Tutorium

Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

13. Termin: (12) Partielle Korrelation

Sean Mulready

Ersatztermine für den 12.07.:

- Mo, 10.07., 11-13 Uhr: **G22A, R105**
- Di, 11.07., 09-11 Uhr: **G22A, R110**

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.

1. Erläutern Sie die Motivation zur Bestimmung bedingter und partieller Korrelationen.

- Korrelation impliziert keine Kausalität: Nur weil die Veränderung der Werte zweier ZV x und y scheinbar in Verbindung steht (korreliert), muss nicht das eine die Ursache für das andere sein
- Kausalität: Koinzidenz (Korrelation) mit zeitlicher Rangfolge (Ursache immer vor der Wirkung)
- Es könnte auch noch eine oder mehrere Zufallsvariablen geben, die die Korreltation der Werte verursachen
- bedingte und partielle Korrelation um zu schauen: Gibt es wirklich einen direkten Zusammenhang zwischen x und y?
- wir betrachten im folgenden ein Modell mit drei ZV (x, y, z) und der Frage: besteht ein direkter Zusammenhang zwischen x und y, wie es die Daten vermuten lassen? Oder liegt die Korrelation doch an z?
- Wie können wir den Einfluss von z auf die Korrelation von x und y "herausrechnen" ?
- Zufallsvariablen hier mit lateinischen Buchstaben bezeichnet!

2. Definieren Sie die Begriffe der bedingten Kovarianz und der bedingten Korrelation.

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.

3. Geben Sie das Theorem zu bedingter Korrelation und Korrelationen bei Normalverteilung an.

Theorem (Bedingte Korrelation und Korrelationen bei Normalverteilung)

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

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

Bemerkungen

- Mit den Stichprobenkorrelationen $r_{x,y}, r_{x,z}, r_{y,z}$ ergibt sich ein entsprechender Schätzer für $\rho(x, y|z)$ als

$$r_{x,y|z} = \frac{r_{x,y} - r_{x,z} r_{y,z}}{\sqrt{(1 - r_{x,z}^2)\sqrt{(1 - r_{y,z}^2)}}} \tag{3}$$

4. Definieren Sie den Begriff der partiellen Korrelation.

Definition (Partielle Korrelation)

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

$$\begin{aligned} x &= \beta_0^{x,z} + \beta_1^{x,z} z \\ y &= \beta_0^{y,z} + \beta_1^{y,z} z \end{aligned} \tag{4}$$

mit Residualvariablen

$$\begin{aligned} e^{x,z} &= x - \beta_0^{x,z} - \beta_1^{x,z} z \\ e^{y,z} &= y - \beta_0^{y,z} - \beta_1^{y,z} z \end{aligned} \tag{5}$$

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

$$\rho(x, y \setminus z) := \rho(e^{x,z}, e^{y,z}). \tag{6}$$

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

5. Definieren Sie den Begriff der partiellen Stichprobenkorrelation.

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

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

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

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

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

Bemerkungen

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

6. Geben Sie das Theorem zu bedingter und partieller Korrelation bei Normalverteilung wieder.

Theorem (Bedingte und Partielle Korrelation bei Normalverteilung)

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

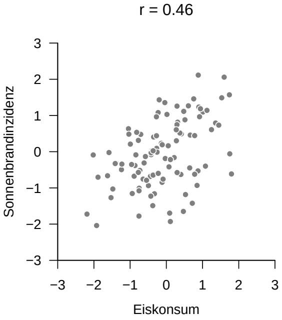
$$\rho(x, y|z) = \rho(x, y \setminus z) \tag{7}$$

7. Erläutern Sie die Auswertung einer partiellen Korrelation anhand eines Anwendungsbeispiels.

Im *Magdeburger Tageblatt für total seriöse Recherche und Wissenschaft* steht folgende Schlagzeile:

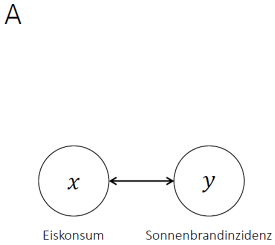
Gar nicht cool: Verursacht Eiscreme Hautkrebs?

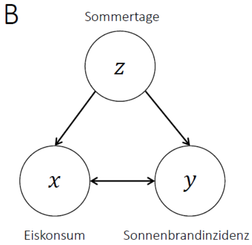
Daten zeigen *eindeutigen* Zusammenhang!



Anwendungsbeispiel

- Die erhobenen Daten zum jährlichen Eiskonsum und der jährlichen Sonnenbrandinzidenz stehen scheinbar in direkter Beziehung. Je höher der Eiskonsum in einem Land in einem Jahr, desto höher auch die Sonnenbrandinzidenz.
- Ist also der Eiskonsum die *Ursache* für den Sonnenbrand oder umgekehrt?
- Oder gibt es einen anderen Einfluss, dessen Höhe sowohl den Eiskonsum, als auch den Sonnenbrand beeinflusst?
- Idee: Könnte die Anzahl der Sommertage* sowohl auf den Eiskonsum, als auch die Sonnenbrandinzidenz wirken?





- Wie können wir jetzt den tatsächlichen Zusammenhang von Eiskonsum x und Sonnenbrandinzidenz y, also ohne den Einfluss von Sommertagen z untersuchen?
- *Sommertage: Tage deren maximale Temperatur über 25°C liegt

Anwendungsbeispiel

Vorgehen: Wir berechnen die partielle Korrelation. Das können wir auf verschiedene Arten machen. Aus den Folien der Vorlesung und dem Script liegen uns genug Daten vor um die bedingte und damit in diesem Falle auch die partielle Korrelation zu berechnen.(Aus dem Theorem zu bedingter und partieller Korrelation bei Normalverteilung wissen wir, dass $\rho(x, y|z) = \rho(x, y \setminus z)$, wenn unsere ZV multivariat normalverteilt sind)

- Wir wollen die bedingte Korrelation berechnen, gegeben als

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

- $\mathbb{S}(y|z)$ und $\mathbb{S}(x|z)$ können wir auch schreiben als $\sqrt{\mathbb{C}(y|z)}$ und $\sqrt{\mathbb{C}(x|z)}$
- Nun brauchen wir nur noch die Kovarianzmatrix $\Sigma_{x,y|z}$, welche sich berechnen lässt durch

$$\Sigma_{x,y|z} = \Sigma_{x,y} - \Sigma_{(x,y),z}\Sigma_z^{-1}\Sigma_{z,(x,y)} \tag{9}$$

Anwendungsbeispiel

- Wir nehmen den Kovarianzmatrixparamter von (x, y, z) wie im Script angegeben

$$\Sigma := \begin{pmatrix} 1.0 & 0.5 & 0.9 \\ 0.5 & 1.0 & 0.5 \\ 0.9 & 0.5 & 1.0 \end{pmatrix} \tag{10}$$

- Damit können wir jetzt die Kovarianzmatrix $\Sigma_{x,y|z}$ berechnen:

$$\begin{aligned} \Sigma_{x,y|z} &:= \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix} - \begin{pmatrix} 0.9 \\ 0.5 \end{pmatrix} (1.0)^{-1} \begin{pmatrix} 0.9 & 0.5 \end{pmatrix} \\ &:= \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix} - \begin{pmatrix} 0,81 & 0,45 \\ 0,45 & 0,25 \end{pmatrix} \\ &:= \begin{pmatrix} 0.19 & 0.05 \\ 0.05 & 0.75 \end{pmatrix} \end{aligned} \tag{11}$$

- Mit der Kovarianzmatrix $\Sigma_{x,y|z}$ können wir die bedingte und in diesem Falle auch die partielle Korrelation berechnen

$$\rho(x, y|z) := \frac{\mathbb{C}(x, y|z)}{\sqrt{\mathbb{C}(x, x|z)}\sqrt{\mathbb{C}(y, y|z)}} := \frac{0.05}{\sqrt{0.19}\sqrt{0.75}} \approx 0.13 \tag{12}$$

- Betrachten wir aus der obigen Matrix Σ nur die (Ko)Varianzen von x und y, so ergibt sich $\rho(x, y) = 0.5$

Anwendungsbeispiel

Was schließen wir daraus: Betrachten wir den tatsächlichen Zusammenhang von Eiskonsum x und Sonnenbrandinzidenz y *ohne* den Einfluss der Sommertage z, so ergibt sich, dass wir höchstens von einem schwachen Zusammenhang von Eiskonsum und Sonnenbrandinzidenz ausgehen können. Der tatsächliche Einfluss auf Eiskonsum x und Sonnenbrandinzidenz y kommt nicht von gegenseitig, sondern durch die Anzahl der Sommertage z.