Analysis of Longitudinal Data SoSe 2016 **Blatt 1**

Prof. Dr. Sonja Greven und Dipl. Stat. Jona Cederbaum 19.04.2016

Dieses Aufgabenblatt soll Sie mit der Struktur und Besonderheiten longitudinaler Daten vertraut machen, sowie deren graphische Darstellungsmöglichkeiten in `R` vermitteln. Die zu bearbeitenden Aufgaben beziehen sich auf die Inhalte der ersten und zweiten Vorlesungsfolien.

Hinweis: Aufgrund der Fülle an Stoff ist die Übung so konzipiert, dass vorausgesetzt wird, dass zumindest versucht wurde die Aufgaben **im Voraus** zu bearbeiten.

Aufgabe 1:

In dieser Aufgabe beschäftigen wir uns mit dem Datensatz `rats`. Lesen Sie hierzu zunächst die Beschreibung des Datensatzes (auf der Homepage) durch.

a) Laden Sie den Datensatz von der Homepage herunter und lesen Sie ihn in `R` ein. Beachten Sie dabei, dass NAs als solche erkannt werden (*Hinweis:* Option `na.strings()` verwenden). Wandeln Sie die Variable `GROUP` in eine Faktorvariable mit entsprechenden Labels für die drei Experimentalgruppen um. Wandeln Sie auch die Variable `SUBJECT` in eine Faktorvariable um und werfen Sie einen ersten Blick auf den Datensatz.

b) Verwenden Sie nun die Funktion `reshape()`, um den Datensatz `rats` so umzuformatieren, dass pro Tier eine Zeile im Datensatz steht und nennen Sie den umformatierten Datensatz `rats.wide`. Benutzen Sie `rats.wide` um einen Überblick über die Ausfallraten in den verschiedenen Experimentalgruppen zu gewinnen.

i) Folgen die Ausfälle (`NAs`) einem bestimmten Mechanismus?

ii) Für wie viele Tiere in den einzelnen Experimentalgruppen liegen alle sechs oder nur fünf, vier, drei, etc. Messungen vor? Stellen Sie die Anzahlen über die Zeit grafisch dar. Lässt sich ein Muster erkennen?

c) Mithilfe des `groupedData`-Formats im `R`-Paket `nlme` lassen sich viele Aufrufe für Schätzungen und Plots für longitudinale Daten vereinfachen. Machen Sie sich mit dem Format vertraut und legen Sie, ausgehend von den Originaldaten, einen neuen Datensatz `rats2` im `groupedData`-Format an. Plotten Sie nun die Verläufe der einzelnen Tiere getrennt nach Experimentalgruppen.
Hinweis: `?plot.nmGroupedData` könnte hilfreich sein.

d) Fitten Sie nun ein lineares Modell für alle Ratten (pooled) mit `TIME` als Einflussgröße und visualisieren Sie die Verläufe der Residuen für die einzelnen Tiere (mithilfe des Pakets `lattice`). Sie können hierzu den auf der Homepage verfügbaren Code verwenden.

i) Was fällt Ihnen bei Betrachtung des Plots auf?

ii) Sind die Annahmen des gewöhnlichen linearen Modells erfüllt?

Aufgabe 2:

In dieser Aufgabe beschäftigen wir uns mit dem Datensatz cd4 (auf der Homepage). Dieser Datensatz umfasst 2 376 Beobachtungen der Zahl der CD4-Zellen im Blut von 369 HIV-infizierten/AIDS-kranken Männern vor und nach dem Zeitpunkt, zu dem erstmals HIV-Antikörper im Blut nachgewiesen wurden (*Serokonversion*). Die Zahl der CD4-Zellen dient hierbei als Biomarker für den Zustand des Immunsystems. Das Hauptinteresse liegt darin, die Form des Verlaufs des CD4-Gehalts über die Zeit hinweg zu bestimmen.

a) Lesen Sie den Datensatz in `R` ein, wandeln Sie die Variablen `drug` und `ID` in Faktorvariablen um (mit entsprechenden Labels für die Variable `drug`) und verschaffen Sie sich einen ersten Überblick über die Daten.

b) Plotten Sie nun die Verläufe der einzelnen Patienten und schätzen sie eine glatte Mittelwertsfunktion mithilfe des auf der Homepage verfügbaren Codes. Wieso sollte die glatte Mittelwertskurve nur als exploratives Tool verwendet werden?

c) Wiederholte Messungen eines Subjekts sind in der Regel korreliert, was sich in den Residuen widerspiegelt. Berechnen Sie diese, indem Sie die glatte Mittelwertskurve aus b) von den Beobachtungen abziehen.

i) Bilden Sie nun mithilfe des auf der Homepage verfügbaren Codes für jedes Subjekt die paarweisen Produkte der Residuen als einen Schätzer der Kovarianz.

ii) Plotten Sie nun alle paarweisen Produkte in einem Scatterplot gegen die Distanz der entsprechenden Beobachtungen und legen Sie mithilfe der Funktion `loess()` eine glatte Mittelwertskurve durch den Scatterplot. Was können Sie dem Plot entnehmen?

d) Betrachten Sie folgendes Modell für die cd4-Daten

$$\texttt{CD4}_{ij} = \alpha_d z_{ij} + \alpha_{\bar{d}}(1 - z_{ij}) + \beta_d z_{ij} t_{ij} + \beta_{\bar{d}}(1 - z_{ij}) t_{ij} + \varepsilon_{ij}, \tag{1}$$

wobei $\texttt{CD4}_{ij}$ die $j$te CD4-Messung an Subjekt i zum Zeitpunkt t_{ij} ist und $z_{ij} = 1$, falls keine Drogen genommen wurden und 0 sonst.

i) Fitten Sie ein lineares Modell mithilfe von generalized least-squares (Funktion `gls()` im Paket `nlme`) unter Unabhängigkeitsannahme aller Beobachtungen.
Hinweis: Verwenden Sie hierfür die Spezifikation `correlation=NULL`.

ii) Es ist anzunehmen, dass die Residuen eines Subjekts, $\varepsilon_{ij}, j = 1, \ldots, n_i$, nicht unabhängig sind. Nehmen Sie für eine neue Schätzung an, dass alle Residuen von Subjekt i gleich stark korreliert sind - unabhängig vom zeitlichen Abstand - und vergleichen Sie anschließend die Koeffizientenschätzer und die Standardfehler der beiden Schätzungen. Was fällt auf?
Hinweis: Verwenden Sie `correlation=corCompSymm(form=~ 1|ID)` für die zweite Schätzung.

iii) Welche weiteren Annahmen könnte man sinnvollerweise für die Korrelationsstruktur der Residuen treffen?