Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

Prof. Dr. Dirk Ostwald

(12) Partielle Korrelation

Motivation

Bedingte Korrelation

Partielle Korrelation

Selbstkontrollfragen

Motivation

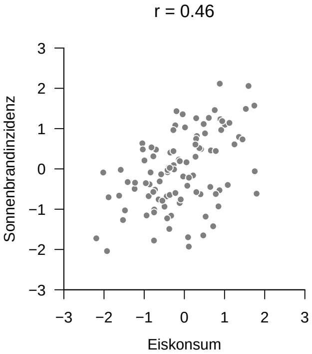
Bedingte Korrelation

Partielle Korrelation

Selbstkontrollfragen

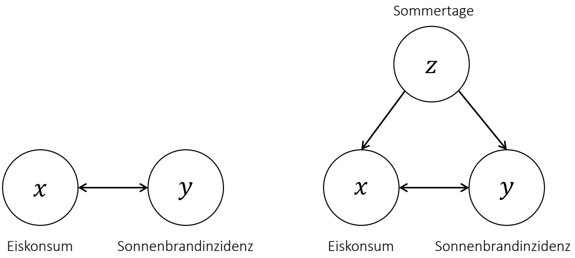
Motivation

Jährlicher Eiskonsum und jährliche Sonnenbrandinzidenz



- Korrelation impliziert keine Kausalität.
- Kausalität wird zumeist als Koinzidenz mit zeitlicher Rangfolge modelliert.
- Einstiege in die kausale Inferenz geben z.B. Pearl (2000) und Imbens and Rubin (2015).

Jährlicher Eiskonsum und jährliche Sonnenbrandinzidenz



- Korrelation von Eiskonsum und Sonnenbrandinzidenz nach Korrektur für Sommertage?
- "Herausrechnen" des Einflusses von z auf die Kovariation von x und y?
- In diesem Abschnitt bezeichnen wir Zufallsvariablen mit kleinen lateinischen Buchstaben!

$\Rightarrow$ Bedingte Korrelation und Partielle Korrelation im Falle dreier Zufallsvariablen.

Motivation

Bedingte Korrelation

Partielle Korrelation

Selbstkontrollfragen

Bedingte Korelation

Definition (Bedingte Kovarianz und bedingte Korrelation)

Gegeben seien drei Zufallsvariablen x, y, z einer gemeinsamen Verteilung $\mathbb{P}_{x,y,z}(x, y, z)$. Weiterhin sei $\mathbb{P}_{x,y|z}(x, y)$ die bedingte Verteilung von x und y gegeben z. Dann heißt die Kovarianz von x und y in der Verteilung $\mathbb{P}_{x,y|z}(x, y)$ die *bedingte Kovarianz von x und y gegeben z* und wird mit $\mathbb{C}(x, y|z)$ bezeichnet. Weiterhin seien $\mathbb{P}_{x,y|z}(x)$ und $\mathbb{P}_{x,y|z}(y)$ die marginalen Verteilungen von x und y gegeben z, respektive, und $\mathbb{S}(x|z)$ und $\mathbb{S}(y|z)$ seien die Standardabweichungen von x und y hinsichtlich $\mathbb{P}_{x,y|z}(x)$ und $\mathbb{P}_{x,y|z}(y)$, respektive. Dann heißt die Korrelation von x und y in der Verteilung $\mathbb{P}_{x,y|z}(x, y)$,

$$\rho(x, y|z) := \frac{\mathbb{C}(x, y|z)}{\mathbb{S}(x|z)\mathbb{S}(y|z)} \tag{1}$$

die *bedingte Korrelation von x und y gegeben z*

Bemerkungen

- Die bedingte Kovarianz zweier ZVen ist die Kovarianz zweier ZVen in einer bedingten Verteilung
- Die bedingte Korrelation zweier ZVen ist die Korrelation zweier ZVen in einer bedingten Verteilung.
- Durch Vertauschen der Variablennamen kann man analog $\rho(y, z|x)$ und $\rho(x, z|y)$ definieren.

Beispiel

Die Zufallsvariablen x, y, z seien multivariat normalverteilt. Wir wollen die bedingte Korrelation von x und y gegeben z bestimmen. Für $v := (x, y, z)^T$ gelte also, dass

$$v \sim N(\mu, \Sigma) \tag{2}$$

mit

$$\mu := \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix} \text{ und } \Sigma := \begin{pmatrix} \sigma_x^2 & \sigma_{x,y}^2 & \sigma_{x,z}^2 \\ \sigma_{y,x}^2 & \sigma_y^2 & \sigma_{y,z}^2 \\ \sigma_{z,x}^2 & \sigma_{z,y}^2 & \sigma_z^2 \end{pmatrix} \tag{3}$$

Um die Kovarianzmatrix der bedingten Verteilung von x und y gegeben z zu bestimmen, definieren wir zunächst

$$\Sigma_{x,y} := \begin{pmatrix} \sigma_x^2 & \sigma_{x,y}^2 \\ \sigma_{y,x}^2 & \sigma_y^2 \end{pmatrix}, \Sigma_z := \begin{pmatrix} \sigma_z^2 \end{pmatrix} \text{ und } \Sigma_{(x,y),z} := \Sigma_{z,(x,y)}^T := \begin{pmatrix} \sigma_{x,z}^2 \\ \sigma_{y,z}^2 \end{pmatrix}, \tag{4}$$

so dass

$$\Sigma = \begin{pmatrix} \Sigma_{x,y} & \Sigma_{(x,y),z} \\ \Sigma_{z,(x,y)} & \Sigma_z \end{pmatrix} \tag{5}$$

Mit dem Theorem zu bedingten Normalverteilungen (vgl. (4) Normalverteilungen) ist dann die Kovarianzmatrix des Zufallsvektors (x, y) gegeben durch

$$\Sigma_{x,y|z} = \Sigma_{x,y} - \Sigma_{(x,y),z}\Sigma_z^{-1}\Sigma_{z,(x,y)}. \tag{6}$$

Beispiel (fortgeführt)

Mit den Eigenschaften der multivariaten Normalverteilung gilt dann, dass die Diagonaleinträge von $\Sigma_{x,y|z}$ den bedingten Varianzen von x und y gegeben z entsprechen und dass der Nichtdiagonaleintrag die bedingte Kovarianz von x und y gegeben z ist. In anderen Worten gilt

$$\Sigma_{x,y|z} = \begin{pmatrix} \mathbb{C}(x,x|z) & \mathbb{C}(x,y|z) \\ \mathbb{C}(y,x|z) & \mathbb{C}(y,y|z) \end{pmatrix}. \tag{7}$$

Die bedingte Korrelation $\rho(x,y|z)$ von x und y gegeben z ergibt sich dann aus den Einträgen von $\Sigma_{x,y|z}$ gemäß

$$\rho(x,y|z) = \frac{\mathbb{C}(x,y|z)}{\sqrt{\mathbb{C}(x,x|z)}\sqrt{\mathbb{C}(y,y|z)}} \tag{8}$$

Für

$$\Sigma := \begin{pmatrix} 1.0 & 0.5 & 0.9 \\ 0.5 & 1.0 & 0.5 \\ 0.9 & 0.5 & 1.0 \end{pmatrix} \tag{9}$$

ergibt sich beispielsweise

$$\rho(x,y) = 0.50 \text{ und } \rho(x,y|z) \approx 0.13. \tag{10}$$

Beispiel (fortgeführt)

```r
# Bedingte Korrelation bei Normalverteilung
S         = matrix(c( 1,.5,.9,                                   # \Sigma
                     .5, 1,.5,
                     .9,.5, 1), nrow  = 3, byrow = TRUE)
rho_xy    = S[1,2]/(sqrt(S[1,1])*sqrt(S[2,2]))                   # \rho(x,y)
S_xy_z    = S[1:2,1:2] -  S[1:2,3] %*% solve(S[3,3]) %*%S[3,1:2] # \Sigma_{x,y|z}
rho_xy_z  = S_xy_z[1,2]/(sqrt(S_xy_z[1,1])*sqrt(S_xy_z[2,2]))    # \rho(x,y|z)

# Ausgabe
cat("rho(x,y)   :"  , rho_xy,
    "\nrho(x,y|z) :", rho_xy_z)
```

```
> rho(x,y)   : 0.5
> rho(x,y|z) : 0.132
```

Theorem (Bedingte Korrelation und Korrelationen bei Normalverteilung)

x, y, z seien drei gemeinsam multivariat normalverteilte Zufallsvariablen. Dann gilt

$$\rho(x, y|z) = \frac{\rho(x, y) - \rho(x, z)\rho(y, z)}{\sqrt{(1 - \rho(x, z)^2)}\sqrt{(1 - \rho(y, z)^2)}} \tag{11}$$

Bemerkungen

- $\rho(x, y|z)$ kann bei Normalverteilung aus den Korrelationen $\rho(x, y), \rho(x, z), \rho(y, z)$ berechnet werden.
- Ein entsprechender Schätzer für $\rho(x, y|z)$ ergibt sich mit den Stichprobenkorrelationen $r_{x,y}, r_{x,z}, r_{y,z}$ als

$$r_{x,y|z} = \frac{r_{x,y} - r_{x,z}r_{y,z}}{\sqrt{(1 - r_{x,z}^2)}\sqrt{(1 - r_{y,z}^2)}} \tag{12}$$

- Mit $\rho(x, y|z) = \rho(x, y \setminus z)$ bei Normalverteilung gilt die Formel auch für die partielle Korrelation.

Bedingte Korrelation

<u>Beweis</u>

Ohne Beschränkung der Allgemeinheit betrachten wir den Fall eines standardisierten multivariate normalverteilten Zufallsvektors $v := (x, y, z)^T$ mit Kovarianzmatrixparameter

$$\Sigma := \begin{pmatrix} 1 & \rho(x,y) & \rho(x,z) \\ \rho(y,x) & 1 & \rho(y,z) \\ \rho(z,x) & \rho(z,y) & 1 \end{pmatrix}. \tag{13}$$

Wir definieren nun zunächst

$$\Sigma_{x,y} := \begin{pmatrix} 1 & \rho(x,y) \\ \rho(y,x) & 1 \end{pmatrix}, \Sigma_z := \begin{pmatrix} 1 \end{pmatrix} \text{ und } \Sigma_{(x,y),z} := \Sigma_{z,(x,y)}^T := \begin{pmatrix} \rho(x,z) \\ \rho(y,z) \end{pmatrix}, \tag{14}$$

so dass

$$\Sigma = \begin{pmatrix} \Sigma_{x,y} & \Sigma_{(x,y),z} \\ \Sigma_{z,(x,y)} & \Sigma_z \end{pmatrix}. \tag{15}$$

Mit dem Theorem zu bedingten Normalverteilungen (vgl. (4) Normalverteilungen) ist dann die Kovarianzmatrix des Zufallsvektors (x, y) gegeben durch

$$\Sigma_{x,y|z} = \Sigma_{x,y} - \Sigma_{(x,y),z}\Sigma_z^{-1}\Sigma_{z,(x,y)}. \tag{16}$$

Bedingte Korrelation

Beweis (fortgeführt)

Es ergibt sich also

$$
\begin{aligned}
\begin{pmatrix} \sigma^2_{x,x|z} & \sigma^2_{x,y|z} \\ \sigma^2_{y,x|z} & \sigma^2_{y,y|z} \end{pmatrix}
&= \begin{pmatrix} 1 & \rho(x,y) \\ \rho(y,x) & 1 \end{pmatrix} - \begin{pmatrix} \rho(x,z) \\ \rho(y,z) \end{pmatrix} \begin{pmatrix} 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho(x,z) & \rho(y,z) \end{pmatrix} \\
&= \begin{pmatrix} 1 & \rho(x,y) \\ \rho(y,x) & 1 \end{pmatrix} - \begin{pmatrix} \rho(x,z)\rho(x,z) & \rho(x,z)\rho(y,z) \\ \rho(y,z)\rho(x,z) & \rho(y,z)\rho(y,z) \end{pmatrix} \\
&= \begin{pmatrix} 1 - \rho(x,z)^2 & \rho(x,y) - \rho(x,z)\rho(y,z) \\ \rho(y,x) - \rho(y,z)\rho(x,z) & 1 - \rho(y,z)^2 \end{pmatrix}.
\end{aligned}
\tag{17}
$$

Damit folgt dann direkt

$$
\rho(x,y|z) = \frac{\sigma^2_{x,y|z}}{\sqrt{\sigma^2_{x,x|z}}\sqrt{\sigma^2_{y,y|z}}} = \frac{\rho(x,y) - \rho(x,z)\rho(y,z)}{\sqrt{1-\rho(x,z)^2}\sqrt{1-\rho(y,z)^2}}.
\tag{18}
$$

□

Motivation

Bedingte Korrelation

Partielle Korrelation

Selbstkontrollfragen

Partielle Korrelation

Definition (Partielle Korrelation)

x, y, z seien Zufallsvariablen mit linear-affinen Abhängigkeiten zwischen x und z sowie zwischen y und z,

$$
\begin{aligned}
x &:= \beta_0^{x,z} + \beta_1^{x,z} z, \\
y &:= \beta_0^{y,z} + \beta_1^{y,z} z
\end{aligned}
\tag{19}
$$

mit Residualvariablen

$$
\begin{aligned}
e^{x,z} &:= x - \beta_0^{x,z} - \beta_1^{x,z} z, \\
e^{y,z} &:= y - \beta_0^{y,z} - \beta_1^{y,z} z.
\end{aligned}
\tag{20}
$$

Dann ist die *partielle Korrelation von x und y mit auspartialisiertem z* definiert als

$$
\rho(x, y \setminus z) := \rho(e^{x,z}, e^{y,z}). \tag{21}
$$

Bemerkungen

- $e^{x,z}$ ist die Zufallsvariable x, aus der der Einfluss von z "herausgerechnet" wurde.
- $e^{y,z}$ ist die Zufallsvariable y, aus der der Einfluss von z "herausgerechnet" wurde.
- $\rho(x, y \setminus z)$ ist also die Korrelation von x und y, aus denen jeweils der Einfluss von z "herausgerechnet" wurde.

Definition (Partielle Stichprobenkorrelation)

x, y, z seien Zufallsvariablen mit linear-affinen Abhängigkeiten zwischen y und z sowie zwischen x und z wie in der Definition der partiellen Korrelation. Weiterhin seien

- $\{(x_i, y_i, z_i)\}_{i=1,\dots,n}$ eine Menge von Realisierungen des Zufallsvektors $(x, y, z)^T$,
- $\hat{\beta}_0^{x,z}, \hat{\beta}_1^{x,z}$ die Ausgleichsgeradenparameter für $\{(x_i, z_i)\}_{i=1,\dots,n}$,
- $\hat{\beta}_0^{y,z}, \hat{\beta}_1^{y,z}$ die Ausgleichsgeradenparameter für $\{(y_i, z_i)\}_{i=1,\dots,n}$.

Schließlich seien für $i = 1, \dots, n$

- $e_i^{x,z} := x_i - \hat{\beta}_0^{x,z} - \hat{\beta}_1^{x,z} z_i$
- $e_i^{y,z} := y_i - \hat{\beta}_0^{y,z} - \hat{\beta}_1^{y,z} z_i$

die Residualwerte der jeweiligen Ausgleichsgeraden. Dann heißt die Stichprobenkorrelation der Wertemenge $\{(e_i^{y,z}, e_i^{x,z})\}_{i=1,\dots,n}$ *partielle Stichprobenkorrelation der x_i und y_i mit auspartialisierten z_i.*

Bemerkungen

- Die partielle Stichprobenkorrelation wird als Schätzer der partiellen Korrelation genutzt.

Theorem (Bedingte und Partielle Korrelation bei Normalverteilung)

x, y, z seien drei gemeinsam multivariat normalverteilte Zufallsvariablen. Dann gilt

$$\rho(x, y|z) = \rho(x, y \setminus z) \tag{22}$$

Bemerkungen

- Wir verzichten auf einen Beweis.
- Generell sind bedingte und partielle Korrelationen nicht identisch.
- Für Details, siehe zum Beispiel Lawrance (1976) und Baba, Shibata, and Sibuya (2004).

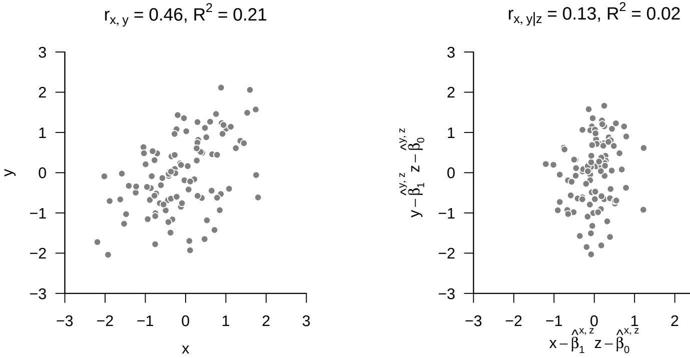
Partielle Korrelation

Beispiel

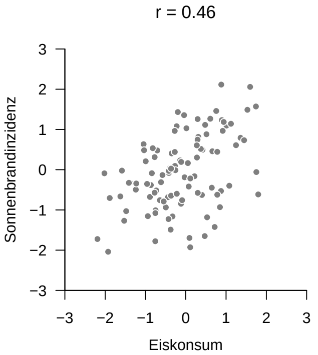
```
# Modellformulierung und Datenrealisierung
library(MASS)                                      # Multivariate Normalverteilung
set.seed(1)                                        # reproduzierbare Daten
S     = matrix(c( 1,.5,.9,                         # Kovarianzmatrixparameter \Sigma
                 .5, 1,.5,
                 .9,.5, 1),nrow=3,byrow=TRUE)
n     = 1e6                                        # Anzahl Realisierungen  von v := (x,y,z)^T
xyz   = mvrnorm(n,rep(0,3),S)                      # Realisierungen         von v := (x,y,z)^T

# Partielle Stichprobenkorrelation als Residualstichprobenkorrelation
bars  = apply(xyz, 2, mean)                        # Stichprobenmittel
s     = apply(xyz, 2, sd)                          # Stichprobenstandardabweichungen
c     = cov(xyz)                                   # Stichprobenkovarianzen
b_xz1 = c[1,3]/c[3,3]                              # beta_1 (x,z)
b_xz0 = bars[1] - b_xz1*bars[3]                    # beta_0 (x,z)
b_yz1 = c[2,3]/c[3,3]                              # beta_1 (y,z)
b_yz0 = bars[2] - b_yz1*bars[3]                    # beta_0 (y,z)
e_xz  = xyz[,1] - b_xz1*xyz[,3] - b_xz0            # Residualwerte e^{x,z}
e_yz  = xyz[,2] - b_yz1*xyz[,3] - b_yz0            # Residualwerte e^{y,z}
pr_e  = cor(e_xz,e_yz)                             # \rho(x,y\z)

# Partielle Stichprobenkorrelation aus Stichprobenkorrelationen
r       = cor(xyz)                                 # Stichprobenkorrelationsmatrix
pr_r_n = r[1,2]-r[1,3]*r[2,3]                      # \rho(x,y\z) Formel Zähler
pr_r_d = sqrt((1-r[1,3]^2)*(1-r[2,3]^2))           # \rho(x,y\z) Formel Nenner
pr_r    = pr_r_n/pr_r_d                            # \rho(x,y\z)

# partielle Stichprobenkorrelation aus Toolbox
library(ppcor)                                     # Laden der Toolbox
pr_t    = pcor(xyz)                                # \rho(x,y\z),\rho(x,z\y),\rho(y,z\x)

# Ausgabe
cat("r(x,y)                          :"  , r[1,2],
    "\nr(x,y/z) aus Residuenkorrelation :" , pr_e,
    "\nr(x,y/z) aus Korrelationen       :" , pr_r,
    "\nr(x,y/z) aus Toolbox             :" , pr_t$estimate[1,2])
```

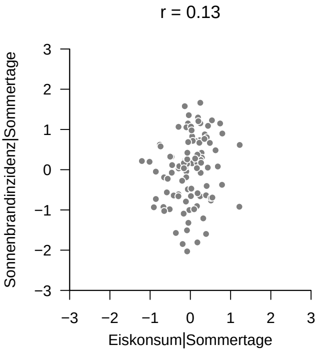
```
> r(x,y)                            : 0.5
> r(x,y/z) aus Residuenkorrelation : 0.133
> r(x,y/z) aus Korrelationen       : 0.133
> r(x,y/z) aus Toolbox             : 0.133
```

Partielle Korrelation



Partielle Korrelation





Motivation

Bedingte Korrelation

Partielle Korrelation

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie die Motivation zur Bestimmung bedingter und partieller Korrelationen.
2. Definieren Sie die Begriffe der bedingten Kovarianz und der bedingten Korrelation.
3. Geben Sie das Theorem zu bedingter Korrelation und Korrelationen bei Normalverteilung an.
4. Definieren Sie den Begriff der partiellen Korrelation.
5. Definieren Sie den Begriff der partiellen Stichprobenkorrelation.
6. Geben Sie das Theorem zu bedingter und partieller Korrelation bei Normalverteilung wieder.
7. Erläutern Sie die Auswertung einer partiellen Korrelation anhand eines Anwendungsbeispiels.

References

Baba, Kunihiro, Ritei Shibata, and Masaaki Sibuya. 2004. "Partial Correlation and Conditional Correlation as Measures of Conditional Independence." *Australian <Html_ent Glyph="@amp;Äscii="&/> New Zealand Journal of Statistics* 46 (4): 657–64. https://doi.org/10.1111/j.1467-842X.2004.00360.x.

Imbens, Guido, and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Academic Press.

Lawrance, A. J. 1976. "On Conditional and Partial Correlation." *The American Statistician* 30 (3): 146. https://doi.org/10.2307/2683864.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K. ; New York: Cambridge University Press.