Allgemeines Lineares Modell

BSc Psychologie, SoSe 2024

Joram Soch

(2) Korrelation

Grundbegriffe der Korrelation

Korrelation und Bestimmtheitsmaß

Anwendung/Praxis

Selbstkontrollfragen

Grundbegriffe der Korrelation

Korrelation und Bestimmtheitsmaß

Anwendung/Praxis

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie

Mehr Therapiestunden

⇒ Höhere Wirksamkeit?

Unabhängige Variable

- Anzahl Therapiestunden

Abhängige Variable

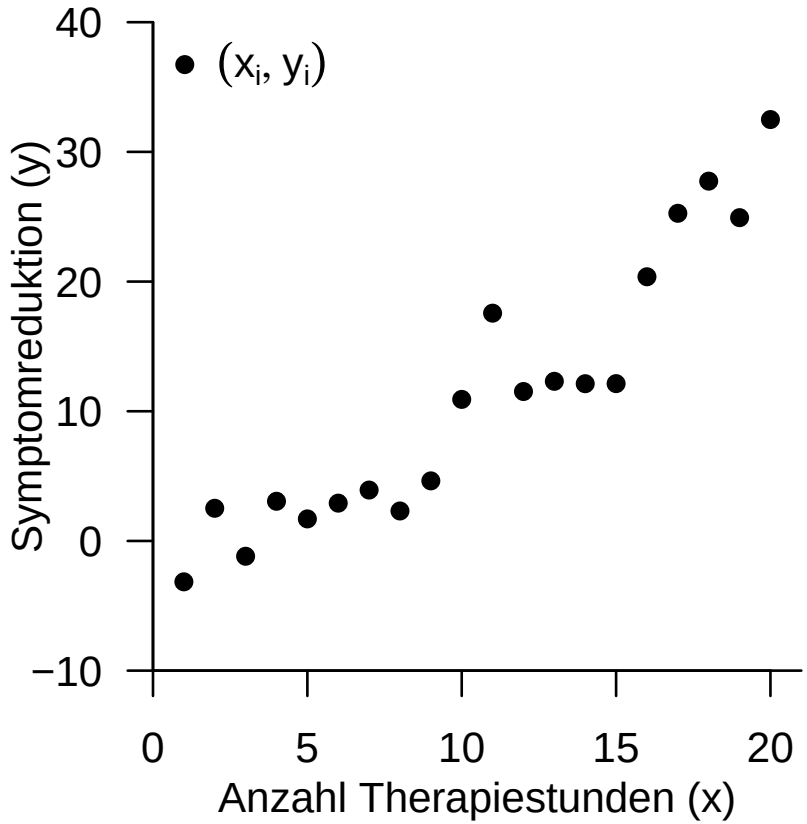
- Symptomreduktion

Beispieldatensatz

$i = 1, \ldots, 20$ 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 bzw. υ bezeichnen.

Bemerkungen

- $\rho(\xi, \upsilon)$ wird auch *Korrelationskoeffizient* von ξ und υ genannt.
- Wir haben bereits gesehen, dass $-1 \leq \rho(\xi, \upsilon) \leq 1$ 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 $\rho(\xi, \upsilon) = 0$.
- Wenn $\rho(\xi, \upsilon) = 0$ ist, sind aber ξ und υ nicht notwendigerweise unabhängig voneinander.

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 = "Daten/Korrelation_Simulation.csv"                    # Dateipfad generieren
D     = read.table(fname, sep = ",", header = TRUE)           # Beispieldatensatz als Dataframe laden
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
```

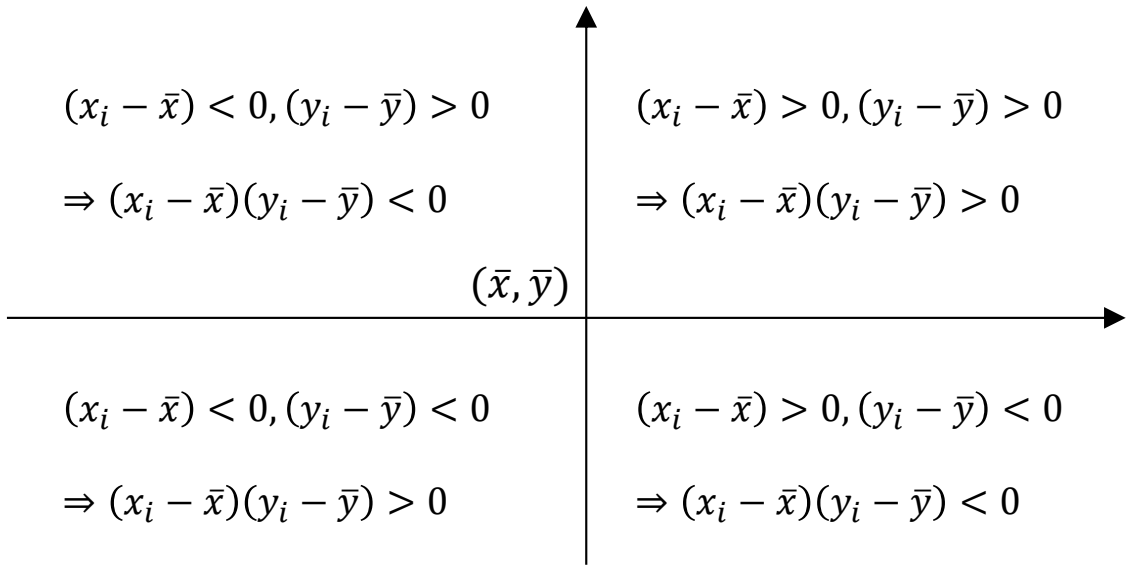
```
> [1] 0.938
```

```r
# automatische Berechnung mit der R-Funktion 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

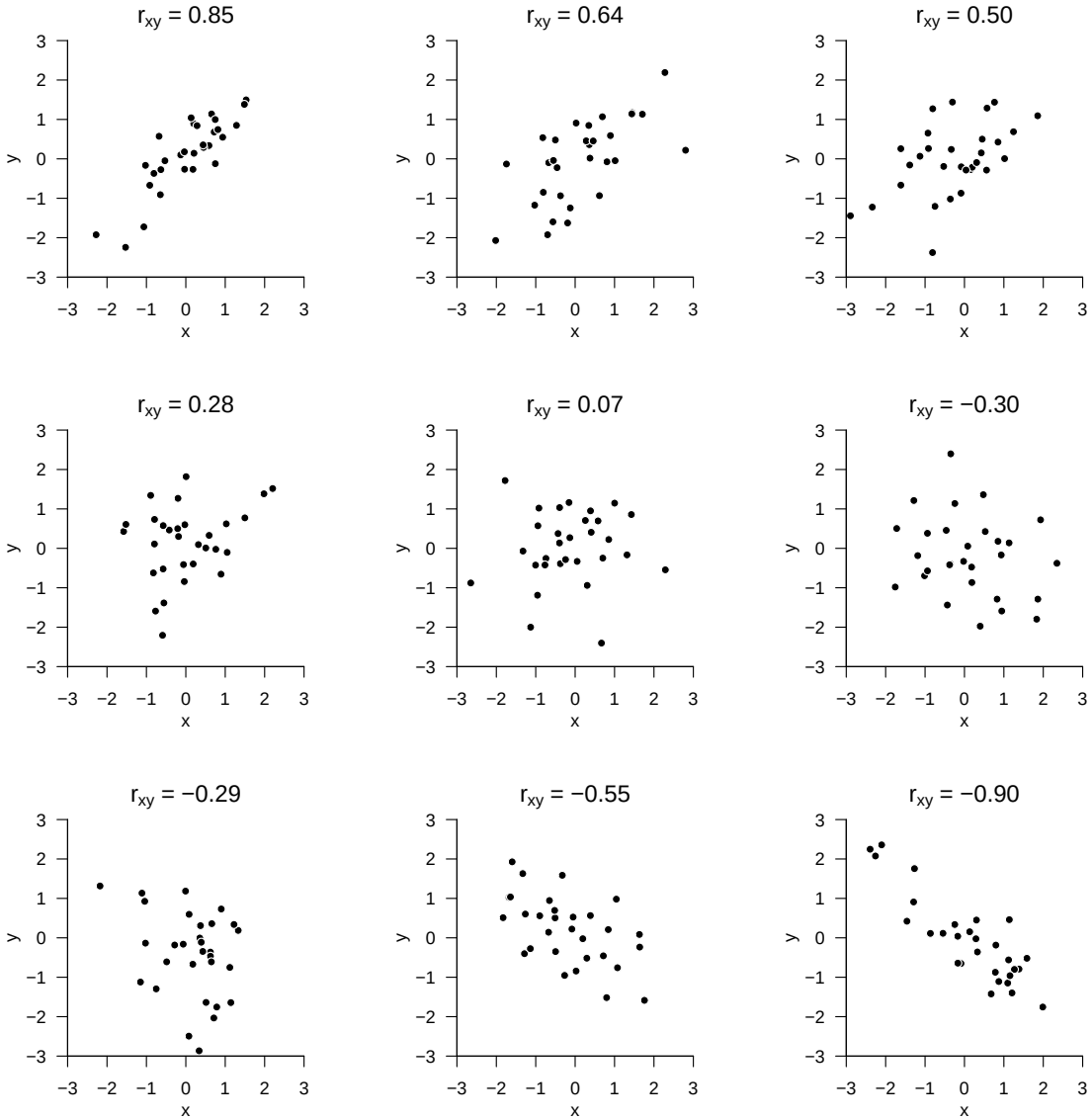


häufige richtungsgleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ positive Korrelation

häufige richtungsungleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ negative Korrelation

keine häufigen richtungsgleichen oder richtungsungleichen Abweichungen $\Rightarrow$ keine Korrelation

Beispiele


r_{xy} = 0.85
r_{xy} = 0.64
r_{xy} = 0.50
r_{xy} = 0.28
r_{xy} = 0.07
r_{xy} = −0.30
r_{xy} = −0.29
r_{xy} = −0.55
r_{xy} = −0.90
x
y

Theorem (Korrelation und linear-affine Abhängigkeit)

ξ und υ seien zwei Zufallsvariablen mit positiver Varianz. Dann gilt: Ihre Korrelation beträgt

$$\rho(\xi, \upsilon) = 1 \text{ oder } \rho(\xi, \upsilon) = -1 \tag{6}$$

genau dann, wenn eine lineare-affine Abhängigkeit der folgenden Form zwischen ξ und υ besteht:

$$\upsilon = \beta_0 + \beta_1 \xi \text{ mit } \beta_0, \beta_1 \in \mathbb{R} \, . \tag{7}$$

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

$$\upsilon = \beta_0 + \beta_1 \xi \Leftrightarrow -\beta_0 + \upsilon = \beta_1 \xi \Leftrightarrow \xi = -\frac{\beta_0}{\beta_1} + \frac{1}{\beta_1} \upsilon \Leftrightarrow \xi = \tilde{\beta}_0 + \tilde{\beta}_1 \upsilon \, , \tag{8}$$

  wobei wie folgt ersetzt wurde:

$$\tilde{\beta}_0 = -\frac{\beta_0}{\beta_1} \text{ und } \tilde{\beta}_1 = \frac{1}{\beta_1} \, . \tag{9}$$

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 Erwartungswert, Varianz und Standardabweichung gilt, dass

$$\mathbb{E}(\upsilon) = \beta_0 + \beta_1 \mathbb{E}(\xi) \text{ und } \mathbb{V}(\upsilon) = \beta_1^2 \mathbb{V}(\xi) \text{ und somit } \mathbb{S}(\upsilon) = \pm \beta_1 \mathbb{S}(\xi) \; . \tag{10}$$

Wegen $\mathbb{V}(\xi) > 0$ und $\mathbb{V}(\upsilon) > 0$ gilt damit $\beta_1 \neq 0$. Weiterhin gilt:

$$\begin{aligned} \upsilon - \mathbb{E}(\upsilon) &= \beta_0 + \beta_1 \xi - \mathbb{E}(\upsilon) \\ &= \beta_0 + \beta_1 \xi - \beta_0 - \beta_1 \mathbb{E}(\xi) \\ &= \beta_1 \xi - \beta_1 \mathbb{E}(\xi) \\ &= \beta_1 (\xi - \mathbb{E}(\xi)) \; . \end{aligned} \tag{11}$$

Für die Kovarianz von ξ und υ ergibt sich also:

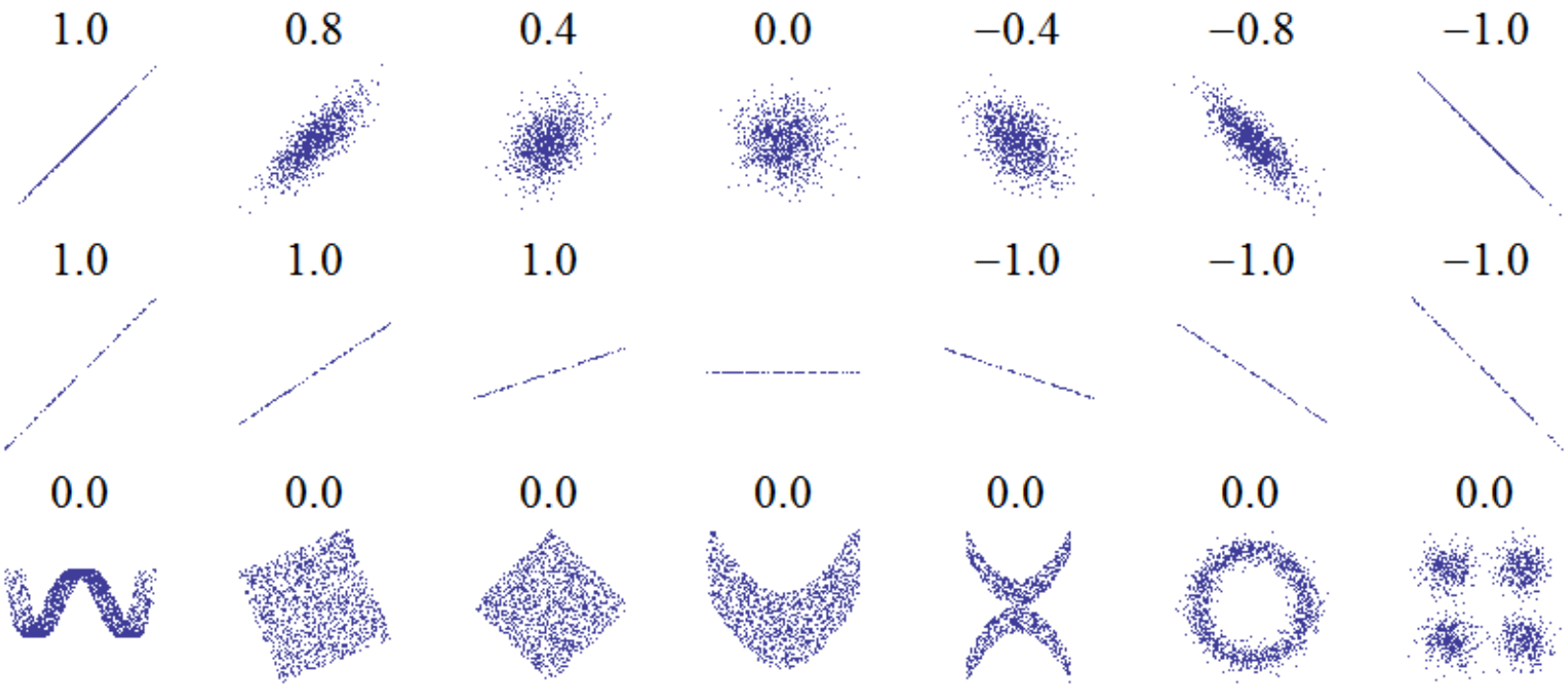
$$\begin{aligned} \mathbb{C}(\xi, \upsilon) &= \mathbb{E}\left((\upsilon - \mathbb{E}(\upsilon))(\xi - \mathbb{E}(\xi))\right) \\ &= \mathbb{E}\left(\beta_1 (\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))\right) \\ &= \beta_1 \mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right) \\ &= \beta_1 \mathbb{V}(\xi) \; . \end{aligned} \tag{12}$$

Damit ergibt für die Korrelation von ξ und υ:

$$\rho(\xi, \upsilon) = \frac{\mathbb{C}(\xi, \upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)} = \pm \frac{\beta_1 \mathbb{V}(\xi)}{\mathbb{S}(\xi)\beta_1 \mathbb{S}(\xi)} = \pm \frac{\beta_1 \mathbb{V}(\xi)}{\beta_1 \mathbb{V}(\xi)} = \pm 1. \tag{13}$$

□

Beispiele



(Quelle: *Wikimedia Commons*: "Correlation_examples.png"; Lizenz: gemeinfrei.)

Grundbegriffe der Korrelation

Korrelation und Bestimmtheitsmaß

Anwendung/Praxis

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 $R^2 = 0.25$, ist $r_{xy} = -0.5$, dann ist $R^2 = 0.25$.

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

$\Rightarrow$ *Per se* 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, ein wichtiges 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

$$f_{\hat{\beta}} : \mathbb{R} \to \mathbb{R}, x \mapsto f_{\hat{\beta}}(x) := \hat{\beta}_0 + \hat{\beta}_1 x. \tag{14}$$

Dann werden für $i = 1, ..., n$

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{15}$$

die durch die Ausgleichsgerade *erklärten Werte* genannt und

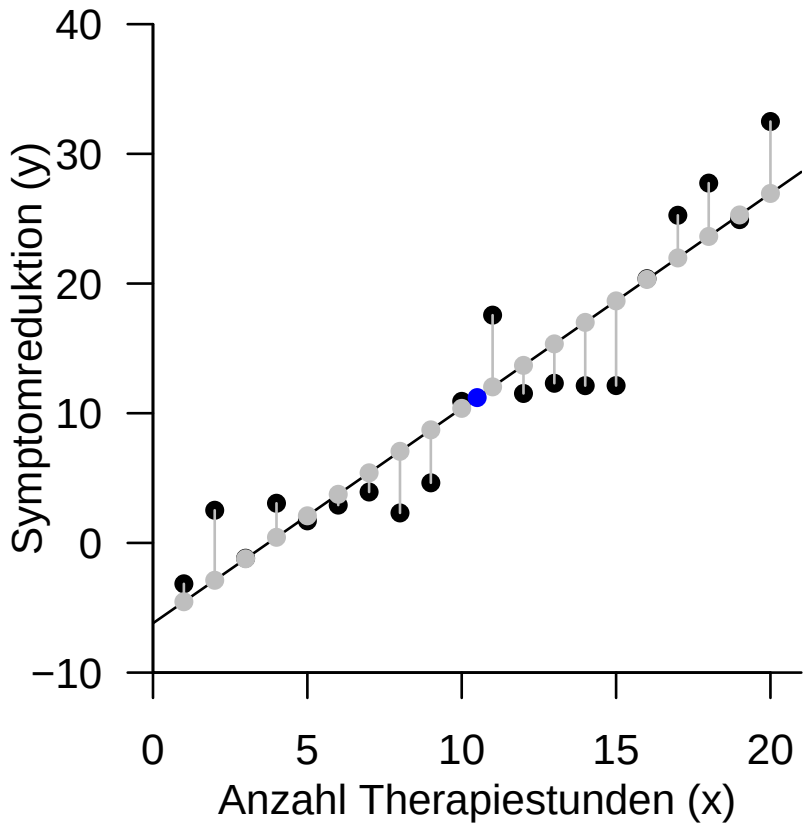
$$\hat{\varepsilon}_i := y_i - \hat{y}_i \tag{16}$$

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 geschätzter Datenvorhersage und beobachteten Datenwerten.

Erklärte Werte und Residuen einer Ausgleichsgerade



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

Motivation

Die in einer abhängigen Variable enthaltene Gesamtvarianz lässt sich in verschiedene Beiträge partitionieren: denjenigen Teil der Varianz, der sich mit Rückgriff auf unabhängige Variablen erklären lässt (*erklärte Varianz*), und denjenigen Teil der Varianz, der aus dem Beitrag von Zufallsvariablen resultiert (*nicht-erklärte Varianz*).

Dieser Gedanke bildet die Grundlage für die Methode der Quadratsummenzerlegung. Wir werden diese im Folgenden zunächst am Beispiel der einfachen linearen Regression mit Ausgleichsgerade betrachten und das Konzept später im Rahmen von einfaktorieller und zweifaktorieller Varianzanalysen wieder aufgreifen.

Definition (Quadratsummen)

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

$$\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i \text{ und } \hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, \text{ für } i = 1, ..., n \tag{17}$$

das Stichprobenmittel der y-Werte und die durch die Ausgleichsgerade erklärten Werte. Dann sind

$\text{SQT} := \sum_{i=1}^{n} (y_i - \bar{y})^2$ die *totale Quadratsumme*

$\text{SQE} := \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ die *erklärte Quadratsumme*

$\text{SQR} := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ die *residuelle Quadratsumme*

Bemerkungen

- SQT wird auch als *total sum of squares* (TSS) bezeichnet.
- SQE wird auch als *explained sum of squares* (ESS) bezeichnet.
- SQR wird auch als *residual sum of squares* (RSS) bezeichnet.

Bemerkungen (fortgeführt)

- SQT repräsentiert die Gesamtstreuung 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:

$$\text{SQR} := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f_{\hat{\beta}}(x_i))^2 := \sum_{i=1}^{n} \hat{\varepsilon}_i^2. \tag{18}$$

  $\Rightarrow$ Große Werte von SQR repräsentieren große Abweichungen der erklärten von den beobachteten y_i-Werten.

  $\Rightarrow$ Kleine Werte von SQR repräsentieren geringe Abweichungen der erklärten von den beobachteten y_i-Werten.
- SQR ist also ein Maß für die Güte der Beschreibung der Datenmenge durch die Ausgleichsgerade.

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 SQT, SQE und SQR die totale, die erklärte und die residuelle Quadratsumme. Dann gilt

$$\text{SQT} = \text{SQE} + \text{SQR} \tag{19}$$

Bemerkungen

- Die totale Quadratsumme entspricht der Summe aus erklärter Quadratsumme und residueller Quadratsumme.

Beweis

Wir schreiben zunächst die totale Quadratsumme aus:

$$\begin{aligned}
\mathsf{SQT} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^{n}\left((y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\right) \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \mathsf{SQE} + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \mathsf{SQR}\ .
\end{aligned} \tag{20}$$

Es verbleibt also zu zeigen, dass der mittlere Term Null ist:

$$2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0\ . \tag{21}$$

Beweis (fortgeführt)

Unter Gebrauch des Ausdrucks $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ für die erklärten Werte einer Ausgleichsgerade und nach Einsetzen des Ausgleichsgeradenparameters $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (siehe Einheit (1) in *Allgemeines Lineares Modell*) ergibt sich dann für diese Summe zunächst:

$$
\begin{aligned}
\sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) \\
&= \sum_{i=1}^{n} 2(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y}) \\
&= \sum_{i=1}^{n} 2\left((y_i - \bar{y}) - (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})\right)(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}) \\
&= 2\sum_{i=1}^{n} \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\right) \hat{\beta}_1 (x_i - \bar{x}) \\
&= 2\sum_{i=1}^{n} \left((y_i - \hat{y}_i)\hat{\beta}_1 (x_i - \bar{x}) - \hat{\beta}_1 (x_i - \bar{x})\hat{\beta}_1 (x_i - \bar{x})\right) \\
&= 2\left(\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \hat{y}_i)(x_i - \bar{x}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})\right) .
\end{aligned}
\tag{22}
$$

Beweis (fortgeführt)

Hierin erkennen wir nun Vielfache der Stichprobenkovarianz c_{xy} und der Stichprobenvarianz s_x^2:

$$\begin{aligned}\sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2\left(\hat{\beta}_1(n-1)c_{xy} - \hat{\beta}_1^2(n-1)s_x^2\right) \\ &= 2(n-1)\left(\hat{\beta}_1 c_{xy} - \hat{\beta}_1^2 s_x^2\right) \ .\end{aligned} \tag{23}$$

Einsetzen des Ausgleichsgeradenparameters $\hat{\beta}_1 = c_{xy}/s_x^2$ ergibt schließlich:

$$\begin{aligned}\sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2(n-1)\left(\left(\frac{c_{xy}}{s_x^2}\right)c_{xy} - \left(\frac{c_{xy}}{s_x^2}\right)^2 s_x^2\right) \\ &= 2(n-1)\left(\frac{c_{xy}^2}{s_x^2} - \frac{c_{xy}^2}{s_x^2}\right) \\ &= 2(n-1)\cdot 0 \\ &= 0 \ .\end{aligned} \tag{24}$$

□

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 zugehörige erklärte Quadratsumme SQE und totale Quadratsumme SQT heißt

$$R^2 := \frac{SQE}{SQT} \tag{25}$$

Bestimmtheitsmaß oder *Determinationskoeffizient*.

Bemerkungen

- Es gilt $R^2 = 0$ genau dann, wenn $SQE = 0$ ist.

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

  $\Rightarrow$ $R^2 = 0$ beschreibt also den Fall der schlechtestmöglichen Erklärung der Daten durch die Ausgleichsgerade.

- Es gilt $R^2 = 1$ genau dann, wenn $SQE = SQT$ ist.

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

  $\Rightarrow$ $R^2 = 1$ beschreibt also den Fall, dass sämtliche Datenvariabilität durch die Ausgleichsgerade erklärt wird.

- Man sagt, dass R^2 "der Anteil der durch die Ausgleichsgerade erklärten Varianz an der gesamten Varianz der beobachteten Daten" ist.

Theorem (Bestimmtheitsmaß und residuelle Quadratsumme)

Für einen Datensatz $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ seien R^2 das Bestimmtheitsmaß sowie SQR und SQT die residuelle bzw. totale Quadratsumme. Dann lässt sich R^2 alternativ ausdrücken als

$$R^2 = 1 - \frac{SQR}{SQT} \ . \tag{26}$$

Bemerkungen

- Es gilt $R^2 = 0$ genau dann, wenn $SQR = SQT$ ist.
- Es gilt $R^2 = 1$ genau dann, wenn $SQR = 0$ ist.

Beweis

Mit der Definition von R^2 und dem Theorem zur Quadratsummenzerlegung bei Ausgleichsgerade gilt

$$R^2 := \frac{SQE}{SQT} = \frac{SQT - SQR}{SQT} = \frac{SQT}{SQT} - \frac{SQR}{SQT} = 1 - \frac{SQR}{SQT} \ . \tag{27}$$

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

$$\mathsf{R}^2 = r_{xy}^2. \tag{28}$$

Bemerkungen

- Bei zwei Variablen entspricht das Bestimmtheitsmaß dem Quadrat der Stichprobenkorrelation.
- Mit $-1 \leq r_{xy} \leq 1$ folgt aus dem Theorem direkt, dass $0 \leq \mathsf{R}^2 \leq 1$.

Beweis

Wir halten zunächst fest, dass mit

$$\bar{\hat{y}} := \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 + \hat{\beta}_1\bar{x} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x} = \bar{y} \tag{29}$$

folgt, dass

$$\begin{aligned}
\mathsf{SQE} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1\bar{x})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_1(x_i - \bar{x}))^2 \\
&= \hat{\beta}_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \ .
\end{aligned} \tag{30}$$

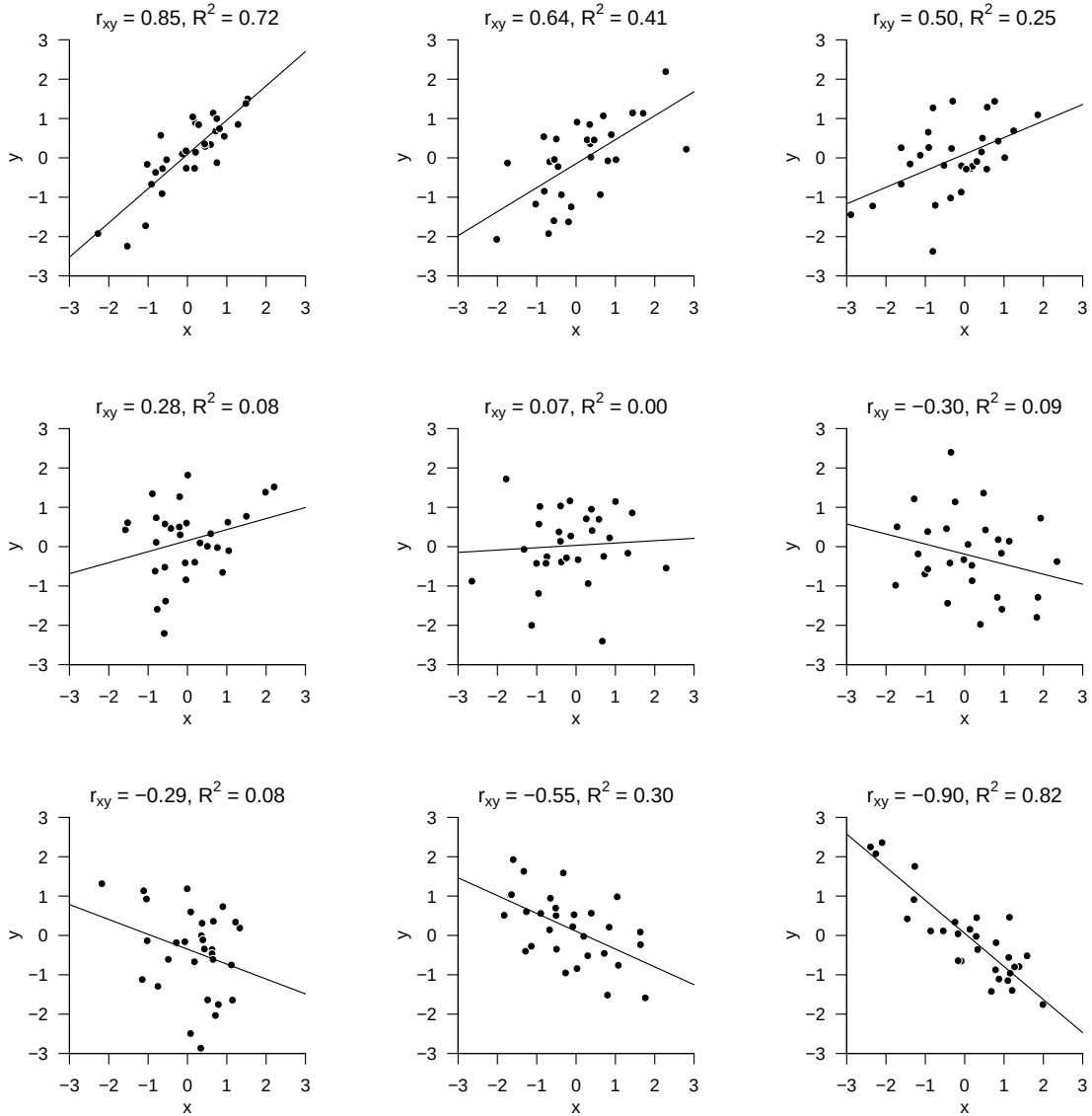
Beweis (fortgeführt)

Damit ergibt sich dann

$$
\begin{aligned}
\mathsf{R}^2 &= \frac{\mathsf{SQE}}{\mathsf{SQT}} \\
&= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \hat{\beta}_1^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \frac{c_{xy}^2}{s_x^4} \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \frac{c_{xy}^2}{s_x^4} \frac{s_x^2}{s_y^2} \\
&= \frac{c_{xy}^2}{s_x^2 s_y^2} \\
&= \left(\frac{c_{xy}}{s_x s_y}\right)^2 \\
&= r_{xy}^2 \, .
\end{aligned}
\tag{31}
$$

□

Beispiele


r_{xy} = 0.85, R^2 = 0.72
r_{xy} = 0.64, R^2 = 0.41
r_{xy} = 0.50, R^2 = 0.25
r_{xy} = 0.28, R^2 = 0.08
r_{xy} = 0.07, R^2 = 0.00
r_{xy} = −0.30, R^2 = 0.09
r_{xy} = −0.29, R^2 = 0.08
r_{xy} = −0.55, R^2 = 0.30
r_{xy} = −0.90, R^2 = 0.82
x
y

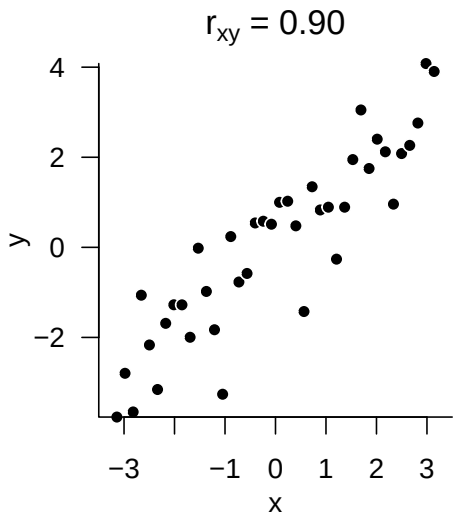
Grundbegriffe der Korrelation

Korrelation und Bestimmtheitsmaß

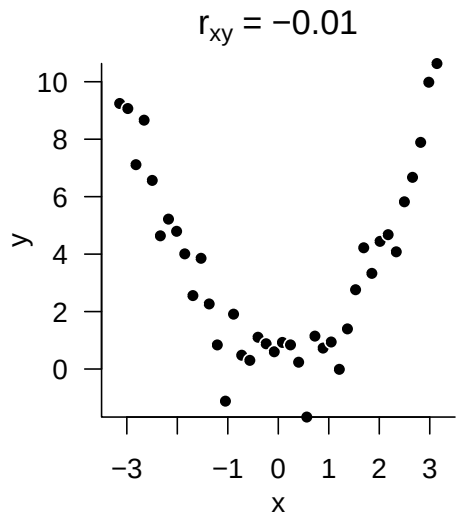
Anwendung/Praxis

Selbstkontrollfragen

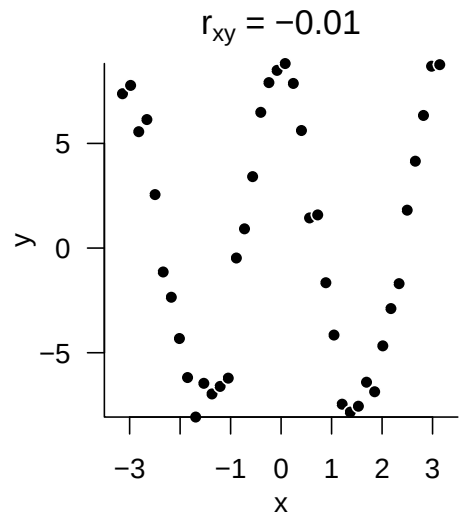
Funktionale Abhängigkeiten und Stichprobenkorrelation



$$y_i = x_i + \varepsilon_i$$



$$y_i = x_i^2 + \varepsilon_i$$



$$y_i = 8\cos(2x_i) + \varepsilon_i$$

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

Theorem (Stichprobenkorrelation bei linear-affinen Transformationen)

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

$$(\tilde{x}_i, \tilde{y}_i) = (a_x x_i + b_x, a_y y_i + b_y), a_x, a_y \neq 0. \tag{32}$$

Dann gilt

$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \tag{33}$$

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

$$
\begin{aligned}
r_{\tilde{x}\tilde{y}} &:= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i-\bar{\tilde{x}})(\tilde{y}_i-\bar{\tilde{y}})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{x}_i-\bar{\tilde{x}}\right)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{y}_i-\bar{\tilde{y}}\right)^2}} \\
&= \frac{\sum_{i=1}^{n}(a_x x_i + b_x - (a_x\bar{x}+b_x))(a_y y_i + b_y - (a_y\bar{y}+b_y))}{\sqrt{\sum_{i=1}^{n}(a_x x_i + b_x - (a_x\bar{x}+b_x))^2}\sqrt{\sum_{i=1}^{n}(a_y y_i + b_y - (a_y\bar{y}+b_y))^2}} \\
&= \frac{a_x a_y \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{a_x^2\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{a_y^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (34) \\
&= \frac{a_x a_y}{|a_x||a_y|}\,\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\
&= \frac{a_x a_y}{|a_x||a_y|}\,\frac{c_{xy}}{s_x s_y} \\
&= \frac{a_x a_y}{|a_x||a_y|}\, r_{xy}.
\end{aligned}
$$

Nach Durchspielen aller möglichen Vorzeichenfälle ergibt sich:

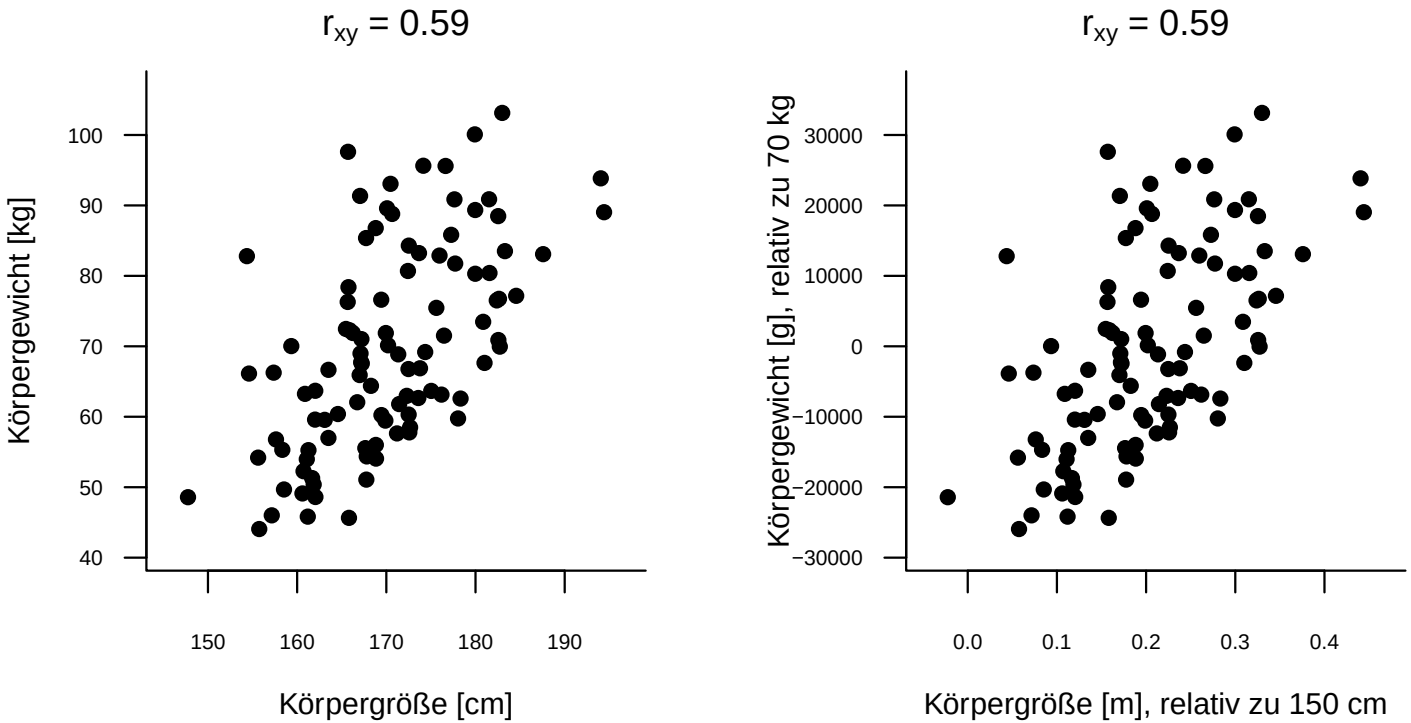
$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \qquad (35)$$

□

Beispiel: Körpergröße und Körpergewicht in unterschiedlichen Einheiten ($n = 100$)

$$y_i = -100 + 1\,x_i + \varepsilon_i \quad \text{mit} \quad \varepsilon_i \sim N(0, 12^2) \quad \text{für} \quad i = 1, \ldots, n \tag{36}$$

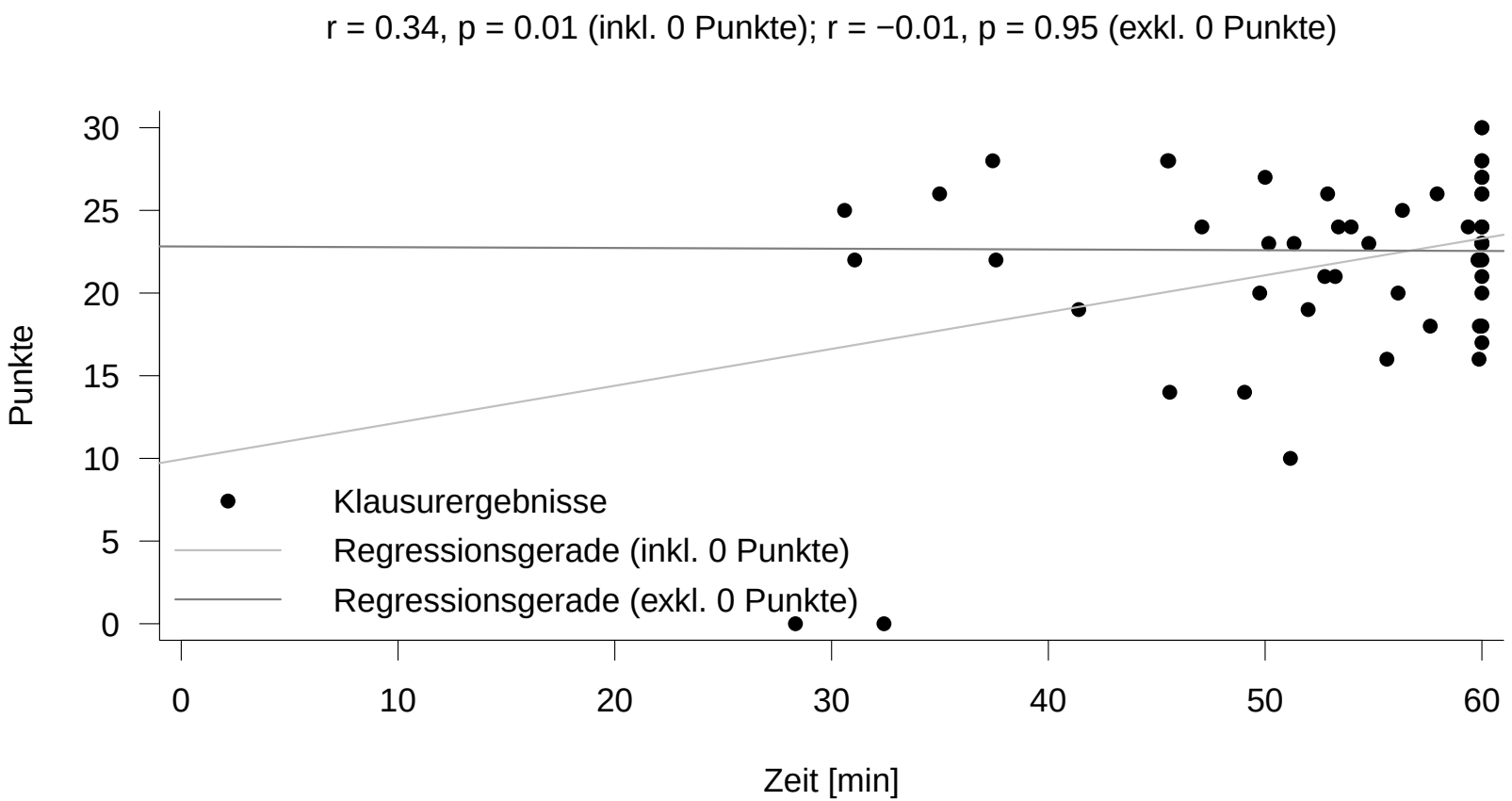
$$\tilde{x}_i = (x_i - 150)/100 \quad \text{und} \quad \tilde{y}_i = (y_i - 70) \cdot 1000 \quad \text{für} \quad i = 1, \ldots, n \tag{37}$$



- $(x_i, y_i),\ (\tilde{x}_i, \tilde{y}_i)$

Sensitivität der Stichprobenkorrelation gegenüber Ausreißern

Beispiel: Klausurzeit und Punktzahl, ALM-Klausur, SoSe 2024 ($n = 56$)



Grundbegriffe der Korrelation

Korrelation und Bestimmtheitsmaß

Anwendung/Praxis

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 zum Zusammenhang von Korrelation und linear-affiner Abhängigkeit wieder.
5. Geben Sie die Definitionen von erklärten Werten und Residuen einer Ausgleichsgerade wieder.
6. Geben Sie das Theorem zur Quadratsummenzerlegung bei einer Ausgleichsgerade wieder.
7. Erläutern Sie die intuitiven Bedeutungen von SQT, SQE und SQR.
8. Geben Sie die Definition des Bestimmtheitsmaßes R^2 wieder.
9. Geben Sie das Theorem zum Zusammenhang von R^2 und residueller Quadratsumme wieder.
10. Geben Sie das Theorem zum Zusammenhang von R^2 und Stichprobenkorrelation wieder.
11. Erläutern Sie die Bedeutung von hohen und niedrigen Werten von R^2 im Lichte der Ausgleichsgerade.
12. Geben Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen wieder.
13. Erläutern Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen am Beispiel.