



Tutorium Allgemeines Lineares Modell

BSc Psychologie SoSe 2022

4. Termin: Korrelation

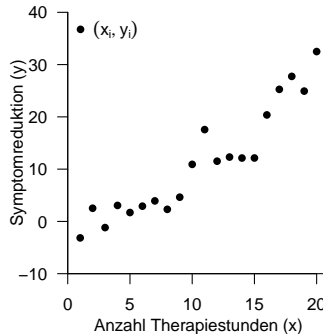
Belinda Fleischmann

Ersatztermin für 16. Juni

- Vorziehen auf Montag, 13. Juni nachmittags?
- oder Doppelstunde am Donnerstag (23. Juni) die Woche drauf?
- Alternativen?

Follow-up letzte Wochen

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

Follow-up letzte Wochen

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	$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, u. (Wir betrachten Y_i als Zufallsvariable mit Normalverteilung)	$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$
Korrelation	$\rho(X, Y) := \frac{C(X, Y)}{S(X)S(Y)}$ (Wir betrachten Y_i und X_i als Zufallsvariablen mit Varianzen $V(X)$ und $V(Y)$ und Kovarianz $C(X, Y)$)	$r_{xy} := \frac{c_{xy}}{s_x s_y}, R^2$

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.

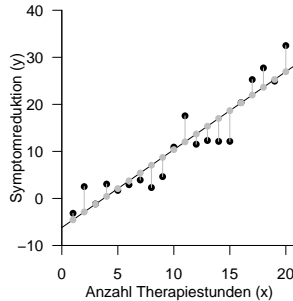
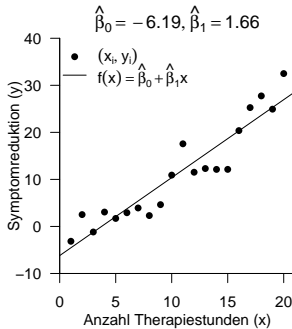
Follow-up letzte Wochen

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



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

Selbstkontrollfragen - Regression

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 Bestimmtheitsmaß.
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 Regressionanalyse das Ergebnis einer Korrelationsanalyse errechnet werden kann.

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

Definition (Korrelation)

Die *Korrelation* zweier Zufallsvariablen X und Y ist definiert als

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

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

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

Definition (Stichprobenkorrelation)

$\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}$ sei eine Wertemenge. Weiterhin seien:

- Die Stichprobenmittel der x_i und y_i definiert als

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

- Die Stichprobenstandardabweichungen x_i und y_i definiert als

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

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

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

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

$$r_{xy} := \frac{c_{xy}}{s_x s_y} \quad (5)$$

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

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

- Die Stichprobenkorrelation ist die standardisierte Stichprobenkovarianz (c_{xy}).
- c_{xy} misst die insgesamt (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 (richtungsgleich oder richtungsungleich)
 - \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. Berechnen Sie die Korrelation von Anzahl der Therapiestunden und Symptomreduktion anhand der Daten in Beispieldatensatz.csv.

```
# Laden des Beispieldatensatzes
fname = file.path(getwd(), "Daten", "Beispieldatensatz.csv")
D      = read.table(fname, sep = ",", header = TRUE)
x_i    = D$x_i                      # x_i Werte
y_i    = D$y_i                      # y_i Werte
n      = length(x_i)                # n (= length(y_i))

# "Manuelle" Berechnung der Stichprobenkorrelation
x_bar  = (1/n)*sum(x_i)              # Stichprobenmittel x
y_bar  = (1/n)*sum(y_i)              # Stichprobenmittel y
s_x    = sqrt(1/(n-1)*sum((x_i - x_bar)^2)) # Stichprobenstandardabweichung x
s_y    = sqrt(1/(n-1)*sum((y_i - y_bar)^2)) # Stichprobenstandardabweichung y
c_xy   = 1/(n-1) * sum((x_i - x_bar) * (y_i - y_bar)) # Stichprobenkovarianz
r_xy_man = c_xy/(s_x * s_y)          # Stichprobenkorrelation

# Automatische Berechnung mit cor()
r_xy_aut = cor(x_i, y_i)             # Stichprobenkorrelation

# Ausgabe
cat("r (manuell)" , r_xy_man,
    "\n r (automatisch)", r_xy_aut)

> r (manuell) 0.938
> r (automatisch) 0.938
```

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

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

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

Dann gilt

$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \quad (7)$$

6. 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)

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

Definition (Erklärte Werte und Residuen einer Ausgleichsgerade)

Gegeben seien eine Wertemenge $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ und die zu dieser Wertemenge gehörende Ausgleichsgerade

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

Dann werden für $i = 1, \dots, n$

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (9)$$

die durch die Ausgleichsgerade *erklärten Werte* genannt und

$$\hat{\varepsilon}_i := y_i - \hat{y}_i \quad (10)$$

die *Residuen* der Ausgleichsgerade genannt.

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

Theorem (Quadratsummenzerlegung bei Ausgleichsgerade)

Für eine Wertemenge $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ und ihre zugehörige Ausgleichsgerade $f_{\hat{\beta}}$ seien für

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

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

$$\begin{aligned} \text{SQT} &:= \sum_{i=1}^n (y_i - \bar{y})^2 && \text{die Total Sum of Squares} \\ \text{SQE} &:= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 && \text{die Explained Sum of Squares} \\ \text{SQR} &:= \sum_{i=1}^n (y_i - \hat{y}_i)^2 && \text{die Residual Sum of Squares} \end{aligned}$$

Dann gilt

$$\text{SQT} = \text{SQE} + \text{SQR} \quad (12)$$

9. 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
 - ⇒ Große Werte von SQE repräsentieren eine große absolute Steigung der y_i mit den x_i
 - ⇒ 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.
 - ⇒ Große Werte von SQR repräsentieren große Abweichungen der erklärten von den beobachteten y -Werten
 - ⇒ 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.

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

Definition (Bestimmtheitsmaß R^2)

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

$$R^2 := \frac{\text{SQE}}{\text{SQT}} \quad (13)$$

Bestimmtheitsmaß oder Determinationskoeffizient.

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

Theorem (Stichprobenkorrelation und Bestimmtheitsmaß)

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

$$R^2 = r_{xy}^2. \quad (14)$$

12. 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.

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 Bestimmtheitsmaß.

Teil 1/2

```
# Laden des Beispieldatensatzes
fname = file.path(getwd(), "Daten", "Beispieldatensatz.csv")
D      = read.table(fname, sep = ",", header = TRUE)
n      = length(D$x_i)                # n (= length(y_i))

# Stichprobenstatistiken (incl. Korrelation)
x_bar  = (1/n)*sum(D$x_i)              # Stichprobenmittel x
y_bar  = (1/n)*sum(D$y_i)              # Stichprobenmittel y
s_x    = sqrt(1/(n-1)*sum((D$x_i - x_bar)^2)) # Stichprobenstandardabweichung x
s_y    = sqrt(1/(n-1)*sum((D$y_i - y_bar)^2)) # Stichprobenstandardabweichung y
c_xy   = 1/(n-1) * sum((D$x_i - x_bar) * (D$y_i - y_bar)) # Stichprobenkovarianz
r_xy   = c_xy/(s_x * s_y)
```

13. Berechnen Sie in einem R-Skript R^2 für die Daten in der Datei Beispieldatensatz.csv ...

Teil 2/2

```
# Erklärte Werte der Ausgleichsgeraden (y_hat)
beta_1_hat = c_xy / s_x^2          # beta_0 (Parameter der Ausgleichsgerade)
beta_0_hat = y_bar - beta_1_hat * x_bar # beta_1 (Parameter der Ausgleichsgerade)
D$y_hat = beta_0_hat + beta_1_hat * D$x_i # Funktionswerte der Ausgleichsgeraden/erklärte Werte (y_hat)

# Berechnung R^2 nach Definition
SQT = sum((D$y_i - y_bar)^2)        # Total sum of squares
SQE = sum((D$y_hat - y_bar)^2)      # Explained sum of squares
R_sqr_def = SQE/SQT                # Bestimmtheitsmaß

# Berechnung R^2 über Stichprobenkorrelationskoeffizienten
R_sqr = r_xy^2                     # Bestimmtheitsmaß

# Ausgabe
cat("r_xy", r_xy, "\n SQT", SQT, "\n SQE", SQE,
    "\n R_sqr wie Def", R_sqr_def, "\n R_sqr mit r_xy", R_sqr)

> r_xy 0.938
> SQT 2076
> SQE 1826
> R_sqr wie Def 0.879
> R_sqr mit r_xy 0.879
```

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

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

$$Y = \beta_0 + \beta_1 X \text{ mit } \beta_0, \beta_1 \in \mathbb{R} \quad (15)$$

zwischen X und Y , wenn

$$\rho(X, Y) = 1 \text{ oder } \rho(X, Y) = -1 \quad (16)$$

gilt.

15. Geben Sie die Definition der Regressionsgerade zweier Zufallsvariablen wieder.

Definition (Regressionsgerade zweier Zufallsvariablen)

X und Y seien zwei Zufallsvariablen. Dann heißt

$$Y = \beta_0 + \beta_1 X \text{ mit} \quad (17)$$

mit

$$\beta_1 := \frac{\mathbb{C}(X, Y)}{\mathbb{V}(X)} \text{ und } \beta_0 := \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) \quad (18)$$

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

$$E := Y - \beta_0 - \beta_1 X \quad (19)$$

heißt die *Residualvariable*.

16. Geben Sie das Theorem zur Optimalität der Regressionsgerade zweier Zufallsvariablen wieder.

Theorem (Optimalität der Regressionsgerade zweier Zufallsvariablen)

Unter allen Geraden der Form

$$Y = \beta_0 + \beta_1 X \quad (20)$$

ist die Gerade mit

$$\beta_1 := \frac{\mathbb{C}(X, Y)}{\mathbb{V}(X)} \text{ und } \beta_0 := \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) \quad (21)$$

diejenige, für die

$$\tilde{q} : \mathbb{R}^2 \rightarrow \mathbb{R}, (\beta_0, \beta_1) \mapsto \tilde{q}(\beta_0, \beta_1) := \mathbb{E} \left((Y - (\beta_0 + \beta_1 X))^2 \right) \quad (22)$$

ein Minimum hat.

17. Geben Sie das Theorem zum Zusammenhang von Korrelation und Regression an.

Theorem (Zusammenhang von Korrelation und Regression)

X und Y seien zwei Zufallsvariablen,

$$Y = \beta_0 + \beta_1 X \text{ mit } \beta_1 := \frac{\mathbb{C}(X, Y)}{\mathbb{V}(X)} \text{ und } \beta_0 := \mathbb{E}(Y) - \tilde{\beta}_1 \mathbb{E}(X) \quad (23)$$

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

$$X = \tilde{\beta}_0 + \tilde{\beta}_1 Y \text{ mit } \tilde{\beta}_1 := \frac{\mathbb{C}(X, Y)}{\mathbb{V}(Y)} \text{ und } \tilde{\beta}_0 := \mathbb{E}(X) - \tilde{\beta}_1 \mathbb{E}(Y) \quad (24)$$

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

$$\beta_1 \tilde{\beta}_1 = \frac{\mathbb{C}(X, Y)}{\mathbb{V}(X)} \frac{\mathbb{C}(X, Y)}{\mathbb{V}(Y)} = \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} = \rho(X, Y)^2. \quad (25)$$

18. Erläutern Sie, wie aus den Ergebnissen einer Regressionanalyse das Ergebnis einer Korrelationsanalyse errechnet werden kann.

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

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.