

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“)

Navigationssymbole

Navigationssymbole

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.

Navigationssymbole

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

Wir betrachten gruppierte Daten mit Gruppenindex i :

$$Y_{ij} = x'_{ij}\beta + \gamma_i + \epsilon_{ij} \quad i = 1, \dots, g; \quad j = 1, \dots, n_i \quad (10.1)$$

$$\epsilon \sim N(0, \sigma^2 I) \quad (10.2)$$

$$\gamma_i \sim N(0, \sigma_\gamma^2) \quad (10.3)$$

γ_i und ϵ unabhängig

γ_i : Zufällige Effekte (Random Intercept)

Navigationssymbole

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

$$Y_{ij} = x'_{ij}\beta + \delta_{ij} \quad (10.4)$$

$$\delta_{ij} = \epsilon_{ij} + \gamma_i \quad (10.5)$$

$$\text{Var}(\delta_{ij}) = \sigma^2 + \sigma_\gamma^2 \quad (10.6)$$

$$\text{cov}(\delta_{i_1j_1}, \delta_{i_2j_2}) = \sigma_\gamma^2 \quad (10.7)$$

$$\text{cov}(\delta_{i_1j_1}, \delta_{i_2j_2}) = 0 \text{ für } i_1 \neq i_2 \quad (10.8)$$

Darstellung als allgemeines lineares Modell:

$$Y = X\beta + \delta$$

$$\delta \sim N(0, \sigma^2 I + \text{diag}[\sigma_\gamma^2 e_i e_i'])$$

$$e_i := \text{1-Vektor der Länge } n_i$$

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

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \epsilon_{ij}$$

$$\beta_{0i} = \beta_0 + \beta_1 \cdot I_m(i)$$

$$\beta_{1i} = \beta_2 + \beta_3 \cdot I_m(i)$$

Ein hierarchisches Modell für longitudinale Daten Stufe 1

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 γ_i sind unabhängig, und $\gamma_i \sim N(\mathbf{0}, G)$, $i = 1, \dots, g$,

G ist die Kovarianzmatrix der zufälligen Effekte innerhalb einer Einheit.

β_1 und β_2 : eigentlich interessierende Populationseffekte

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

$$\mathbf{Y}_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \quad (10.9)$$

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

Ein lineares Modell für die subjektspezifischen Regressionskoeffizienten β_i :

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

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

Substitution von (10.10) in (10.9) ergibt

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

mit $X_i = Z_iK_i$, das **lineare gemischte Modell** mit **fixed effects** (festen Effekten) β und **random effects** (Zufallseffekten) γ_i .

Annahme:

$$\left. \begin{array}{l} \gamma_i \sim N(\mathbf{0}, G), \epsilon_i \sim N(\mathbf{0}, \Sigma_i) \\ \gamma_1, \dots, \gamma_g, \epsilon_1, \dots, \epsilon_g \text{ unabhängig.} \end{array} \right\} \Rightarrow Y_i \sim N(X_i\beta, Z_iGZ_i' + \Sigma_i) \text{ (marginales Modell)} \tag{10.12}$$

Das lineare gemischte Modell (LMM) in allgemeiner Darstellung

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

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

- 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:

$$R = \text{diag}(\Sigma_i)$$

Im Varianzkomponentenmodell gilt:

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

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

Marginales und bedingtes (conditionales) Modell

Marginales Modell:

$$Y = X\beta + \delta \tag{10.15}$$

$$\delta = Z\gamma + \epsilon \tag{10.16}$$

$$\delta \sim N(0, R + ZGZ') \tag{10.17}$$

Bedingtes Modell:

$$Y|\gamma \sim N(X\beta + Z\gamma, R) \tag{10.18}$$

$$\gamma \sim N(0, G) \tag{10.19}$$

Inferenz im gemischten linearen Modell II

Ist ϑ bekannt, so ist der MLE von β bedingt auf ϑ (gewichteter KQ-Schätzer:)

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

Einsetzen liefert die Profil-Log-Likelihood:

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

Inferenz im gemischten linearen Modell I

Die Inferenz erfolgt zunächst mit Hilfe des marginalen Modells:

Sei ϑ ein Vektor aller Parameter, die in G und R vorkommen.

ϑ und β können nach der Maximum - Likelihood Methode geschätzt werden:

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

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

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

ML und REML-Schätzer

Maximieren von (10.22) bezüglich ϑ liefert ML-Schätzer. Da dieser nicht erwartungstreu ist, verwendet man häufig den sogenannten restringierten ML-Schätzer:

Dieser maximiert

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

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

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

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

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 $(\hat{\beta}_j - \beta_j)/s.e.(\hat{\beta}_j)$ durch eine t -Verteilung approximieren und die zugehörigen Freiheitsgrade geeignet schätzen.

