Aufgabe 1:

(Simulation von fehlenden Daten)

(a) Erzeugen Sie komplette Daten aus einem LMM ohne fixe Einflussgrößen mit zufälligem Achsenabschnitt für $N = 10$ Personen und $n = 2$ Beobachtungen pro Person. Erstellen Sie auf diese Weise den Datensatz `sim1`. Speichern Sie auch die erzeugten zufälligen Intercepts $b_i, i = 1, \ldots, N$, für jede Person.

(b) Sei R_i ein Indikator für die i'te Person, d.h. mit $R_i = 1$ wenn Y_{i2} beobachtet wurde und $R_i = 0$ wenn Y_{i2} fehlt. Von Interesse ist es die Wahrscheinlichkeit $\pi_i = P(R_i = 1|Y_{i1}, Y_{i2}, X)$ für fehlende Werte bzw. Dropout zum Zeitpunkt 2 mit Hilfe von logistischer Regression zu modellieren. Im Fall ohne Kovariablen wäre ein mögliches Modell: $R_i \sim \text{Bernoulli}(\pi_i)$ mit

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 Y_{i1} + \beta_2 Y_{i2}.$$

Geben Sie die Restriktionen für $(\beta_0, \beta_1, \beta_2)$ an wenn der Dropout

(i) MCAR
(ii) MAR
(iii) NMAR

ist.

(c) Erzeugen Sie für jede der $N = 10$ Personen eine Realisation des Indikators R_i für $\beta_0 = \beta_1 = 0$ und $\beta_2 = 1$. Falls $R_i = 0$ setzten Sie $Y_{i2} =$ `NA`. Erstellen Sie auf diese Weise den Datensatz `sim2`.

(d) Wiederholen Sie Teil (b) in dem Sie das Modell

$$\text{logit}\, P(R_i = 1|b_i) = \alpha_0 + \alpha_1 b_i,$$

mit $\alpha_0 = -1$ und $\alpha_1 = 1$ benutzen. Erstellen Sie auf diese Weise den Datensatz `sim3`. Um was für einen Dropout-Mechanismus handelt es sich?

(e) Fitten Sie für jedes der drei oben erstellten `data.frames sim1, sim2` und `sim3` ein LMM mit Random-Intercept.
Hinweis: Konvertieren Sie die Daten mit der Funktion `reshape` in Long-Format und benutzen Sie die Option `na.action=na.exclude` in der Funktion `lme` um fehlende Y_{i2} zu ignorieren.

(f) Simulieren Sie 500 vollständige Datensätze wie oben und erzeugen Sie daraus jeweils Datensätze mit Dropouts nach den obigen 2 Mechanismen. Schätzen Sie jeweils ein LMM mit Random-Intercept und vergleichen Sie die resultierenden Verteilungen der Schätzer für den

Intercept, für die Varianz des zufälligen Intercepts und die Varianz der Fehler zwischen den 3 verschiedenen Dropout-Prozessen.