta}_1\bar{x} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x} = \bar{y} \tag{23}
$$

folgt, dass

$$
\begin{aligned}
\mathsf{SQE} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1\bar{x})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_1(x_i - \bar{x}))^2 \\
&= \hat{\beta}_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2
\end{aligned} \tag{24}
$$

Beweis

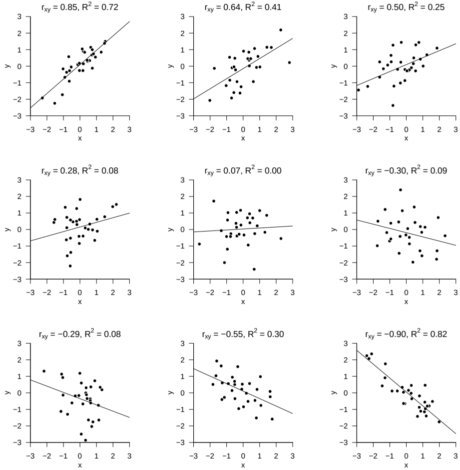
Damit ergibt sich dann

$$
\begin{aligned}
\mathsf{R}^2 &= \frac{\mathsf{SQE}}{\mathsf{SQT}} \\
&= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{c_{xy}^2}{s_x^4} \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{c_{xy}^2}{s_x^4} \frac{s_x^2}{s_y^2} \\
&= \frac{c_{xy}^2}{s_x^2 s_y^2} \\
&= \left(\frac{c_{xy}}{s_x s_y}\right)^2 \\
&= r_{xy}^2 .
\end{aligned}
\tag{25}
$$

□

Korrelation und Bestimmheitsmaß

Beispiele



Grundlagen

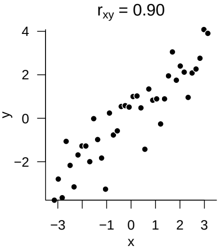
Korrelation und Bestimmtheitsmaß

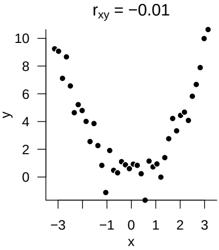
Korrelation und linear-affine Abhängigkeit

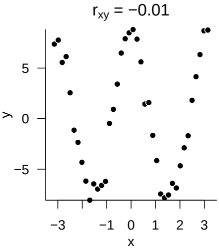
Selbstkontrollfragen

Korrelation und linear-affine Abhängigkeit

Funktionale Abhängigkeiten und Stichprobenkorrelation







$$\upsilon_i = x_i + \varepsilon_i \qquad \upsilon_i = x_i^2 + \varepsilon_i \qquad \upsilon_i = 8\cos(2x_i) + \varepsilon_i$$

$$\varepsilon_i \sim N(0,1)$$

Theorem (Korrelation und linear-affine Abhängigkeit)

ξ und υ seien zwei Zufallsvariablen mit positiver Varianz. Dann besteht genau dann eine lineare-affine Abhängigkeit der Form

$$\upsilon = \beta_0 + \beta_1 \xi \text{ mit } \beta_0, \beta_1 \in \mathbb{R} \tag{26}$$

zwischen ξ und υ, wenn

$$\rho(\xi, \upsilon) = 1 \text{ oder } \rho(\xi, \upsilon) = -1 \tag{27}$$

gilt.

Bemerkungen

- Die linear-affine Abhängigkeit $\upsilon = \beta_0 + \beta_1 \xi$ impliziert eine linear-affine Abhängigkeit $\xi = \tilde{\beta}_0 + \tilde{\beta}_1 \upsilon$, denn

$$\upsilon = \beta_0 + \beta_1 \xi \Leftrightarrow -\beta_0 + \upsilon = \beta_1 \xi \Leftrightarrow \xi = -\frac{\beta_0}{\beta_1} + \frac{1}{\beta_1}\upsilon \Leftrightarrow \xi = \tilde{\beta}_0 + \tilde{\beta}_1 \upsilon \tag{28}$$

mit

$$\tilde{\beta}_0 = -\frac{\beta_0}{\beta_1} \text{ und } \tilde{\beta}_1 = \frac{1}{\beta_1}. \tag{29}$$

Korrelation und linear-affine Abhängigkeit

Beweis

Wir beschränken uns auf den Beweis der Aussage, dass aus $\upsilon = \beta_0 + \beta_1\xi$ folgt, dass $\rho(\xi,\upsilon) = \pm 1$ ist. Dazu halten wir zunächst fest, dass mit den Theoremen zu den Eigenschaften von Erwartungswert und Varianz gilt, dass

$$\mathbb{E}(\upsilon) = \beta_0 + \beta_1\mathbb{E}(\xi) \text{ und } \mathbb{V}(\upsilon) = \beta_1^2\mathbb{V}(\xi). \tag{30}$$

Wegen $\mathbb{V}(\xi) > 0$ und $\mathbb{V}(\upsilon) > 0$ gilt damit $\beta_1 \neq 0$. Es folgt dann

$$\beta_1 > 0 \Rightarrow \mathbb{S}(\upsilon) = \beta_1\mathbb{S}(\xi) > 0 \text{ und } \beta_1 < 0 \Rightarrow \mathbb{S}(\upsilon) = -\beta_1\mathbb{S}(\xi) > 0. \tag{31}$$

Weiterhin gilt

$$\begin{aligned}
\upsilon - \mathbb{E}(\upsilon) &= \beta_0 + \beta_1\xi - \mathbb{E}(\upsilon) \\
&= \beta_0 + \beta_1\xi - \beta_0 - \beta_1\mathbb{E}(\xi) \\
&= \beta_1\xi - \beta_1\mathbb{E}(\xi) \\
&= \beta_1(\xi - \mathbb{E}(\xi)).
\end{aligned} \tag{32}$$

Für die Kovarianz von ξ und υ ergibt sich also

$$\begin{aligned}
\mathbb{C}(\xi,\upsilon) &= \mathbb{E}\left((\upsilon - \mathbb{E}(\upsilon))(\xi - \mathbb{E}(\xi))\right) \\
&= \mathbb{E}\left(\beta_1(\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))\right) \\
&= \beta_1\mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right) \\
&= \beta_1\mathbb{V}(\xi).
\end{aligned} \tag{33}$$

Damit ergibt für die Korrelation von ξ und υ

$$\rho(\xi,\upsilon) = \frac{\mathbb{C}(\xi,\upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)} = \pm\frac{\beta_1\mathbb{V}(\xi)}{\mathbb{S}(\xi)\beta_1\mathbb{S}(\xi)} = \pm\frac{\beta_1\mathbb{V}(\xi)}{\beta_1\mathbb{V}(\xi)} = \pm 1. \tag{34}$$

Grundlagen

Korrelation und Bestimmtheitsmaß

Korrelation und linear-affine Abhängigkeit

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition der Korrelation zweier Zufallsvariablen wieder.
2. Geben Sie die Definitionen von Stichprobenmittel, -standardabweichung, -kovarianz und -korrelation wieder.
3. Erläutern Sie anhand der Mechanik der Kovariationsterme, wann eine Stichprobenkorrelation einen hohen absoluten Wert annimmt, einen hohen positiven Wert annimmt, einen hohen negativen Wert annimmt und einen niedrigen Wert annimmt.
4. Geben Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen wieder.
5. Erläutern Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen.
6. Geben Sie die Definitionen von erklärten Werten und Residuen einer Ausgleichsgerade wieder.
7. Geben Sie das Theorem zur Quadratsummenzerlegung bei einer Ausgleichsgerade wieder.
8. Erläutern Sie die intuitiven Bedeutungen von SQT, SQE und SQR.
9. Geben Sie die Definition des Bestimmtheitsmaßes R^2 wieder.
10. Geben Sie das Theorem zum Zusammenhang von Stichprobenkorrelation und Bestimmtheitsmaß wieder.
11. Erläutern Sie die Bedeutung von hohen und niedrigen R^2 Werten im Lichte der Ausgleichsgerade.
12. Geben Sie das Theorem zum Zusammenhang von Korrelation und linear-affiner Abhängigkeit wieder.