Tutorium

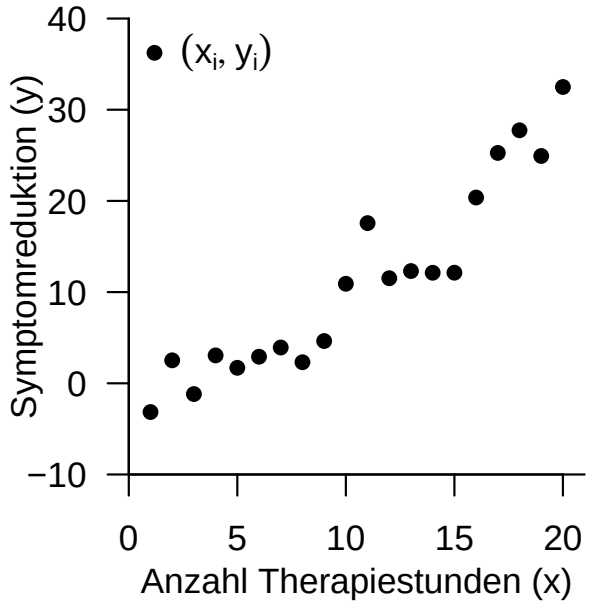
Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

4. Termin: (2) Korrelation

Sean Mulready

Mal angenommen, wir haben Daten von 20 Proband:innen erhoben. Dann haben wir als Ausgangspunkt die beobachteten Werte für y_i



Jetzt wollen wir die Daten beschreiben. Genauer gesagt, wollen wir beschreiben und quantifizieren, welcher Zusammenhang zwischen "Anzahl Therapiestunden" (x) und "Symptomreduktion" (y) bestehen könnte. Wie können wir vorgehen? Anders gefragt, welche Maße können wir (bisher) bestimmen?

Was können wir bestimmen, um den Zusammenhang quantitativ zu beschreiben?

- Ausgleichsgerade
- Einfache lineare Regression
- Korrelation und Bestimmtheitsmaß

Was können wir bestimmen, um den Zusammenhang quantitativ zu beschreiben?

Modell	Modellannahmen	Was wir auf Basis der beobachteten Daten (x_i, y_i) bestimmen können
Ausgleichsgerade	Die Ausgleichsgerade mit Funktionswerten $f_\beta(x) = \beta_0 + \beta_1 x_i$ minimiert die Summe der quadrierten Abweichungen $q(\beta)$	$\hat{\beta}_0, \hat{\beta}_1, q(\hat{\beta}_0, \hat{\beta}_1)$
Einfache lineare Regression	$\upsilon_i \sim N(\beta_o + \beta_1 x_i, \sigma^2)$, u.v. (Wir betrachten υ_i als Zufallsvariable mit Normalverteilung)	$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$
Korrelation	$\rho(\xi, \upsilon) := \frac{\mathbb{C}(\xi, \upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)}$ (Wir betrachten υ_i *und* ξ_i als Zufallsvariablen mit Varianzen $\mathbb{V}(\xi)$ und $\mathbb{V}(\upsilon)$ und Kovarianz $\mathbb{C}(\xi, \upsilon)$)	$r_{xy} := \frac{c_{xy}}{s_x s_y}$, R^2

Anmerkungen:

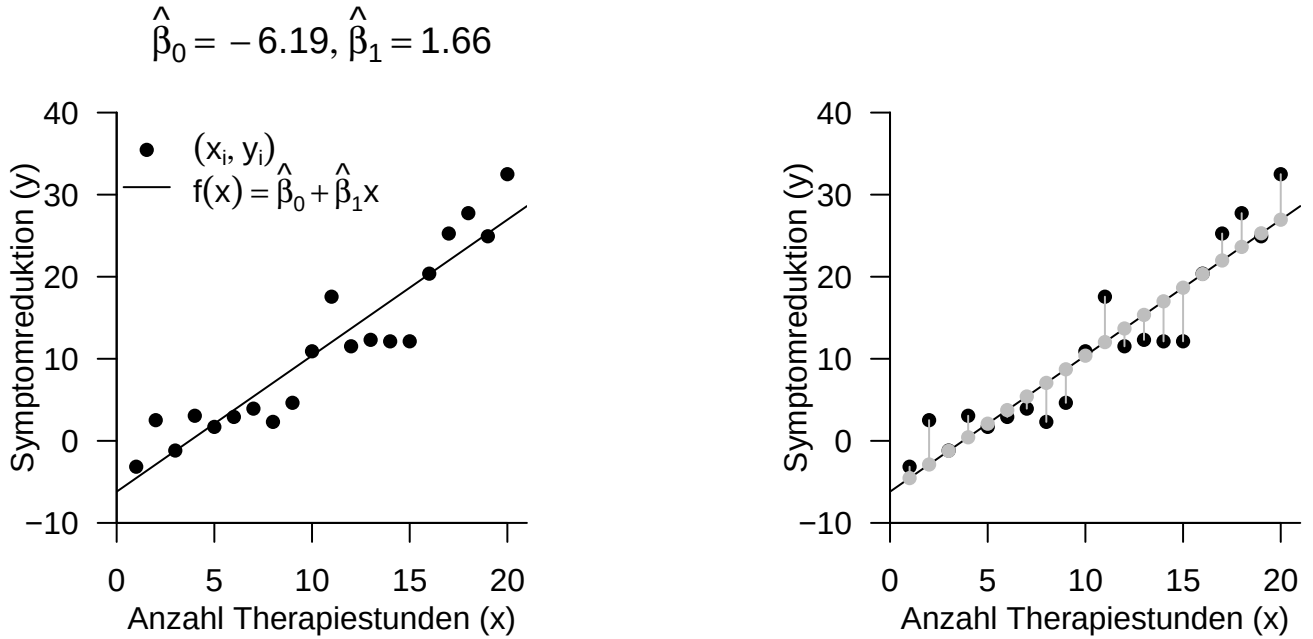
- Beobachtungen, Messungen oder eine Stichprobe sind konkret vorliegende Datenwerte, die eine Zufallsvariable annehmen kann. Wir nennen einzelne Werte, die eine Zufallsvariable annehmen kann, Realisierungen der Zufallsvariable.

Angewendet auf unseren Beispieldatensatz:

Parameterschätzer für Ausgleichsgeraden: $\hat{\beta}_0 = -6.2$, $\hat{\beta}_1 = 1.7$, $q(\hat{\beta}) = 250$

Parameterschätzer für einfachen linearen Regression: $\hat{\beta}_0 = -6.2$, $\hat{\beta}_1 = 1.7$, $\hat{\sigma}^2 = 3.54$

Korrelation: $r_{xy} = 0.938$, $R^2 = 0.88$



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

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.

1. Geben Sie die Definition der Korrelation zweier Zufallsvariablen wieder.

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.

- Aus Unabhängigkeit folgt auch Unkorreliertheit
- ABER: Aus Unkorreliertheit folgt nicht automatisch Unabhängigkeit!

2. Geben Sie die Definitionen von Stichprobenmittel, -standardabweichung, -kovarianz und -korrelation wieder.

Definition (Stichprobenkorrelation)

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^2$ 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 $(x_1, y_1), \ldots, (x_n, y_n)$ definiert als

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

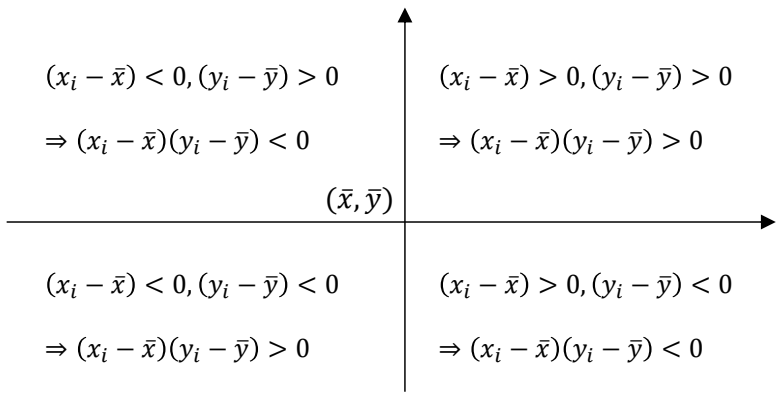
Dann ist die *Stichprobenkorrelation* der $(x_1, y_1), \ldots, (x_n, y_n)$ definiert als

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

und wird auch *Stichprobenkorrelationskoeffizient* genannt.

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.

Kovariationsterme: $(x_i - \bar{x})(y_i - \bar{y})$



- Die Stichprobenkorrelation ist die standardisierte Stichprobenkovarianz (c_{xy}).
- c_{xy} misst die insgesamte (aufsummierte) gemeinsame Abweichung der Beobachtungspunkte von ihren Stichprobenmitteln. Für jeden Beobachtungspunkt wird diese gemeinsame Abweichung als Produkt der Abweichung der x_i und y_i von den jeweiligen Stichprobenmitteln errechnet.
  - Wenn beide Beobachtungspunkte positiv, oder beide Beobachtungspunkte negativ, also richtungsgleich von ihren Mittelwerten abweichen, wird dieses Produkt positiv.
  - Wenn beide Beobachtungspunkte in konträre, oder richtungsungleiche Richtungen von ihren Mittelwerten abweichen, wird dieses Produkt negativ.
- Häufige Abweichungen der x_i und y_i von ihren Mittelwerten
  $\Rightarrow$ hohe absolute Korrelation
  - 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 Abweichungen der x_i und y_i von ihren Mittelwerten $\Rightarrow$ niedrige absolute Korrelation

4. Geben Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen wieder.

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

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

Dann gilt

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

5. Erläutern Sie das Theorem zur Stichprobenkorrelation bei linear-affinen Transformationen.

- Der Betrag der Stichprobenkorrelation ändert sich bei linear-affiner Datentransformation nicht.
- Man sagt, dass die Stichprobenkorrelation im Gegensatz zur Stichprobenkovarianz *maßstabsunabhängig* ist.
- Das heißt, der Betrag der Stichprobenkorrelation bleibt unverändert, wenn wir die Werte linear-affin transformieren (z.B. Stunden $\rightarrow$ Minuten, Grad Celcius $\rightarrow$ Grad Fahrenheit)

6. Geben Sie die Definitionen von erklärten Werten und Residuen einer Ausgleichsgerade wieder.

Definition (Erklärte Werte und Residuen einer Ausgleichsgerade)

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

$$f_{\hat{\beta}} : \mathbb{R} \to \mathbb{R}, x \mapsto f_{\hat{\beta}}(x) := \hat{\beta}_0 + \hat{\beta}_1 x \tag{8}$$

Dann werden für $i = 1, ..., n$

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{9}$$

die durch die Ausgleichsgerade *erklärten Werte* genannt und

$$\hat{\varepsilon}_i := y_i - \hat{y}_i \tag{10}$$

die *Residuen* der Ausgleichsgerade genannt.

7. Geben Sie das Theorem zur Quadratsummenzerlegung bei einer Ausgleichsgerade wieder.

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

$$\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i \text{ und } \hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, \text{ für } i = 1, ..., n \tag{11}$$

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

$$\text{SQT} = \text{SQE} + \text{SQR} \tag{12}$$

8. Erläutern Sie die intuitiven Bedeutungen von SQT, SQE und SQR.

- 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.

  $\Rightarrow$ 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.

9. Geben Sie die Definition des Bestimmtheitsmaßes R^2 wieder.

Definition (Bestimmtheitsmaß R^2)

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

$$R^2 := \frac{SQE}{SQT} \tag{13}$$

Bestimmtheitsmaß oder *Determinationskoeffizient*.

10. Geben Sie das Theorem zum Zusammenhang von Stichprobenkorrelation und Bestimmtheitsmaß wieder.

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

$$R^2 = r_{xy}^2. \tag{14}$$

11. Erläutern Sie die Bedeutung von hohen und niedrigen R^2 Werten im Lichte der Ausgleichsgerade.

- 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 $R^2 = 1$ 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.
- Man sagt, dass "R^2 die durch die Ausgleichsgerade erklärte Varianz an der Gesamtvarianz ist.

12. Geben Sie das Theorem zum Zusammenhang von Korrelation und linear-affiner Abhängigkeit wieder.

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

$$\upsilon = \beta_0 + \beta_1 \xi \text{ mit } \beta_0, \beta_1 \in \mathbb{R} \tag{15}$$

zwischen ξ und υ, wenn

$$\rho(\xi, \upsilon) = 1 \text{ oder } \rho(\xi, \upsilon) = -1 \tag{16}$$

gilt.