Allgemeines Lineares Modell

BSc Psychologie SoSe 2022

Prof. Dr. Dirk Ostwald

Datum	Einheit	Thema
08.04.2022	Grundlagen	(1) Regression
	Osterpause	
22.04.2022	Grundlagen	(2) Korrelation
29.04.2022	Grundlagen	(3) Matrizen
06.05.2022	Grundlagen	(4) Normalverteilungen
13.05.2022	Theorie	(5) Modellformulierung
20.05.2022	Theorie	(6) Modellschätzung
27.05.2022	Theorie	(7) Modellevaluation
03.06.2021	Anwendung	(8) Studiendesign
10.06.2021	Anwendung	(9) T-Tests
17.06.2021	Anwendung	(10) Einfaktorielle Varianzanalyse
24.06.2022	Anwendung	(11) Zweifaktorielle Varianzanalyse
01.07.2022	Anwendung	(12) Multiple Regression
11.07.2022	Q & A	Online 14 - 17 Uhr
14.07.2022	Klausur	G16-H5 11 - 12 Uhr
März 2023	Klausurwiederholungstermin	

(12) Multiple Regression

Überblick

Faktorielle und Parametrische ALM Designs

Faktorielle ALM Designs

- Designmatrizen mit 1en und 0en, manchmal -1en.
- Betaparameter repräsentieren Gruppenerwartungswerte.
- Betaparameterschätzer repräsentieren Gruppenstichprobenmittel.
- $\Rightarrow$ T-Tests, Einfaktorielle Varianzanalyse, Mehrfaktorielle Varianzanalyse

Parametrische ALM Designs

- Designmatrizen besitzen Spalten mit kontinuierlichen reellen Werten.
- Die Designmatrixsspalten werden *Regressoren*, *Prädiktoren*, oder *Kovariaten* genannt.
- Betaparameter repräsentieren Steigungsparameter.
- Betaparameterschätzer ergeben sich als normalisierte Regressor-Daten Kovarianzen.
- Es besteht ein enger Bezug zur Theorie der Korrelation.
- $\Rightarrow$ Einfache lineare Regression, Multiple lineare Regression

Faktoriell-parametrische ALM Designs

- Designmatrizen mit mehreren faktoriellen und parametrischen Werten.
- Die parametrischen Regressoren werden oft als kontrollierte Kovariaten betrachtet.
- $\Rightarrow$ Kovarianzanalyse

Überblick

ALM Designs als Hypothesentestverfahren*

Testen von Unterschiedshypothesen

- T-Tests
- Einfaktorielle Varianzanalyse
- Mehrfaktorielle Varianzanalyse
- Kovarianzanalyse

Testen von Zusammenhangshypothesen

- Einfache lineare Regression/Korrelation
- Multiple lineare Regression/Multiple Korrelation

*Diese Sichtweise durch den Lehrenden nicht favorisiert.

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Ausblick

Selbstkontrollfragen

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Ausblick

Selbstkontrollfragen

Anwendungsszenario

- Generalisierung der einfachen linearen Regression zu mehr als einer unabhängigen Variable.
- Eine univariate abhängige Variable bestimmt an randomisierten experimentellen Einheiten.
- Zwei oder mehr "kontinuierliche" unabhängige Variablen.
- Die unabhängigen Variablen heißen Regressoren, Prädiktoren, Kovariaten oder Features.

Ziele

- Quantifizierung des Erklärungspotentials der Variation der UVs durch die Variation der AVs.
- Quantifizierung des Einflusses einzelner UVs auf die AV im Kontext anderer UVs.
- Prädiktion von AV Werten aus UV Werten nach Parameterschätzung.

Anwendungsbeispiel

- BDI Differenzwerte in Abhängigkeit von Therapiedauer und Alter

Anwendungsszenario

Beispieldatensatz

$n = 100$

ID	Age	Therapy	BDI
1	50	16	9
2	38	13	9
3	46	16	10
4	62	17	3
5	25	23	38
6	34	23	29
7	36	24	31
8	36	18	21
9	57	20	20
10	46	17	16
11	59	21	18
12	54	22	24
13	27	14	20
14	56	18	12
15	41	20	31
16	46	23	30
17	23	15	23
18	36	22	27
19	44	19	23
20	70	20	11
21	72	13	-3
22	57	16	12
23	67	16	3
24	41	19	23
25	44	14	15

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Ausblick

Selbstkontrollfragen

Modellformulierung

Definition (Modell der multiplen Regression)

y_i mit $i = 1, ..., n$ sei die Zufallsvariable, die den $i$ten Wert einer abhängigen Variable modelliert. Dann hat das *Modell der multiplen Regression* die strukturelle Form

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, ..., n \text{ und } \sigma^2 > 0, \quad (1)$$

wobei $x_{ij} \in \mathbb{R}$ mit $1 \le i \le n$ und $1 \le j \le p$ den $i$ten Wert der $j$te unabhängigen Variable bezeichnet. Die unabhängigen Variablen werden auch *Regressoren*, *Prädiktoren*, *Kovariaten* oder *Features* genannt. Mit

$$x_i := (x_{i1}, ..., x_{ip})^T \in \mathbb{R}^p \text{ und } \beta := (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p \quad (2)$$

hat das Modell der multiplen Regression die Datenverteilungsform

$$y_i \sim N(\mu_i, \sigma^2) \text{ u.i.v. für } i = 1, ..., n, \text{ wobei } \mu_i := x_i^T\beta. \quad (3)$$

In diesem Zusammenhang wird $x_i \in \mathbb{R}^p$ auch als $i$ter *Featurevektor* bezeichnet. Die Designmatrixform des Modells der multiplen Regression schließlich ist gegeben durch

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (4)$$

mit

$$y := (y_1, ..., y_n)^T, X := (x_{ij})_{1 \le i \le n, 1 \le j \le p} \in \mathbb{R}^{n \times p}, \beta := (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p \text{ und } \sigma^2 > 0. \quad (5)$$

Bemerkung

- Das Modell der multiplen Regression und die allgemeine Form des ALMs sind identisch.

Beispieldatensatzerzeugung

```
# Datensimulation
library(MASS)                                          # Multivariate Normalverteilung
set.seed(10)                                           # reproduzierbare Daten
n             = 100                                    # Anzahl Datenpunkte
p             = 3                                      # Anzahl Parameter
x_1           = round(runif(n,20,80))                  # Regressorwerte Alter
x_2           = round(runif(n,12,24))                  # Regressorwerte Therapiedauer
X             = matrix(c(rep(1,n),x_1,x_2), nrow = n) # Designmatrix
I_n           = diag(n)                                # Identitätsmatrix
beta          = matrix(c(5,-.5,2), nrow = p)           # Betaparametervektor
sigsqr        = 10                                     # Varianzparameter
y             = mvrnorm(1, X %*% beta, sigsqr*I_n)     # eine Realisierung eines n-dimensionalen ZVs

# Dataframeformatierung
library(writexl)                                       # Excel Output
D            = data.frame("ID" = 1:n)                  # Dataframe Initialisierung und ID Variable
D$Age        = x_1                                     # Alter
D$Therapy    = x_2                                     # Therapiedauer
D$BDI        = y                                       # PrePost-BDI Differenzwerte

# Datenspeicherung
write_xlsx(D, file.path(getwd()       , "12_Daten", "12_Multiple_Regression_Daten.xlsx"))
write.csv( D, file = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv"))
```

Beispieldatenvisualisierung

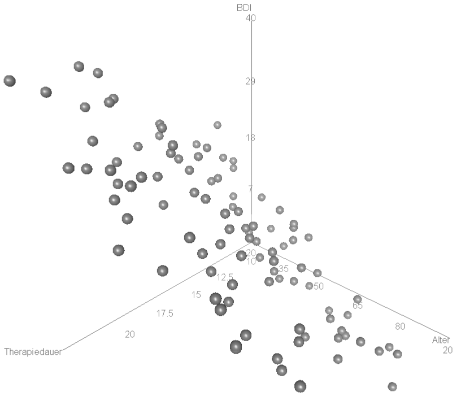
```
# Dateneinlesen
fname      = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)

# Open GL Visualisierung  mit car package, siehe ?scatter3d für Details
library(car)
scatter3d(
D$Age,
D$BDI,
D$Therapy,
xlab        = "Alter",
ylab        = "BDI",
zlab        = "Therapiedauer",
point.col   = "gray40",
axis.col    = rep("black",3),
axis.scales = T,
axis.ticks  = T,
surface     = F)
```

```
> Lade nötigen Namensraum: rgl

> Lade nötigen Namensraum: mgcv
```

Beispieldatenvisualisierung



Modellschätzung

Überblick

Der Betaparameterschätzer hat bekanntlich die Form

$$\hat{\beta} := (X^T X)^{-1} X^T y \tag{6}$$

Dabei quantifizieren in sehr grober Auflösung

- $X^T y \in \mathbb{R}^p$ die Kovariation der Regressoren mit den Daten und
- $X^T X \in \mathbb{R}^{p \times p}$ die Kovariation der Regressoren untereinander.

Damit ergibt sich für die Betaparameterschätzer also eine Interpretation als "regressorkovarianznormalisierte Regressordatenkovariation,"

$$\hat{\beta} \approx \text{Regressorkovarianz}^{-1} \cdot \text{Regresssordatenkovarianz} \tag{7}$$

Im Folgenden wollen wir diese Intuition am Beispiel einer einfachen multiplen Regression mit einem Interzeptregressor und zwei unabhängigen Variablen Regressoren vertiefen, wobei die betreffenden Kovariationen einmal durch Stichprobenkorrelationen und einmal durch partielle Stichprobenkorrelationen quantifiziert werden sollen.

Theorem (Betaparameterschätzer und Korrelationen)

Gegeben sei ein multiples Regressionsmodel der Form

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}. \tag{8}$$

Dann gilt

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\ \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\ \frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_2}} \end{pmatrix}, \tag{9}$$

wobei für die y_i, x_{i1} und x_{i2} mit $i = 1, \ldots, n$ $\bar{\cdot}$, $s.$ und $r_{\cdot,\cdot}$ die entsprechenden Stichprobenmittel, Stichprobenstandardabweichungen, und Stichprobenkorrelationen bezeichnen.

Bemerkung

- In Bezug auf die Regressoren sind die Begriffe Stichprobenmittel, Stichprobenstandardabweichung, und Stichprobenkorrelation lediglich formal gemeint, nach Voraussetzung des ALMs sind die Regressorenwerte keine Realisierungen von Zufallsvariablen.

Beweis

Wir erinnern zunächst daran, dass die Form des Betaparameterschätzers bekanntlich zum System der Normalengleichungen äquivalent ist (vgl. (6) Modellschätzung),

$$\hat{\beta} = (X^T X)^{-1} X^T y \Leftrightarrow X^T X \hat{\beta} = X^T y. \tag{10}$$

Ausschreiben des Normalengleichungssystems für den hier betrachteten ALM Spezialfall ergibt dann zunächst

$$
\begin{aligned}
& X^T X \hat{\beta} = X^T y \\
\Leftrightarrow & \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
\Leftrightarrow & \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1} x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2} x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix} \\
\Leftrightarrow & \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1} x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2} x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix}
\end{aligned}
$$

Modellschätzung

Beweis (fortgeführt)

und damit

$$
\begin{aligned}
& X^T X \hat{\beta} = X^T y \\
\Leftrightarrow & \begin{pmatrix} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} \\ \hat{\beta}_0 \sum_{i=1}^n x_{i2} + \hat{\beta}_1 \sum_{i=1}^n x_{i1} x_{i2} + \hat{\beta}_2 \sum_{i=1}^n x_{i2}^2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix}
\end{aligned}
$$

Aus der Gleichung der ersten Vektorkomponenten folgt dann direkt die Form von $\hat{\beta}_0$ mit

$$
\begin{aligned}
\sum_{i=1}^n y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} \\
\Leftrightarrow \frac{1}{n} \sum_{i=1}^n y_i &= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \frac{1}{n} \sum_{i=1}^n x_{i2} \\
\Leftrightarrow \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2
\end{aligned} \tag{11}
$$

Modellschätzung

Beweis (fortgeführt)

Einsetzen dieser Form von $\hat{\beta}_0$ in die Gleichung der zweiten Vektorkomponenten ergibt dann

$$
\begin{aligned}
\hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\
(\bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2) \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\
\bar{y}\sum_{i=1}^n x_{i1} - \hat{\beta}_1\bar{x}_1 \sum_{i=1}^n x_{i1} - \hat{\beta}_2\bar{x}_2 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\
\hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 - \hat{\beta}_1\bar{x}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} - \hat{\beta}_2\bar{x}_2 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n y_i x_{i1} - \bar{y}\sum_{i=1}^n x_{i1} \\
\hat{\beta}_1 \left(\sum_{i=1}^n x_{i1}^2 - \bar{x}_1 \sum_{i=1}^n x_{i1}\right) + \hat{\beta}_2 \left(\sum_{i=1}^n x_{i1}x_{i2} - \bar{x}_2 \sum_{i=1}^n x_{i1}\right) &= \sum_{i=1}^n y_i x_{i1} - \bar{y}\sum_{i=1}^n x_{i1}
\end{aligned}
$$

Beweis (fortgeführt)

Im Beweis des Theorems zur Ausgleichsgerade (vgl. (1) Regression) haben wir gesehen, dass

$$
\begin{aligned}
\sum_{i=1}^n x_{i1}x_{i1} - \bar{x}_1 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1) \\
\sum_{i=1}^n x_{i1}x_{i2} - \bar{x}_2 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\
\sum_{i=1}^n y_i x_{i1} - \bar{y} \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n (y_i - \bar{y})(x_{i1} - \bar{x}_1)
\end{aligned}
\tag{12}
$$

Modellschätzung

Beweis (fortgeführt)

Es ergibt sich also, dass

$$
\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1) + \hat{\beta}_2 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) &= \sum_{i=1}^{n}(y_i - \bar{y})(x_{i1} - \bar{x}_1) \\
\hat{\beta}_1 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1)}{n-1} + \hat{\beta}_2 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{i1} - \bar{x}_1)}{n-1}
\end{aligned}
\tag{13}
$$

Mit den Definitionen von Stichprobenstandardabweichung und -korrelation folgt dann weiter

$$
\begin{aligned}
\hat{\beta}_1 s_{x_1} s_{x_1} + \hat{\beta}_2 c_{x_1,x_2} &= c_{y,x_1} \\
\hat{\beta}_1 \frac{s_{x_1} s_{x_1}}{s_y s_{x_1}} + \hat{\beta}_2 \frac{c_{x_1,x_2}}{s_y s_{x_1}} &= \frac{c_{y,x_1}}{s_y s_{x_1}} \\
\hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{c_{x_1,x_2}}{s_y s_{x_1}} &= r_{y,x_1} \\
\hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{c_{x_1,x_2} s_{x_2}}{s_y s_{x_1} s_{x_2}} &= r_{y,x_1} \\
\hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{s_{x_2}}{s_y} r_{x_1,x_2} &= r_{y,x_1}
\end{aligned}
\tag{14}
$$

Beweis (fortgeführt)

Definition von

$$b_j := \frac{s_{x_j}}{s_y}, j = 1, 2 \tag{15}$$

erlaubt dann die Schreibweise

$$b_1 + b_2 r_{x_1,x_2} = r_{y,x_1}. \tag{16}$$

Schließlich folgt analog durch Vertauschen der Subskripte aus der Gleichung der dritten Vektorkomponenten

$$b_1 r_{x_1,x_2} + b_2 = r_{y,x_2} \tag{17}$$

Insgesamt haben wir also gesehen, dass die Definition des Betaparameterschätzers im vorliegenden ALM Spezialfall ergibt, dass mit

$$\hat{\beta}_j = b_j \frac{s_y}{s_{x_j}}, j = 1, 2 \tag{18}$$

gilt, dass

$$\begin{aligned} r_{y,x_1} &= b_1 + b_2 r_{x_1,x_2} \\ r_{y,x_2} &= b_1 r_{x_1,x_2} + b_2 \end{aligned} \tag{19}$$

Modellschätzung

Beweis (fortgeführt)

Damit folgt aus der zweiten Gleichung dann sofort

$$b_2 = r_{y,x_2} - b_1 r_{x_1,x_2}. \tag{20}$$

Einsetzen in die erste Gleichung ergibt dann

$$\begin{aligned}
b_1 + (r_{y,x_2} - b_1 r_{x_1,x_2}) r_{x_1,x_2} &= r_{y,x_1} \\
\Leftrightarrow b_1 + r_{y,x_2} r_{x_1,x_2} - b_1 r_{x_1,x_2}^2 &= r_{y,x_1} \\
\Leftrightarrow r_{y,x_2} r_{x_1,x_2} + b_1 \left(1 - r_{x_1,x_2}^2\right) &= r_{y,x_1} \\
\Leftrightarrow b_1 \left(1 - r_{x_1,x_2}^2\right) &= r_{y,x_1} - r_{y,x_2} r_{x_1,x_2} \\
\Leftrightarrow b_1 &= \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2}
\end{aligned} \tag{21}$$

Modellschätzung

Beweis (fortgeführt)

Für b_2 ergibt sich damit weiterhin

$$
\begin{aligned}
& b_2 = r_{y,x_2} - b_1 r_{x_1,x_2} \\
\Leftrightarrow\ & b_2 = r_{y,x_2} - \left(\frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}\right) r_{x_1,x_2} \\
\Leftrightarrow\ & b_2 = \frac{r_{y,x_2}\left(1 - r^2_{x_1,x_2}\right)}{1 - r^2_{x_1,x_2}} - \frac{r_{y,x_1} r_{x_1,x_2} - r_{y,x_2} r^2_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \\
\Leftrightarrow\ & b_2 = \frac{r_{y,x_2} - r_{y,x_2} r^2_{x_1,x_2} - r_{y,x_1} r_{x_1,x_2} + r_{y,x_2} r^2_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \\
\Leftrightarrow\ & b_2 = \frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}
\end{aligned}
\tag{22}
$$

Damit folgen dann aber

$$
\begin{aligned}
\hat{\beta}_1 &= b_1 \frac{s_y}{s_{x_1}} = \left(\frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}\right) \frac{s_y}{s_{x_1}} \\
\hat{\beta}_2 &= b_2 \frac{s_y}{s_{x_2}} = \left(\frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}\right) \frac{s_y}{s_{x_2}}
\end{aligned}
\tag{23}
$$

und es ist alles gezeigt. $\square$

Modellschätzung

Anwendungsbeispiel

```
# Dateneinlesen
fname       = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv")
D           = read.table(fname, sep = ",", header = TRUE)          # Datensatz

# Modellschätzung
y           = D$BDI                                                # Abhängige Variable
n           = length(y)                                            # Anzahl Datenpunkte
X           = matrix(c(rep(1,n), D$Age, D$Therapy), nrow = n)      # Desigmatrix
beta_hat    = solve(t(X) %*% X) %*% t(X) %*% y                     # Betaparameterschätzer
eps_hat     = y - X %*% beta_hat                                   # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat) /(n-p)                       # Varianzparameterschätzer

# Betaparameterschätzer aus Stichprobenmittel, -standardabweichungen und -korrelationen
y12         = cbind(y,X[,-1])                                      # y,x_1,x_2 Matrix
bars        = apply(y12, 2, mean)                                  # Stichprobenmittel
s           = apply(y12, 2, sd)                                    # Stichprobenstandardabweichungen
r           = cor(y12)                                             # Stichprobenkorrelationen
beta_hat_1 = (r[1,2] - r[1,3]*r[2,3])/(1 - r[2,3]^2)*(s[1]/s[2]) # \hat{\beta}_1
beta_hat_2 = (r[1,3] - r[1,2]*r[2,3])/(1 - r[2,3]^2)*(s[1]/s[3]) # \hat{\beta}_2
beta_hat_0 = bars[1] - beta_hat_1*bars[2] - beta_hat_2*bars[3]    # \hat{\beta}_0

# Ausgabe
cat("beta_hat ALM-Schätzer          :"  , beta_hat,
    "\nbeta_hat Deskriptivstatistiken :", c(beta_hat_0,beta_hat_1,beta_hat_2))
```

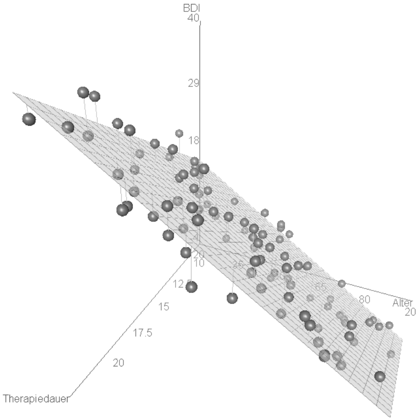
```
> beta_hat ALM-Schätzer          : 5.42 -0.481 1.91
> beta_hat Deskriptivstatistiken : 5.42 -0.481 1.91
```

Beispieldatenvisualisierung

```
# Dateneinlesen
fname      = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)         # Datensatz

# Open GL Visualisierung  mit car package, siehe ?scatter3d für Details
library(car)
scatter3d(
D$Age,
D$BDI,
D$Therapy,
xlab        = "Alter",
ylab        = "BDI",
zlab        = "Therapiedauer",
point.col   = "gray40",
axis.col    = rep("black",3),
axis.scales = T,
axis.ticks  = T,
surface     = T,
surface.col = "gray70",
neg.res.col = "gray70",
pos.res.col = "gray70")
```

Beispieldatenvisualisierung



Theorem (Betaparameterschätzer und partielle Korrelationen)

Gegeben sei ein multiples Regressionsmodel der Form

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}. \tag{24}$$

Dann gilt

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\ r_{y,x_1|x_2} \sqrt{\frac{1-r^2_{y,x_2}}{1-r^2_{x_1,x_2}}} \frac{s_y}{s_{x_1}} \\ r_{y,x_2|x_1} \sqrt{\frac{1-r^2_{y,x_1}}{1-r^2_{x_2,x_1}}} \frac{s_y}{s_{x_2}} \end{pmatrix}, \tag{25}$$

wobei für $1 \leq k, l \leq 2$ und $i = 1, ..., n$

- $r_{y,x_k|x_l}$ die partielle Stichprobenkorrelation der y_i und x_{ik} gegeben die x_{il} ist,
- r_{y,x_k} die Stichprobenkorrelation der y_i und x_{ik} ist, und
- r_{x_k,x_l} die Stichprobenkorrelation der x_{ik} und x_{il} ist.

Modellschätzung

Bemerkungen

- Im Allgemeinen gilt für $1 \leq i, l \leq k$, dass $\hat{\beta}_k \neq r_{y,x_k|x_l}$.
- Betaparameterschätzer sind also im Allgemeinen keine partiellen Stichprobenkorrelationen.
- $\hat{\beta}_k = r_{y,x_k|x_l}$ für $1 \leq i, l \leq k$ gilt genau dann, wenn $s_y = s_{x_1} = s_{x_2}$ und zudem
  - $r_{y,x_k} = r_{x_k,x_l} = 0$, wenn also die Stichprobenkorrelationen der Daten und der Werte des zweiten Regressors, sowie die Stichprobenkorrelation der Werte der beiden Regressoren gleich Null sind. Dies kann der Fall sein, wenn einer der Regressoren die Daten "sehr gut erklärt" und der andere Regressor von dem ersten "sehr verschieden" ist.
  - $|r_{y,x_l}| = |r_{x_k,x_l}|$, wenn also die obige Stichprobenkorrelationen dem Betrage nach gleich sind. Dies ist vermutlich selten der Fall.

Beweis

Wir betrachten $\hat{\beta}_1$, das Resultat für $\hat{\beta}_2$ folgt dann durch Vertauschen der Indizes. Wir haben in vorherigem Theorem gesehen, dass

$$\hat{\beta}_1 = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \frac{s_y}{s_{x_1}} \tag{26}$$

Weiterhin haben wir in (2) Korrelation gesehen, dass unter der Annahme der multivariaten Normalverteilung von y, x_1, x_2 ein Schätzer für die partielle Kkorrelation von y und x_1 gegeben x_2 durch

$$r_{y,x_1|x_2} = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{\sqrt{1 - r^2_{y,x_2}}\sqrt{1 - r^2_{x_1,x_2}}} \tag{27}$$

gegeben ist. Für $\hat{\beta}_1$ ergibt sich somit

$$\begin{aligned}
\hat{\beta}_1 &= \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \left(1 - r^2_{x_1,x_2}\right) \hat{\beta}_1 &= \left(r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}\right) \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \frac{1 - r^2_{x_1,x_2}}{\sqrt{1 - r^2_{y,x_2}}\sqrt{1 - r^2_{x_1,x_2}}} \hat{\beta}_1 &= \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{\sqrt{1 - r^2_{y,x_2}}\sqrt{1 - r^2_{x_1,x_2}}} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \frac{1 - r^2_{x_1,x_2}}{\sqrt{1 - r^2_{y,x_2}}\sqrt{1 - r^2_{x_1,x_2}}} \hat{\beta}_1 &= r_{y,x_1|x_2} \frac{s_y}{s_{x_1}}
\end{aligned} \tag{28}$$

Beweis

und damit weiter

$$
\begin{aligned}
\hat{\beta}_1 &= r_{y,x_1|x_2} \frac{\sqrt{1 - r_{y,x_2}^2}\sqrt{1 - r_{x_1,x_2}^2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \hat{\beta}_1 &= r_{y,x_1|x_2} \frac{\sqrt{1 - r_{y,x_2}^2}\sqrt{1 - r_{x_1,x_2}^2}}{\left(\sqrt{1 - r_{x_1,x_2}^2}\right)^2} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \hat{\beta}_1 &= r_{y,x_1|x_2} \frac{\sqrt{1 - r_{y,x_2}^2}}{\sqrt{1 - r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \hat{\beta}_1 &= r_{y,x_1|x_2} \sqrt{\frac{1 - r_{y,x_2}^2}{1 - r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}}
\end{aligned} \tag{29}
$$

Modellschätzung

Anwendungsbeispiel

```
# Dateneinlesen
fname      = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)         # Datensatz

# Modellschätzung
y          = D$BDI                                               # Abhängige Variable
n          = length(y)                                           # Anzahl Datenpunkte
X          = matrix(c(rep(1,n), D$Age, D$Therapy), nrow = n)     # Desigmatrix
p          = ncol(X)                                             # Anzahl Parameter
beta_hat   = solve(t(X) %*% X) %*% t(X) %*% y                    # Betaparameterschätzer
eps_hat    = y - X %*% beta_hat                                  # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat) /(n-p)                     # Varianzparameterschätzer

# Betaparameterschätzer aus partiellen Korrelationen und Korrelationen
library(ppcor)                                                   # partielle Korrelationentoolbox
y12        = cbind(y,X[,-1])                                     # y,x_1,x_2 Matrix
bars       = apply(y12, 2, mean)                                 # Stichprobenmittel
s          = apply(y12, 2, sd)                                   # Stichprobenstandardabweichungen
r          = cor(y12)                                            # Stichprobenkorrelationen
pr         = pcor(y12)                                           # partielle Stichprobenkorrelationen
pr         = pr$estimate                                         # partielle Stichprobenkorrelationen
beta_hat_1 = pr[1,2]*sqrt((1-r[1,3]^2)/(1-r[2,3]^2))*(s[1]/s[2]) # \hat{\beta}_1
beta_hat_2 = pr[1,3]*sqrt((1-r[1,2]^2)/(1-r[3,2]^2))*(s[1]/s[3]) # \hat{\beta}_2
beta_hat_0 = bars[1] - beta_hat_1*bars[2] - beta_hat_2*bars[3]   # \hat{\beta}_0

# Ausgabe
cat("Korrelationen  r(y,x_1),r(y,x_2),r(x_1,x_2)        :"  , c(r[1,2],r[1,3],r[2,3]),
    "\nPartielle Korrelationen r(y,x_1|x_2), r(y,x_2|x_1) :", c(pr[1,2],pr[1,3]),
    "\nbeta_hat ALM Schätzer                              :", beta_hat,
    "\nbeta_hat aus partieller Korrelation                :", c(beta_hat_0,beta_hat_1,beta_hat_2))
```

```
> Korrelationen  r(y,x_1),r(y,x_2),r(x_1,x_2)        : -0.726 0.644 -0.0268
> Partielle Korrelationen r(y,x_1|x_2), r(y,x_2|x_1) : -0.927 0.909
> beta_hat ALM Schätzer                              : 5.42 -0.481 1.91
> beta_hat aus partieller Korrelation                : 5.42 -0.481 1.91
```

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Ausblick

Selbstkontrollfragen

Parameterinferenz | T-Tests

Zur Erinnerung (vgl. (7) Modellevaluation)

Theorem (T-Teststatistik)

Es sei

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{30}$$

das ALM in generativer Form. Weiterhin seien

$$\hat{\beta} := (X^T X)^{-1} X^T y \text{ und } \hat{\sigma}^2 := \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} \tag{31}$$

die Betaparameter- und Varianzparameterschätzer, respektive. Schließlich sei für einen *Kontrastgewichtsvektor* $c \in \mathbb{R}^p$ und einen *Nullhypothesenbetaparameter* $\beta_0 \in \mathbb{R}^p$ die *T-Teststatistik* definiert als

$$T := \frac{c^T \hat{\beta} - c^T \beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}. \tag{32}$$

Dann gilt

$$T \sim t(\delta, n - p) \text{ mit } \delta := \frac{c^T \beta - c^T \beta_0}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}} \tag{33}$$

Parameterinferenz | T-Tests

Einige mögliche Kontrastgewichtsvektoren und Nullhypothesen im Anwendungsbeispiel:

$$
\begin{array}{lll}
c = (1,0,0)^T & H_0 : \beta_1 = 0 & H_A : \beta_1 \neq 0 \\
c = (0,1,0)^T & H_0 : \beta_2 = 0 & H_A : \beta_2 \neq 0 \\
c = (0,0,1)^T & H_0 : \beta_3 = 0 & H_A : \beta_3 \neq 0 \\
c = (0,1,-1)^T & H_0 : \beta_2 - \beta_3 = 0 & H_A : \beta_2 - \beta_3 \neq 0 \\
\ldots & \ldots & \ldots
\end{array}
$$

Parameterinferenz | T-Tests

```r
# Dateneinlesen
fname      = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)         # Datensatz

# Modellschätzung
y          = D$BDI                                            # Abhängige Variable
n          = length(y)                                        # Anzahl Datenpunkte
X          = matrix(c(rep(1,n), D$Age, D$Therapy), nrow = n)  # Desigmatrix
p          = ncol(X)                                          # Anzahl Parameter
beta_hat   = solve(t(X) %*% X) %*% t(X) %*% y                 # Betaparameterschätzer
eps_hat    = y - X %*% beta_hat                               # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat) /(n-p)                  # Varianzparameterschätzer

# Modellevaluation | Parameterinferenz
C          = cbind(diag(p), matrix(c(0,1,-1), nrow = 3))      # Kontrastgewichtsvektoren
ste        = rep(NaN, ncol(C))                                # Konstraststandardfehler
tee        = rep(NaN, ncol(C))                                # T-Statistiken
pvals      = rep(NaN, ncol(C))                                # p-Werte
for(i in 1:ncol(C)){
    c        = C[,i]                                          # Kontrastgewichtsvektor
    t_num    = t(c)%*%beta_hat                                # Zähler der T-Statistik
    ste[i]   = sqrt(sigsqr_hat*t(c)%*%solve(t(X)%*%X)%*%c)    # Kontraststandardfehler/Nenner der T-Statistik
    tee[i]   = t_num/ste[i]                                   # T-Statistik
    pvals[i] = 2*(1 - pt(abs(tee[i]),n-p))                    # p-Wert
}
# Ausgabe
R            = data.frame(c(beta_hat, t(C[,4]%*%beta_hat)),ste, tee, pvals)
rownames(R)  = c("(Intercept)", "Age", "Therapy", "Age-Therapy")
colnames(R)  = c("Estimate", "Std. Error", "t value", "Pr(>|t|)")
print(R)
```

```
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)    5.422     1.9024    2.85  0.00534
> Age           -0.481     0.0198  -24.33  0.00000
> Therapy        1.912     0.0893   21.41  0.00000
> Age-Therapy   -2.393     0.0909  -26.32  0.00000
```

Modellinferenz | F-Tests

Zur Erinnerung (vgl. (7) Modellevaluation)

Theorem (F-Statistik)

Für $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ und $\sigma^2 > 0$ sei ein ALM der Form

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{34}$$

mit der Partitionierung

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, X_1 \in \mathbb{R}^{n \times p_1}, X_2 \in \mathbb{R}^{n \times p_2}, \text{ und } \beta := \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \beta_1 \in \mathbb{R}^{p_1}, \beta_2 \in \mathbb{R}^{p_2}, \tag{35}$$

mit $p = p_1 + p_2$ gegeben. Schließlich sei

$$K := \begin{pmatrix} 0_{p_1} \\ 1_{p_2} \end{pmatrix} \in \mathbb{R}^p \tag{36}$$

ein Kontrastgewichtsvektor. Dann gilt

$$F \sim f(\delta, p_2, n-p) \text{ mit } \delta := \frac{K^T\beta \left(K^T(X^TX)^{-1}K\right)^{-1} K^T\beta}{\sigma^2} \tag{37}$$

Modellinferenz | F-Tests

$p_1 := 1$

```
# Dateneinlesen
fname       = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv")
D           = read.table(fname, sep = ",", header = TRUE)      # Datensatz

# Modellevaluation
y           = D$BDI                                            # Abhängige Variable
n           = length(y)                                        # Anzahl Datenpunkte
X           = matrix(c(rep(1,n), D$Age, D$Therapy), nrow = n)  # Desigmatrix vollständiges Modell
p           = ncol(X)                                          # Anzahl Parameter vollständiges Modell
p_1         = 1                                                # Anzahl Parameter reduziertes Modell
p_2         = p - p_1                                          # Anzahl zusätzlicher Parameter im vollst. Modell
X_1         = X[,1:p_1]                                        # Designmatrix reduzierters Modell
beta_hat_1  = solve(t(X_1)%*%X_1)%*%t(X_1)%*%y                 # Betaparameterschätzer reduziertes Modell
beta_hat    = solve(t(X) %*%X )%*%t(X) %*%y                    # Betaparameterschätzer vollständiges Modell
eps_hat_1   = y-X_1%*%beta_hat_1                               # Residuenvektor reduziertes Modell
eps_hat     = y - X%*%beta_hat                                 # Residuenvektor vollständiges Modell
eh1_eh1     = t(eps_hat_1) %*% eps_hat_1                       # RQS reduziertes Modell
eh_eh       = t(eps_hat) %*% eps_hat                           # RQS vollständiges Modell
sigsqr_hat  = eh_eh/(n-p)                                      # Varianzparameterschätzer vollst. Modell
f           = ((eh1_eh1-eh_eh)/p_2)/sigsqr_hat                 # F-Statistik
pval        = 1 - pf(f,p_2,n-p)                                # p-Wert

# Ausgabe
cat("F-statistic:", f, "on", p_2, "and", n-p, "DF", "p-value: ", paste(pval))
```

```
> F-statistic: 540 on 2 and 97 DF p-value:  0
```

Modellformulierung, Modellschätzung und Modellevaluation mit R

```
fname  = file.path(getwd(), "12_Daten", "12_Multiple_Regression_Daten.csv") # Datensatzdatei
D      = read.table(fname, sep = ",", header = TRUE)                        # Datensatzeinlesen
alm    = lm(BDI ~ Age + Therapy, data = D)                                  # Modellformulierung und Modellschätzung
summary(alm)
```

```
>
> Call:
> lm(formula = BDI ~ Age + Therapy, data = D)
>
> Residuals:
>    Min     1Q Median     3Q    Max
> -7.178 -2.165  0.438  2.585  7.119
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)   5.4225     1.9024    2.85   0.0053 **
> Age          -0.4815     0.0198  -24.33   <2e-16 ***
> Therapy       1.9119     0.0893   21.41   <2e-16 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 3.07 on 97 degrees of freedom
> Multiple R-squared:  0.918,   Adjusted R-squared:  0.916
> F-statistic:  540 on 2 and 97 DF,  p-value: <2e-16
```

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Ausblick

Selbstkontrollfragen

Ausblick

Allgemeines Lineares Modell SoSe 2023

- Konfidenzintervalle
- Allgemeine Kontrasttheorie
- Kovarianzanalyse

Weiterführende Theorie des Allgemeinen Linearen Modells

Relaxation der Unabhängigkeitsannahme der Fehlerterme

$\Rightarrow$ Generalized Least Squares, Repeated-Measures Designs, ...

Modellierung von Beta- und Varianzparametern als Zufallsvariablen

$\Rightarrow$ Hierarchische lineare Modelle, linear mixed models, Bayesian estimation, Varianzkomponentenschätzung, ...

Nichtlineare Transformationen von Erwartungswertparametern

$\Rightarrow$ Generalisierte lineare Modelle, logistische Regression, neuronale Netze, ...

Multivariate Erweiterung der Datenvariable

$\Rightarrow$ Multivariate ALMs, Faktoranalyse, Strukturgleichungsmodelle, ...

Zeitliche Erweiterung der Datenvariable

$\Rightarrow$ Linear Gaussian State Space Models, Kalman Filter, Bayesian Filtering, ...

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Ausblick

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie das Anwendungsszenario und die Ziele der multiplen Regression.
2. Definieren Sie das Modell der multiplen Regression.
3. Erläutern Sie die Begriffe Regressor, Prädiktor, Kovariate und Feature im Rahmen der multiplen Regression.
4. Erläutern Sie, warum $\hat{\beta} \approx$ Regressorkovarianz^{-1}Regressordatenkovarianz gilt.
5. Erläutern Sie den Zusammenhang zwischen Betaparameterschätzern und partieller Korrelation in einem multiplen Regressionmodell mit Interzeptprädiktor und zwei kontinuierlichen Prädiktoren anhand der Formel

$$\hat{\beta}_1 = r_{y,x_1|x_2} \sqrt{\frac{1 - r^2_{y}, x_2}{1 - r^2_{x_1,x_2}}} \frac{s_y}{s_{x_1}}. \tag{38}$$

6. $X \in \mathbb{R}^{n \times 2}$ sei die Designmatrix eines multiplen Regressionsmodells mit zwei Prädiktoren und Betaparametervektor $\beta := (\beta_1, \beta_2)^2$. Geben Sie den Kontrastgewichtsvektor an, um die Nullhypothese $H_0 : \beta_1 = \beta_2$ mithilfe der T-Statistik zu testen.
7. Simulieren Sie einen Datensatz eines multiplen Regressionsmodells mit Interzept und zwei kontinuierlichen Regressoren $x_1, x_2 \in \mathbb{R}^n$, wobei $x_{i2} := ax_{i1} + \xi_i$ mit $\xi_i \sim N(0, \sigma^2_\xi)$ für $i = 1, ..., n$ sein soll. Wählen Sie für die Simulation des Datensatzes $y \in \mathbb{R}^n$ den wahren, aber unbekannten, Betaparametervektor $\beta = (0, 1, 0)^T$ und testen Sie die Nullhypothesen $H_0 : \beta_j = 0$ für $j = 0, 1, 2$. Erläutern Sie Ihre Ergebnisse. Wiederholen Sie Analyse für den wahren, aber unbekannten, Betaparametervektor $\beta = (0, 0, 1)^T$.