Analyse Longitudinaler Daten SoSe 2016 **Blatt 5**

Prof. Dr. Sonja Greven und Dipl. Stat. Jona Cederbaum 21.06.2016

Auf diesem Aufgabenblatt wiederholen und vertiefen wir Modelldiagnose und Modellwahl im linearen gemischten Modell (Vorlesungsfolien 7) und beginnen mit der Betrachtung nicht normalverteilter longitudinaler Daten und ihrer Modellierungsmöglichkeiten (Vorlesungsfolien 8).

Aufgabe 1: Modelldiagnose

In dieser Aufgabe beschäftigen wir uns mit der Diagnose von geschätzten linearen gemischten Modellen, um die getroffenen Annahmen und Spezifikationen zu überprüfen.

(a) Laden Sie den Datensatz `vitamin` von der Homepage herunter und lesen Sie die Beschreibung. Verschaffen Sie sich einen ersten Überblick.

(b) Schätzen Sie ein lineares gemischtes Modell (m_RIRS) mit zufälligen Interzepts und zufälligen Steigungen für jedes Kind und festen Effekten für `time` und für die Interaktion von `group` und `time`. Nehmen Sie an, dass keine serielle Korrelation vorliegt.
Hinweis: Um keinen Haupteffekt für `group` zu schätzen, verwenden Sie in der `formula`: `group*time - group`.

(c) Wieso braucht man hier keinen Haupteffekt für `group` zu betrachten?

(d) Was erhalten Sie, wenn Sie `predict(m_RIRS)` aufrufen und was liefert Ihnen `predict(m_RIRS,level=0)`?

(e) Sie möchten nun den Modellfit bewerten. Dafür ist es üblich die Residuen gegen die Kovariablen zu plotten. Welche beiden Modellschwächen lassen prinzipiell sich dadurch aufzeigen?

(f) Betrachten Sie nun die populations-spezifischen Residuen $r_{ij} = y_{ij} - x_{ij}^\top \hat{\beta}$ und plotten Sie die Residuen gegen die Kovariable `time`. Interpretieren Sie den Plot.

(g) Welchen weiteren Plot könnten Sie betrachten, um Fehlspezifikationen des Mittelwertes zu überprüfen?

(h) Wieso ist die Betrachtung eines Quantil-Quantil Plots für die Residuen r_{ij} ungeeignet?

(i) Welche Alternativen zu den Residuen r_{ij} könnte man betrachten?

Aufgabe 2: Modellwahl

In dieser Aufgabe geht es darum lineare gemischten Modellen zu vergleichen und das passendere auszuwählen.

(a) In Aufgabe 1 haben wir gesehen, dass der Mittelwert nicht gut spezifiziert wurde, daher kann es Sinn machen, anstelle von `time` die transformierte Variable `log(time)` zu verwenden. Schätzen Sie im Folgenden ein Modell `m_RIRSlog`, das mit dem Modell aus Aufgabe 1 bis auf die Transformation von `time` identisch ist. Verwenden Sie zur Schätzung **ML** (statt REML) und schätzen Sie auch das Modell `m_RIRS` noch einmal **mit ML**.
Hinweis: Beachten Sie, dass auch die zufälligen Steigungen mit der Transformation angepasst werden.

(b) Wie lassen sich Modell `m_RIRS` und Modell `m_RIRSlog` vergleichen, d.h. wie kann man auswählen welches der beiden Modelle geeigneter ist? Zu welcher Entscheidung kommen Sie bei ihrer Modellselektion?

(c) Welche Schwierigkeit würde auftreten, wenn die beiden obigen Modelle, die wir vergleichen wollen, mit REML geschätzt würden?

(d) Welche Annahme trifft man bei der Betrachtung des marginalen, sowie des konditionalen AIC und was lässt sich daraus für ihre Verwendung folgern?

(e) Betrachten Sie im Folgenden erneut das Modell `m_RIRSlog`, allerdings ohne zufällige Steigungen. Was für eine Entscheidung würden Sie erwarten, wenn Sie das Modell mit zufälligen Steigungen und das Modell ohne zufällige Steigungen mit dem marginalen AIC vergleichen würden?

(f) Verwenden Sie stattdessen das konditionale AIC (aus `R`-package `cAIC4`) um die beiden Modelle zu vergleichen. Zu welcher Entscheidung kommen Sie?
Hinweis: Um die Funktion `cAIC{cAIC4}` verwenden zu können, schätzen Sie die Modelle mit der Funktion `lmer` aus Paket `lme4` (vgl. Vorlesungsfolien). Betrachten Sie die Hilfe `?lme4`, um die wesentlichen Unterschiede zur bereits bekannten Funktion `lme` zu verstehen.

Aufgabe 3: Nicht normalverteilte longitudinale Daten

Im Folgenden gehen wir weg von der Normalverteilungsannahme der $\mathbf{Y}_i$ und lassen zu, dass die Daten aus einer Verteilung der Exponentialfamilie stammen. Dies erfolgt analog zum Übergang von linearen Modellen zu generalisierten linearen Modellen (GLMs).

(a) Überlegen Sie sich als ersten Schritt ein paar eigene Beispiele für nicht **normalverteilte longitudinale** Daten (nicht die aus der Vorlesung!).

(b) Im Datensatz `epil` im `R`-Paket `MASS` liegen die Anzahlen an epileptischen Anfällen für 59 Epileptiker über die Zeit vor. Die Probanden wurden einer Behandlungsgruppe mit dem Medikament Progabide und einer Placebogruppe randomisiert zugeteilt.

(i) Welches Modell würden Sie wählen, wenn sie sich für individuelle Prognosen für die Probanden interessieren?

(ii) Welches Modell würden Sie wählen, wenn sie sich für den Einfluss des Medikaments Progabide auf Populationsniveau interessieren?

(iii) Sei $\hat{\beta}$ der geschätzte Einfluss von Progabide aus dem Modell in (i). Was müssten Sie bei der Interpretation von $\hat{\beta}$ berücksichtigen? Entspricht die Interpretation der des Effekts von Progabide aus dem Modell in (ii)?