

Aufgabe 1:

In der Amsterdamer Wachstums- und Gesundheitsstudie wurde das Gesamt-Serumcholesterol (TC) von 147 Probanden insgesamt 6 Mal pro Proband gemessen. Zu Beginn der Studie 1977 ($t = 0$) waren alle Probanden 13 Jahre alt. Die ersten vier Messungen wurden im Jahresabstand, die letzten beiden Messungen 1985 und 1991 durchgeführt. Die Daten sind in der Datei `amsterdam.txt` von der Webseite der Veranstaltung erhältlich – die Datei `amsterdam-description.txt` enthält eine genauere Beschreibung der Daten und Kovariablen.

- (a) Erstellen Sie mit der Funktion `xypplot` einen Spaghettiplot der TC-Messungen für die vier möglichen Kombinationen von `smoker` und `gender`. Benutzen Sie `bwplot`, um für die beiden Geschlechter Boxplots der TC-Messungen über die Zeit zu erstellen. Kommentieren Sie die Plots.
- (b) Berechnen Sie für jeden Zeitpunkt Mittelwerte und Standardabweichungen der TC-Messungen für die beiden Geschlechter.
- (c) Berechnen Sie die empirischen Korrelationen zwischen den Messzeitpunkten. Kommentieren Sie im Hinblick auf die empirischen Korrelation Vor- und Nachteile der Korrelationsstrukturen `independence`, `unstructured`, `exchangeable`/compound symmetry und `stat_M_dep`/banded Toeplitz als Arbeitskorrelationsmatrizen bei einer GEE-Modellierung der Daten. Geben Sie für jede Korrelationsstruktur auch die Anzahl zu schätzenden Parameter an.
- (d) Fitten Sie GEE-Modelle, bei denen nur Intercept sowie die Haupteffekte für `t`, `fitness`, `bfitness`, `gender` und `smoker` eingehen. Erstellen Sie eine Tabelle, in der Sie Regressionskoeffizienten und robuste Standardfehler für jede der Korrelationsstrukturen aus Teil (c) angeben. Kommentieren Sie die Unterschiede in $\hat{\beta}$ bzw. $\hat{se}(\hat{\beta})$.
- (e) Passen Sie an die Daten ein LME-Modell mit Random-Intercept und Random-Slope an. Geben Sie die Modellformel und die entsprechenden Verteilungsannahmen für die vorkommenden Zufallsvariablen an. Identifizieren Sie aus dem Output die Schätzer aller auftretenden Parameter.
- (f) Testen Sie, ob der Random-Slope benötigt wird. Was müssen Sie bei diesem Test beachten? Wie wäre eine Ablehnung der Nullhypothese inhaltlich zu interpretieren?
- (g) Skizzieren Sie mit Hilfe eines Plots von `t` gegen die TC-Werte, was positive bzw. negative Korrelation zwischen den beiden Random-Effects bedeutet.
- (h) Wieso liefern ein LME mit Random-Intercept und ein GEE mit Sandwich-Schätzung (= robust) basierend auf einer `exchangeable`-Arbeitskovarianz fast die gleichen Ergebnisse für

$\hat{\beta}$, nicht aber für deren Standardfehler? Warum sind die Ergebnisse für die modellbasierte (=naive) Schätzung so ähnlich?

Aufgabe 2:

Für die Amsterdam-Daten ist außerdem von Interesse ob die Studienteilnehmer in eine kardiovaskuläre Risikogruppe fallen. Für jeden Zeitpunkt t ist der kardiovaskuläre Risikoindikator Y_{it}^* (0=klein, 1=groß) von Individuum i durch $Y_{it}^* = I(Y_{it} > q_{0.75}(t))$ gegeben, wobei $q_{0.75}(t)$ das 75%-Quantil der beobachteten TC-Messungen zum Zeitpunkt t ist.

Führen Sie diese Dichotomisierung für die Amsterdam-Daten durch und fügen Sie diese Größe als `tc.dich` zu den Daten hinzu.

Hinweis: Zur Kontrolle enthält die Datei `tc.dich.txt` die gewünschten Werte.

Im Folgenden sollen Sie die Auswirkung 3 verschiedener Missing-Mechanismen an diesen Daten ausprobieren:

- (a) **Szenario 1:** Das für die Cholesterol-Messungen zuständige Labor arbeitete schlecht. Es zeigt sich dass im Durchschnitt etwa ein Drittel der Proben falsch analysiert wurde und als fehlende Messung eingestuft werden muss.
- (b) **Szenario 2:** Kinder und Jugendliche, die als Risikopatienten eingestuft wurden ($Y_{it_{c,i}}^* = 1$), werden zur Behandlung an entsprechende Spezialisten verwiesen und im Folgenden ($t > t_{c,i}$) nicht mehr im Rahmen der vorliegenden Studie betrachtet, weil durch die entsprechende Medikation das natürliche Entwicklungsmuster gestört wird.
- (c) **Szenario 3:** Mit einer Wahrscheinlichkeit von 0.8 fallen Kinder und Jugendliche, deren TC-Wert zum Zeitpunkt t über dem entsprechenden 75%-Quantil liegt, bereits vor der Messung zum Zeitpunkt t aus der Studie, weil ihre besorgte Eltern sie in hermetisch abgeriegelte Diät-Camps haben einliefern lassen, in denen sie den Rest ihrer traurigen Jugend ohne Frühstückseier und Mayonnaise verbringen müssen.

Überlegen Sie für jedes der 3 Szenarien um was für einen Missing-Mechanismus es sich handelt. Erzeugen Sie nach den hier beschriebenen Mechanismen aus den echten Daten Datensätze mit fehlenden Werten und vergleichen Sie die Punktschätzer der festen Effekte und der Varianzparameter auf Basis dieser Datensätze mit denen aus den echten Daten.

Aufgabe 3:

Benutzen Sie im folgenden wieder den kompletten Datensatz. Von Interesse ist nun die Frage, wovon es abhängt ob Studienteilnehmer in die kardiovaskuläre Risikogruppe fallen.

- (a) Fitten Sie ein gemischtes logistisches Regressionsmodell mit festen Effekten (nur Haupteffekte) für `t`, `gender` und `smoker` sowie einem Random-Intercept.

Hinweis: Benutzen Sie `glmmML()` aus dem gleichnamigen Paket oder `lmer()` aus dem Paket `lme4`.

- (b) Geben Sie die entsprechende Modellformel für dieses Modell an und identifizieren Sie aus dem Output die Schätzer für alle relevanten Größen. Interpretieren Sie die Regressionskoeffizienten für `smoker` und `gender` auf der Odds-Ratio-Skala und geben Sie für diese Kovariablen auch 95%-Konfidenzintervalle an.

- (c) Fitten Sie ein binäres GEE mit Intercept und Haupteffekten für `t`, `gender` und `smoker`. Benutzen Sie eine unstrukturierte Arbeitskorrelationsmatrix. Geben Sie die Regressionskoeffizienten und robuste 95%-Konfidenzintervalle für `smoker` und `gender` auf der Odds-Ratio-Skala an und interpretieren Sie diese.
- (d) Betrachten Sie den Unterschied zwischen der Interpretation der Odds-Ratios mit GEEs und GLMMs. In welcher Situation ist welche Methodik von Vorteil?