

## Multivariate Regression

### Aufgabe 1:

In dieser Aufgabe wird ein Datensatz mit 3167 Wettkampfergebnissen der weltbesten Zehnkämpfer aus den Jahren 1998 bis 2009 betrachtet (siehe [www.decathlon2000.ee](http://www.decathlon2000.ee)). Der Datensatz enthält 3167 Zeilen mit 'rohen' Resultaten der Athleten in den 10 Disziplinen. Die rohen Resultate wurden mit Hilfe der Formeln von [www.iaaf.org](http://www.iaaf.org) zusätzlich in das im Zehnkampf übliche Punktesystem umgewandelt. Ziel ist es, ein multivariates Regressionsmodell anzupassen. Dazu werden als multivariate Zielgröße die Ergebnisse in 100m, Weitsprung, 400m und 110m Hürden betrachtet. Als Kovariablen werden Jahr und Monat des Wettkampfes sowie das Alter eines Athleten in das Modell aufgenommen. In R ergibt sich folgende Modellformel:

```
lm(cbind(M100, LJ, M400, MH110) ~ month + year + age, data=decathlon)
```

Der R-Output dieses multivariaten linearen Regressionsmodells befindet sich auf Seite 2-3. Die folgenden Teilaufgaben beziehen sich auf diesen Output.

- (a) Geben Sie die zugrunde liegende Modellformel in Matrixschreibweise an. Erklären Sie jeweils die einzelnen Komponenten.
- (b) Wie lautet die geschätzte Parametermatrix  $\hat{\mathbf{B}}$ ? Interpretieren Sie die Elemente dieser Matrix.
- (c) Wie würde der KQ-Schätzer eines univariaten Regressionsmodells lauten, wenn als Zielgröße nur Weitsprung in Abhängigkeit von Jahr, Monat und Alter untersucht wird?
- (d) Übersetzen Sie die folgenden Fragestellungen in Hypothesen der Form  $H_0 : \mathbf{CBD} = \mathbf{\Gamma}$ .
  - i. Trägt das Modell grundsätzlich zur Erklärung der Zielgrößen bei?
  - ii. Gibt es einen signifikanten Zusammenhang zwischen dem Alter eines Athleten und den Zielgrößen 100m, Weitsprung, 400m und 110m Hürden?
  - iii. Gibt es einen signifikanten Unterschied zwischen dem Zusammenhang von Monat, von Jahr und von Alter mit den vier Responsevariablen?

Geben Sie für jede Fragestellung die Matrizen  $\mathbf{C}$ ,  $\mathbf{D}$  und  $\mathbf{\Gamma}$  an.

- (e) Welche der folgenden Hypothesen könnten mit univariaten Regressionsmodellen überprüft werden? Für welche wäre ein multivariates Modell nötig?
  - i. Gibt es einen signifikanten Zusammenhang zwischen dem 100m-Lauf und dem Monat eines Wettkampfes?
  - ii. Gibt es einen signifikanten Unterschied zwischen dem Zusammenhang von Alter vs. 100m und dem Zusammenhang von Alter vs. 400m?
  - iii. Gibt es einen signifikanten Zusammenhang zwischen den Weitsprung-Ergebnissen und den drei Kovariablen Jahr, Monat und Alter?

- (f) Der folgende R-Output zeigt die Ergebnisse der Tests auf einzelne Variablen im Modell. Beschreiben Sie zunächst im Allgemeinen die Struktur eines solchen Tests auf einzelne Variablen (Hypothesen, Teststatistik und Ablehnbereich). Interpretieren Sie dann für jede Kovariable getrennt das Ergebnis.

```
Analysis of Variance Table

            Df    Wilks approx F num Df den Df    Pr(>F)
(Intercept)  1 0.00391   201378      4   3160 < 2.2e-16 ***
month        1 0.98981         8      4   3160 1.605e-06 ***
year         1 0.99518         4      4   3160 0.004164 **
age          1 0.95637        36      4   3160 < 2.2e-16 ***
Residuals    3163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## R-Modelloutput

```
Response M100 :
Call:
lm(formula = M100 ~ month + year + age, data = decathlon)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-197.131  -40.374   -1.127   40.969  208.470
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2762.3468   628.2478   4.397 1.13e-05 ***
month        -0.7089    0.6476  -1.095  0.27374
year         -0.9810    0.3134  -3.130  0.00176 **
age           0.8198    0.3151   2.602  0.00932 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 61.02 on 3163 degrees of freedom
Multiple R-squared:  0.00571,    Adjusted R-squared:  0.004767
F-statistic: 6.055 on 3 and 3163 DF,  p-value: 0.0004163
```

```
Response LJ :
Call:
lm(formula = LJ ~ month + year + age, data = decathlon)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-220.804  -51.648   -3.296   49.540  273.128
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1353.8735   766.7895   1.766  0.0776 .
month         3.1971    0.7904   4.045 5.36e-05 ***
year        -0.3178    0.3825  -0.831  0.4062
age          3.1526    0.3846   8.198 3.52e-16 ***
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.48 on 3163 degrees of freedom  
Multiple R-squared: 0.02672, Adjusted R-squared: 0.0258  
F-statistic: 28.94 on 3 and 3163 DF, p-value: < 2.2e-16

Response M400 :

Call: lm(formula = M400 ~ month + year + age, data =  
decathlon)

Residuals:

| Min      | 1Q      | Median | 3Q     | Max     |
|----------|---------|--------|--------|---------|
| -253.156 | -44.775 | 2.295  | 46.051 | 217.322 |

Coefficients:

|             | Estimate  | Std. Error | t value | Pr(> t )     |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | 3136.5579 | 675.8570   | 4.641   | 3.61e-06 *** |
| month       | -0.8083   | 0.6967     | -1.160  | 0.246046     |
| year        | -1.1693   | 0.3371     | -3.468  | 0.000531 *** |
| age         | 0.1286    | 0.3390     | 0.379   | 0.704466     |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.64 on 3163 degrees of freedom  
Multiple R-squared: 0.00428, Adjusted R-squared: 0.003336  
F-statistic: 4.532 on 3 and 3163 DF, p-value: 0.003555

Response MH110 :

Call:  
lm(formula = MH110 ~ month + year + age, data = decathlon)

Residuals:

| Min     | 1Q     | Median | 3Q    | Max    |
|---------|--------|--------|-------|--------|
| -322.79 | -43.62 | 1.77   | 44.27 | 191.22 |

Coefficients:

|             | Estimate  | Std. Error | t value | Pr(> t )   |
|-------------|-----------|------------|---------|------------|
| (Intercept) | 1164.9826 | 670.8829   | 1.736   | 0.0826 .   |
| month       | 0.2334    | 0.6915     | 0.337   | 0.7358     |
| year        | -0.2058   | 0.3347     | -0.615  | 0.5387     |
| age         | 3.2795    | 0.3365     | 9.747   | <2e-16 *** |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.16 on 3163 degrees of freedom  
Multiple R-squared: 0.02954, Adjusted R-squared: 0.02862  
F-statistic: 32.09 on 3 and 3163 DF, p-value: < 2.2e-16