Aufgabe 1:

Betrachten Sie das multiple Regressionsmodell

$$Y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

mit $x_{1i} = i, x_{2i} = i \bmod 2, \epsilon_i \sim N(0, 1/3)$ unabhängig, $\beta_0 = 7, \beta_1 = 0.01$ und $\beta_2 = 1$.

(a) Simulieren Sie drei Datensätze nach diesem Modell mit $n = 10, n = 100$ und $n = 1000$. Berechnen Sie für jeden dieser Datensätze die Parameterschätzungen für $\beta_0, \beta_1, \beta_2$ und σ^2 und diskutieren Sie die Ergebnisse. Erstellen Sie auch Residualplots und aussagekräftige Graphiken, in denen Sie jeweils Y_i mit $\hat{Y}_i$ vergleichen.

(b) Wie lauten die Ergebnisse, wenn man den Einfluss von x_{1i} ignoriert? Wie, wenn man den von x_{2i} ignoriert? Erstellen Sie auch hier wieder entsprechende Graphiken.

(c) Betrachten Sie nun verschiedene Modellabweichungen von (1) mit $n = 1000$:

i) Setzen Sie in obigen Datensatz für eine Beobachtung Y_i einen extremen Wert (Ausreißer) ein

ii) Simulieren Sie einen neuen Datensatz nach (1), jetzt jedoch mit auf $[-1, 1]$ gleichverteilten Fehlern ϵ_i.

iii) Simulieren Sie einen Datensatz nach (1), jetzt jedoch mit heteroskedastischen Fehlern $\epsilon_i \sim N(0, 2 \cdot i/300)$.

iv) Simulieren Sie einen Datensatz nach Modell (1), wobei Sie den Term $\beta_1 \cdot x_{1i}$ durch $\beta_1 \cdot x_{1i}^2/100$ ersetzen.

Berechnen Sie Parameterschätzungen unter Annahme des Modells (1) und zeigen Sie, welche diagnostischen Residualplots auf die obigen Modellabweichungen hindeuten. Vergleichen Sie schließlich wieder Y_i mit $\hat{Y}_i$ anhand von Graphiken.

Aufgabe 2:

Datensatz 5[1] auf der Übungshomepage enthält eine Untersuchung von 24 Paaren für welche Einflussfaktoren auf die Anzahl Spermien des männlichen Partners untersucht worden sind. Folgenden Variablen sind im Datensatz enthalten:

Variable	Beschreibung
`count`	gemessene Anzahl Spermien der männlichen Versuchsperson
`f.age`	Alter Frau (in Jahren)
`f.weight`	Gewicht Frau (in kg)
`f.height`	Größe Frau (in cm)
`m.age`	Alter Mann (in Jahren)
`m.weight`	Gewicht Mann (in kg)
`m.height`	Größe Mann (in cm^2)
`m.vol`	Testikelgröße Mann

(a) Sie möchten nun den Zusammenhang zwischen den diversen potentiellen Einflussfaktoren und der Anzahl Spermien untersuchen. Passen Sie hierzu ein Modell mit `count` als Zielvariable und allen Einflussvariablen an. Erzeugen sie sich hierzu zunächst einen Datensatz der keine fehlenden Werte (`NA`) enthält. Speichern Sie das Ergebnis im Objekt `lm.spermc`.
Hinweis: Nutzen Sie die Funktion `na.omit`.

(b) Erkären Sie die Idee, die dem variance inflation factor zugrunde liegt und berechnen Sie diesen für das vorliegende Modell.

(c) Betrachten Sie die standardmäßig ausgegebenen Diagnostik-Plots für das Modell aus (a):

```
layout(matrix(1:4, nrow = 2))
lm.spermc <- lm(count ~ ., data = sc2)
plot(lm.spermc)
```

Geben Sie stichpunktartig an, welche Größen die jeweilige Grafik angibt und welchen Nutzen Sie für die Modelldiagnose hat. Welche Beobachtung ist auf jeder Graphik auffällig?

(d) Berechnen Sie das *Leverage* dieses Punktes sowie die dazugehörige *Cook's Distance* (Achten Sie dabei darauf, dass die Indizierung in den Plots sich auf den Originaldatensatz bezieht). Geben Sie hierfür zunächst die allgemeinen Definitionen der beiden Kenngrößen an und gehen Sie kurz darauf ein wie man diese interpretiert. Sind die berechneten Werte für die Beobachtung aus Aufgabe (c) auffällig?

(e) Passen Sie nun ein Modell ohne diesen Punkt an. Stellen Sie die Koeffizientenschätzer sowie Konfidenzintervalle dieses Modells den Schätzern und KIs des Modells aus (a) graphisch gegenüber. Welches Modell würden Sie bevorzugen?

[1] Backer, R.R. and M.A. Bellis (1993). Human sperm competition: ejaculate adjustment by males and the function of masturbation.