

Tutorium 1

Analyse longitudinaler Daten

Prof. Dr. Sonja Greven, Almond Stöcker,
Johanna Völkl

Material: Alexander Bauer

16. Mai 2017

- 1 Longitudinale Daten
- 2 Das LLMM
- 3 Anwendung in R

Longitudinale Daten

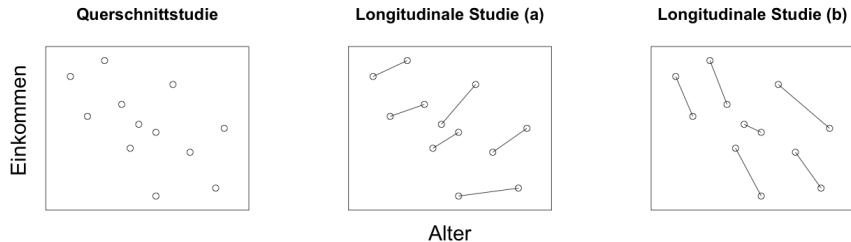
- 1 Longitudinale Daten
- 2 Das LLMM
- 3 Anwendung in R

Longitudinale Daten

Longitudinale Daten:

- Mehrere Beobachtungen pro Subjekt über die Zeit
- Beobachtungen pro Subjekt sind korreliert
- Häufig fehlende Daten, z.B. wegen Dropouts
- **Balancierte Daten:**
Gleiche Anzahl an Beobachtungen n_i zu gleichen Zeitpunkten
 $t_{ij} = t_j, j = 1, \dots, n_i$ für alle Subjekte i (und keine fehlenden Werte)
- **Äquidistante Daten:**
Beobachtungszeitpunkte liegen alle gleich weit auseinander:
 $d = t_{j+1} - t_j$

Longitudinale Daten



- ⇒ Unterscheidung zwischen Querschnitts- und longitudinalen Effekten
- ⇒ Hier: Kohorten- und Alterseffekt

Longitudinale Daten

Vorteile longitudinaler Analysen:

- Mögliche Trennung von Querschnitts- und longitudinalen Effekten
- Effizientere Schätzer als in Querschnitt-Designs
⇒ weniger Personen benötigt
- Zusätzlicher Schutz vor Confounding
(Subjekte bilden ihre eigene Kontrolle)
- Modellierung individueller Verläufe
(dadurch auch Prognosen für bestimmte Subjekte möglich)

Das LLMM

- 1 Longitudinale Daten
- 2 Das LLMM**
- 3 Anwendung in R

Das LLMM

Was passiert, wenn man Korrelation der Beobachtungen ignoriert?

- Inferenz nicht mehr gültig, da Modellannahmen verletzt
- Schätzer sind weniger effizient

Notation:

- N : Anzahl der Subjekte
- n_i : Anzahl der Beobachtungen für Subjekt i , $i = 1, \dots, N$
- $n = \sum_{i=1}^N n_i$: Anzahl Beobachtungen über alle Subjekte
- $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$: Zufallsvektor der Beobachtungen von Subjekt i

Das LLMM

Das **Longitudinal Linear Mixed Model** (LLMM):

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \\ \mathbf{b}_i &\stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{D}), & i = 1, \dots, N, \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i), \\ \mathbf{b}_1, \dots, \mathbf{b}_N, \boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N &\text{ sind unabhängig,} \end{aligned}$$

mit

- Designmatrizen \mathbf{X}_i ($n_i \times p$) und \mathbf{Z}_i ($n_i \times q$)
- Fixed effects $\boldsymbol{\beta}$ ($p \times 1$) und Random effects \mathbf{b}_i ($q \times 1$)
- Fehler $\boldsymbol{\epsilon}_i$ ($n_i \times 1$)

Anmerkung:

$\mathbf{b}_1, \dots, \mathbf{b}_N$ (zwischen Subjekten) sind unabhängig
 b_{i1}, \dots, b_{iq} (innerhalb eines Subjekts) i.A. nicht (\mathbf{D} keine Diagonalmatrix)!

Das LLMM

Beobachtungsebene:

$$\begin{aligned}
 Y_{ij} &= \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i + \epsilon_{ij} \\
 &\stackrel{\text{z.B.}}{=} \beta_0 + \beta_1 x_{ij} + b_{i0} + b_{i1} z_{ij} + \epsilon_{ij} \\
 \mathbf{b}_i &\stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}_q, \mathbf{D}), \quad \epsilon_i \sim \mathcal{N}(\mathbf{0}_{n_i}, \boldsymbol{\Sigma}_i)
 \end{aligned}$$

Subjektebene:

$$\begin{aligned}
 \mathbf{Y}_i &= \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i \\
 \mathbf{b}_i &\stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}_q, \mathbf{D}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}_{n_i}, \boldsymbol{\Sigma}_i)
 \end{aligned}$$

Matrixebene:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

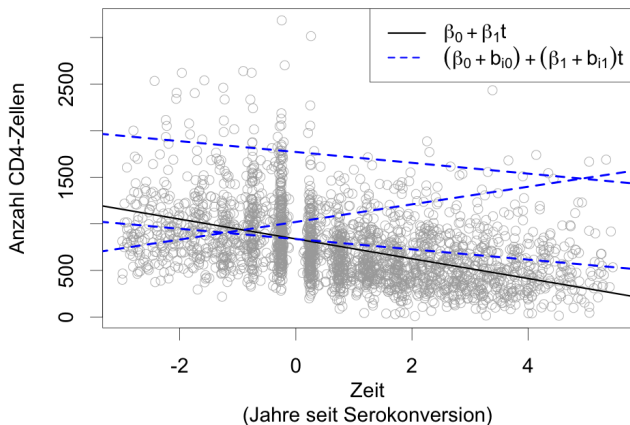
$$\begin{bmatrix} \mathbf{b} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0}_{Nq} \\ \mathbf{0}_n \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0}_{Nq \times n} \\ \mathbf{0}_{n \times Nq} & \mathbf{R} \end{bmatrix} \right)$$

mit Block-diagonalen $\mathbf{G} = \text{diag}(\mathbf{D}, \dots, \mathbf{D})$, $\mathbf{R} = \text{diag}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N)$

Das LLMM

Beispiel: CD4-Daten

Globaler Trend vs. individuelle Verläufe



Das LLMM

Random effects:

- Idee: Auffangen von Effekten nicht gemessener / messbarer subjekt-spezifischer Kovariablen
 - ⇒ Beispiel: Random Intercept zur Berücksichtigung individueller CD4-Niveaus
- Zentrale Annahme: **Random effects assumption**

$$\forall i, j : \mathbb{E}(b_i | x_{ij}) = 0$$

⇒ Unabhängigkeit der b_i von den Kovariablen

(Genauere Erläuterungen zur REA erfolgen später)

Konditionale vs. Marginale Betrachtung

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, \quad \mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}_q, \mathbf{D}), \quad \epsilon_i \sim \mathcal{N}_{n_i}(\mathbf{0}_{n_i}, \boldsymbol{\Sigma}_i)$$

Konditionaler EW

Marginaler EW

$$\mathbf{E}(\mathbf{Y}_i | \mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$$

$$\mathbf{E}(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$$

Konditionale Varianz

Marginale Varianz

$$\text{Cov}(\mathbf{Y}_i | \mathbf{b}_i) = \boldsymbol{\Sigma}_i$$

$$\text{Cov}(\mathbf{Y}_i) = \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_i$$

⇒ Marginal: Prognosen nicht für spezifische Personen

⇒ Konditional: Prognosen für spezifische Personen

Anwendung in R

1 Longitudinale Daten

2 Das LLMM

3 Anwendung in R

Anwendung in R

Die reshape-Funktion:

long-Format

SUBJECT	GROUP	RESPONSE	TIME
10	1	71.3	50
10	1	78.5	60
10	1	82.5	70
1	2	69.3	50
1	2	73.2	60
1	2	77.4	70

wide-Format

SUBJECT	GROUP	RESPONSE.50	RESPONSE.60	RESPONSE.70
10	1	71.3	78.5	82.5
1	2	69.3	73.2	77.4

Anwendung in R

Die reshape-Funktion:

Benötigte Argumente:

- `direction = "wide"`:
 - `v.names`: Vektor der Namen der zeitvariierenden Variablen
 - `timevar`: Zeitvariable
 - `idvar`: ID-Variable
- `direction = "long"`:
 - `varying`: Vektor der Spaltennamen im long-Format, welche die zeitvariierenden Variablen enthalten
Alternativ: Vektor mit Spaltenindizes statt den Namen
 - `timevar`: Name der zu erstellenden Zeitvariable
 - `idvar`: ID-Variable

Anwendung in R

Gemischte Modelle in R:

- `lme` (`nlme`):
`model <- lme(y ~ x, random = ~ 1 + x | ID, ...)`
- `lmer` (`lme4`):
`model <- lmer(y ~ x + (1 + x | ID), ...)`
 - Vergleich `nlme` vs. `lme4`: siehe `?lme4`
- `gam` (`mgcv`):
`model <- gam(y ~ x + s(ID, bs = "re") + s(ID, x, bs = "re"), method = "REML", ...)`
 - Für Einbeziehung glatter Effekte
 - Nur Einbezug unabhängiger Random Effects möglich

Anwendung in R

Gemischte Modelle in R:

- `gamm` (`mgcv`):

```
model <- gamm(y ~ x + s(ID, bs = "re") +  
s(ID, x, bs = "re"), method = "REML", ...)
```

- Flexiblere Random effect- und Korrelationsstrukturen (z.B. AR(1)-Fehler) möglich durch Basierung auf `gam` und `lme(r)`
- Modell besteht aus zwei Komponenten (`model$lme`, `model$gam`)
- „`gamm` assumes that you know what you are doing!“ (siehe `?gamm`)