Prof. Dr. Sonja Greven,
Dipl. Stat. Jona Cederbaum
Institut für Statistik

Aufgabe	1	2	3	4	5	6	$\sum$
Punkte							

Klausur in Analyse longitudinaler Daten

Sommersemester 2016, 22. Juli 2016

Name (in Großbuchstaben): _____

Matrikelnummer: _____

Studiengang: _____

Ich bestätige, dass ich die unten stehenden Hinweise zur Kenntnis genommen und die Klausurangabe auf Vollständigkeit überprüft habe. Ich bin mit der Veröffentlichung meines Klausurergebnisses (Matrikelnummer, Note) als Aushang im Institut für Statistik, Ludwigstr. 33 einverstanden. (Falls nicht, den letzten Satz bitte streichen.)

Unterschrift: _____

Hinweise zur Klausur (1–8):

1. Überprüfen Sie zunächst, ob Ihre Klausurangabe vollständig ist. Die Klausurangabe sollte aus **7 Blättern** mit **6 Aufgaben** bestehen.
2. Die Bearbeitungszeit beträgt **90 Minuten**. In den ersten 30 Minuten und in den letzten 15 Minuten ist keine vorzeitige Abgabe vorgesehen.
3. Es können maximal **85 Punkte** erreicht werden. Ein nachvollziehbarer Lösungsweg ist Voraussetzung zum Erlangen der vollen Punktzahl. Das Lösen der Bonusaufgabe (Aufgabe 6) ist optional. Auch ohne die Bonusaufgabe kann die volle Punktzahl erreicht werden.
4. Es sind folgende Hilfsmittel zugelassen:
   - Taschenrechner
   - 2 DIN-A4-Blätter mit handschriftlichen Notizen auf Vorder- und Rückseite
   - bei Bedarf ein Wörterbuch
5. Verwenden Sie für Ihre Lösungen ausschließlich die Klausurangabe. Bei Bedarf können zusätzliche Blätter bei den Klausuraufsichten erfragt werden. Schreiben Sie auf jedes Blatt Ihrem Namen und Ihre Matrikelnummer.
6. Legen Sie einen Lichtbildausweis und einen aktuell gültigen Studienausweis bereit.
7. Bei Unterschleif erfolgt eine Meldung an das Prüfungsamt. Sie sind verpflichtet, durch Ihr Verhalten jegliche Missverständnisse diesbezüglich auszuschließen.
8. Verlassen Sie den Prüfungsraum erst, nachdem Sie der Aufsicht die Klausur persönlich übergeben haben. Für den Eingang der kompletten Klausur bei der Aufsicht sind Sie selbst verantwortlich. Bitte verlassen Sie nach der Klausur den Hörsaaltrakt zügig und leise, damit Sie die Teilnehmer anderer Klausuren nicht stören.

Viel Erfolg!

Aufgabe 1 14 Punkte

Das folgende Modell entstammt einer Studie zur Behandlung von Adipositas (Fettleibigkeit). Die Adipositas-Patienten wurden dabei in zwei verschiedene Kliniken (`clinic`) eingeteilt, in denen sie über einen Zeitraum von 60 Tagen behandelt wurden. Damit die Patienten nicht zu weit entfernt von ihrer Familie untergebracht sind, wurden sie nach Wohnort (Stadt/Land) auf die beiden Kliniken verteilt. In beiden Kliniken wurde täglich (`day`) der Gesundheitszustand jedes Patienten in Form eines Scores (`score`) erhoben. Man interessiert sich insbesondere dafür, ob es einen Unterschied hinsichtlich der Score-Entwicklung zwischen den beiden Kliniken gibt.

```
model1 <- lme(score ~ day * clinic, random = ~ 1 + day | id, data = adipos)
```

a) (1P) Was für eine Art Modell wurde hier geschätzt?

b) (2P) Wieso wäre die Verwendung eines fixed effects Modells hier nicht sinnvoll?

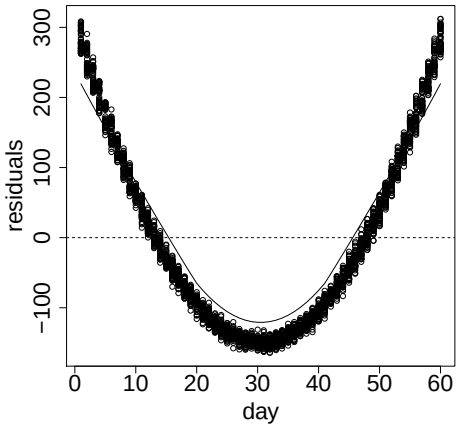
c) (3P) Wieso kann man nicht davon ausgehen, dass in dem spezifizierten Modell die random effects assumption bezüglich der Kovariable `clinic` erfüllt ist? Begründen Sie Ihre Antwort anhand eines passenden Beispiels.

d) (1P) Welche Annahme wird in dem Modell über den Zusammenhang von zufälligen Interzepts und zufälligen Steigungen getroffen?

e) (5P) Mit welchem Test könnten Sie testen, ob sich die Annahme über den Zusammenhang der zufälligen Interzepts und der zufälligen Steigungen vereinfachen lässt? Geben Sie auch die Nullhypothese H_0 und die Verteilung der Teststatistik unter H_0 an. Benennen Sie dabei alle verwendeten Komponenten.

f) (2P) Was können Sie aus dem folgenden Plot hinsichtlich der Modelldiagnose ablesen? Begründen Sie **kurz** ihre Antwort.

```
r <- resid(model1, level = 0)
plot(adipos$day, r, xlab = "day", ylab = "residuals")
lines(lowess(adipos$day, r))
abline(h = 0, lty = 2)
```



Lösung 1

Aufgabe 2 27 Punkte

In den letzten Jahren kam immer mehr sogenanntes "Superfood" auf den Markt, das besonders gesund sein soll. Chiasamen beispielsweise sollen aufgrund ihres hohen Gehalts an Omega3-Fettsäuren insbesondere Thrombosen entgegenwirken. Um den tatsächlichen Nutzen vom täglichen Verzehr von Chiasamen wissenschaftlich zu überprüfen, lässt das Verbraucherschutzministerium eine Studie durchführen, bei der die Probanden randomisiert in zwei Gruppen eingeteilt werden. Über einen Zeitraum von 2 Jahren sollten die Probanden in Gruppe 1 (`group=0`) keine Chiasamen verzehren und die in Gruppe 2 (`group=1`) täglich Chiasamen verzehren. Ab einem Monat nach Beginn wurde monatlich (`month`) bei allen Probanden erhoben, ob es zu einer Thrombose (`thrombosis`) gekommen ist oder nicht (ja/nein). Das angewandte Modell lieferte unter anderem den folgenden Output:

```
model3 <- glmmPQL(fixed = thrombosis ~ month + group,
          random = ~ 1 | id, family = binomial, data = dat)

summary(model3)
Linear mixed-effects model fit by maximum likelihood
 Data: dat
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | id
        (Intercept)  Residual
StdDev:   0.8639935 0.9511475

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: thrombosis ~ month + group
                 Value  Std.Error   DF    t-value p-value
(Intercept)  0.0332044 0.09474085 4599   0.350476  0.7260
month        0.0447339 0.05913039 4599   0.756531  0.4494
group       -1.9518854 0.14313599  198 -13.636580  0.0000

Number of Observations: 4800
Number of Groups: 200
```

a) (13P) Schreiben Sie das Modell für diese konkrete Anwendung hin. Spezifizieren und benennen Sie dabei **alle** Komponenten und schreiben Sie **alle** Annahmen hin. *Hinweis:* Falls nicht anders angegeben, wurde die natürliche Linkfunktion verwendet.

b) (6P) Interpretieren Sie den Schätzer für die Kovariable `group`.

c) (2P) Aus welchem Grund muss das Schätzkriterium bei der Schätzung von generalisierten linearen gemischten Modellen approximiert werden?

d) (2P) Aus welchem Grund werden in dem Modelloutput kein AIC, BIC und kein Wert für die Log-Likelihood ausgegeben?

e) (3P) Das Ministerium interessiert sich vor allem für die erwartete Veränderung der Chance eine Thrombose zu bekommen in der Population, wenn alle täglich Chiasamen zu sich nehmen würden. Die beiden Statistiker, die die Studie auswerten, kommen dabei zu unterschiedlichen Ergebnissen: Der eine Statistiker behauptet, dass die durchschnittliche logarithmierte Chance an Thrombose zu erkranken für einen festen Monat um ca. 1.952 sinkt, wenn täglich Chiasamen verzehrt werden. Der zweite Statistiker ist der Ansicht, dass der Effekt kleiner ist und die logarithmierte Chance um weniger als 1.952 sinkt (ceteris paribus). Welcher der beiden hat Recht? Begründen Sie Ihre Antwort.

f) (1P) Was für ein Modell hätte stattdessen verwendet werden können, um zu bestimmen wie groß die Chance eine Thrombose zu bekommen in der Population ist, wenn alle täglich Chiasamen zu sich nehmen würden?

Lösung 2

Aufgabe 3 8 Punkte

a) (1P) Aus welchem Grund werden die zufälligen Effekte im linearen gemischten Modell nicht auf Basis der marginalen Log-Likelihood vorhergesagt?

b) (1P) Beschreiben Sie **kurz**, wie sich das marginale Modell bei normalverteilten Daten aus dem konditionalen Modell herleiten lässt?

c) (2P) Wie verändert sich die Streuung der Prädiktionen der zufälligen Effekte $\boldsymbol{b}_i$ im Vergleich zu der Streuung, die man erhält, wenn die $\boldsymbol{b}_i$ stattdessen als feste Effekte geschätzt werden (in Worten)? Wie nennt man den Effekt?

d) (4P) Betrachten Sie die kompakte Matrixdarstellung des linearen gemischten Modells:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}.$$

Wie sieht die Kovariablenmatrix, $\boldsymbol{Z}$, der zufälligen Effekte für alle Individuen gemeinsam aus und welche Dimension hat sie? Benennen Sie alle Komponenten.

Lösung 3

Aufgabe 4 16 Punkte

Die Körpergröße (`height`) von Jungen und Mädchen wird im Rahmen einer Wachstumsstudie ab dem 10. Lebensjahr monatlich drei Jahre lang von den Eltern der Kinder gemessen. Eigentlich sollen die Eltern die Größe ihrer Kinder jeden Monat (`month`) messen. Manche Eltern sind allerdings nicht ganz konsequent und messen ihre Kinder nicht regelmäßig, weshalb die Messungen im letzten Studienjahr im Studienzentrum vorgenommen werden, in der dann alle Kinder gemeinsam gemessen werden.

a) (2P) Handelt es sich bei den Daten um balancierte Daten? Begründen Sie **kurz** Ihre Antwort.

b) (3P) Um einen ersten Eindruck von den Daten zu bekommen, möchten Sie die Mittelwertskurve über alle Kinder darstellen.
Schlagen Sie eine explorative Methode vor, wie sie in dieser Datensituation vorgehen könnten. Wieso müssen Sie hier darauf achten, dass die Methode nur als exploratives Tool angewendet werden sollte?

Das folgende Modell wird für die Analyse der Daten verwendet:

```
model2 <- lme(height ~ sex * month, random = ~1|child, data = growth),
```

wobei `sex` eine Dummyvariable für das Geschlecht der Kinder und `child` eine faktor-kodierte ID-Variable ist.

c) (3P) Nennen Sie drei Eigenschaften der marginalen Korrelation zwischen zwei Messungen an einem Kind in `model2`.

d) (1P) Wie groß ist die Korrelation zwischen zwei Messungen am gleichen Kind bedingt auf den zufälligen Interzept in `model2`?

e) (2P) Handelt es sich bei den fehlenden Werten um Dropout? Begründen Sie **kurz** Ihre Antwort.

f) (3P) Von welchem Missing Mechanismus ist hier auszugehen? Begründen Sie Ihre Antwort.

g) (1P) Mit welcher Schätzmethode wurden die Varianzparameter in ihrem Modell geschätzt?

h) (1P) Erklären Sie **kurz**, weshalb sich die Schätzer der festen Effekte unterscheiden können, je nachdem, ob die Varianzparameter mit Maximum Likelihood oder mit Restricted Maximum Likelihood geschätzt wurden.

Lösung 4

Aufgabe 5	15 Punkte

Im Folgenden sehen Sie die Satzanfänge von fünf verschiedenen Aussagen (A-E). Ergänzen Sie jeden Satzanfang passend mit einer der drei angegebenen Möglichkeiten. Es ist jeweils genau eine Aussage richtig. Markieren Sie die passende Ergänzung auf Ihrer Angabe (keine Begründung notwendig). Sie erhalten für jede richtige Antwort 3 Punkte, für jede falsche Antwort 1.5 Minuspunkte. Sie können in dieser Aufgabe maximal 15 und minimal 0 Punkte erreichen.

A. Bei der Maximium Likelihood Schätzung in linearen gemischten Modellen wird wie folgt vorgegangen:

   (1) Die festen Effekte und die Varianzparameter werden simultan geschätzt, indem die Residual Maximum Likelihood maximiert wird.

   (2) Zuerst werden die festen Effekte unter einer Unabhängigkeitsannahme geschätzt und dann in die Schätzer der Varianzparameter eingesetzt. Durch die zuvorige Schätzung der festen Effekte verliert man Freiheitsgerade, was dazu führt, dass der Maximum Likelihood Schätzer für die Varianzparameter verzerrt ist.

   (3) Die Varianzparameter werden durch Maximierung der Profile-Likelihood geschätzt und die Schätzer werden dann eingesetzt, um die Schätzer der festen Effekte zu erhalten.

B. Inwiefern stellt das lineare gemischte Modell (LMM) einen ganz besonderen Spezialfall des generalisierten linearen gemischten Modells (GLMM) dar?

   (1) Im Gegensatz zum GLMM gibt es im LMM eine analytische Form für die Schätzer der festen Effekte und der Varianzkomponenten.

   (2) Im LMM können zusätzliche Annahmen über den Fehlerterm getroffen werden und man geht nicht immer, wie im GLMM, von conditional independence aus.

   (3) Im LMM implizieren sich die konditionale und die marginale Sichtweise gegenseitig.

C. Häufig kommt es bei der Erhebung longitudinaler Daten zu fehlenden Werten, was bei der Wahl der Analysemethoden beachtet werden sollte. Gibt es fehlende Werte, so

   (1) sollten nur die Subjekte in die Analyse eingehen, für die keine fehlenden Werte vorliegen.

   (2) sind die Schätzer basierend auf generalisierten Schätzgleichungen (GEE) nur konsistent, wenn die Wahrscheinlichkeit für das Fehlen nicht von den Responsewerten abhängt.

   (3) stellen lineare gemischte Modelle generell keine valide Methode für die Analyse dar.

D. Ein großer Vorteil von generalisierten Schätzgleichung gegenüber Maximum Likelihood Methoden ist, dass auch bei Missspezifikation der Korrelationsstruktur

   (1) die Schätzer der festen Effekte und der Sandwichschätzer der Kovarianz der festen Effekte konsistente Schätzer sind.

   (2) die Schätzer der festen und der zufälligen Effekte konsistente Schätzer sind.

   (3) die Schätzer der festen Effekte konsistente und effiziente Schätzer sind.

E. Ein Arbeitsamts erhebt am Ende jedes Monats, wie viele Bewerbungen Langzeitarbeitslose in dem Monat versendet haben. Statt Bewerbungen zu schreiben, sind ein paar Bewerber unerlaubterweise in Urlaub gefahren. Sie behalten lieber für sich, dass sie keine Bewerbung versendet haben und füllen daher den Fragebogen des Arbeitsamts nicht aus. Welcher Missing Mechanismus liegt hier vor?

   (1) MCAR

   (2) MAR

   (3) NMAR

Bonusaufgabe (Aufgabe 6) (5P)

Der beste lineare unverzerrte Prädiktor (BLUP) kann auf mehrere Arten hergeleitet werden. Nennen Sie die beiden in der Vorlesung besprochenen Herleitungen und skizzieren Sie **eine** der Herleitungen. Benennen Sie dabei alle verwendeten Komponenten.

Lösung 6