Allgemeines Lineares Modell

BSc Psychologie SoSe 2022

Prof. Dr. Dirk Ostwald

(2) Korrelation

Grundlagen

Korrelation und Bestimmtheitsmaß

Korrelation und lineare Abhängigkeit

Korrelation und Regression

Partielle Korrelation

Selbstkontrollfragen

Grundlagen

Korrelation und Bestimmtheitsmaß

Korrelation und lineare Abhängigkeit

Korrelation und Regression

Partielle Korrelation

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie

Mehr Therapiestunden

⇒ Höhere Wirksamkeit?

Unabhängige Variable

- Anzahl Therapiestunden

Abhängige Variable

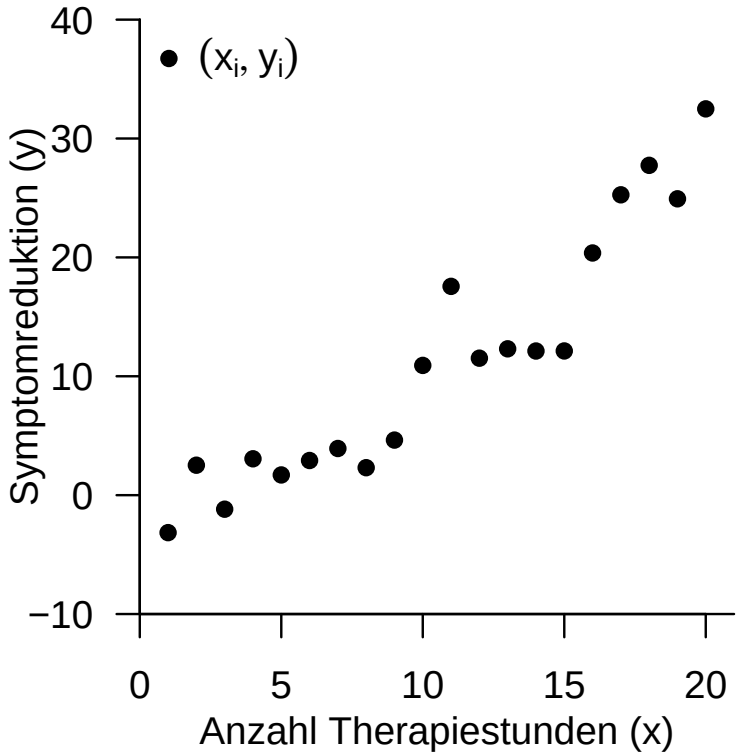
- Symptomreduktion

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 X und Y ist definiert als

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

wobei $\mathbb{C}(X, Y)$ die Kovarianz von X und Y und $\mathbb{V}(X)$ und $\mathbb{V}(Y)$ die Varianzen von X und Y, respektive, bezeichnen.

Bemerkungen

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

Definition (Stichprobenkorrelation)

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}$ sei eine Wertemenge. 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

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

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

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

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

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

und wird auch *Stichprobenkorrelationskoeffizient* genannt.

Beispiel

```
# 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
```

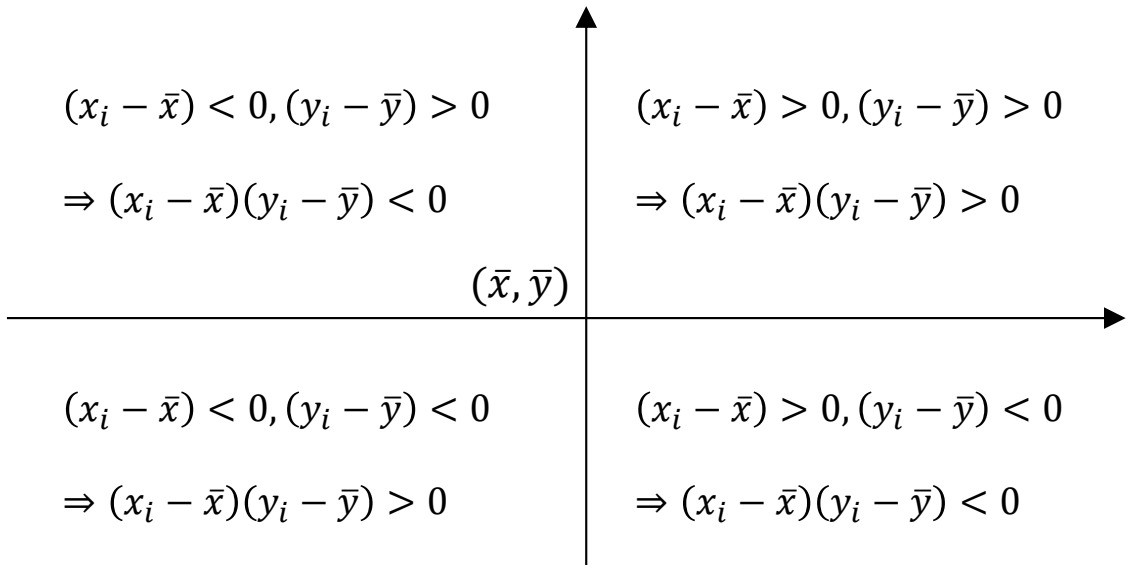
```
> [1] 0.938
```

```
# 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

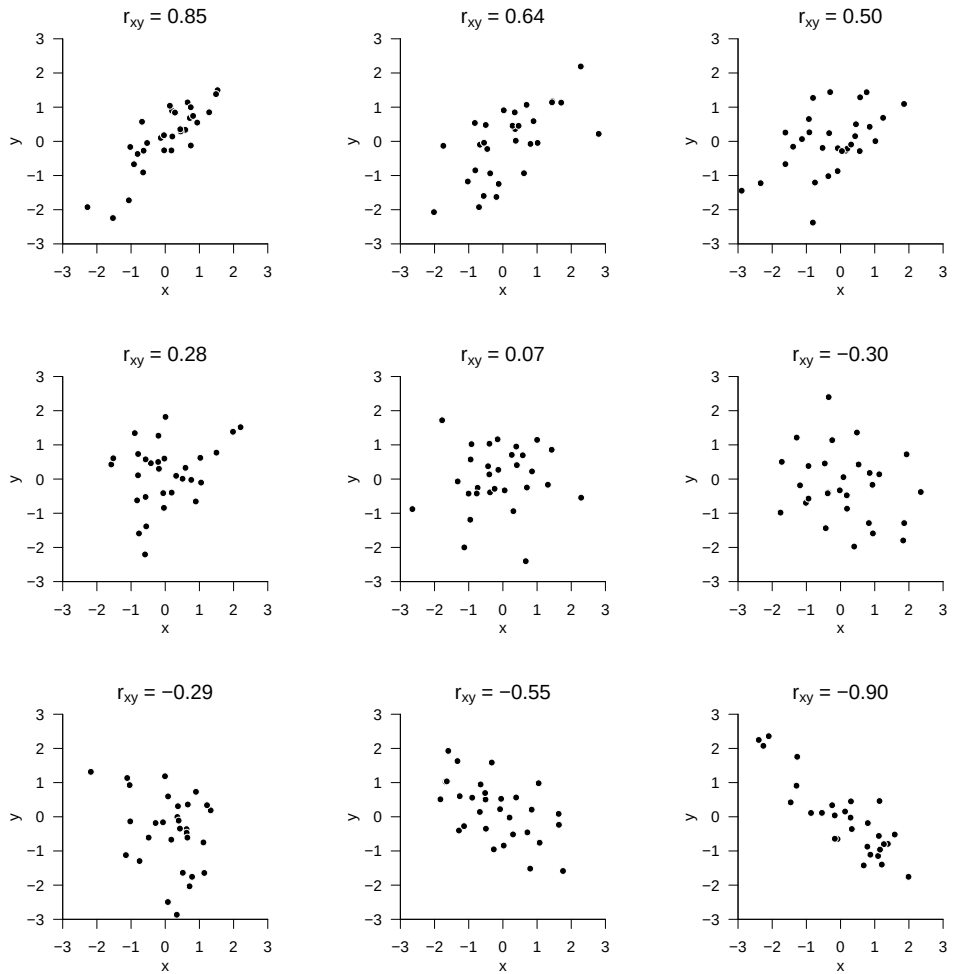


Häufige richtungsgleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ Positive Korrelation

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

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

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 (Stichprobenkorrelation bei linear-affinen Transformationen)

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

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

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

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

Also folgt, durch Durchspielen aller möglichen Vorzeichenfälle, dass

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

□

Grundlagen

Korrelation und Bestimmtheitsmaß

Korrelation und lineare Abhängigkeit

Korrelation und Regression

Partielle Korrelation

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.

$\Rightarrow$ *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 eine Wertemenge $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ und die zu dieser Wertemenge 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 $i = 1, ..., n$

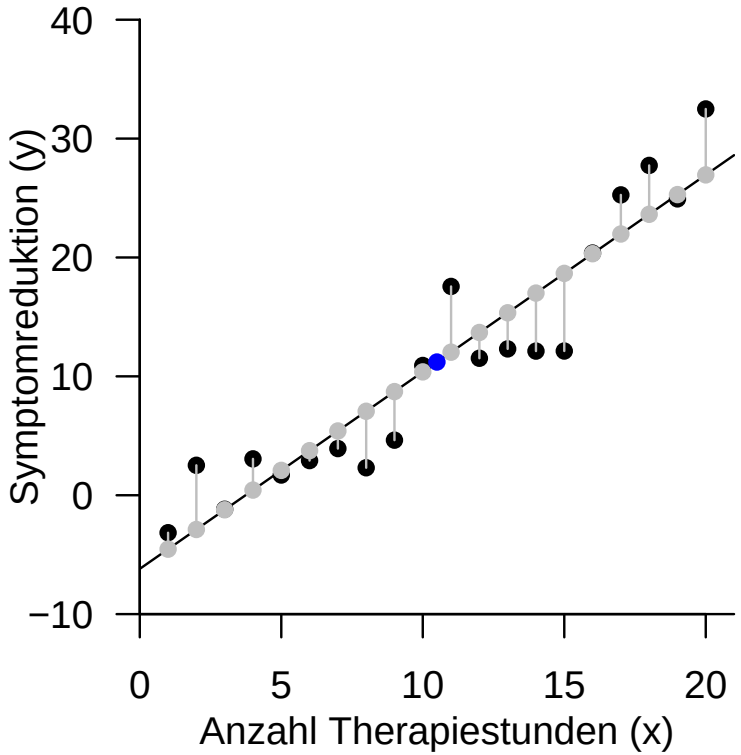
$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{11}$$

die durch die Ausgleichsgerade *erklärten Werte* genannt und

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

die *Residuen* der Ausgleichsgerade genannt.

Erklärte Werte und Residuen



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

Theorem (Quadratsummenzerlegung bei Ausgleichsgerade)

Für eine Wertemenge $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ und ihre 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 y-Werte und die durch die Ausgleichsgerade erklärten Werte, respektive. Weiterhin seien

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

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

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

Dann gilt

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

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- und y-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 von den beobachteten y-Werten

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

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

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

und damit auch

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

sowie

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

aus

Beweis (fortgeführt)

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

□

Definition (Bestimmtheitsmaß R^2)

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

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

Bestimmtheitsmaß oder *Determinationskoeffizient*.

Theorem (Stichprobenkorrelation und Bestimmtheitsmaß)

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

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

Korrelation und Bestimmtheitsmaß

Bemerkungen

- Mit $-1 \leq r_{xy} \leq 1$ folgt aus dem Theorem direkt, dass $0 \leq R^2 \leq 1$.
- 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 einer denkbar schlechten Erklärung der Daten durch die Ausgleichsgerade.
- Es gilt $R^2 = 1$ genau dann, wenn SQE $=$ SQT ist.

  $\Rightarrow$ Für $R^2 = 0$ ist also die Gesamtstreuung gleich der durch die Ausgleichsgerade erklärten Streuung.

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

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

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

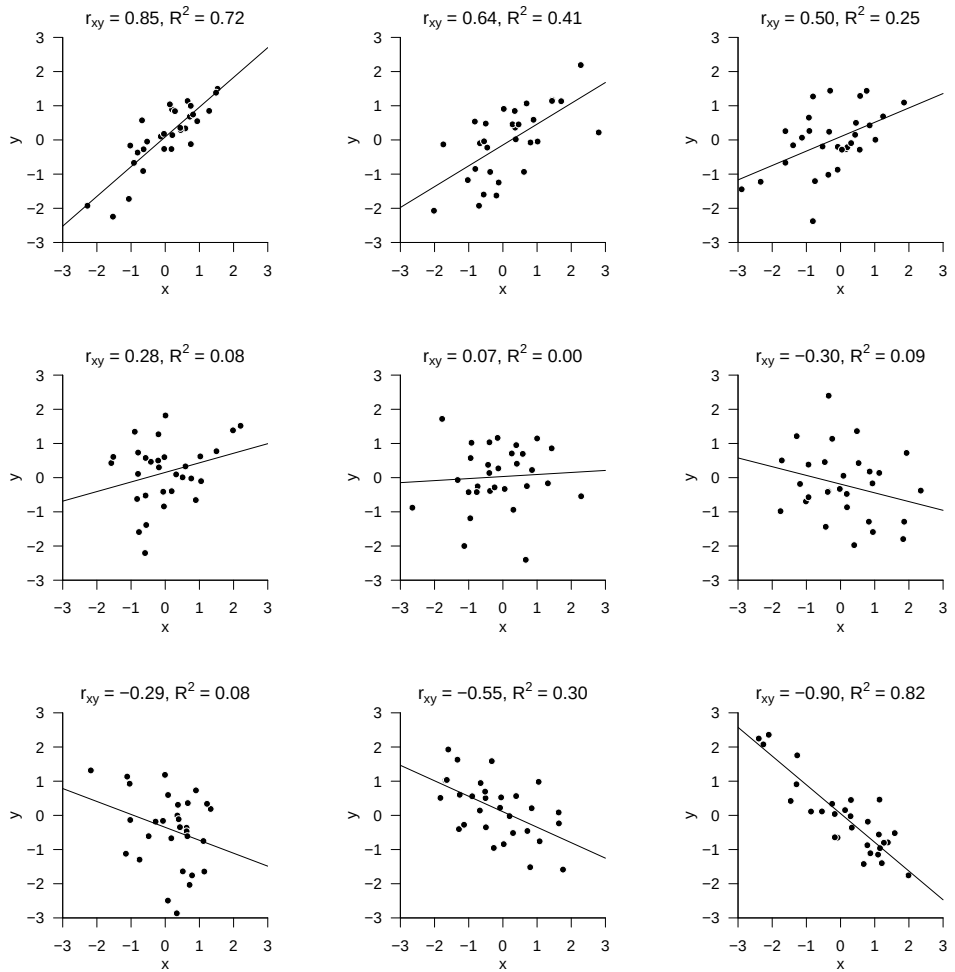
Beweis

Damit ergibt sich dann

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

□

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

Grundlagen

Korrelation und Bestimmtheitsmaß

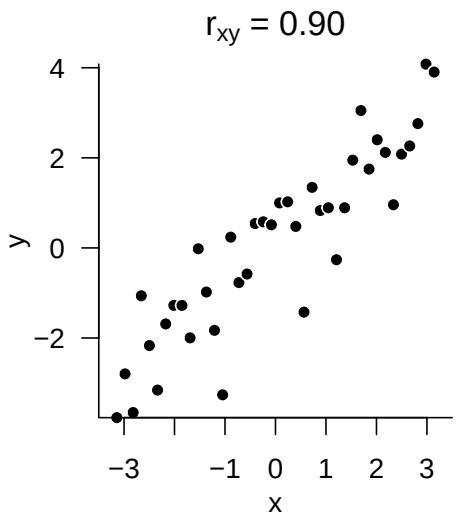
Korrelation und lineare Abhängigkeit

Korrelation und Regression

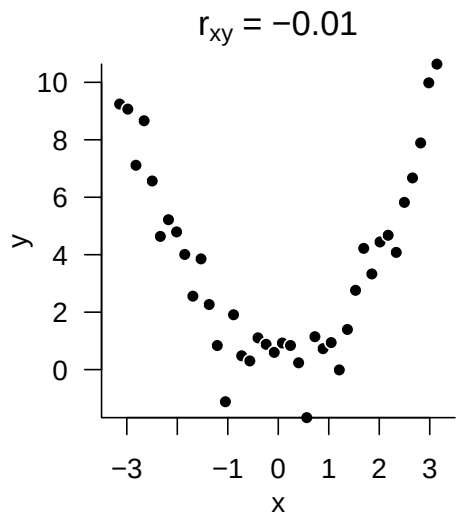
Partielle Korrelation

Selbstkontrollfragen

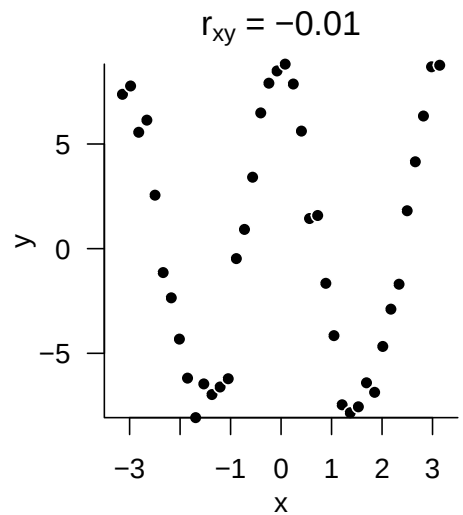
Funktionale Abhängigkeiten und Stichprobenkorrelation



$$y_i = x_i + \varepsilon_i$$



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



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

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

Theorem (Korrelation und linear-affine Abhängigkeit)

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

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

zwischen X und Y, wenn

$$\rho(X, Y) = 1 \text{ oder } \rho(X, Y) = -1 \tag{27}$$

gilt.

Bemerkungen

- Die lineare Abhängigkeit $Y = \beta_0 + \beta_1 X$ impliziert eine lineare Abhängigkeit $X = \tilde{\beta}_0 + \tilde{\beta}_1 Y$, denn

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

  mit

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

Beweis

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

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

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

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

Weiterhin gilt

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

Für die Kovarianz von X und Y ergibt sich also

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

Damit ergibt für die Korrelation von X und Y

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

Grundlagen

Korrelation und lineare Abhängigkeit

Korrelation und Regression

Korrelation und Bestimmtheitsmaß

Partielle Korrelation

Selbstkontrollfragen

Überblick

Der fundamentale Unterschied zwischen "Korrelation" und "Regression" ist, dass

- bei Korrelation sowohl die UV (die x's) als auch die AV (die y's) als Zufallsvariablen modelliert werden,
- bei Regression dagegen lediglich die AV als Zufallsvariable modelliert wird und die UV als vorgegeben gilt.

Dieser Tatsache unbenommen, kann man auf gegebene Daten prinzipiell natürlich sowohl "Korrelation" als auch "Regression" anwenden. Das Ergebnis einer Regressionsanalyse lässt sich in das Ergebnis einer Korrelationsanalyse umrechnen. Die zusätzlich Durchführung einer Korrelationsanalyse bei durchgeführter Regressionsanalyse erzeugt kein mehr an Information oder Verständnis über den Zusammenhang von UV und AV.

Für ein tieferes Verständnis dieser Zusammenhänge ist ein Regressionsmodell nötig, indem auch die UV eine Zufallsvariable ist. In Abgrenzung zum Modell der einfachen linearen Regression, in dem die UV keine Zufallsvariable ist, bezeichnen wir dieses Modell als *Regression*. Letztlich gerät die Terminologie hier an eine Grenze und es muss jeweils geprüft bzw. geschlossen werden, welches Modell Datenanalysten nun tatsächlich vorschwebt.

Definition (Regressionsgerade zweier Zufallsvariablen)

X und Y seien zwei Zufallsvariablen. Dann heißt

$$Y = \beta_0 + \beta_1 X \text{ mit} \tag{35}$$

mit

$$\beta_1 := \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)} \text{ und } \beta_0 := \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) \tag{36}$$

die *Regressionsgerade der Zufallsvariablen X auf Y*, β_0 und β_1 heißen die zugehörigen *Regressionskoeffizienten*, und die Zufallsvariable

$$E := Y - \beta_0 - \beta_1 X \tag{37}$$

heißt die *Residualvariable*.

Bemerkungen

- X und Y sind Zufallsvariablen, β_0 und β_1 sind keine Zufallsvariablen.

Theorem (Optimalität der Regressionsgerade zweier Zufallsvariablen)

Unter allen Geraden der Form

$$Y = \beta_0 + \beta_1 X \tag{38}$$

ist die Gerade mit

$$\beta_1 := \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)} \text{ und } \beta_0 := \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) \tag{39}$$

diejenige, für die

$$\tilde{q} : \mathbb{R}^2 \to \mathbb{R}, (\beta_0, \beta_1) \mapsto \tilde{q}(\beta_0, \beta_1) := \mathbb{E}\left((Y - (\beta_0 + \beta_1 X)^2\right) \tag{40}$$

ein Minimum hat.

Beweis

Wir halten zunächst fest, dass

$$\begin{aligned}\tilde{q}(\beta_0, \beta_1) &= \mathbb{E}\left(Y - \beta_0 - \beta_1 X\right) \\ &= \mathbb{E}\left(Y - \beta_1 X - \beta_0 + \beta\mathbb{E}(X) - \beta_1\mathbb{E}(X) + \mathbb{E}(Y) - \mathbb{E}(Y)\right) \\ &= \mathbb{E}\left((Y - \mathbb{E}(Y)) - \beta_1(X - \mathbb{E}(X)) + (\mathbb{E}(Y) - \beta_1\mathbb{E}(X) - \beta_0)\right)\end{aligned} \tag{41}$$

Ausmultiplizieren und Anwendung des Theorems zu den Eigenschaften des Erwartungswerts ergibt dann

$$\tilde{q}(\beta_0, \beta_1) = \mathbb{V}(Y) + \beta_1^2\mathbb{V}(X) - 2\beta_1\mathbb{C}(X, Y) + \left(\mathbb{E}(Y) - \beta_1\mathbb{E}(X) - \beta_0\right)^2 \tag{42}$$

Berechnen der partiellen Ableitungen von $\tilde{q}$ hinsichtlich von β_0 und β_1 ergibt dann

$$\frac{\partial}{\partial\beta_0}\tilde{q}(\beta_0, \beta_1) = -2\left(\mathbb{E}(Y) - \beta_1\mathbb{E}(X) - \beta_0\right) \tag{43}$$

und

$$\frac{\partial}{\partial\beta_1}\tilde{q}(\beta_0, \beta_1) = 2\beta_1\mathbb{V}(X) - 2\mathbb{C}(X, Y) - 2\mathbb{E}(X)\left(\mathbb{E}(Y) - \beta_1\mathbb{E} - \beta_0\right) \tag{44}$$

Nullsetzen von (43) ergibt dann als notwendige Bedingungen für ein Minimum von $\tilde{q}$

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\tilde{q}(\beta_0^*, \beta_1^*) = 0 \Leftrightarrow \mathbb{E}(Y) - \beta_1^*\mathbb{E}(X) - \beta_0^* = 0 \\ \frac{\partial}{\partial\beta_1}\tilde{q}(\beta_0^*, \beta_{1*}) = 0 \Leftrightarrow 2\beta_1^*\mathbb{V}(X) - 2\mathbb{C}(X, Y) - 2\mathbb{E}(X)\left(\mathbb{E}(Y) - \beta_1^*\mathbb{E} - \beta_0^*\right) = 0\end{aligned} \tag{45}$$

Die erste Gleichung impliziert dann für die zweite Gleichung, dass

$$2\beta_1^*\mathbb{V}(X) - 2\mathbb{C}(X, Y) = 0 \Leftrightarrow \beta_1^* = \frac{\mathbb{C}(X, Y)}{\mathbb{V}(X)} \tag{46}$$

Einsetzen in die erste Gleichung ergibt dann

$$\mathbb{E}(Y) - \beta_1^*\mathbb{E}(X) - \beta_0^* = 0 \Leftrightarrow \beta_0^* = \mathbb{E}(Y) - \beta_1^*\mathbb{E}(X) \tag{47}$$

Theorem (Zusammenhang von Korrelation und Regression)

X und Y seien zwei Zufallsvariablen,

$$Y = \beta_0 + \beta_1 X \text{ mit } \beta_1 := \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)} \text{ und } \beta_0 := \mathbb{E}(Y) - \tilde{\beta}_1 \mathbb{E}(X) \tag{48}$$

sei die Regressionsgerade der Zufallsvariablen Y bezüglich der Zufallsvariablen X mit den Regressionskoeffizienten β_0 und β_1 und

$$X = \tilde{\beta}_0 + \tilde{\beta}_1 Y \text{ mit } \tilde{\beta}_1 := \frac{\mathbb{C}(X,Y)}{\mathbb{V}(Y)} \text{ und } \tilde{\beta}_0 := \mathbb{E}(X) - \tilde{\beta}_1 \mathbb{E}(Y) \tag{49}$$

sei die Regressionsgerade der Zufallsvariablen X bezüglich der Zufallsvariablen Y mit den Regressionskoeffizienten $\tilde{\beta}_0$ und $\tilde{\beta}_1$. Dann gilt

$$\beta_1 \tilde{\beta}_1 = \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)} \frac{\mathbb{C}(X,Y)}{\mathbb{V}(Y)} = \frac{\mathbb{C}(X,Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} = \rho(X,Y)^2. \tag{50}$$

Bemerkungen

- $\rho(X,Y)$ kann aus den Regressionskoeffizienten von X auf Y und von Y auf X errechnet werden.

Definition (Stichprobenregressionsgerade)

$(X_1, Y_1), \ldots, (X_n, Y_n)$ sei eine Stichprobe von zweidimensionale Zufallsvektoren mit identischen unabhängigen Verteilungen. Weiterhin sei für $i = 1, \ldots, n$

$$Y_1 = \beta_0 + \beta_1 X_1 \text{ mit } \beta_1 := \frac{\mathbb{C}(X_1, Y_1)}{\mathbb{V}(X_1)} \text{ und } \beta_0 := \beta_1 \mathbb{E}(X_1) + \mathbb{E}(Y_1) \tag{51}$$

die Regressionsgerade der Zufallsvariablen Y_1 bezüglich der Zufallsvariablen X_1 mit den Regressionskoeffizienten β_0 und β_1. Schließlich seien

- $\bar{x}$ und $\bar{y}$ die Stichprobenmittel von Realisierungen der Komponenten der Stichprobe,
- s_X^2 und s_Y^2 die Stichprobenvarianzen von Realisierungen der Komponenten der Stichprobe und
- $c_{X,Y}$ die Stichprobenkovarianz von Realisierungen der Stichprobe.

Dann heißt für $x \in \mathbb{R}$

$$y = b_0 + b_1 x \text{ mit } b_1 := \frac{c_{X,Y}}{s_X^2} \text{ und } b_0 := \bar{y} - b_1 \bar{x} \tag{52}$$

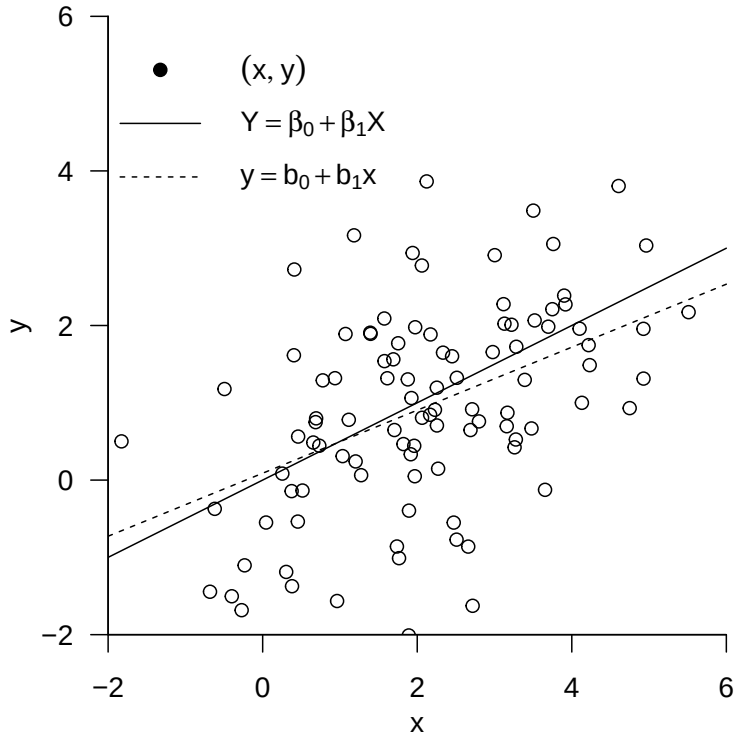
die Regressionsgerade der y-Werte bezüglich der x_i Werte in der Stichprobe.

Simulation einer Regressionsgerade

```
library(MASS)                                      # multivariate Normalverteilungen

# Modellformulierung
n       = 1e2                                       # Anzahl an Stichprobenvektoren
C_XY    = 1                                        # Kovarianz von X und Y
EX      = 2                                        # Erwartungswert von X
EY      = 1                                        # Erwartungswert von Y
VX      = 2                                        # Varianz von X
VY      = 2                                        # Varianz von Y
beta_1  = C_XY/VX                                  # Regressionskoeffizient
beta_0  = -beta_1*EX + EY                          # Regressionskoeffizient

# Realisierungsgeneration
mu      = c(EX, EY)                                # Erwartungswertparameter
Sigma   = matrix(c(VX, C_XY, C_XY,VY), nrow = 2) # Kovarianzmatrixparameter
xy      = mvrnorm(n, mu, Sigma)

# Stichprobenstatistiken
x_bar   = mean(xy[,1])                             # Stichprobenmittel  der x_1,...,x_n
y_bar   = mean(xy[,2])                             # Stichprobenmittel  der y_1,...,y_n
s2X     = var(xy[,1])                              # Stichprobenvarianz der x_1,...,x_n
s2Y     = var(xy[,2])                              # Stichprobenvarianz der y_1,...,y_n
c_xy    = cov(xy[,1],xy[,2])                       # Stichprobenkovarianz

# Stichprobenregressionsgeradenparameter
b_1     = c_xy/s2X                                 # Stichprobenregressionskoeffizient
b_0     = -b_1*x_bar + y_bar                       # Stichprobenregressionskoeffizient
```

```
> beta_0   : 0
> beta_1   : 0.5
> b_0      : 0.136
> b_1      : 0.514
```

Simulation einer Regressionsgerade



Grundlagen

Korrelation und lineare Abhängigkeit

Korrelation und Regression

Korrelation und Bestimmtheitsmaß

Partielle Korrelation

Selbstkontrollfragen

Grundlagen

Korrelation und lineare Abhängigkeit

Korrelation und Regression

Korrelation und Bestimmtheitsmaß

Partielle Korrelation

Selbstkontrollfragen

Grundlagen

Korrelation und Bestimmtheitsmaß

Korrelation und lineare Abhängigkeit

Korrelation und Regression

Partielle Korrelation

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. Berechnen Sie die Korrelation von Anzahl der Therapiestunden und Symptomreduktion anhand der Daten in Beispieldatensatz.csv.
5. Geben Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen wieder.
6. Erläutern Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen.
7. Geben Sie die Definitionen von erklärten Werten und Residuen einer Ausgleichsgerade wieder.
8. Geben Sie das Theorem zur Quadratsummenzerlegung bei einer Ausgleichsgerade wieder.
9. Erläutern Sie die intuitiven Bedeutungen von SQT, SQE und SQR.
10. Geben Sie die Definition des Bestimmtheitsmaßes R^2 wieder.
11. Geben Sie das Theorem zum Zusammenhang von Stichprobenkorrelation und Bestimmtheitsmaß wieder.
12. Erläutern Sie die Bedeutung von hohen und niedrigen R^2 Werten im Lichte der Ausgleichsgerade.
13. Berechnen Sie in einem R-Skript R^2 für die Daten in der Datei Beispieldatensatz.csv anhand der Definition von R^2. Überprüfen Sie Ihr Ergebnis anhand des Theorems zum Zusammenhang von Stichprobenkorrelation und Bestimmheitsmaß.
14. Geben Sie das Theorem zum Zusammenhang von Korrelation und linear-affiner Abhängigkeit wieder.
15. Geben Sie die Definition der Regressionsgerade zweier Zufallsvariablen wieder.
16. Geben Sie das Theorem zur Optimalität der Regressionsgerade zweier Zufallsvariablen wieder.
17. Geben Sie das Theorem zum Zusammenhang von Korrelation und Regression an.
18. Erläutern Sie, wie aus den Ergebnissen einer Regressionananlyse das Ergebnis einer Korrelationsanalyse errechnet werden kann.