Gemischte Modelle

September 2015

Sonja Greven (V) und Sarah Brockhaus (Ü)
Institut für Statistik
Ludwig-Maximilians-Universität München
http://www.statistik.lmu.de/institut/ag/fda/mixedmodels_2015/

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. (2009)**. *Regression: Modelle, Methoden und Anwendungen (2. Auflage).* Springer. - Begleitend zur Vorlesung. Erhältlich als Ebook bei der Universitätsbibliothek: `https://opacplus.ub.uni-muenchen.de/search?bvnr=BV035722952`
- 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`.

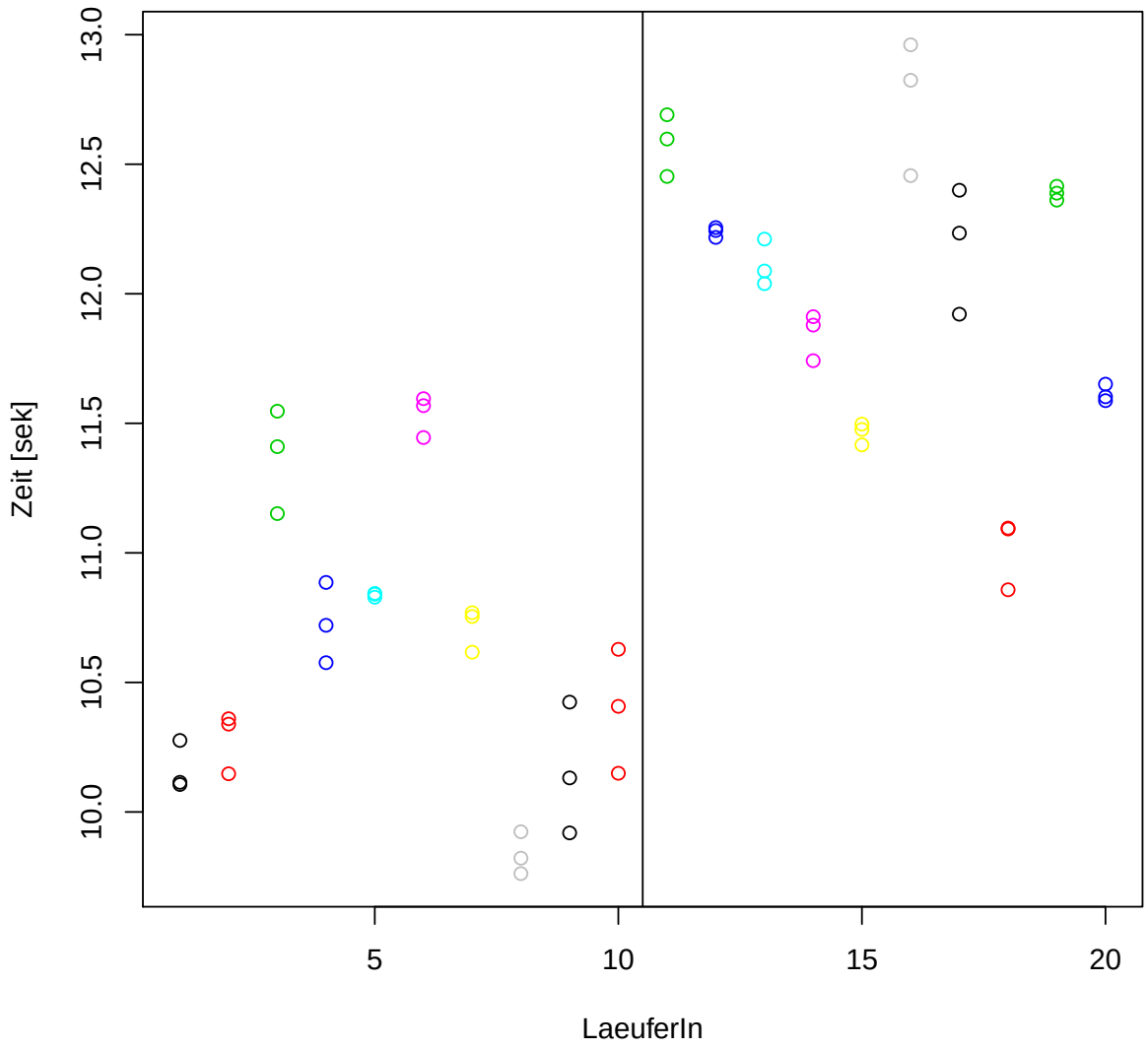
Inhalt der Vorlesung Gemischte Modelle

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

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 μ_i (d.h. ein Wert pro Person), wäre ein sinnvolles Modell (iid = unabhängig identisch verteilt)

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

da die Durchschnittszeiten einzelner Person um das Geschlechtsmittel β_{g_i}, $g_i \in \{1, 2\}$, variieren.

Die beobachteten Zeiten pro Lauf streuen um die individuelle Durchschnittszeit:

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

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

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

Ein Blick auf das Modell

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

Die Effekte b_i sind

- die „Fehler"terme in (1)
- Teil des Erwartungswertes in (2)
- sogenannte **zufällige Effekte** in Modell (3). Sie spiegeln hier wieder, dass die LäuferInnen aus einer Population kommen (in der wir die individuellen Durchschnittszeiten als normalverteilt um das Geschlechtsmittel annehmen).

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

Interpretation der Parameter

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

- β_1, β_2 die Durchschnittszeiten für Männer / Frauen (Populationsparameter)
- b_i die Abweichung der Durchschnittszeit von Person i vom Mittel der Geschlechtsgruppe (individuelle Effekte)
- τ^2 die Varianz der Durchschnittszeiten pro Geschlechtsgruppe (in den zwei Gruppen als gleich **angenommen**)
- ε_{ij} die Abweichung der j-ten Laufzeit von der Durchschnittszeit für Person i
- σ^2 die Varianz der persönlichen Laufzeiten (für alle i gleich **angenommen**)

Bedingte Sicht

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

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

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

b_i modelliert individuelle Effekte im Erwartungswert (EW) analog zum linearen Modell mit festen Effekten

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

Aber:

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

Marginale Sicht

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

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

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

b_i induziert eine Kovarianzstruktur analog zum allgemeinen linearen Modell

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

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. (z.B. Patienten über die Zeit)
- Clusterdaten / gruppierte Daten: Gruppen (Cluster) mit mehreren Beobachtungen pro Gruppe. (z.B. Läuferdaten, Daten für Familien)
- Hierarchische Daten: Gruppierte Daten mit mehreren geschachtelten Ebenen. (z.B. Schüler in Klassen in Schulen, Patienten in Ärzten in Krankenhäusern)
- Gekreuzte Designs: Gruppierte Daten mit mehreren gekreuzten Ebenen. (z.B. alle Personen bearbeiten die gleichen Aufgaben)

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

Das allgemeine lineare gemischte Modell (LMM)

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

$$\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times p}{\boldsymbol{X}}\ \underset{p\times 1}{\beta} + \underset{n\times r}{\boldsymbol{Z}}\ \underset{r\times 1}{\boldsymbol{b}} + \underset{n\times 1}{\varepsilon}\,. \tag{6}$$

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

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

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

Zu den Verteilungsannahmen

- Die **Normalverteilungsannahme der Fehler** ε 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** $\boldsymbol{b}$ ist die **Normalverteilungsannahme** nicht zwingend. Alternative Verteilungen, z.B. Mischungsverteilungen, möglich. I.d.R. werden dann die Algorithmen zur Berechnung der Schätzer komplexer.

- ε und $\boldsymbol{b}$ sind als unabhängig angenommen.

Vorteile der Analyse mit gemischten Modellen

Zufällige Effekte können als Platzhalter für die Effekte von unbeobachteten oder unzureichend gemessenen Kovariablen dienen, die **Korrelation zwischen Beobachtungen an den gleichen Beobachtungseinheiten** verursachen.

Im Gegensatz zum linearen Regressionsmodell mit unabhängigen Fehlern führt die Berücksichtigung dieser Korrelation

- zu einer **verbesserten Schätzgenauigkeit** (kleineren wahren Standardfehlern) $\rightarrow$ **Übung**
- zu validen modellbasierten Standardfehlern und damit Konfidenzintervallen und Tests $\rightarrow$ **Übung**

Vorteile der Analyse mit gemischten Modellen

Im Gegensatz zum linearen Regressionsmodell mit festen Effekten für die Beobachtungseinheiten

- können Effekte für Kovariablen (z.B. Geschlecht), die nur zwischen Beobachtungseinheiten variieren, geschätzt werden.
- werden die festen Effekte wegen der kleineren Anzahl von Modellparametern effizienter geschätzt

Im Gegensatz zum linearen Regressionsmodell mit allgemeiner Kovarianz

- erlauben die Vorhersagen für die zufälligen Effekte **individuelle Prognosen**.

Weitere Annahmen in gemischten Modellen

Auch wenn die zufälligen Effekte die Effekte von unbeobachteten Kovariablen auffangen, so können Sie **nicht** Confounding durch solche Kovariablen verhindern.

Dies liegt daran, dass die zufälligen Effekte (marginal, siehe später), die Kovarianzstruktur beeinflussen, nicht jedoch den Erwartungswert.

Streng genommen, lautet in (7) die Annahme an ε und $\boldsymbol{b}$:

$$\mathrm{E}(\varepsilon|\boldsymbol{X}, \boldsymbol{b}) = \mathbf{0}, \qquad \mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \mathbf{0}$$

(Regressionsannahme und Random effects-Annahme), so dass $\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\beta$.
Dies ist insbesondere verletzt, wenn Confounding vorliegt.

Beispiel Confounding

Betrachte das Modell

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + z_{ij}\gamma + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

y_{ij} misst den Schulerfolg von Kind j in Schule i; z_{ij} ist a) Förderbedarf des Kindes oder b) Familieneinkommen. Beispiele für mögliches Confounding wären

a) Bei Förderbedarf wählen Eltern Schulen mit unterstützender Schulkultur, die sich auch im Schulerfolg wiederspiegelt (Random effects-Annahme verletzt).

b) Familieneinkommen korreliert mit Sprachkenntnissen, Ziele der Eltern für ihre Kinder etc., die mit Schulerfolg korrelieren (Regressionsannahme verletzt).

Ein Modell mit festen Effekten b_i würde den Bias durch a), jedoch nicht den Bias durch b) verhindern. Ziel muss auf jeden Fall sein, in $\boldsymbol{x}_{ij}$ mögliche Confounder möglichst gut abzubilden.

Frei nach Clarke, Crawford, Steele & Vignoles (2010): *The Choice Between Fixed and Random Effects Models: Some Considerations for Educational Research.* IZA Discussion Paper No. 5287

Beispiel 100-Meter-Lauf

Das Modell für die 100-Meter-Lauf-Daten

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

lässt sich in die Form

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

bringen mit

$$\boldsymbol{y} = (y_{11}, \ldots, y_{20,3}) \qquad \boldsymbol{b} = (b_1, \ldots, b_{20})$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 \end{pmatrix}' \qquad \boldsymbol{G} = \tau^2 \boldsymbol{I}_{20}$$

$$\beta = (\beta_1, \beta_2)' \qquad \varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{20,3})$$

$$\boldsymbol{Z} = \begin{pmatrix} 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ & & & \ddots & & & \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 1 \end{pmatrix}' \qquad \boldsymbol{R} = \sigma^2 \boldsymbol{I}_{60}.$$

Spezialfall Longitudinal- und Clusterdaten

Wiederholte Beobachtungen y_{ij} der Zielvariablen $\rightarrow$ **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 $(\boldsymbol{x}'_{ij}, \boldsymbol{z}'_{ij})'$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$.

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

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

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

Annahmen:

- Unabhängig und identisch normalverteilte zufällige Effekte $\boldsymbol{b}_i \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{D})$,
- unabhängig und normalverteilte Fehler $\varepsilon_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$,
- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \varepsilon_1, \ldots, \varepsilon_N$ unabhängig,
- $\boldsymbol{D}$ positiv semi-definit und $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_N$ positiv definit.

Spezialfall Longitudinal- und Clusterdaten

$$\boldsymbol{y}_i \quad = \quad \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N.$$

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ entspricht den (festen) **Populationseffekten.**
- $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})'$ entspricht den (zufällligen) **subjekt- / clusterspezifischen Effekten.**
- $\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \vdots \\ \boldsymbol{x}'_{in_i} \end{pmatrix}$, $\boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{z}'_{i1} \\ \vdots \\ \boldsymbol{z}'_{in_i} \end{pmatrix}$ 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 $\boldsymbol{z}_{ij}$ in $\boldsymbol{x}_{ij}$ enthalten sind, lassen sich die $\boldsymbol{b}_i$ mit $\mathrm{E}(\boldsymbol{b}_i) = \boldsymbol{0}$ als **individuelle Abweichungen** vom Populationsmittel interpretieren.

Spezialfall Longitudinal- und Clusterdaten

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

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

mit

- $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$ und $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_N')'$ der Länge $n = \sum_{i=1}^{N} n_i$,
- β der Länge p,
- $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_N')'$ der Länge $r = Nq$,
- $\boldsymbol{X} = (\boldsymbol{X}_1' | \ldots | \boldsymbol{X}_N')'$ der Dimension $n \times p$,
- $\boldsymbol{Z} = \text{blockdiag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ der Dimension $n \times Nq$,
- $\boldsymbol{G} = \text{blockdiag}(\boldsymbol{D}, \ldots, \boldsymbol{D}, \ldots, \boldsymbol{D})$ der Dimension $Nq \times Nq$,
- $\boldsymbol{R} = \text{blockdiag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_i, \ldots, \boldsymbol{\Sigma}_N)$ 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:

$$\begin{aligned} y_{ijk} &= \beta_0 + b_i + b_{ij} + \varepsilon_{ijk}, \\ & \quad b_i \overset{iid}{\sim} N(0, \tau_1^2),\ \varepsilon_{ijk} \overset{iid}{\sim} N(0, \tau_2^2),\ b_i, \varepsilon_{ijk} \text{ unabh.} \end{aligned}$$

wobei

- y_{ijk}: der Lesescore des k-ten Kindes in der j-ten Klasse der i-ten Schule
- β_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
- ε_{ijk}: die Abweichung der Lesefähigkeit von Kind k von der mittleren Lesefähigkeit seiner Klasse.

Spezialfall hierarchische Struktur: Beispiel

Das Modell lässt sich wieder in die allgemeine LMM-Form

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

bringen mit

- $\boldsymbol{y} = (y_{1,1,1}, \ldots, y_{11,3,30})'$ und $\varepsilon = (\varepsilon_{1,1,1}, \ldots, \varepsilon_{11,3,30})'$ der Länge $n = 875$,
- $\boldsymbol{X} = (1, \ldots, 1)'$ der Dimension 875×1,
- $\beta = \beta_0$,
- $\boldsymbol{b} = (b_1, \ldots, b_{11} | b_{1,1}, \ldots, b_{11,3})'$ der Länge $r = 11 + 29 = 40$,
- $\boldsymbol{Z} = (\text{blockdiag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{11}) | \text{blockdiag}(\boldsymbol{Z}_{1,1}, \ldots, \boldsymbol{Z}_{11,3}))$ der Dimension 875×40, wobei $\boldsymbol{Z}_i$ bzw. $\boldsymbol{Z}_{ij}$ Einservektoren sind mit Länge gleich der Anzahl der Schüler (in Klasse j) in Schule i,
- $\boldsymbol{G} = \text{blockdiag}(\tau_1^2 \boldsymbol{I}_{11}, \tau_2^2 \boldsymbol{I}_{29})$ der Dimension 40×40,
- $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_{875}$ der Dimension 875×875.

Spezialfall gekreuzte Struktur: Beispiel

In einem Phonetik-Experiment sprechen 9 Subjekte 140 Worte, in denen *s*- und *sch*-Laute vorkommen, je 5 mal. Ein akustischer Index y misst, ob der gesprochene Laut einem *s* oder *sch* näher ist.

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

$$y_{ijk} = \boldsymbol{x}_j'\boldsymbol{\beta} + b_i + c_j + d_{ij} + \varepsilon_{ijk},$$

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

$$b_i, c_j, d_{ij}, \varepsilon_{ijk} \text{ unabhängig},\ i = 1, \ldots, 9;\ j = 1, \ldots, 140;\ k = 1, \ldots, 5,$$

mit

- $\boldsymbol{\beta}$ Effekte von Wortmerkmalen wie Betonung etc.
- 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}.

Spezialfall gekreuzte Struktur: Beispiel

Das Modell lässt sich wieder in die allgemeine LMM-Form

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

bringen mit

- $\boldsymbol{y} = (y_{1,1,1}, \ldots, y_{9,140,5})'$ und $\varepsilon = (\varepsilon_{1,1,1}, \ldots, \varepsilon_{9,140,5})'$ der Länge $n = 9 \cdot 140 \cdot 5 = 6300$,
- $\boldsymbol{X}$ der Dimension $6300 \times p$ enthält $\boldsymbol{x}_j$ zeilenweise,
- $\boldsymbol{Z} = (\boldsymbol{I}_9 \otimes \boldsymbol{1}_{140\times 5} | \boldsymbol{1}_9 \otimes \boldsymbol{I}_{140} \otimes \boldsymbol{1}_5 | \boldsymbol{I}_{9\times 140} \otimes \boldsymbol{1}_5)$ der Dimension 6300×1409, wobei $\boldsymbol{1}_l$ der Einservektor der Länge l ist und $\otimes$ das Kroneckerprodukt.
- $\boldsymbol{b} = (b_1, \ldots, b_9 | c_1, \ldots, c_{140} | d_{1,1}, \ldots, d_{9,140})'$ der Länge $r = 9 + 140 + 9 \cdot 140 = 1409$,
- $\boldsymbol{G} = \text{blockdiag}(\tau_b^2 \boldsymbol{I}_9, \tau_c^2 \boldsymbol{I}_{140}, \tau_d^2 \boldsymbol{I}_{9\times 140})$ der Dimension 1409×1409,
- $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_{6300}$ der Dimension 6300×6300.

Die Kovarianzstruktur

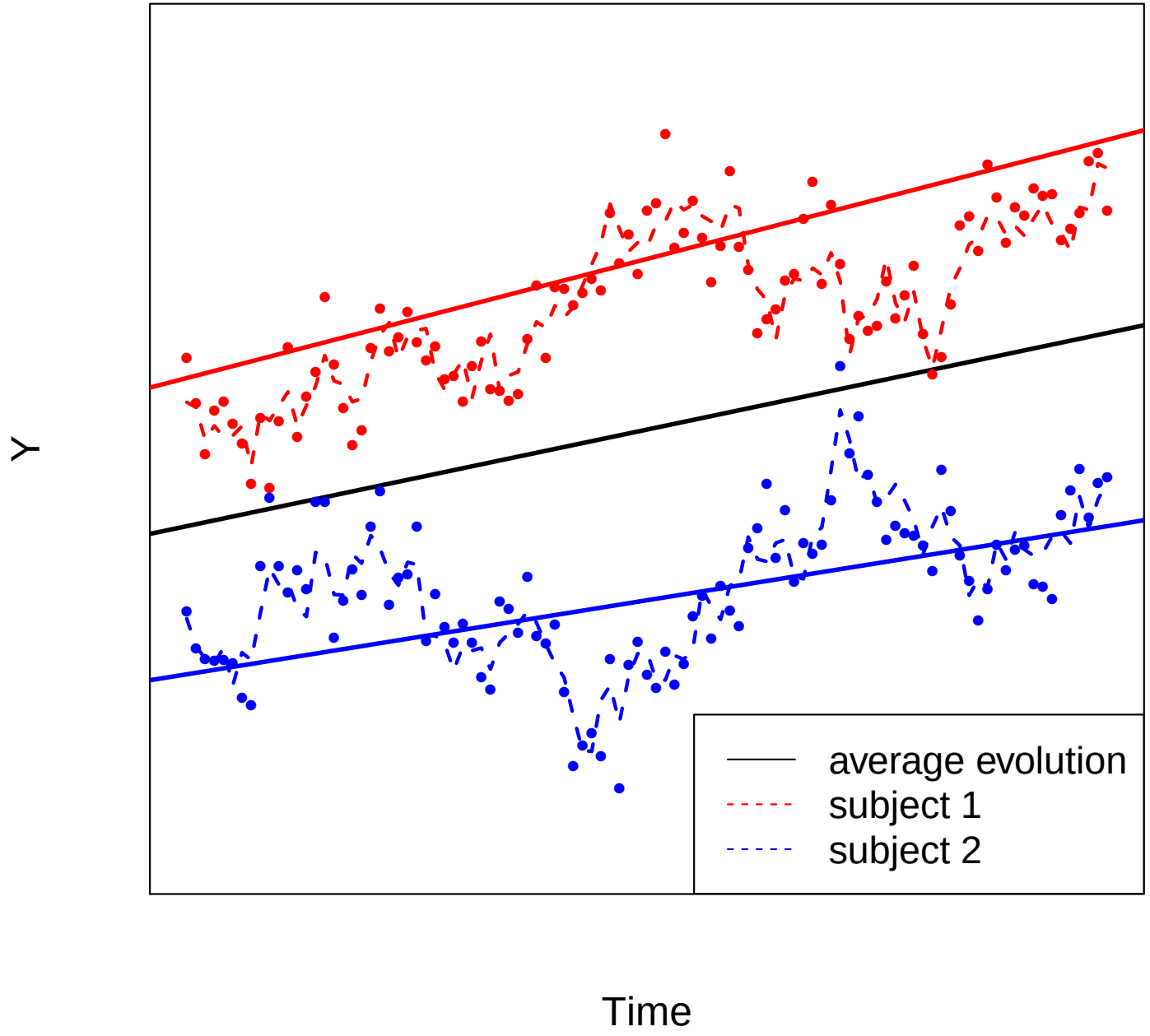
- $\boldsymbol{G}$ und $\boldsymbol{R}$ modellieren die Abhängigkeitsstruktur von $\boldsymbol{b}$ bzw. ε. Zusätzliche Annahmen (z.B. Diagonalmatrix) - unabhängig für $\boldsymbol{G}$ und $\boldsymbol{R}$ - ergeben Modelle verschiedener Komplexität und Flexibilität.

- $\boldsymbol{G}$, $\boldsymbol{R}$ und $\boldsymbol{Z}$ implizieren zusammen die Kovarianzstruktur für $\boldsymbol{y}$, $\boldsymbol{V} = \text{Cov}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$.

- $\boldsymbol{Z}\boldsymbol{b}$ zusammen mit $\boldsymbol{G}$ modelliert Unterschiede zwischen Beobachtungseinheiten - z.B. zwischen Schülern und Klassen.

- ε ist der Fehlerterm. $\boldsymbol{R}$ fängt möglicherweise verbleibende Autokorrelation auf, die nicht durch $\boldsymbol{Z}\boldsymbol{b}$ erklärt wird.

Die Kovarianzstruktur - longitudinales Beispiel



Conditional Independence Model

Die stärkste Annahme für die Fehler ist, dass sie unabhängig und identisch normalverteilt sind, $\boldsymbol{R} = \sigma^2 \boldsymbol{I_n}$ oder

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n \quad \Leftrightarrow \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

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

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

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

Spezialfall Random Intercept Modell

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

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

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

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

$$\text{Cov}(y_{ij}, y_{ik}) = \tau^2 + \sigma^2 \delta_{jk}$$

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

(mit Kronecker-Delta $\delta_{jk} = 1$ für $j = k$, $\delta_{jk} = 0$ sonst.)
Block-konstante Korrelationsstruktur der Zielvariablen (***Compound Symmetry***).
ρ groß wenn interindividuelle Varianz groß relativ zur intraindividuellen Varianz.

Random Intercept-Random Slope Modell

Beim ***Random Intercept-Random Slope Modell***

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

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

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

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

Nicht-longitudinales Beispiel: t_{ij} Dosis eines Medikaments - b_{1i} berücksichtigt individuelle Unterschiede in der Reaktion auf das Medikament.

Allgemeines R

Manchmal ist die Annahme $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$ zu vereinfachend.

Autokorrelation
Bei Longitudinaldaten ($\rightarrow$ ALD) z.B. wird $\boldsymbol{R}$ block-diagonal gewählt, mit (bei balanzierten Daten) unstrukturierten Kovarianzen $\boldsymbol{\Sigma}_i$ oder

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

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

Häufig nur entweder unabhängiger oder autokorrelierter Fehler gut schätzbar.

Heteroskedastizität
Bei den 100-Meter-Lauf-Daten könnte man z.B. zulassen, dass die Varianz $\sigma^2_{g_i}$ der individuellen Zeiten vom Geschlecht abhängt.

Konditionale und marginale Perspektive

Konditionale oder bedingte Perspektive auf das gemischte Modell:

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

Interpretation: Die zufälligen Effekte sind individuelle Effekte (von 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:

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

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

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

  die Dichte einer Normalverteilung (NV) mit EW $\boldsymbol{X}\beta$ und Kovarianz $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$ ist.
- Im LMM, wenn $\boldsymbol{y}|\boldsymbol{b}$ normalverteilt ist, lässt sich diese Integration analytisch durchführen. Dass dies für andere Verteilungen der Exponentialfamilie nicht geht, ist ein wesentlicher Grund dafür, dass die Inferenz für **generalisierte lineare gemischte Modelle** (GLMMs) schwieriger ist als für LMMs.

Konditionale und marginale Perspektive

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

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

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

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

Konsequenzen für die Schätzung

- Hauptinteresse nur an festen Effekten $\boldsymbol{\beta}$
  $\Rightarrow$ Verwendung des marginalen Modells.

- Interesse an festen und zufälligen Effekten $\boldsymbol{\beta}$, $\boldsymbol{b}$ sowie den Varianz-/Kovarianzkomponenten in der Kovarianzmatrix $\boldsymbol{D}$
  $\Rightarrow$ Verwendung der zweistufigen Darstellung.