

Bernd Klaus (bernd.klaus@imise.uni-leipzig.de)
Verena Zuber (verena.zuber@imise.uni-leipzig.de)

<http://uni-leipzig.de/~zuber/teaching/ws09/r-kurs/>

1 Aufgabe: R als Taschenrechner, Matrixmultiplikation

Starten Sie R.

(a) Erzeugen Sie folgende Matrizen:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{pmatrix} \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

(b) Berechnen Sie:

- Die Determinante der Matrix A . Was kann man aus dieser Determinante schließen?
- Invertieren und transponieren Sie A .

(c) Multiplizieren Sie die erste Zeile der Matrix A mit der zweiten Spalte der Matrix:

- elementweise
- als Matrixmultiplikation
- transponieren Sie abschließend beide Vektoren vor der Matrixmultiplikation

(d) Berechnen Sie den *Least Squares* Schätzer (mehr dazu im Kapitel über lineare Regression):

$$beta = (A^t A)^{-1} A^t y$$

2 Aufgabe: Sigmoidfunktion

Die Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ gemäß

$$f(x) = \frac{1}{1 + \exp(-x)}$$

bezeichnet man als *Sigmoidfunktion*. Sie spielt z.B. bei der Bestimmung von Reaktionsgleichgewichten eine wichtige Rolle. Eine Implementierung findet sich in dem R-Paket **e1071**.

- (a) Finden Sie heraus, wie man Werte dieser Funktion in R mit Hilfe des Pakets **e1071** berechnen lassen kann.
- (b) Bestimmen Sie die Funktionswerte für die Stellen $-2, -1.8, -1.6, \dots, +2$ und speichern Sie diese in einem Vektor **sig** ab.

3 Aufgabe: Umgang mit einem kleinen Datensatz

- (a) Lesen Sie den Datensatz *Patienten.csv* mit der Funktion `read.csv` ein.
- (b) Überprüfen Sie, ob Sie die Daten wirklich als Datensatz eingelesen haben.
- (c) Welche Variablen gibt es und welche Werte nehmen die Variablen an?
- (d) Besteht ein fehlender Wert beim Gewicht? Wenn ja, ersetzen Sie diesen durch den Mittelwert der gegebenen anderen Variablenwerte.
- (e) Berechnen Sie die mittlere Größe und Gewicht der Patienten.

4 Aufgabe: Datensatz OECD

Der Datensatz OECD enthält Variablen (Stand 2009), die das Wohlergehen von Kindern in den Mitgliedsstaaten messen sollen. Abgefragt wurde:

- Einkommen: das durchschnittliche Einkommen der Eltern in US Dollar
- Armut: der Anteil [in Prozent] an Kindern in einem armen Elternhaus
- Bildung: der Anteil [in Prozent] an Kindern, die ohne eine Grundausstattung (Bücher, Schreibtisch, Computer, Internet) für Bildung auskommen
- WenigRaum: der Anteil [in Prozent] an Kindern, die auf zu wenig Raum wohnen
- Umwelt: der Anteil [in Prozent] an Kindern, die unter schlechten Umweltbedingungen leben
- Lesen: mittlerer PISA-Score zur Lesefähigkeit
- Geburtsgewicht: der Anteil [in Prozent] an Kindern, die bei der Geburt weniger als 2.5kg wiegen
- Säuglingssterblichkeit: Säuglingssterblichkeit (<1 Jahr) [x in Tausend]
- Sterblichkeit: Sterblichkeit (<20 Jahre) [x in 100 000]
- Selbstmord: Selbstmord von Jugendlichen im Alter von 15 bis 19 [x in 100 000]
- Bewegung: der Anteil [in Prozent] an 11, 13 und 15 jährigen Jugendlichen, die sich regelmäßig bewegen
- Rauchen: der Anteil [in Prozent] an 15 jährigen Jugendlichen, die mindestens einmal die Woche rauchen
- Alkohol: der Anteil [in Prozent] an 13-15 jährigen Jugendlichen, die mindestens zweimal betrunken waren
- Bullying: der Anteil [in Prozent] an Kindern, die angeben in der Schule bedroht zu werden
- Schule: der Anteil [in Prozent] an Kindern, die angeben die Schule zu mögen

- (a) Lesen Sie den Datensatz *oecd* mit der Funktion `data<-read.csv(file="Daten/oecd", header=TRUE)` ein und überprüfen Sie die Dimension der Daten.
- (b) Berechnen Sie die Mittelwerte und Varianzen der einzelnen Variablen mit dem geeigneten `apply` Befehl.
- (c) Überprüfen Sie, ob die Niederlande in der Länderliste des Datensatzes auftaucht. Gibt es auch einen Eintrag für China? (Benutzen sie die R-Hilfe, um herauszufinden wie man auf die Ländernamen zugreifen kann.)
- (d) In welchem Land waren die meisten Jugendlichen mindestens zweimal betrunken? Wie hoch ist der maximale Prozentsatz?
- (e) In welchem Land ist die Säuglingssterblichkeit am geringsten? Wie hoch ist sie in diesem Land?
- (f) In welchen Ländern ist der Prozentsatz an Jugendlichen, die sich regelmäßig bewegen kleiner als der Durchschnitt?
- (g) Für welche Länder werden besonders viele Kinder in der Schule bedroht? Als Indikator für "besonders viel" soll ein Bullying Wert gelten, der mindestens eine Standardabweichung (standard deviation) vom Mittelwert aller Ländern entfernt ist.
- (h) Erstellen Sie einen neuen Datensatz, der aufsteigend nach dem Einkommen geordnet ist. Speichern Sie diesen in einer neuen *.csv* Datei.