1 Multiple Regression

1.1 Mathematische Grundlagen

Im Allgemeinen Linearen Modell (ALM) wird die in einer gemessenen Variable enthaltene Variabilität in verschiedene Varianzquellen zerlegt, wobei Fehlerterme (= Differenzen zwischen Datenpunkten und Modellvorhersage) als multivariat normalverteilt mit Erwartungswertparameter Null und einem sphärischen Kovarianzmatrixparameter angenommen werden:

$$y = X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I_n)\ . \tag{1}$$

Hierbei ist y ein beobachteter n-dimensionaler Datenvektor, X eine festgelegte $n \times p$ Designmatrix, ε ein nicht-beobachtbarer n-dimensionaler Fehlervektor und I_n die n-dimensionale Einheitsmatrix.

Die wahren, aber unbekannten Parameter des ALMs, der p-dimensionale Betaparametervektor β und der Varianzparameter $\sigma^2 > 0$, werden typischerweise mithilfe folgender Formeln geschätzt:

$$\begin{aligned} \hat{\beta} &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \\ \hat{\sigma}^2 &= \frac{1}{n-p}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})\ . \end{aligned} \tag{2}$$

Hierbei wird $\hat{\beta} \in \mathbb{R}^p$ als Betaparameterschätzer und $\hat{\sigma}^2 \in \mathbb{R}_{>0}$ als Varianzparameterschätzer bezeichnet.

Zum Zwecke der Modellevaluation (d.h. für Hypothesentests und Konfidenzintervalle) im ALM kann auf Grundlage der Modellparameterschätzer eine T-Statistik wie folgt berechnet werden:

$$T = \frac{c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta_0}{\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} \tag{3}$$

Hierbei sind c ein p-dimensionaler Kontrastgewichtsvektor und β_0 ein Nullparametervektor gemäß Nullhypothese, die zusammen die Nullhypothese $H_0 : c^{\mathrm{T}}\beta = c^{\mathrm{T}}\beta_0$ festlegen, während $\hat{\beta}$ und $\hat{\sigma}^2$ die ALM-Parameterschätzer darstellen.

Gemäß der Theorie der Frequentistischen Inferenz für das ALM folgt diese T-Statistik unter der Nullhypothese einer T-Verteilung mit einem Freiheitsgradparameter, der sich aus den Dimensionen der Designmatrix n und p ergibt:

$$T \sim t(n-p), \quad \text{wenn } H_0 \text{ zutrifft}\ . \tag{4}$$

Vollzieht man einen zweiseitigen T-Test der Nullhypothese, so ergibt sich der p-Wert

$$\text{p-Wert} = 2(1 - \psi(|T|; n - p)) \ , \tag{5}$$

wobei $\psi(x; n)$ die kumulative Verteilungsfunktion (KVF) der t-Verteilung mit Freiheitsgradparameter n ist.

1.2 Analyse in R

Die Datei `FADE_SAME.csv` enthält den in der ersten Seminarsitzung vorgestellten Datensatz. Erklären Sie die Funktion des folgenden R-Codes:

```
# Daten einlesen
fname = 'FADE_SAME.csv'                          # Dateiname
D     = read.csv(fname)                          # Dataframe

# Variablen extrahieren
n     = 259                                      # Anzahl Personen in Studie A
y     = as.matrix(D[1:n,'memory.SAME'])          # Memory-SAME-Score dieser Personen
x1    = as.matrix(D[1:n,'age'])                  # Alter in Jahren
x2    = as.matrix(D[1:n,'memory'])               # Gedächtnisleistung

# Designmatrix erzeugen
p     = 3                                        # Anzahl Spalten der Designmatrix
X     = matrix(rep(1,n))                         # erste Spalte: konstanter Regressor
X     = cbind(X, (x1>50))                        # zweite Spalte: Indikator-Regressor
X     = cbind(X,  x2-mean(x2))                   # dritte Spalte: kontinuierlicher Regressor
print(dim(y))                                    # Überprüfung Datenvektor
print(dim(X))                                    # Überprüfung Designmatrix
```

1.3 Erste Programmieraufgabe

Überprüfen Sie mithilfe eines Hypothesentests, ob ein statistisch signifikanter Unterschied im Memory-SAME-Score zwischen der Gruppe der jungen Erwachsenen (18-35 Jahre) und der Gruppe der älteren Erwachsenen (>50 Jahre) besteht, wenn zugleich der Effekt von Gedächtnisleistung berücksichtigt wird. Gehen Sie dazu wie folgt vor:

- Bestimmen Sie die Parameterschätzer des durch `y` und `X` beschriebenen ALMs mithilfe der obigen Formeln. In R werden Matrizen mit `t()` transponiert, mit `%*%` multipliziert und mit `solve()` invertiert, sodass beispielsweise $(X^{\mathrm{T}}X)^{-1}$ durch `solve(t(X) %*% X)` berechnet wird.
- Legen Sie den Kontrastgewichtsvektor auf $c = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^{\mathrm{T}}$ und den Nullparametervektor auf $\beta_0 = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^{\mathrm{T}}$ fest. Der Spaltenvektor $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^{\mathrm{T}}$ kann in R z.B. durch `matrix(c(1,2,3), ncol = 1)` erzeugt werden.

- Ermitteln Sie den Wert der T-Statistik gemäß der obigen Formel. Hierbei bietet es sich an, den Zähler $c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta_0$ und den Nenner $\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}$ separat zu berechnen und dann zu dividieren. Die Wurzel wird in R mit `sqrt()` berechnet.
- Ermitteln Sie den p-Wert gemäß der obigen Formel. In R heißt die kumulative Verteilungsfunktion der t-Verteilung `pt()`. Informieren Sie sich über die Eingabeparameter dieser Funktion, um den p-Wert korrekt zu bestimmen.
- Geben Sie die Resultate ihrer Analyse aus. Sie sollten folgende Ergebnisse erhalten:

```
Anzahl Datenpunkte       :  259
Anzahl Regressoren       :  3
Betaparameterschätzer    :  -0.153 -0.907 3.005
Varianzparameterschätzer :  0.427
Kontrastgewichtsvektor   :  0 1 0
T-Statistik              :  -10.363
p-Wert                   :  0
```

1.4 Abbildung in R

Im nächsten Schritt sollen die Ergebnisse visualisiert werden. Erklären Sie den folgenden R-Code und die daraus resultierende Abbildung:

```r
# Visualisierung
library(latex2exp)
par(
    family     = "sans",
    pty        = "m",
    bty        = "o",
    lwd        = 1,
    las        = 1,
    mgp        = c(2,1,0),
    yaxs       = "i",
    cex        = 1.2)

# Punktwolken
plot(x2[x1<50], y[x1<50],
    pch        = 16,
    col        = "black",
    xlab       = "Gedächtnisleistung",
    ylab       = "Memory-SAME-Score",
    xlim       = c( 0.5, 1.0),
    ylim       = c(-3.5, 2.5))
points(x2[x1>50], y[x1>50],
    pch        = 16,
    col        = "gray80")

# Regressionsgeraden
y_hat          = X %*% beta_hat
```

```r
lines(x2[x1<50], y_hat[x1<50],
    col         = "black")
lines(x2[x1>50], y_hat[x1>50],
    col         = "gray80")
legend("topleft", c("junge Erwachsene", "ältere Erwachsene"),
    lty         = 0,
    pch         = 16,
    col         = c("black", "gray80"),
    bty         = "n",
    cex         = 1)

# Speichern
dev.copy2pdf(
    file        = "Abbildungen/Multiple_Regression_1.pdf",
    width       = 9,
    height      = 9)
```

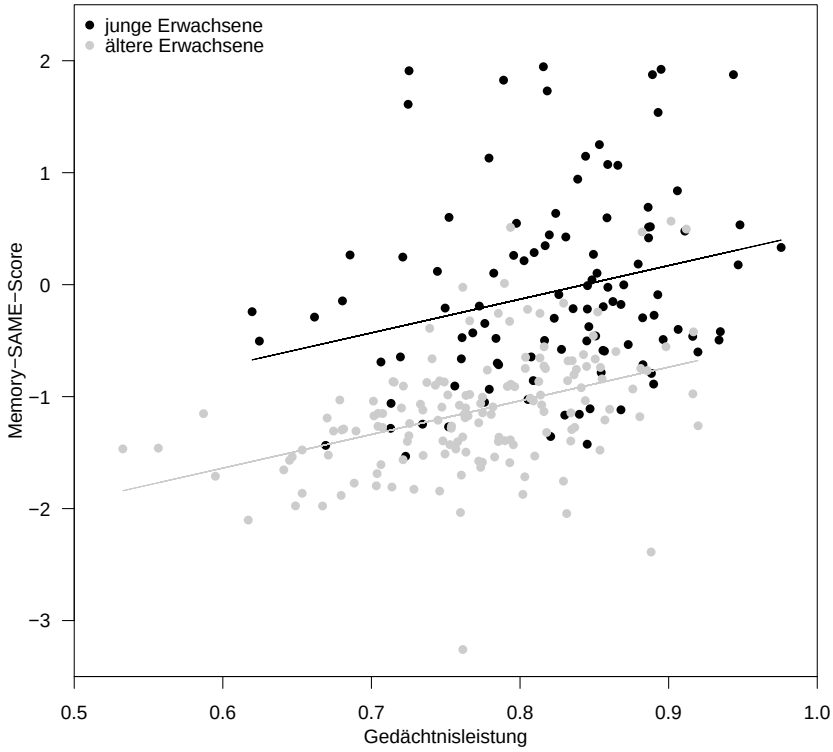

Abbildung 1. Zusammenhang zwischen Memory-SAME-Score, Gedächtnisleistung und Altersgruppe.

1.5 Zweite Programmieraufgabe

Verändern Sie diesen Code nun so, dass den beiden Regressionsgeraden in der Abbildung erlaubt wird, verschiedene Anstiegsparameter zu haben. Gehen Sie dazu wie folgt vor:

- Benutzen Sie die R-Funktion `polyfit()` aus dem Paket `pracma`, um Offset- und Anstiegsparameter separat für die Datensätze der jungen (`x2[x1<50], y[x1<50]`) bzw. älteren (`x2[x1>50], y[x1>50]`) Erwachsenen zu berechnen. Der Grad des zu bestimmenden Polynoms muss auf 1 gesetzt werden.
- Ersetzen Sie die Berechnung der prädizierten Daten `y_hat` im vorangegangenen Code so, dass die Datenvorhersage für junge und ältere Erwachsene separat mit Ausgleichsgeraden der Form `y_hat = b[1]*x2 + b[2]`, aber auf Basis ihrer jeweiligen Offset- und Anstiegsparameter erfolgt.
- Erzeugen Sie die Abbildung wie oben. Sie sollten in etwa folgende Abbildung erhalten:

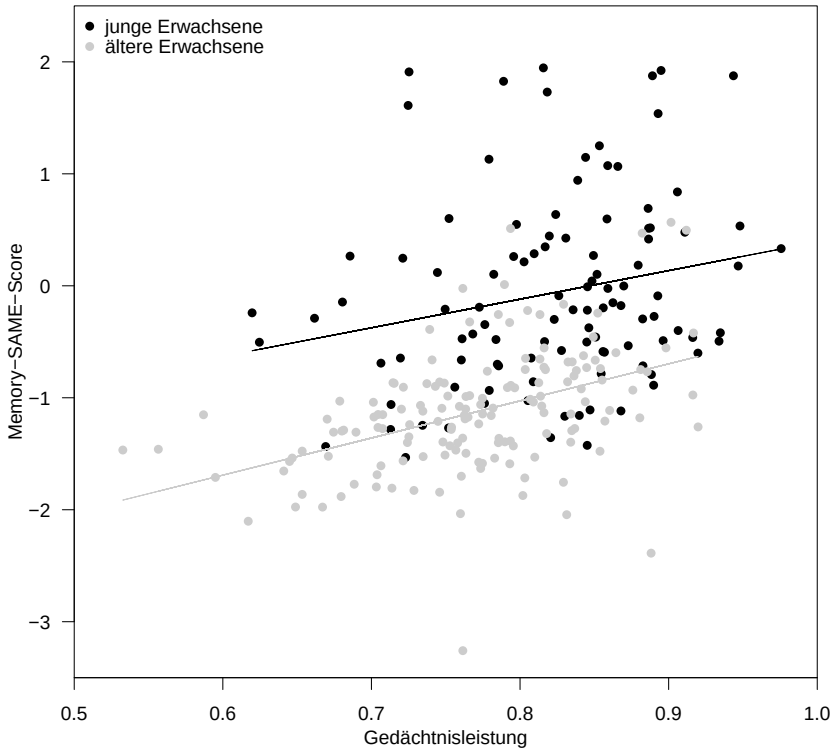

Abbildung 2. Zusammenhang zwischen Memory-SAME-Score, Gedächtnisleistung und Altersgruppe.

1.6 Lückentext

Füllen Sie mit den in der Übung erzielten Ergebnissen den folgenden Lückentext aus und präsentieren Sie die Ergebnisse im Seminar:

Lückentext: Die Regressionskoeffizienten werden auf _____, _____ und _____ geschätzt und der Varianzparameterschätzer ist _____. Dies bedeutet, dass der durchschnittliche Memory-SAME-Score in der Referenzgruppe der jungen Erwachsenen _____ ist, die durchschnittliche Abweichung hiervon in der Gruppe der älteren Erwachsenen _____ beträgt und sich der Score pro Einheit Gedächtnisleistung um _____ erhöht. Die T-Statistik für den Effekt von Altersgruppe (ältere vs. junge Erwachsene) beträgt _____. Die Anzahl der Freiheitsgrade beträgt _____, es ergibt sich ein p-Wert von _____. D.h., der Effekt von Altersgruppe ist _____ (negativ/positiv) und statistisch _____ (signifikant/nicht signifikant).

1.7 Mögliche Klausurfrage

Präsentieren Sie im Seminar folgende Klausurfrage und erklären Sie die richtige Antwort:

Frage: $X \in \mathbb{R}^{n\times 3}$ und $\beta = \begin{pmatrix}\beta_0 & \beta_1 & \beta_2\end{pmatrix}^{\mathrm{T}}$ seien die Designmatrix und der Betaparametervektor eines multiplen Regressionsmodells. Welche Nullhypothese wird durch die T-Statistik mit dem Kontrastgewichtsvektor $c = \begin{pmatrix}0 & 1 & -1\end{pmatrix}^{\mathrm{T}}$ getestet?

a) H_0: $\beta_1 - \beta_2 = 0$

b) H_0: $\beta_1 + \beta_2 = 0$

c) H_0: $\beta_1 = 0 \ \wedge \ \beta_2 = 0$

d) H_0: $\beta_0 = 0 \ \wedge \ \beta_1 = 1 \ \wedge \ \beta_2 = -1$

1.8 Kinderwitz

Was versteht man unter einer Turbine?

Antwort: Nichts, ist ja viel zu laut.