Regression I — Serie 3

1. Es seien

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 2 & 4 \\ -1 & -1 & 0 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 5 \\ 3 \\ -4 \end{bmatrix}.$$

Berechnen Sie mit R die folgenden Ausdrücke, falls sie definiert sind.
Lösen Sie mindestens a) - d) von Hand.

a) $2 \cdot \mathbf{A}$ **b)** $\mathbf{A} + \mathbf{B}$ **c)** $\mathbf{A} \cdot \mathbf{B^T}$ **d)** $\mathbf{A} \cdot \mathbf{x}$
e) $\mathbf{A} \cdot \mathbf{B}$ **f)** $\mathbf{B^T} \cdot \mathbf{y}$ **g)** $\mathbf{A} \cdot \mathbf{A^T}$ **h)** $\mathbf{A^T} \cdot \mathbf{A}$
i) $\mathbf{x^T} \cdot \mathbf{x}$ **j)** $\mathbf{x} \cdot \mathbf{x^T}$

2. Regressionsmodell in Matrix-Schreibweise, Simulation
Wir betrachten ein multiples lineares Modell in Matrix-Schreibweise:
$\underline{Y} = \boldsymbol{x}\underline{\beta} + \underline{E}$ mit den Koeffizienten $\beta_0 = 10$, $\beta_1 = 5$, $\beta_2 = -2$.

Die erklärenden Variablen eines entsprechenden Datensatzes für die Regressionsrechnung seien in folgender Tabelle gegeben.

i	$x_i^{(1)}$	$x_i^{(2)}$
1	0	4
2	1	1
3	2	0
4	3	1
5	4	4

a) Berechnen Sie die "wahren Werte" der Regressionsfunktion $\mathcal{E}\langle Y_i \rangle = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)}$.
Hinweis: Fügen Sie "vorne" an die Tabelle eine Spalte mit Einsen an.
Setzen Sie: `t.beta <- c(10,5,-2)`

b) Erzeugen Sie zufällige, normalverteilte Fehler $E_i \sim \mathcal{N}(0, \sigma^2 = 1)$ und addieren Sie diese zu den wahren Werten, um damit beobachtete Werte zu simulieren. Berechnen Sie die geschätzten Koeffizienten mit der Funktion `lm()`.

c) Auf diese Weise können Sie jetzt 100 geschätzte Koeffizienten-Vektoren erzeugen und deren Verteilung grafisch darstellen: Koeffizienten als Funktion der Simulationsnummer, Streudiagramme von $\widehat{\beta}_1$ vs. $\widehat{\beta}_0$; $\widehat{\beta}_2$ vs. $\widehat{\beta}_0$; $\widehat{\beta}_2$ vs. $\widehat{\beta}_1$.
R-Hinweise: Die 100 Simulationen erzeugen Sie am elegantesten wie folgt:

1. Fehlermatrix E der Dimension (5×100) erzeugen. Pro Spalte sind die Fehler eines Experimentes enthalten.
`t.E <- matrix(rnorm(500),ncol=100)`

2. Eine Matrix mit 100 identischen Spalten erzeugen, welche pro Spalte jeweils die "wahren Werte" `t.y` enthält:
   ```
   t.Y <- matrix(rep(t.y,100),nrow=5,byrow=F)
   ```
3. Daraus die simulierten Beobachtungen berechnen.
   ```
   t.Y <- t.E + t.Y
   ```
4. Eine Resultatmatrix der Dimension 100×3 definieren.
   Entweder mit einer for-Schleife die 100 Experimente auswerten und die Koeffizienten pro Experiment in einer Zeile der Resultatmatrix speichern,
   ```
   r.coef <- matrix(nrow=100,ncol=3)
   for (i in 1:100)
   {
   r.coef[i,] <- lm(t.Y[,i]~t.x[,2]+t.x[,3])$coefficients
    },
   ```
   oder das Ganze etwas eleganter mit `apply` lösen
   ```
   r.coef <- t(apply(t.Y,2,FUN=function(y) lm...
   ```

3. Der Datensatz `antkoerp` enthält die leicht abgeänderten Daten zum Beispiel der Antikörper-Produktion aus dem Skript (1.h). Die Variablen sind:

`raddos`	Dosis von Co^{60} Gamma-Strahlen
`zeit`	Anzahl Tage zwischen der Bestrahlung und der Injektion eines Öls
`y`	Menge der produzierten Antikörper

a) Betrachten Sie das Modell $y_i = \alpha + \beta\ \texttt{raddos}_i + E_i$. Ist β signifikant von 0 verschieden? Ist das Modell gut?

b) Betrachten Sie das quadratische Modell $Y_i = \beta_0 + \beta_1\ \texttt{raddos}_i + \beta_2\ (\texttt{raddos}_i)^2 + E_i$. Vergleichen Sie die Resultate mit den Resultaten aus a). Welches Modell passt besser?
R-Hinweis: Den Term $\beta_2\ (\texttt{raddos}_i)^2$ kann man mit `I(raddos^2)` ins Modell nehmen.

c) Sind die Fehler normalverteilt? Erzeugen Sie ein Normalverteilungs-Diagramm (normal plot).

d) Zeichnen Sie in einem Streudiagramm `y` und die mit dem quadratischen Modell geschätzten Werte gegen `raddos`. Wie gut passt der Fit?