Tutorium 5

Analyse longitudinaler Daten

Prof. Dr. Sonja Greven, Dipl. Stat. Jona Cederbaum,
Alexander Bauer

28. Juni 2016

Übersicht

Modellwahl

Modellwahl

Ziel: Wahl des optimalen Modells durch Vergleich der (nicht notwendigerweise genesteten) Modelle M_1 und M_2

Informationskriterien:

- $AIC = -2l + 2df$
- $BIC = -2l + ln(n)df$

  mit maximaler log-Likelihood l, Fallzahl n und Anzahl Parameter df
- Zu treffende Entscheidungen:

  1) Verwendung welcher Likelihood?

  2) Wie definiert man die Parameterzahl?
  (Keine eindeutige Definition von df in Gemischten Modellen!)

Modellwahl

Informationskriterien im LLMM:

- Das **marginale** AIC (mAIC):
  - Verwendung der marginalen Likelihood
  - $df \overset{z.B.}{=}$ Anzahl aller Parameter in θ
  - Annahme: zwei unabhängige Beobachtungen entstammen gleicher marginaler Verteilung, teilen aber nicht die gleichen random effects

    $\Rightarrow$ Fokus auf populationsspezifische fixed effects

    $\Rightarrow$ Benutzung nur mit ML-Likelihood!
- Das **konditionale** AIC (cAIC):
  - Verwendung der konditionalen Likelihood
  - df wird geschätzt (random effects $\widehat{=}$ zwischen 0 und Nq Parametern)

    (z.B. per parametrischem Bootstrap - siehe Efron (2004))
  - Annahme: zwei unabhängige Beobachtungen entstammen gleicher konditionaler Verteilung und teilen die gleichen random effects.

    $\Rightarrow$ Fokus auf Subjekte bzw. random effects
- **Beachten:** `AIC(lme)` in `R` gibt mAIC zurück!

GLMMs & Marginale Modelle

Übersicht

Drei Ansätze zur Erweiterung von GLMs auf longitudinale Daten:

- **Marginale Modelle:** Modellierung der marginalen Korrelation und/oder implizite Modellierung der Korrelation durch robuste Standardfehler (Schätzmethode GEE)
- **Gemischte Modelle:** Korrelation wird durch Annahme modelliert, dass hinter Beobachtungen eines Subjekts ein gleicher Prozess steht
- **Transition-/Markov-Modelle:** Beobachtungen sind korreliert, weil die Vergangenheit (z.B. letzte q Beobachtungen) die aktuellen Werte beeinflusst

  Beispiel: Modellierung der Infektionswahrscheinlichkeit abhängig davon, ob Person beim letzten Arztbesuch eine Infektion hatte

$$\text{logit} P(Y_{ij} = 1 | Y_{ij-1}, \ldots, Y_{i1}) = \beta_0 + \beta_1 x_{ij} + \gamma Y_{ij-1}$$

GLMMs

Notwendige Komponenten für **Definition eines GLMM:**

1) **Verteilungsannahme:**

$$Y_{ij}|\mathbf{b}_i \overset{u.}{\sim} \text{Expo-Fam.}(\theta_{ij}, \phi)$$

$$\widehat{=} \qquad f(Y_{ij} = y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) = \exp\left\{\frac{y_{ij}\theta_{ij} - \phi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}$$

2) **Systematische Komponente:** Spezifizierung des linearen Prädiktors

$$\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i, \qquad \mathbf{b}_i \overset{iid}{\sim} \mathcal{N}_q(\mathbf{0}, \mathbf{D})$$

3) **Linkfunktion:** Wahl der Linkfunktion $g(\cdot)$

$$g(\mu_{ij}) = \eta_{ij}$$

mit $\mu_{ij} = \mathbb{E}(Y_{ij}|\mathbf{b}_i)$ dem konditionalen Erwartungswert

GLMMs

Schätzung im GLMM:
Schwierigkeit: Sehr komplexe Likelihood

$$L(\beta, \mathbf{D}, \phi) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f(\mathbf{y}_{ij}|\mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i$$

Mögliche numerische Approximationen:

i) Approximation der Daten: **Penalisierte Quasi-Likelihood** (PQL)
   - Ansatz: Darstellung von Y_{ij} durch Taylor-Approximation
     $\Rightarrow$ Schätzung per Schaukel-Algorithmus
     (Äquivalent zu LMM-Schätzung auf Pseudo-Daten Y_{ij}^*)
   - Kein LQ-Test/AIC, da Verwendung von Pseudo-Likelihood!
   - Besser, je größer n_i und je näher Y_{ij} an NV

GLMMs

Schätzung im GLMM:

Schwierigkeit: Sehr komplexe Likelihood

$$L(\beta, \mathbf{D}, \phi) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f(\mathbf{y}_{ij}|\mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i$$

Mögliche numerische Approximationen:

ii) Approximation des Integranden: **Laplace-Approximation**

- Ansatz: Darstellung des Integranden als $\int_{\mathbb{R}^q} \exp\{Q_i(\mathbf{b}_i)\} d\mathbf{b}_i$

  $\Rightarrow$ Approximation von $Q_i(\mathbf{b}_i)$ durch Taylor-Approximation

  $\Rightarrow$ Schätzung per Schaukel-Algorithmus

- führt zu schneller Schätzung

- Besser, je größer n_i und je „weniger diskret“ Y_{ij}

GLMMs

Schätzung im GLMM:
Schwierigkeit: Sehr komplexe Likelihood

$$L(\beta, \mathbf{D}, \phi) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f(\mathbf{y}_{ij}|\mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i$$

Mögliche numerische Approximationen:

iii) Approximation des Integrals: **Gauß-Quadratur**

- Likelihood-Teile: $f(\mathbf{y}_i|\beta, \mathbf{D}, \phi) = \int_{\mathbb{R}^q} \prod f(y_{ij}|\mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i$
- Ansatz: $\int_{\mathbb{R}^q} f(\mathbf{z})\phi(\mathbf{z})d\mathbf{z} \approx \sum_{l=1}^{Q} w_l f(\mathbf{z}_l)$
  mit $f(\mathbf{z})$ bekannt, MNV-Dichte $\phi(\mathbf{z})$, Gewichten w_l, Stützstellen $\mathbf{z}_l$
- je höher Q, desto genauere Approximation
  ($Q = 1 \rightarrow$ Laplace-Approximation)

Marginale Modelle / GEE

Alternativer Ansatz zu GLMMs: **Marginale Modelle**

Komponenten eines marginalen Modells:

1) Spezifizierung der marginalen Erwartung $\mu_{ij} = \mathbb{E}(Y_{ij})$:

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$$

2) Spezifizierung der marginalen Varianz in Abhängigkeit von μ_{ij}:

$$\text{Var}(Y_{ij}) = \phi v(\mu_{ij})$$

3) Spezifizierung der Korrelation zwischen den Y_{ij} als Funktion ρ in Abhängigkeit von einem (zu schätzenden) Parameter $\boldsymbol{\alpha}$:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha})$$

Marginale Modelle / GEE

Beispiele für Spezifizierung der Korrelationsstruktur:

- $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$ mit $\alpha \leq 1$
- $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk}$ (unstrukturiert)

Marginale Modelle:

- Explizite Modellierung der Korrelation, keine Random Effects
- Problem: Gemeinsame Verteilung der $Y_{i1}, \ldots, Y_{in_i}$ teilweise sehr komplex / nicht voll spezifiziert

  $\Rightarrow$ Alternative zu aufwendigen ML-Verfahren: **Schätzung per GEE**

Marginale Modelle / GEE

Generalized Estimating Equations (Schätzmethode):

- Fokus auf Modellierung des marginalen Erwartungswerts

  $\Rightarrow$ keine Spezifizierung der gemeinsamen Verteilung notwendig
- Erweiterung der Quasi-Likelihood Methode für korrelierte Messungen

  $\Rightarrow$ Schätzung auf Basis einer *Arbeitskorrelation*

  $\Rightarrow$ Konsistente Schätzung auch bei falscher Korrelationsstruktur

GEE-Ansatz:

- Erinnerung: GLS-Optimierungskriterium zur Schätzung von β

$$\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{X}_i\beta)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta)$$

Marginale Modelle / GEE

GEE-Ansatz:

- GEE: Minimierung von

$$\sum_{i=1}^{N}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}))^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})),$$

mit Arbeitskovarianz $\mathbf{V}_i$ und

$$\mu_{ij} = \mu_{ij}(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta}).$$

$\Rightarrow$ Score-Gleichungen:

$$\sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

Parameterinterpretation

Parameterinterpretation

Im Folgenden **Betrachtung von Gemischten Modellen der Form:**

1) Verteilungsannahme:

$$Y_{ij}|\mathbf{b}_i \overset{u.}{\sim} \text{siehe folgende Folien}$$

2) Systematische Komponente:

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij} + b_i$$
$$b_i \overset{u.i.v.}{\sim} \mathcal{N}(0, \tau^2)$$

3) Linkfunktion:

$$g(\mu_{ij}) = \eta_{ij}$$
$$\text{mit } g(\cdot) = \text{siehe folgende Folien}$$

Parameterinterpretation

Interpretation im LMM:

Betrachte Modell mit $Y_{ij}|\mathbf{b}_i \overset{u.}{\sim} \mathcal{N}(\mu_{ij}, \sigma^2)$ und $g(\mu_{ij}) = \mu_{ij}$.

Konditionale Erwartung:

$$\mathbb{E}_{Y|b}(Y_{ij}|b_i) = \beta_0 + \beta_1 x_{ij} + b_i$$

Marginale Erwartung:

$$\begin{aligned}\mathbb{E}_Y(Y_{ij}) &= \mathbb{E}_b(\mathbb{E}_{Y|b}(Y_{ij}|b_i)) \\ &= \beta_0 + \beta_1 x_{ij} + \mathbb{E}_b(b_i) \\ &= \beta_0 + \beta_1 x_{ij}\end{aligned}$$

$\Rightarrow$ Parameter lassen sich im LMM sowohl subjekt-spezifisch als auch populationsspezifisch interpretieren!

Parameterinterpretation

Interpretation im GLMM: *Beispiel Logit-Link*

Betrachte Modell mit $Y_{ij}|\mathbf{b}_i \overset{u.}{\sim} B(\mu_{ij})$ und $g(\mu_{ij}) = \text{logit}(\mu_{ij})$.

Konditionale Erwartung:

$$P(Y_{ij} = 1|b_i) = \mathbb{E}_{Y|b}(Y_{ij}|b_i) = \frac{\exp(\beta_0 + \beta_1 x_{ij} + b_i)}{\exp(\beta_0 + \beta_1 x_{ij} + b_i) + 1}$$

Marginale Erwartung:

$$\begin{aligned} P(Y_{ij} = 1) = \mathbb{E}_b(\mathbb{E}_{Y|b}(Y_{ij}|b_i)) &= \mathbb{E}_b\left[\frac{\exp(\beta_0 + \beta_1 x_{ij} + b_i)}{\exp(\beta_0 + \beta_1 x_{ij} + b_i) + 1}\right] \\ &\neq \frac{\exp(\beta_0 + \beta_1 x_{ij} + \mathbb{E}_b(b_i))}{\exp(\beta_0 + \beta_1 x_{ij} + \mathbb{E}_b(b_i)) + 1} = \frac{\exp(\beta_0 + \beta_1 x_{ij})}{\exp(\beta_0 + \beta_1 x_{ij}) + 1} \end{aligned}$$

$\Rightarrow$ i.A. nur subjekt-spezifische Interpretation möglich und **keine marginale Interpretation**!

Parameterinterpretation

Interpretation im GLMM: *Beispiel Log-Link*

Betrachte Modell mit $Y_{ij}|\mathbf{b}_i \overset{u.}{\sim} Po(\mu_{ij})$ und $g(\mu_{ij}) = \log(\mu_{ij})$.

Konditionale Erwartung:

$$\mathbb{E}_{Y|b}(Y_{ij}|b_i) = \exp(\beta_0 + \beta_1 x_{ij} + b_i)$$

Marginale Erwartung:

$$\begin{aligned}
\mathbb{E}_Y(Y_{ij}) &= \mathbb{E}_b(\mathbb{E}_{Y|b}(Y_{ij}|b_i)) \\
&= \exp(\beta_0 + \beta_1 x_{ij}) \cdot \mathbb{E}_b(\exp(b_i))) \\
&= \exp(\beta_0 + \beta_1 x_{ij}) \cdot \exp(\tau^2/2), \qquad \text{da } b_i \overset{u.i.v.}{\sim} \mathcal{N}(0, \tau^2) \\
&= \exp((\beta_0 + \tau^2/2) + \beta_1 x_{ij}) = \exp(\beta_0^* + \beta_1 x_{ij})
\end{aligned}$$

$\Rightarrow$ Log-Link ist eine Ausnahme! Hier ist für **alle Variablen, die nur als fixed und nicht als random effect aufgenommen wurden**, auch eine marginale Interpretation möglich!

Parameterinterpretation

Interpretation im Marginalen Modell:

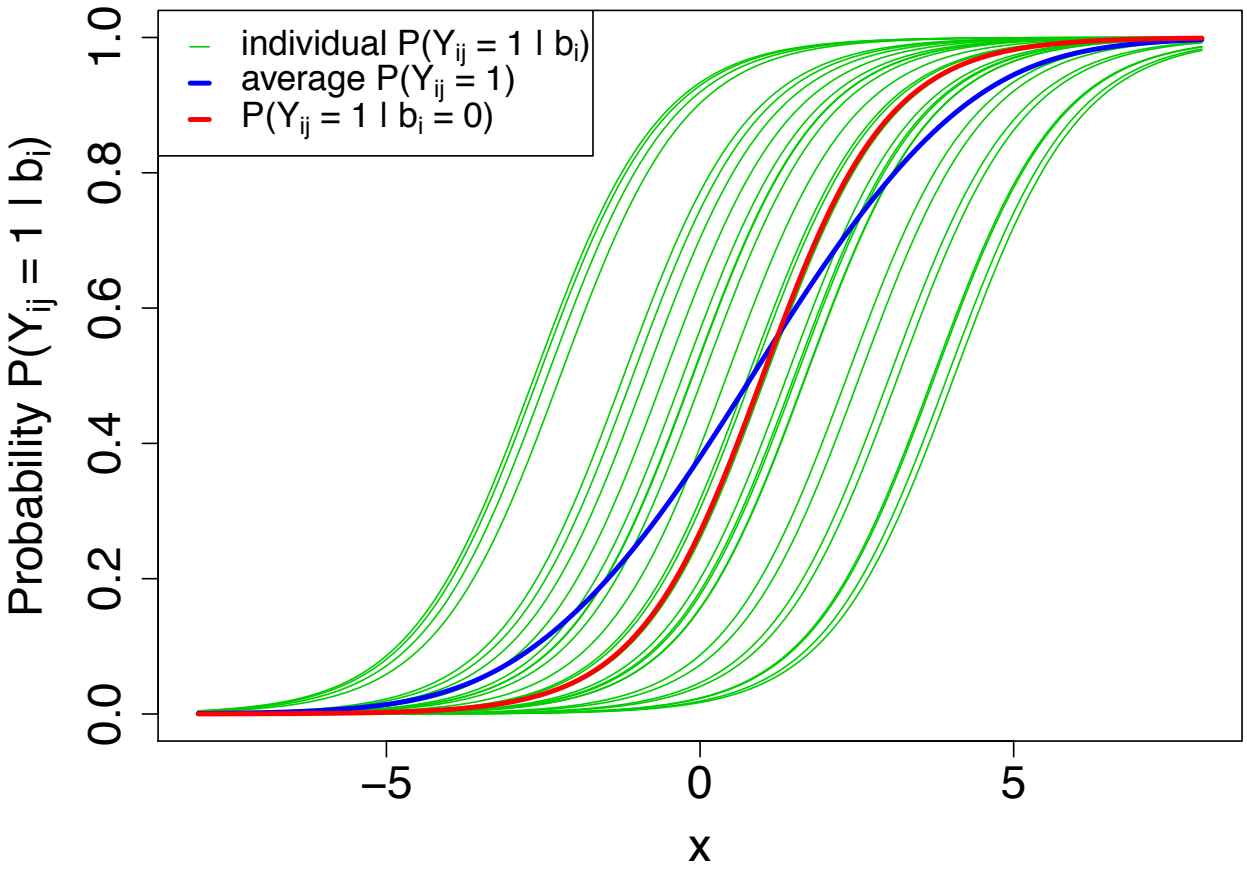
- Keine Unterscheidung zwischen konditionaler und marginaler Betrachtung möglich
- Marginales Modell = GLM mit spezifischer Korrelationsstruktur

  $\Rightarrow$ Interpretation analog zum GLM!

$\Rightarrow$ Populationsspezifische Interpretation der Parameter!

Parameterinterpretation

GLMMs vs. Marginale Modelle: *Beispiel logistische Regression*



⇒ Populationsspezifischer Effekt hier deutlich schwächer als subjekt-spezifische Effekte!

Parameterinterpretation

Fazit: Passendes Modell **abhängig von Fragestellung**

Beispiel: Untersuchung, ob eine neu entwickelte Behandlung einer Krankheit vorbeugt (logistische Regression)

Fragestellung 1: Verringert die Behandlung eines spezifischen Patienten dessen Erkrankungsrisiko? (Wichtig für behandelnden Arzt)
$\Rightarrow$ GLMM

Fragestellung 2: Wenn jeder in der Population das Treatment bekäme, wie würde sich das im Mittel auf das Erkrankungsrisiko auswirken? (Wichtig für Epidemiologen)
$\Rightarrow$ Marginales Modell

GLMMs & Marginale Modelle in R

GLMMs & Marginale Modelle in R

GLMMs:

- `glmer` {`lme4`}: Laplace-Approximation.
  bei `nAGQ` > 1 adaptive Gauß-Quadratur.
- `glmmPQL` {`MASS`}: PQL-Approximation (basierend auf `lme`).
  Wird intern von `gamm` {`mgcv`} verwendet.

Marginale Modelle:

- `gee` {`gee`}