

# Gemischte Modelle

Oktober 2013

Sonja Greven (V) und Sarah Brockhaus (Ü)

Institut für Statistik

Ludwig-Maximilians-Universität München

[http://www.statistik.lmu.de/institut/ag/fda/mixedmodels\\_2013/](http://www.statistik.lmu.de/institut/ag/fda/mixedmodels_2013/)



Mit Dank an Susanne Konrath und Fabian Scheipl für Material vergangener Jahre.

# Literatur Gemischte Modelle

- Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley.
- Diggle, P. J.; Heagerty, P.; Liang, K. L.; Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.
- **Fahrmeir, L., Kneib, T. und Lang, S. (2007)**. *Regression: Modelle, Methoden und Anwendungen*. Springer. - Begleitend zur Vorlesung. Erhältlich als Ebook bei der Universitätsbibliothek: <http://www.ub.uni-muenchen.de/emedien/ebooks/>
- McCulloch, C. E.; Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley.
- **Pinheiro, J. C.; Bates, D. M. (2000)**. *Mixed-Effects Models in S and S-PLUS*. Springer, New York. - Praxisorientierte Einführung in die Analyse gemischter Modelle und ausführliche Beschreibung des R-Pakets `nlme` für LMMs.
- Ruppert, D.; Wand, M. P.; Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press. - Verbindung gemischte Modelle und Penalisierung.
- Verbeke, G.; Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.
- Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC. - Verbindung gemischte Modelle und Penalisierung, R-Paket `mgcv`.

# Inhalt der Vorlesung Gemischte Modelle

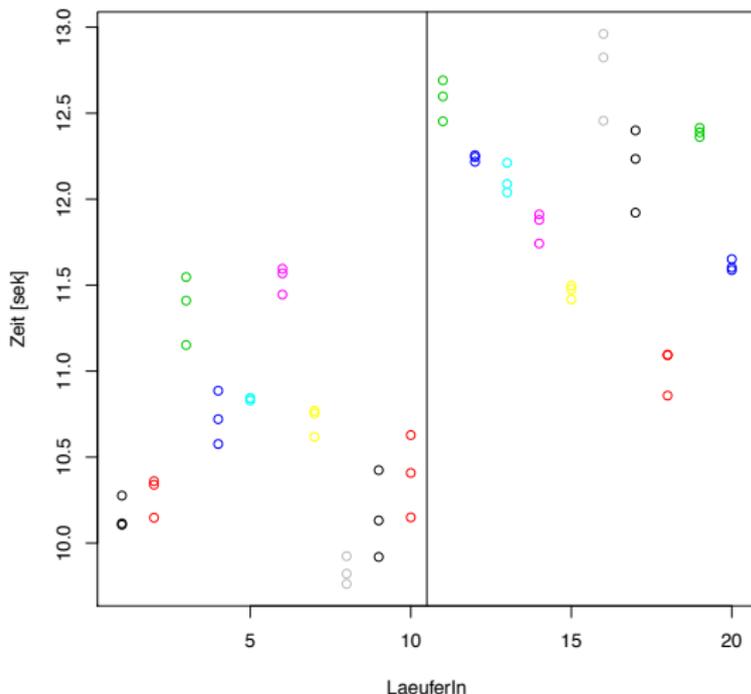
- 1 Das lineare gemischte Modell
- 2 Likelihood-Schätzung für lineare gemischte Modelle
- 3 Likelihood-Inferenz im linearen gemischten Modell
- 4 Bayes-Schätzung für lineare gemischte Modelle
- 5 Additive gemischte Modelle
- 6 Das generalisierte lineare gemischte Modell
- 7 Likelihood-Schätzung für generalisierte lineare gemischte Modelle

# Inhalt

- 1 Das lineare gemischte Modell
  - Motivation
  - Das allgemeine lineare gemischte Modell (LMM)
  - Spezialfälle
  - Die Kovarianzstruktur
  - Konditionale und Marginale Perspektive
- 2 Likelihood-Schätzung für lineare gemischte Modelle
- 3 Likelihood-Inferenz im linearen gemischten Modell
- 4 Bayes-Schätzung für lineare gemischte Modelle
- 5 Additive gemischte Modelle
- 6 Das generalisierte lineare gemischte Modell
- 7 Likelihood-Schätzung für generalisierte lineare gemischte Modelle

# Motivation: 100-Meter-Lauf

Drei 100-Meter-Läufe mit 10 Männern und 10 Frauen. Die Zeiten sehen so aus:



Wie könnte man diese Daten mit bekannten Methoden (lineare Modelle) modellieren? → **Diskussion**

# Motivation: 100-Meter-Lauf

Wir wollen ein Modell, das uns ermöglicht:

- die Schätzung des Geschlechtseffekts (Populationsparameter)
- die Schätzung der LäuferInneneffekte (individuelle Effekte)
- die Schätzung der Korrelationsstruktur
- valide Inferenz.

# 100-Meter-Lauf: Ein lineares gemischtes Modell

Überlegung: Hätten wir pro Person ihre Durchschnittszeit  $\mu_i$  (d.h. ein Wert pro Person), wäre ein sinnvolles Modell (iid = unabhängig identisch verteilt)

$$\mu_i = \beta_{g_i} + b_i, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2), \quad (1)$$

da die Durchschnittszeiten einzelner Person um das Geschlechtsmittel variieren.

Die beobachteten Zeiten pro Lauf streuen um die individuelle Durchschnittszeit:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \quad (2)$$

Zusammen genommen ergibt sich damit das Modell ( $\perp$  = unabhängig)

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2), \quad b_i \perp \varepsilon_{ij}. \quad (3)$$

# Ein Blick auf das Modell

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \quad (3)$$

Die Effekte  $b_i$  sind

- die „Fehler“terme in (1)
- aber Teil des Erwartungswertes in (2)
- sogenannte **zufällige Effekte**. Sie spiegeln hier wieder, dass die LäuferInnen aus einer Population kommen (in der wir die individuellen Durchschnittszeiten als normalverteilt annehmen).

Ein Modell mit zufälligen und festen Effekten nennt man **gemischtes Modell**.

# Interpretation der Parameter

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \quad (3)$$

- $\beta_1, \beta_2$  die Durchschnittszeiten für Männer / Frauen (**Populationsparameter**)
- $b_i$  die Abweichung der Durchschnittszeit von Person  $i$  vom Mittel der Geschlechtsgruppe (**individuelle Effekte**)
- $\tau^2$  die Varianz der Durchschnittszeiten pro Geschlechtsgruppe (in den zwei Gruppen als gleich **angenommen**)
- $\varepsilon_{ij}$  die Abweichung der  $j$ -ten Laufzeit von der Durchschnittszeit für Person  $i$
- $\sigma^2$  die Varianz der persönlichen Laufzeiten (für alle  $i$  gleich **angenommen**)

## Bedingte Sicht

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \quad (3)$$

Betrachte die Verteilung von  $y_{ij}$  bedingt auf  $b_i$  ( $\rightarrow$  **überlegen**):

$$y_{ij}|b_i \stackrel{iid}{\sim} N(\beta_{g_i} + b_i, \sigma^2)$$

$b_i$  modelliert den **Erwartungswert** (EW) analog zum linearen Modell mit festen Effekten

$$y_{ij} = \beta_{g_i} + \beta_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2).$$

Aber:

- Die Geschlechtseffekte sind identifizierbar (kein Confounding).
- Das Modell ist bei 1-2 Messungen für einige  $i$  schätzbar (mehr später).
- Da wir im Wesentlichen  $\tau^2$  schätzen (zur Vorhersage der  $b_i$ ; mehr später), wächst die Zahl der Parameter nicht mit der Anzahl der LäuferInnen.

# Marginale Sicht

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \quad (3)$$

Betrachte die marginale Verteilung von  $y_{ij}$ : ( $\rightarrow$  **überlegen**):

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} \stackrel{iid}{\sim} N\left( \begin{pmatrix} \beta_{g_i} \\ \beta_{g_i} \\ \beta_{g_i} \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2 + \sigma^2 \end{pmatrix} \right) \quad (4)$$

$b_i$  modelliert die **Kovarianzstruktur** analog zum allgemeinen linearen Modell

$$y_{ij} = \beta_{g_i} + \varepsilon_{ij}, \quad \boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i). \quad (5)$$

Hier:

- Begründung für eine mögliche Kovarianzstruktur  $\boldsymbol{\Sigma}_i$  (auch unbalanziert).
- Vorhersage individueller Effekte möglich (mehr später).

(3) impliziert das allgemeine lineare Modell (5), aber nicht umgekehrt!

# Wozu gemischte Modelle?

Anhand des Beispiels zeigt sich bereits: Gemischte Modelle werden gerne verwendet für die Analyse korrelierter Daten. Z.B.

- **Longitudinaldaten:** Wiederholte Beobachtungen in zeitlicher Abfolge an denselben Subjekten/Beobachtungseinheiten.
- **Clusterdaten / gruppierte Daten:** Gruppen (Cluster) mit mehreren Beobachtungen pro Gruppe.
- **Hierarchische Daten:** Gruppierte Daten mit mehreren geschachtelten Ebenen.
- **Gekreuzte Designs:** Gruppierte Daten mit mehreren gekreuzten Ebenen.

Daten vom gleichen Subjekt/Cluster/Beobachtungseinheit sind sich tendenziell ähnlicher als Daten verschiedener Subjekte/Cluster/Beobachtungseinheiten.

# Das allgemeine lineare gemischte Modell (LMM)

**Definition:** In allgemeiner Form ist das **lineare gemischte Modell** gegeben durch

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times p}{\mathbf{X}} \underset{p \times 1}{\boldsymbol{\beta}} + \underset{n \times s}{\mathbf{Z}} \underset{s \times 1}{\mathbf{b}} + \underset{n \times 1}{\boldsymbol{\varepsilon}}. \quad (6)$$

Mit  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$ ,  $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$  und der Annahme, dass die zufälligen Effekte  $\mathbf{b}$  und die Fehler  $\boldsymbol{\varepsilon}$  **unabhängig** sind, ist die **Verteilungsannahme** gegeben durch

$$\begin{pmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N \left( \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix} \right). \quad (7)$$

Die Kovarianzmatrizen  $\mathbf{G}$  für  $\mathbf{b}$  und  $\mathbf{R}$  für  $\boldsymbol{\varepsilon}$  werden als **positiv semi-definit** bzw. **positiv definit** angenommen.

## Zu den Annahmen

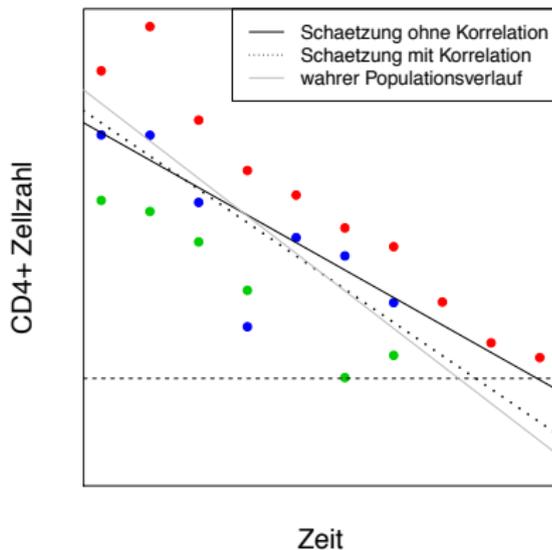
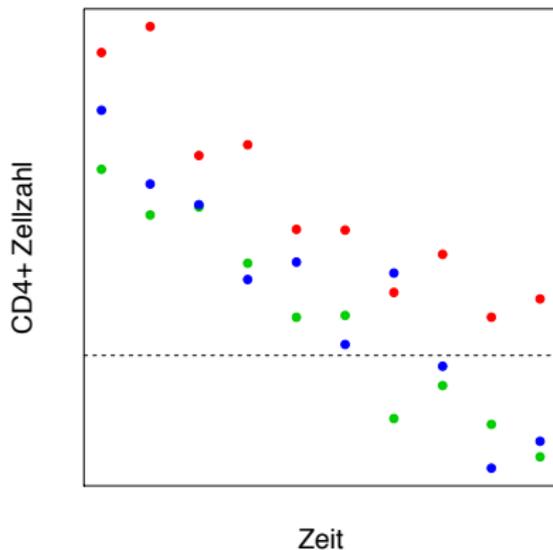
- Die **Normalverteilungsannahme der Fehler**  $\varepsilon$  ist nicht für alle Aussagen zur Inferenz in LMMs notwendig. Da Likelihood-basierte Schätzung üblich ist, nehmen wir sie jedoch in die Definition mit auf.
- Für die **zufälligen Effekte**  $b$  ist die **Normalverteilungsannahme** nicht zwingend. Alternative Verteilungen, z.B. Mischungsverteilungen, möglich. I.d.R. werden dann die Algorithmen zur Berechnung der Schätzer komplexer.

# Vorteile der Analyse mit gemischten Modellen

- Zufällige Effekte können als Platzhalter für die Effekte von unbeobachteten oder unzureichend gemessenen Kovariablen dienen, die **Unterschiede zwischen den Beobachtungen** (z.B. zwischen Gruppen) verursachen.
- Vernachlässigung von Korrelationen führt zu ungünstigen modellbasierten Standardfehlern und damit Konfidenzintervallen und Tests, da gruppierte Daten meist weniger Information enthalten als unabhängige. → **Übung**
- Die Berücksichtigung der Korrelation kann im Vergleich zur Schätzung eines einfachen linearen Modells zu einer **verbesserten Schätzgenauigkeit** (kleineren wahren Standardfehlern) führen. → **Übung**
- Die **Schätzer für die zufälligen Effekte**, erlauben **individuelle Prognosen**, die in einem herkömmlichen Regressionsmodell nicht möglich sind.

# Vorteile der Analyse mit gemischten Modellen

- Besserer (nicht vollständiger) Schutz bei **fehlenden Werten**.



## Beispiel 100-Meter-Lauf

Das Modell für die 100-Meter-Lauf-Daten

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2), \quad b_i \perp \varepsilon_{ij}.$$

lässt sich in die Form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \mathbf{G}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}), \quad \mathbf{b} \perp \boldsymbol{\varepsilon}$$

bringen mit

$$\mathbf{y} = (y_{11}, \dots, y_{20,3})$$

$$\mathbf{b} = (b_1, \dots, b_{20})$$

$$\mathbf{X} = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix}'$$

$$\mathbf{G} = \tau^2 \mathbf{I}_{20}$$

$$\boldsymbol{\beta} = (\beta_1, \beta_2)'$$

$$\boldsymbol{\varepsilon} = (\varepsilon_{11}, \dots, \varepsilon_{20,3})$$

$$\mathbf{Z} = \begin{pmatrix} 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ & & & \ddots & & & \\ 0 & \dots & 0 & 0 & 1 & \dots & 1 \end{pmatrix}'$$

$$\mathbf{R} = \sigma^2 \mathbf{I}_{60}.$$

# Spezialfall Longitudinal- und Clusterdaten

Wiederholte Beobachtungen  $y_{ij}$  der Zielvariablen → **Beispiel in Übung**

- von Subjekt  $i$  zum Zeitpunkt  $t_{ij}$  bei **Longitudinaldaten**
- für das  $j$ -te Objekt aus dem Cluster  $i$  bei **Clusterdaten**

mit jeweils Kovariablenvektor  $(\mathbf{x}'_{ij}, \mathbf{z}'_{ij})'$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ .

Verschiedene **Variabilitätsquellen** in den Daten:

- Zwischen den Subjekten/Clustern, Abweichungen vom Populationsmittel.
- Innerhalb des Subjekts/Clusters, Abweichungen einer Messung vom Mittelwert des entsprechenden Subjekts/Clusters.

## Spezialfall Longitudinal- und Clusterdaten

Das lineare gemischte Modell auf Beobachtungs- bzw. Cluster/Subjekt-Ebene ist

$$\begin{aligned}y_{ij} &= \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i + \varepsilon_{ij}, & j = 1, \dots, n_i, i = 1, \dots, m & \quad \text{bzw.} \\ \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, & i = 1, \dots, m.\end{aligned}$$

wobei  $\mathbf{X}_i$  und  $\mathbf{Z}_i$  die  $n_i$  Zeilen  $\mathbf{x}'_{ij}$  bzw.  $\mathbf{z}'_{ij}$  enthalten und  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ .

### Annahmen:

- Unabhängig und identisch normalverteilte zufällige Effekte  $\mathbf{b}_i \stackrel{iid}{\sim} N(\mathbf{0}, \mathbf{D})$ ,
- unabhängig und normalverteilte Fehler  $\varepsilon_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$ ,
- $\mathbf{b}_1, \dots, \mathbf{b}_m, \varepsilon_1, \dots, \varepsilon_m$  unabhängig,
- $\mathbf{D}$  positiv semi-positiv definit und  $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_m$  positiv definit.

## Spezialfall Longitudinal- und Clusterdaten

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \varepsilon_i, \quad i = 1, \dots, m.$$

- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  entspricht den (festen) **Populationseffekten**.
- $\mathbf{b}_i = (b_{1i}, \dots, b_{qi})'$  entspricht den (zufälligen) **subjekt- / clusterspezifischen Effekten**.
- $\mathbf{X}_i = \begin{pmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{in_i} \end{pmatrix}$ ,  $\mathbf{Z}_i = \begin{pmatrix} \mathbf{z}'_{i1} \\ \vdots \\ \mathbf{z}'_{in_i} \end{pmatrix}$  sind die **Designmatrizen** für die  $p$  populationsspezifischen bzw.  $q$  subjektspezifischen Kovariablen.

Dabei können die Kovariablen mit  $j$  (bzw.  $t_{ij}$ , zeitvariierend) variieren oder nicht.

Falls die Variablen  $\mathbf{z}_{ij}$  in  $\mathbf{x}_{ij}$  enthalten sind, lassen sich die  $\mathbf{b}_i$  mit  $E(\mathbf{b}_i) = \mathbf{0}$  als **individuelle Abweichungen** vom Populationsmittel interpretieren.

## Spezialfall Longitudinal- und Clusterdaten

Das lineare gemischte Modell für Longitudinal- und Clusterdaten ist ein Spezialfall des allgemeinen LMMs

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \mathbf{G}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}), \quad \mathbf{b} \perp \boldsymbol{\varepsilon}$$

mit

- $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_m)'$  und  $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_m)'$  der Länge  $n = \sum_{i=1}^m n_i$ ,
- $\boldsymbol{\beta}$  der Länge  $p$ ,
- $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_m)'$  der Länge  $s = mq$ ,
- $\mathbf{X} = (\mathbf{X}'_1 | \dots | \mathbf{X}'_m)'$  der Dimension  $n \times p$ ,
- $\mathbf{Z} = \text{blockdiag}(\mathbf{Z}_1, \dots, \mathbf{Z}_m)$  der Dimension  $n \times mq$ ,
- $\mathbf{G} = \text{blockdiag}(\mathbf{D}, \dots, \mathbf{D}, \dots, \mathbf{D})$  der Dimension  $mq \times mq$ ,
- $\mathbf{R} = \text{blockdiag}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_i, \dots, \boldsymbol{\Sigma}_m)$  der Dimension  $n \times n$ .

Die blockdiagonalen Kovarianzmatrizen resultieren aus der Unabhängigkeitsannahme für Beobachtungen an verschiedenen Individuen / Clustern.

## Spezialfall hierarchische Struktur: Beispiel

Die Lesefähigkeit von 875 achtjährigen Schülern in 29 Klassen in 11 Schulen wird anhand eines standardisierten Scores  $y$  gemessen. Mögliches Modell:

$$y_{ijk} = \beta_0 + b_i + b_{ij} + \varepsilon_{ijk},$$

wobei

- $y_{ijk}$ : der Lesescore des  $k$ -ten Kindes in der  $j$ -ten Klasse der  $i$ -ten Schule
- $\beta_0$ : die allgemeine mittlere Lesefähigkeit von Achjährigen
- $b_i$ : die Abweichung der mittleren Lesefähigkeit in Schule  $i$  vom allgemeinen Mittel
- $b_{ij}$ : die Abweichung der mittleren Lesefähigkeit der Klasse  $j$  vom Mittel ihrer Schule  $i$
- $\varepsilon_{ijk}$ : die Abweichung der Lesefähigkeit von Kind  $k$  von der mittleren Lesefähigkeit seiner Klasse.

## Spezialfall gekreuzte Struktur: Beispiel

In einem Phonetik-Experiment sprechen 9 Subjekte 140 Wortkombinationen, in denen *s*- und *sch*-Laute vorkommen, je 5 mal. Ein Index  $y$ , der von der Zungenstellung abhängt, misst, wie stark der Laut *s/sch* ausgedrückt wird.

Mögliches Modell für  $y_{ijk}$  ( $i$ te Person,  $j$ tes Wort,  $k$ te Wiederholung):

$$y_{ijk} = \beta + b_i + c_j + d_{ij} + \varepsilon_{ijk},$$

$$b_i \stackrel{iid}{\sim} N(0, \tau_b^2), \quad c_j \stackrel{iid}{\sim} N(0, \tau_c^2), \quad d_{ij} \stackrel{iid}{\sim} N(0, \tau_d^2), \quad \varepsilon_{ijk} \stackrel{iid}{\sim} N(0, \sigma^2),$$

$$i = 1, \dots, 9; \quad j = 1, \dots, 140; \quad k = 1, \dots, 5.$$

mit

- $\beta$  Gesamtmittel
- zufällige Effekte  $b_i$  für Person  $i$ ,  $c_j$  für Wort  $j$  und  $d_{ij}$  für deren Interaktion.

Ein einfacheres Modell wäre das Modell ohne Interaktion  $d_{ij}$ .

# Die Kovarianzstruktur

- $\mathbf{G}$  und  $\mathbf{R}$  modellieren die Abhängigkeitsstruktur von  $\mathbf{b}$  bzw.  $\varepsilon$ . Zusätzliche Annahmen (z.B. Diagonalmatrix) - unabhängig für  $\mathbf{G}$  und  $\mathbf{R}$  - ergeben Modelle verschiedener Komplexität und Flexibilität.
- $\mathbf{G}$ ,  $\mathbf{R}$  und  $\mathbf{Z}$  implizieren zusammen die Kovarianzstruktur für  $\mathbf{y}$ ,  
$$\mathbf{V} = \text{Cov}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}.$$
- $\mathbf{Z}\mathbf{b}$  zusammen mit  $\mathbf{G}$  modelliert Unterschiede zwischen Beobachtungseinheiten - z.B. bei Longitudinaldaten im individuellen Mittelwert und in der Veränderung über die Zeit.
- $\mathbf{R}$  modelliert die verbleibende Autokorrelation, die nicht durch  $\mathbf{Z}\mathbf{b}$  erklärt wird, sowie möglichen Messfehler.

# Die Kovarianzstruktur

Stochastic components in general linear mixed model

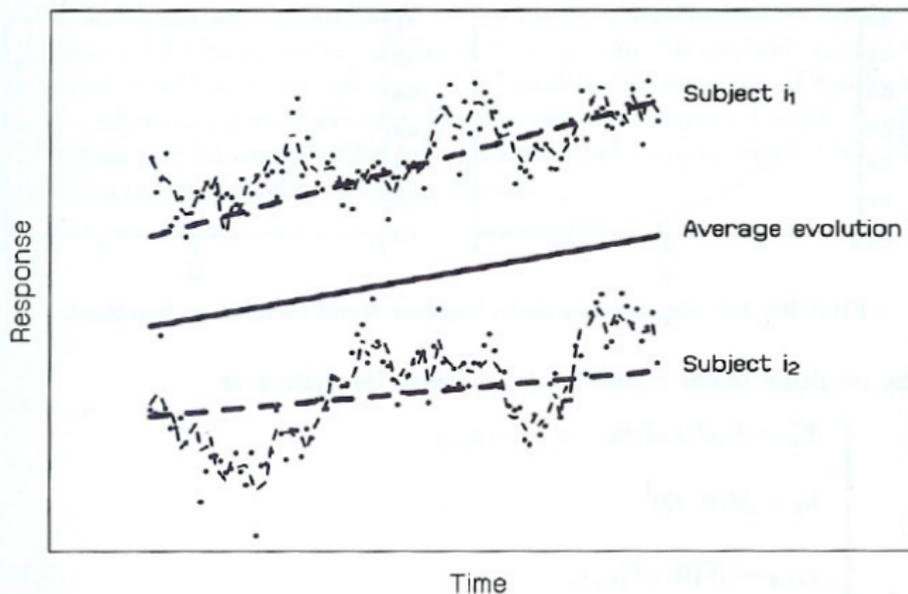


FIGURE 3.1. Graphical representation of the three stochastic components in the general linear mixed model (3.11). The solid line represents the population-average evolution. The lines with long dashes show subject-specific evolutions for two subjects  $i_1$  and  $i_2$ . The residual components of serial correlation and measurement error are indicated by short-dashed lines and dots, respectively.

Quelle: Verbeke & Molenberghs (2000)

# Conditional Independence Model

Die strengste Annahme für die Fehler ist, dass sie unabhängig und identisch normalverteilt sind,  $\mathbf{R} = \sigma^2 \mathbf{I}_n$  oder

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2), i = 1, \dots, n \Leftrightarrow \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Aus der Unabhängigkeit der Fehler folgt die bedingte Unabhängigkeit der  $y_i$  gegeben  $\mathbf{b}$ , also von  $y_i | \mathbf{b}, \dots, y_n | \mathbf{b}$ .

Die **Korrelation** zwischen den Beobachtungen  $y_i$  wird im **Conditional Independence Model** nur durch den Vektor  $\mathbf{b}$  der **zufälligen Effekten** erzeugt.

Werden die zufälligen Effekte zusätzlich unabhängig angenommen, (bei Longitudinal-/Clusterdaten  $\mathbf{D} = \text{diag}(\tau_1^2, \dots, \tau_q^2)$  diagonal) so spricht man von einem **Varianzkomponentenmodell**.

## Spezialfall Random Intercept Modell

Bei Designvektor  $\mathbf{z}'_{ij} = 1$  ergibt sich das **Random Intercept Modell**

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad b_i \stackrel{iid}{\sim} N(0, \tau^2)$$

mit individuellen Interzepten (Beispiel 100-Meter-Lauf mit  $\mathbf{x}'_{ij}\boldsymbol{\beta} = \beta_{g_i}$ ).

In Kombination mit  $\varepsilon_i \sim N(0, \sigma^2 \mathbf{I}_{n_i})$  führt dies zur marginalen Kovarianzstruktur

$$\begin{aligned} \text{Cov}(y_{ij}, y_{ik}) &= \tau^2 + \sigma^2 \delta_{jk} \\ \Rightarrow \text{Corr}(y_{ij}, y_{ik}) &= \frac{\tau^2}{\sigma^2 + \tau^2} =: \rho \geq 0, \quad j \neq k \end{aligned}$$

(mit Kronecker-Delta  $\delta_{jk} = 1$  für  $j = k$ ,  $\delta_{jk} = 0$  sonst.)

Block-konstante Korrelationsstruktur der Zielvariablen (**Compound Symmetry**).  
Groß bei großer Variabilität zwischen Individuen relativ zu innerhalb der Individuen.

# Random Intercept-Random Slope Modell

Beim **Random Intercept-Random Slope Modell**

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}, \quad (b_{0i}, b_{1i})' \stackrel{iid}{\sim} N\left(\mathbf{0}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix}\right)$$

unterscheiden sich individuelle Interzepte und Steigungen über die Zeit. Bei  $\varepsilon_i \sim N(0, \sigma^2 \mathbf{I}_{n_i})$  ergibt sich eine quadratische Varianzfunktion in der Zeit:

$$\begin{aligned} \text{Var}(y_{ij}) &= \tau_1^2 + 2\tau_{12}t_{ij} + \tau_2^2 t_{ij}^2 + \sigma^2 \quad \text{und} \\ \text{Cov}(y_{ij}, y_{ik}) &= \tau_1^2 + \tau_{12}t_{ij} + \tau_{12}t_{ik} + \tau_2^2 t_{ij}t_{ik}, \quad j \neq k. \end{aligned}$$

Bei zusätzlichem quadratischen Term  $b_{2i}t_{ij}^2$  ergibt sich ein Polynom 4. Ordnung.

# Allgemeines $R$

Wenn die Kovarianz nicht allein durch die zufälligen Effekte erklärt werden kann, d.h.  $R = \sigma^2 I_n$  vereinfachend wäre, sollte  $R$  allgemeiner gewählt werden.

Bei Longitudinaldaten z.B. wird  $R$  meist block-diagonal gewählt, mit

$$\text{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = \underbrace{\sigma_1^2 \delta_{jk}}_{\text{Messfehler}} + \underbrace{\sigma_2^2 g(|t_{ij} - t_{ik}|)}_{\text{Autokorrelation}}$$

für eine monoton fallende Funktion  $g(\cdot)$  mit  $g(0) = 1$  und  $\lim_{u \rightarrow \infty} g(u) = 0$ .

Zwei häufig verwendete Familien von Korrelationsfunktionen sind die ( $\rightarrow$  Blatt)

- **Potenz-Exponentialfamilie**  $g(u) = \exp\{-(|u|/\phi)^\kappa\}$  mit  $\phi > 0, 0 < \kappa \leq 2$ .  
 $\kappa = 1$ : Exponential-,  $\kappa = 2$ : Gauss-Korrelationsfunktion.
- zwei-parametrische **Matérn-Familie**.

Auch AR-Prozesse, unstrukturierte Kovarianzen bei balanzierten Daten verwendet.

In Anwendungen **Autokorrelation** häufig durch zufällige Effekte und **Messfehler** dominiert  $\Rightarrow$  Schätzprobleme, falls alle drei Komponenten im Modell.

# Konditionale und Marginale Perspektive

**Konditionale** oder **bedingte Perspektive** auf das gemischte Modell:

$$\mathbf{y}|\mathbf{b} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2\mathbf{R}), \quad \mathbf{b} \sim N(\mathbf{0}, \mathbf{G}). \quad (8)$$

Interpretation: Die zufälligen Effekte sind individuelle Effekte (an Beobachtungseinheiten), die in der Population variieren und regularisiert geschätzt werden.

In dieser **hierarchischen Formulierung** des LMM wird der Erwartungswert von  $y_i$  als Funktion von Populationseffekten und individuellen Effekten modelliert.

**Marginale Perspektive** auf das gemischte Modell:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \quad \text{mit} \quad \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \quad (9)$$

Interpretation: Die zufälligen Effekte induzieren eine Korrelationsstruktur und ermöglichen so eine valide statistische Analyse korrelierter Daten.

In der **marginalen Formulierung** des LMM wird der marginale, über die Population gemittelte Erwartungswert von  $y_i$  als Funktion von Populationseffekten modelliert.

# Konditionale und Marginale Perspektive

- Aus der hierarchischen Darstellung (8) folgt die marginale Darstellung (9).
- Bezeichne mit  $p$  die Dichten der entsprechenden Verteilungen. Dann ist einfach zu zeigen, dass die Dichte der marginalen Verteilung

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{b})p(\mathbf{b})d\mathbf{b}$$

die Dichte einer Normalverteilung (NV) mit EW  $\mathbf{X}\beta$  und Kovarianz  $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$  ist.

- Im LMM, wenn  $\mathbf{y}|\mathbf{b}$  normalverteilt ist, lässt sich diese Integration analytisch durchführen. Dass dies im Allgemeinen nicht geht, ist ein wesentlicher Grund dafür, dass die Inferenz für GLMMs schwieriger ist als für LMMs.

# Konditionale und Marginale Perspektive

- Aus der marginalen Verteilung von  $\mathbf{y}$  alleine folgt nicht die bedingte Verteilung für  $\mathbf{y}$  gegeben  $\mathbf{b}$  und die Verteilung von  $\mathbf{b}$ .
- Das marginale Modell für sich betrachtet nimmt keine zufälligen Effekte an, um Heterogenitäten darzustellen.
- Nicht jede Kovarianz  $\mathbf{V}$  erlaubt hierarchische Interpretation,  $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ .

**Hierarchisches** und **marginale Modell** sind also **nicht äquivalent**. Beispiel:

- Random Intercept Modell konditional:  $\tau^2$  als Varianz muss nicht-negativ sein.
- Marginal Compound Symmetry:  $\tau^2$  als Kovarianz könnte negativ werden.

Dennoch gleiche Interpretation der festen Effekte  $\beta$  in hierarchischer und marginaler Formulierung des LMM. Dies ist aber i.A. für **generalisierte lineare gemischte Modelle** (GLMMs) nicht der Fall! (Mehr dazu in Kapiteln 6+7.)

# Konsequenzen für die Schätzung

- Hauptinteresse nur an festen Effekten  $\beta$   
⇒ Verwendung des marginalen Modells.
- Interesse an festen und zufälligen Effekten  $\beta$ ,  $\mathbf{b}$  sowie den Varianz-/Kovarianzkomponenten in der Kovarianzmatrix  $\mathbf{D}$   
⇒ Verwendung der zweistufigen Darstellung.