Vorlesung Statistik A
Explorative Datenanalyse

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Wintersemester 2026/2027

Partielle Korrelation

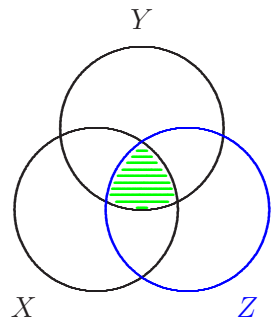
Bei Beschränkung auf eine Betrachtung bivariater Korrelations- und Regressionsanalysen werden die wahren Zusammenhänge möglicherweise falsch eingeschätzt

- Gelegentlich wird ein gefundener korrelativer Zusammenhang zwischen den Merkmalen X und Y als "Scheinkorrelation" bezeichnet, wenn dieser auf eine gemeinsame Abhängigkeit der beiden betrachteten von einem dritten Merkmal Z zurückgeht
- Wird dieser Einfluss des Merkmals Z statistisch eliminiert, so spricht man von partieller Korrelation

Partielle Korrelation: Beispiel

Interkorrelationen und Anteile gemeinsamer Varianz

Merkmale	Korrelation	erklärte Varianz
X vs. Y	$r_{xy} = 0.5$	$r_{xy}^2 = 0.25$
X vs. Z	$r_{xz} = 0.5$	$r_{xz}^2 = 0.25$
Y vs. Z	$r_{yz} = 0.5$	$r_{yz}^2 = 0.25$



Herauspartialisieren

Um den Einfluss des Merkmals Z statistisch zu eliminieren, betrachtet man die Regression der Merkmale X und Y auf Z

Die Residuen

$$x - a_{xz} - b_{xz}\, z \quad \text{und} \quad y - a_{yz} - b_{yz}\, z$$

werden als diejenigen Merkmalsausprägungen interpretiert, aus denen der Einfluss des Merkmals Z herausgerechnet ist (man spricht von "Herauspartialisieren")

Entsprechend betrachtet man im Rahmen einer partiellen Korrelationsanalyse die Korrelation zwischen den Residuen $y - a_{yz} - b_{yz}\, z$ und $x - a_{xz} - b_{xz}\, z$

Der Korrelationskoeffizient $r_{xy.z}$ zwischen $X - a_{xz} - b_{xz}\, Z$ und $Y - a_{yz} - b_{yz}\, Z$ lässt sich in einfacher Weise aus den Interkorrelationen r_{xy}, r_{xz} und r_{yz} berechnen

Berechnung der partiellen Korrelation

Die Berechnung des "partiellen" Korrelationskoeffizienten $r_{xy.z}$ benützt dabei folgende Beziehungen der Kovarianz $\tilde{s}(x, y) = \tilde{s}_{xy}$

- Die Korrelation eines Merkmals X mit einer reellen Konstanten $r \in \mathbb{R}$ ist Null

$$\tilde{s}(x, r) = \tilde{s}(r, x) = 0$$

- Die Kovarianz ist bilinear, d. h. für beliebige Merkmale X, Y, Z und reelle Konstanten p, q gilt

$$\tilde{s}(p \cdot x + q \cdot y,\ z) = p \cdot \tilde{s}(x, z) + q \cdot \tilde{s}(y, z)$$
$$\tilde{s}(z,\ p \cdot x + q \cdot y) = p \cdot \tilde{s}(z, x) + q \cdot \tilde{s}(z, y)$$

Berechnung der partiellen Korrelation

Die Berechnung des "partiellen" Korrelationskoeffizienten $r_{xy.z}$ benützt dabei folgende Beziehungen

- Für die Varianzen der Residuen gilt

$$\tilde{s}^2(x - a_{xz} - b_{xz} \cdot z) = \tilde{s}^2(x) \cdot (1 - r_{xz}^2)$$
$$\tilde{s}^2(y - a_{yz} - b_{yz} \cdot z) = \tilde{s}^2(y) \cdot (1 - r_{yz}^2)$$

- Dies wird nun in die Definition von $r_{xy.z}$ eingesetzt

$$r_{xy.z} = \frac{\tilde{s}(x - a_{xz} - b_{xz} \cdot z,\ y - a_{yz} - b_{yz} \cdot z)}{\tilde{s}(x - a_{xz} - b_{xz} \cdot z) \cdot \tilde{s}(y - a_{yz} - b_{yz} \cdot z)}$$

Berechnung der partiellen Korrelation

Für den Zähler ergibt sich dabei

$$\begin{aligned}
&\tilde{s}(x - a_{xz} - b_{xz} \cdot z,\ y - a_{yz} - b_{yz} \cdot z) = \\
&\quad = \tilde{s}(x, y) - b_{yz} \cdot \tilde{s}(x, z) - b_{xz} \cdot \tilde{s}(y, z) + b_{xz} \cdot b_{yz} \cdot \tilde{s}^2(z) \\
&\quad = \tilde{s}(x, y) - \frac{\tilde{s}(y, z) \cdot \tilde{s}(x, z)}{\tilde{s}^2(z)} - \frac{\tilde{s}(x, z) \cdot \tilde{s}(y, z)}{\tilde{s}^2(z)} + \frac{\tilde{s}(x, z) \cdot \tilde{s}(y, z)}{\tilde{s}^2(z)} \\
&\quad = \tilde{s}(x, y) - \frac{\tilde{s}(y, z) \cdot \tilde{s}(x, z)}{\tilde{s}^2(z)}
\end{aligned}$$

Berechnung der partiellen Korrelation

Hiermit folgt insgesamt

$$r_{xy.z} = \frac{\tilde{s}(x,y)}{\tilde{s}(x) \cdot \tilde{s}(y) \cdot \sqrt{1-r_{xz}^2} \cdot \sqrt{1-r_{yz}^2}} - \frac{\tilde{s}(y,z) \cdot \tilde{s}(x,z)}{\tilde{s}(x) \cdot \tilde{s}(y) \cdot \tilde{s}^2(z) \cdot \sqrt{1-r_{xz}^2} \cdot \sqrt{1-r_{yz}^2}}$$

$$= \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1-r_{xz}^2} \cdot \sqrt{1-r_{yz}^2}}$$

Berechnung der partiellen Korrelation

Für das obige Beispiel ergibt sich

$$
\begin{aligned}
r_{xy.z} &= \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{xz}^2} \cdot \sqrt{1 - r_{yz}^2}} \\
&= \frac{0.5 - 0.5 \cdot 0.5}{\sqrt{1 - 0.5^2} \cdot \sqrt{1 - 0.5^2}} \\
&= \frac{0.25}{0.75} \\
&= 0.33
\end{aligned}
$$

Lernaufwand und Prüfungsleistung

Suppression

Wir betrachten den Zusammenhang zwischen den Merkmalen X (Lernaufwand vor der Prüfung), Y (Prüfungsleistung) und Z (Angst in der Prüfungssituation)

Variablen	Korrelation	erklärte Varianz
X vs. Y	$r_{xy} = +0.2$	$r_{xy}^2 = 0.04$
X vs. Z	$r_{xz} = +0.8$	$r_{xz}^2 = 0.64$
Y vs. Z	$r_{yz} = -0.4$	$r_{yz}^2 = 0.16$

$$r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{xz}^2} \cdot \sqrt{1 - r_{yz}^2}} = \frac{0.2 - 0.8 \cdot (-0.4)}{\sqrt{1 - 0.64} \cdot \sqrt{1 - 0.16}} = +0.95$$

IQ und Hirnvolumen

Semipartielle Korrelation

Wir betrachten Daten aus einer Untersuchung von Willerman et al. (1991) zum Zusammenhang von IQ und der Größe des Gehirns,
`https://dasl.datadescription.com/datafile/iq-brain/`

Für je 20 Männer und Frauen werden neben einem IQ-Wert und dem mittels MRT gemessenen Hirnvolumen auch Körpergröße und -gewicht bestimmt

Da die Größe des Gehirns auch mit der Körpergröße kovariiert, wird dieser Einfluss aus dem Hirnvolumen herauspartialisiert

Da der IQ-Wert unverändert in die Korrelation eingehen sollte, spricht man in diesem Fall von einer *semipartiellen Korrelation*

Für diese semipartielle Korrelation betrachten wir getrennt für Männer und Frauen die Regression des Hirnvolumens auf die Körpergröße und korrelieren das Residuum mit dem IQ-Wert

IQ und Hirnvolumen

Einlesen der Daten und Selegieren der männlichen Personen, wobei `na.omit()` die Fälle mit fehlenden Werten (`NA`) ausschließt

```
dat   <- read.table("brainsize-iq.txt", header = TRUE)
dat_m <- na.omit(dat[dat$Gender == "Male", ])
```

Die Berechnungen ergeben dann

```
with(dat_m, cor(FSIQ, MRI_Count))      # IQ und Hirnvolumen
[1] 0.431

with(dat_m, cor(Height, MRI_Count))    # Groesse und Hirnvolumen
[1] 0.293

lm_m <- lm(MRI_Count ~ Height, dat_m)
cor(dat_m$FSIQ, resid(lm_m))           # semipartielle Korrelation
[1] 0.569
```

IQ und Hirnvolumen

Entsprechend ergibt sich für die Frauen

```
dat_f <- na.omit(dat[dat$Gender == "Female", ])
with(dat_f, cor(FSIQ, MRI_Count))
[1] 0.326

with(dat_f, cor(Height, MRI_Count))
[1] 0.175

lm_f <- lm(MRI_Count ~ Height, dat_f)
cor(dat_f$FSIQ, resid(lm_f))
[1] 0.341
```