Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2014

5 Metrische Einflußgrößen: Polynomiale Regression, Trigonometrische Polynome, Regressionssplines, Transformationen.

6 Modelldiagnose

7 Variablenselektion

Das allgemeine lineare Modell: Gewichtete KQ-Methode, Autokorrelierte und heteroskedastische Störterme

Das logistische Regressionsmodell

Das gemischte lineare Regressionsmodell („Linear mixed Model“)

Beispiel: Studie zur Leseförderung

- Zielgröße: Verbesserung der Lesefähigkeit
- Einflussgrößen: spezielle Förderung
- Störgröße: Ausgangsniveau
- Problem: Versuch wurde klassenweise durchgeführt

Voraussetzung der Unabhängigkeit der Störterme nicht erfüllt
(Cluster Daten)

Abhilfe: Einführung eines Klasseneffekts
Problem: Zu viele Parameter
Abhilfe: Klasseneffekt wird als zufälliger Effekt eingeführt.

Das Modell mit einem einfachen zufälligen Effekt (Varianzkomponenten-Modell, Random - Intercept Modell)

Wir betrachten gruppierte Daten mit Gruppenindex i:

$$
\begin{aligned}
Y_{ij} &= x_{ij}'\beta + \gamma_i + \epsilon_{ij} \quad i = 1, \ldots, g; \;\; j = 1, \ldots, n_i && (10.1)\\
\epsilon &\sim N(0, \sigma^2 I) && (10.2)\\
\gamma_i &\sim N(0, \sigma_\gamma^2) && (10.3)\\
&\quad \gamma_i \text{ und } \epsilon \text{ unabhängig}
\end{aligned}
$$

γ_i: Zufällige Effekte (Random Intercept)

Das marginale Modell

Das obige Modell kann umgeformt werden zu dem **marginalen Modell**

$$
\begin{aligned}
Y_{ij} &= x_{ij}'\beta + \delta_{ij} && (10.4)\\
\delta_{ij} &= \epsilon_{ij} + \gamma_i && (10.5)\\
Var(\delta_{ij}) &= \sigma^2 + \sigma_\gamma^2 && (10.6)\\
cov(\delta_{i_1 j_1}, \delta_{i_1 j_2}) &= \sigma_\gamma^2 && (10.7)\\
cov(\delta_{i_1 j_1}, \delta_{i_2 j_2}) &= 0 \text{ für } i_1 \neq i_2 && (10.8)
\end{aligned}
$$

Darstellung als allgemeines lineares Modell:

$$
\begin{aligned}
Y &= X\beta + \delta\\
\delta &\sim N(0, \sigma^2 I + diag[\sigma_\gamma^2 e_i e_i'])\\
e_i &:= \text{1-Vektor der Länge } n_i
\end{aligned}
$$

Das Modell mit allgemeinen zufälligen Effekten

Beispiel: Patienten mit extremem Übergewicht

- Zielgröße: Gewichtsentwicklung (in nur unregelmäßigen Abständen erhoben)
- Einflussgrößen: Geschlecht, Art der Intervention

Zweistufiges Modell mit individuellem linearem Trend, dessen Steigung vom Geschlecht (Indikator I_m) abhängt

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \epsilon_{ij} \\ \beta_{0i} &= \beta_0 + \beta_1 \cdot I_m(i) \\ \beta_{1i} &= \beta_2 + \beta_3 \cdot I_m(i) \end{aligned}$$

Einsetzen ergibt:

$$Y_{ij} = \beta_0 + \beta_1 \cdot I_m(i) + \beta_2 \cdot t_{ij} + \beta_3 \cdot I_m(i) \cdot t_{ij} + \gamma_{0i} + \gamma_{1i} \cdot t_{ij}$$

Annahme :
γ_{0i} und γ_{1i} zufällige Effekte,
Alle $\boldsymbol{\gamma_i}$ sind unabhängig, und $\boldsymbol{\gamma_i} \sim N(\mathbf{0}, G), i = 1, \ldots, g$,
G ist die Kovarianzmatrix der zufälligen Effekte innerhalb einer Einheit.
β_1 und β_2: eigentlich interessierende Populationseffekte

Ein hierarchisches Modell für longitudinale Daten Stufe 1

Sei $\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ der Vektor der wiederholten Messungen für das i-te Subjekt zu den Zeiten $t_{ij}, j = 1, \ldots, n_i$, für $i = 1, \ldots, g$.

$$\boldsymbol{Y_i} = Z_i \boldsymbol{\beta_i} + \boldsymbol{\varepsilon_i} \tag{10.9}$$

- Z_i eine $(n_i \times q)$-Matrix bekannter Kovariablen, die modelieren, wie sich die Zielgröße für das i-te Subjekt über die Zeit verhält
- $\boldsymbol{\beta_i}$ ein q-dimensionaler Vektor unbekannter subjektspezifischer Regressionskoeffizienten
- $\boldsymbol{\varepsilon_i}$ ein n_i-dimensionaler Vektor mit Residuen für das i-te Individuum
- **Annahme:**
  Alle $\boldsymbol{\varepsilon_i}$ sind unabhängig und $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i), i = 1, \ldots, N$,
  Σ_i unbekannte Kovarianzmatrix.
  (Meist Zusatzannahme: Σ_i hängt von i nur über n_i ab.)

Ein hierarchisches Modell für longitudinale Daten Stufe 2

Ein lineares Modell für die subjektspezifischen Regressionskoeffizienten $\boldsymbol{\beta_i}$:

$$\boldsymbol{\beta_i} = K_i\beta + \boldsymbol{\gamma_i} \tag{10.10}$$

- K_i eine $(q \times p)$-Matrix bekannter Kovariablen
- β ein p-dimensionaler Vektor unbekannter Regressionsparameter
- **Annahme:**
  Alle $\boldsymbol{\gamma_i}$ sind unabhängig, und $\boldsymbol{\gamma_i} \sim N(\mathbf{0}, G), i = 1, \ldots, g$,
  G unbekannte Kovarianzmatrix.

Das lineare gemischte Modell für longitudinale Daten

Substitution von (10.10) in (10.9) ergibt

$$\boldsymbol{Y_i} = X_i\beta + Z_i\boldsymbol{\gamma_i} + \boldsymbol{\varepsilon_i} \tag{10.11}$$

mit $X_i = Z_i K_i$, das **lineare gemischte Modell** mit **fixed effects** (festen Effekten) β und **random effects** (Zufallseffekten) $\boldsymbol{\gamma_i}$.

Annahme:

$$\left.\begin{array}{l} \boldsymbol{\gamma_i} \sim N(\mathbf{0}, G), \boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i) \\ \boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_g}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_g} \text{ unabhängig.} \end{array}\right\} \Rightarrow \begin{array}{l} \boldsymbol{Y_i} \sim N(X_i\beta, Z_i G Z_i' + \Sigma_i) \\ \text{(marginales Modell)} \end{array} \tag{10.12}$$

Das lineare gemischte Modell (LMM) in allgemeiner Darstellung

$$Y = X\beta + Z\gamma + \epsilon \tag{10.13}$$

$$\begin{pmatrix} \gamma \\ \epsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix}\right) \tag{10.14}$$

X und Z	:	feste bekannte Designmatrizen
β	:	Vektor der festen Effekte
γ	:	Vektor der zufälligen Effekte
R	:	Kovarianzmatrix der Störterme
G	:	Kovarianzmatrix der zufälligen Effekte

Bemerkungen:

Das obige Modell ist sehr flexibel und enthält als Spezialfälle das gemischte Modell für Longitudinaldaten und das Varianzkomponenten-Modell.

Im Modell für Longitudinaldaten gilt:

$$R \quad = \quad diag(\Sigma_i)$$

Im Varianzkomponentenmodell gilt:

$$R \quad = \quad \sigma^2 I(\text{ Dimension: Anzahl Beobachtungen } \sum_{i=1}^{g} n_i)$$

$$G \quad = \quad \sigma_\gamma^2 I(\text{ Dimension: g })$$

Marginales und bedingtes (conditionales) Modell

Marginales Modell:

$$
\begin{aligned}
Y &= X\beta + \delta && (10.15)\\
\delta &= Z\gamma + \epsilon && (10.16)\\
\delta &\sim N(0, R + ZGZ') && (10.17)
\end{aligned}
$$

Bedingtes Modell:

$$
\begin{aligned}
Y|\gamma &\sim N(X\beta + Z\gamma, \boldsymbol{R}) && (10.18)\\
\gamma &\sim N(0, \boldsymbol{G}) && (10.19)
\end{aligned}
$$

Inferenz im gemischten linearen Modell I

Die Inferenz erfolgt zunächst mit Hilfe des marginalen Modells:

Sei $\boldsymbol{\vartheta}$ ein Vektor aller Parameter, die in G und R vorkommen.

$\boldsymbol{\vartheta}$ und β können nach der Maximum - Likelihood Methode geschätzt werden:

Als Log-Likelihood ergibt sich (von additiven Konstanten abgesehen):

$$l(\beta, \boldsymbol{\vartheta}) = -\frac{1}{2}\left(\ln|V(\boldsymbol{\vartheta})| + (\boldsymbol{Y} - X\beta)' V^{-1}(\boldsymbol{\vartheta})(\boldsymbol{Y} - X\beta)\right) \qquad (10.20)$$

wobei $V = ZG(\boldsymbol{\vartheta})Z' + R(\boldsymbol{\vartheta})$.

Inferenz im gemischten linearen Modell II

Ist $\boldsymbol{\vartheta}$ bekannt, so ist der MLE von β bedingt auf $\boldsymbol{\vartheta}$ (gewichteter KQ-Schätzer:)

$$\hat{\beta}(\boldsymbol{\vartheta}) = \left(X' V(\boldsymbol{\vartheta})^{-1} X\right)^{-1} X' V^{-1}(\boldsymbol{\vartheta}) Y. \tag{10.21}$$

Einsetzen liefert die Profil-Log-Likelihood:

$$l(\boldsymbol{\vartheta}) = -\frac{1}{2}\left(\ln|V(\boldsymbol{\vartheta})| + (\boldsymbol{Y} - X\beta(\boldsymbol{\vartheta}))' V^{-1}(\boldsymbol{\vartheta})(\boldsymbol{Y} - X\beta(\boldsymbol{\vartheta}))\right) \tag{10.22}$$

ML und REML-Schätzer

Maximieren von (10.22) bezüglich $\boldsymbol{\vartheta}$ liefert ML-Schätzer.
Da dieser nicht erwartungstreu ist, verwendet man häufig den sogenannten restringierten ML-Schätzer:

Dieser maximiert

$$L_R(\boldsymbol{\vartheta}) = l(\boldsymbol{\vartheta}) - \frac{1}{2} ln|X' V(\boldsymbol{\vartheta})^{-1} X| \qquad (10.23)$$

Im einfachen linearen Modell entspricht der REML-Schätzer dem erwartungstreuen Schätzer von σ^2.

Inferenz bezüglich von β im linearen gemischten Modell

Unter dem marginalen Modell (10.11) und bedingt auf $\boldsymbol{\vartheta}$ folgt $\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta})$ einer multivariaten Normalverteilung mit Erwartungswert $\boldsymbol{\beta}$ und Kovarianzmatrix

$$\text{var}(\hat{\boldsymbol{\beta}}) \quad = \quad (X' V^{-1} X)^{-1} \tag{10.24}$$

Da V unbekannt ist, wird es durch den (RE)ML-Schätzer $V(\hat{\vartheta})$ ersetzt.

Zur Konstruktion von Konfidenzintervallen und entsprechenden Tests nimmt man an, dass β asymptotisch normalverteilt ist. Für spezielle Modelle ist dies bewiesen, aber eine allgemeingültige asymptotische Normalverteilungsaussage ist nicht nachgewiesen.

Da die Varianzmatrix V nur geschätzt wird, werden in der Praxis deshalb häufig approximative t-Tests und entsprechende Konfidenzintervalle benutzt, die die Verteilung von $(\hat{\beta}_j - \beta_j)/\hat{s.e.}(\hat{\beta}_j)$ durch eine t-Verteilung approximieren und die zugehörigen Freiheitsgrade geeignet schätzen.