§4 Longitudinaldaten und Repeated Measurements

Wir betrachten nun Versuche mit mehreren Individuen, an denen wiederholt Messungen durchgeführt werden. Wenn diese wiederholten Messungen über die Zeit hinweg gemacht werden, so spricht man von Longitudinaldaten. Wir betrachten zunächst zwei Modelle für Longitudinaldaten, wobei die Zeit die einzige erklärende Variable sein soll. Im Beispiel später werden weitere erklärende Variablen berücksichtigt.

Modell mit zufälligem Achsenabschnitt

$$y_{ij} = \beta_0 + \beta_1 t_j + u_i + e_{ij}, \quad i = 1, \ldots, r,\ j = 1, \ldots, s, n = rs$$

mit

y_{ij}	–	Zufallsvariable mit $\mathrm{E}(y_{ij}) = \beta_0 + \beta_1 t_j$
β_0	–	Achsenabschnitt (Intercept)
β_1	–	Steigungsparameter
t_j	–	j-ter fester, bekannter Zeitpunkt
u_i	–	zufälliger Effekt der i-ten Person, $u_i \sim \mathcal{N}\left(0, \sigma_u^2\right)$
e_{ij}	–	Zufallsfehler, $e_{ij} \sim \mathcal{N}\left(0, \sigma_e^2\right)$

$u_1, \ldots, u_r, e_{11}, \ldots, e_{rs}$ sind stochastisch unabhängig.

Die Zufallsvariable y_{ij} steht für die Beobachtung zum Zeitpunkt t_j an Individuum i. Die Gesamtresiduen teilen sich in zwei Teile auf, der personen-spezifischen Zufallskomponente u_i, die über die Zeit hinweg konstant ist, und der Fehlerkomponente e_{ij}, die über die Zeit variiert. Für die Varianz von y_{ij} gilt

$$\mathsf{Var}(y_{ij}) = \sigma_u^2 + \sigma_e^2$$

und die Beobachtungen an einer Person sind korreliert gemäß

$$\mathsf{Cor}(y_{ij}, y_{ik}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}.$$

Diese *Intraklass-Korrelation* wird interpretiert als der Anteil der Gesamtvariabilität, der auf die Variabilität der Personen zurückzuführen ist. Ein Modell mit zufälligem Achsenabschnitt zwingt die Varianz jeder wiederholten Messung die gleiche zu sein und die Kovarianz zwischen zwei beliebigen Messungen an einer Person ebenfalls gleich zu sein. Diese Kovarianzstruktur wird als *compound symmetry* oder *exchangeable* bezeichnet. Diese Annahmen sind für Longitudinaldaten häufig nicht gerechtfertigt.

Zum Beispiel sind Messungen, die zeitlich näher beianderliegen, häufig stärker korreliert als Messungen, die zeitlich weiter auseinanderliegen. Ein Modell, dass eine realistische Struktur für die Kovarianzen erlaubt, ist das Modell mit zufälligem Achsenabschnitt und zufälligem Steigungsparameter. Modell mit zufälligem Achsenabschnitt und zufälligem Steigungsparameter

$$y_{ij} = \beta_0 + \beta_1 t_j + u_i + v_i t_j + e_{ij}$$

mit

- y_{ij} – Zufallsvariable mit $\mathrm{E}(y_{ij}) = \beta_0 + \beta_1 t_j$
- β_0 – Achsenabschnitt (Intercept)
- β_1 – Steigungsparameter
- t_j – j-ter fester, bekannter Zeitpunkt
- u_i – zufälliger Effekt der i-ten Person, $u_i \sim \mathcal{N}\left(0, \sigma_u^2\right)$
- v_i – zufälliger Steigungsparameter der i-ten Person, $v_i \sim \mathcal{N}\left(0, \sigma_v^2\right)$
- e_{ij} – Zufallsfehler, $e_{ij} \sim \mathcal{N}\left(0, \sigma_e^2\right)$

(u_i, v_i) sind bivariat normalverteilt mit Kovarianz σ_{uv}. Alle anderen Zufallskomponenten sind stochastisch unabhängig.

In diesem Modell gilt für die Varianz von y_{ij}

$$\mathrm{Var}(y_{ij}) = \sigma_u^2 + 2\sigma_{uv}t_j + \sigma_v^2 t_j^2 + \sigma_e^2,$$

die nun nicht mehr konstant für verschiedene Werte von t_j ist. Für die Kovarianz zweier Beobachtungen am selben Individuum gilt:

$$\mathrm{Cov}(y_{ij}, y_{ik}) = \sigma_u^2 + \sigma_{uv}(t_j + t_k) + \sigma_v^2 t_j t_k.$$

Diese Kovarianz ist nun nicht mehr gleich für jedes Paar t_j und t_k.

Die Varianzkomponenten können nun wieder durch die ML-Methode oder die REML-Methode geschätzt werden. Auch hier gilt die Empfehlung die REML-Methode zu benutzen, weil die ML-Methode dazu tendiert, die wahren Varianzkomponenten zu unterschätzen. Ein Modellvergleich kann wiederum mit dem Likelihood-Quotienten-Test durchgeführt werden. Wenn die Modelle durch die REML-Methode geschätzt wurden, so kann der Likelihood-Quotienten-Test für Varianzkomponenten nur durchgeführt werden, wenn in beiden Modellen dieselben festen Effekte vorliegen, siehe Longford (1993).

Beispiel: (Everitt, Hothorn (2010), Proudfoot et al. (2003))
Die Datei *BtheB.txt* enthält die Daten eines randomisierten klinischen Versuches zur Depression. Dabei wurden zwei psychotherapeutische Verfahren, 'Beat the Blues' (BtheB) und 'Treatment as Usual' (TAU), angewendet. Der Datensatz ist auch in dem R Paket `HSAUR2` verfügbar. Eine Zielvariable war *Beck Depression Inventory II* (BDI). BDI wurde pro Patient fünfmal gemessen, vor der Behandlung, 2 Monate, 3 Monate, 5 Monate und 8 Monate nach Behandlung. Neben dem Effekt der Behandlung (`drug`, yes oder no) ist auch noch die Länge der aktuellen depressiven Episode von Interesse (`length`, weniger oder mehr als sechs Monate).

Lesen wir den Datensatz ein und schauen uns die ersten Beobachtungen an.

```
> BtheB <- read.table("BtheB.txt", header=TRUE)
> head(BtheB)
  drug length treatment bdi.pre bdi.2m bdi.3m bdi.5m bdi.8m
1   No    >6m       TAU      29      2      2     NA     NA
2  Yes    >6m     BtheB      32     16     24     17     20
3  Yes    <6m       TAU      25     20     NA     NA     NA
4   No    >6m     BtheB      21     17     16     10      9
5  Yes    >6m     BtheB      26     23     NA     NA     NA
6  Yes    <6m     BtheB       7      0      0      0      0
```

Aus 'weit' mach' 'lang'.

```
> BtheB$subject <- factor(rownames(BtheB))
> nobs <- nrow(BtheB)
> BtheB_long <- reshape(BtheB, idvar = "subject",
+          varying = c("bdi.2m", "bdi.3m", "bdi.5m", "bdi.8m"),
+          direction = "long")
> BtheB_long$time <- rep(c(2, 3, 5, 8), rep(nobs,4))
> head(BtheB_long)
     drug length treatment bdi.pre subject time bdi
1.2m   No    >6m       TAU      29       1    2   2
2.2m  Yes    >6m     BtheB      32       2    2  16
3.2m  Yes    <6m       TAU      25       3    2  20
4.2m   No    >6m     BtheB      21       4    2  17
5.2m  Yes    >6m     BtheB      26       5    2  23
6.2m  Yes    <6m     BtheB       7       6    2   0
```

Es ist hier zu bemerken, dass der Datensatz nun nach den Zeitpunkten und innerhalb der Zeitpunkte nach den Patienten geordnet ist und nicht umgekehrt.

- R Programm: BeattheBlues.R

Im Folgenden betrachten wir ein Beispiel für wiederholte Messungen, bei dem die Zeit keine Rolle spielt.

Beispiel: (Faraway (2006), Crowder und Hand (1990))
Die Datei *Vision.txt* enthält die Daten eines Vesuchs zu Sehstärke von sieben Probanden. Die Zielgröße ist die Zeitdifferenz in Millisekunden zwischen einem Lichtblitz und dem Response im Cortex des Auges. Jedes Auge wird dabei mit vier verschiedenen Brillenglasstärken getestet. Ein Objekt in der Distanz der zweiten Zahl erscheint als wäre es in der Distanz der ersten Zahl.

Lesen wir den Datensatz ein und schauen uns die ersten 10 Beobachtungen an.

```
> vision <- read.table("Vision.txt", header=TRUE)
> vision$power <- as.factor(vision$power)
> vision[1:10,]
   acuity power   eye subject
1     116   6/6  left       1
2     119  6/18  left       1
3     116  6/36  left       1
4     124  6/60  left       1
5     120   6/6 right       1
6     117  6/18 right       1
7     114  6/36 right       1
8     122  6/60 right       1
9     110   6/6  left       2
10    110  6/18  left       2
```

Um einen ersten Eindruck von den Daten zu bekommen, betrachten wir die Abbildung 1. Auffällig ist Proband 6, der einen großen Unterschied zwischen den beiden Augen aufweist. Außerdem könnte die dritte Messung für das linke Auge nicht korrekt sein.

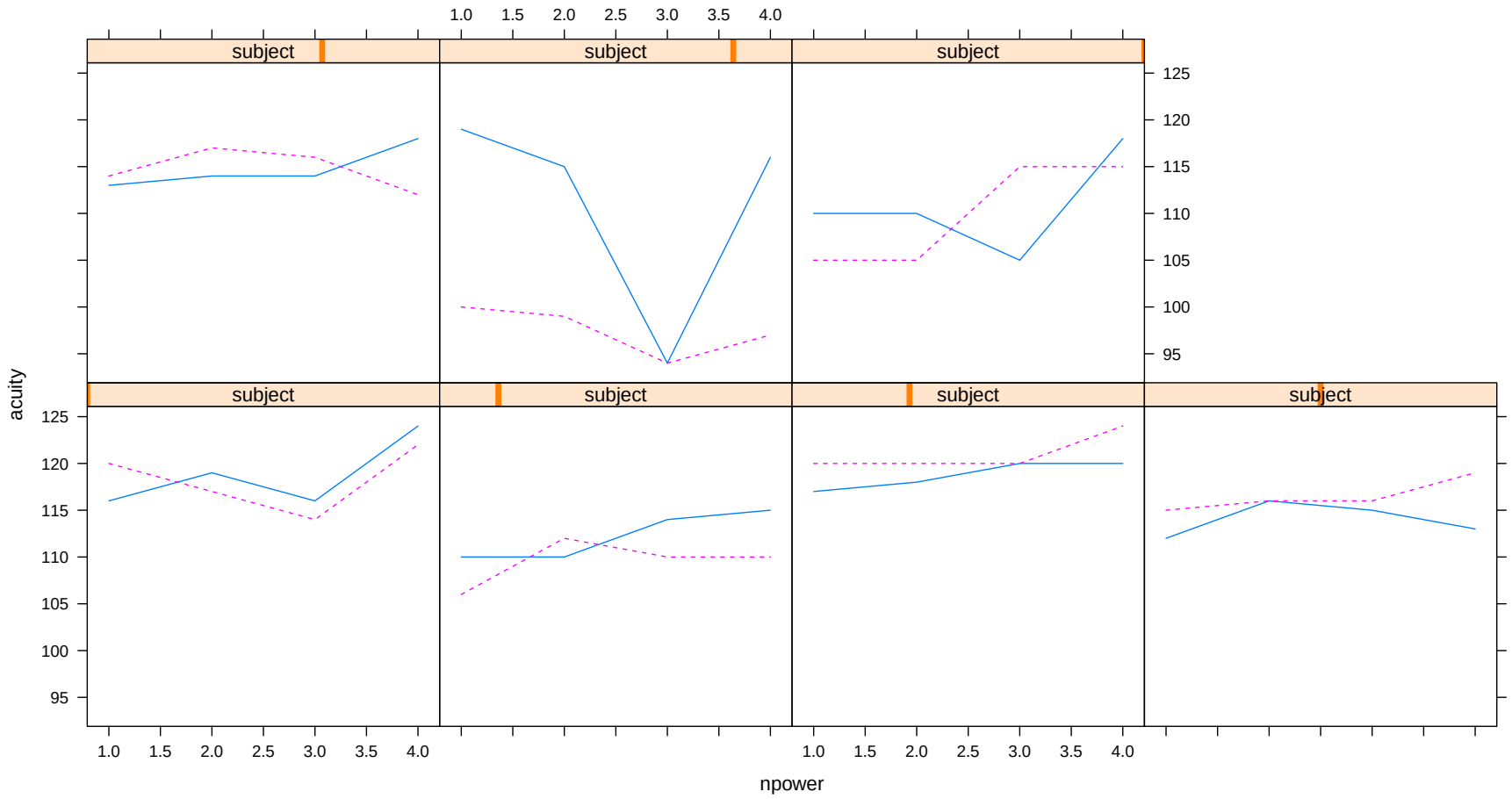

Figure 1: Sehstärkeprofile. Die durchgezogene Linie ist für das linke Auge, die gestrichelte für das rechte Auge.

Wir müssen nun die Daten modellieren. Die Brillenglasstärken (`power`) sind ein fester Effekt. Die Probanden (`subject`) sollten als zufällige Effekte behandelt werden. Da wir nicht glauben, dass es eine konsistente Differenz zwischen dem rechten und dem linken Auge gibt, sollten wir den Faktor Auge hierarchisch innerhalb der Probanden modellieren. Wir betrachten also das folgende hierarchische Modell

$$y_{ijk} = \mu + p_j + s_i + e_{ik} + \epsilon_{ijk},$$

$$i = 1, \ldots, 7 \text{ (subject)}, \; j = 1, \ldots 4 \text{ (power)}, \; k = 1, 2 \text{ (eye)}.$$

Der Effekt p_j ist fest, die anderen Terme sind zufällige mit $s_i \sim \mathcal{N}(0, \sigma_s^2)$, $e_{ik} \sim \mathcal{N}(0, \sigma_e^2)$ und $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

- R Programm: Vision.R

Aufgabe

Die Datei *Phosphat.txt* enthält die Daten von 33 Individuen bezogen auf das anorganische Phosphat im Blutplasma zu verschiedenen Zeitpunkten nach einmaliger Glukosegabe. 20 der Individuen wurden als Kontrolle klassifiziert und 13 als adipös. Passen Sie ein geeignetes Gemischtes Lineares Modell an!