Logistische Regression mit korrelierten Beobachtungen am Beispiel der DHP

Oliver Kuß

Institut für Medizinsche Epidemiologie,
Biometrie und Informatik,
Universität Halle-Wittenberg;

Kolloquium Hannover/Magdeburg,
Hannover, 11. Juni 2002

Gliederung

- Einführung
- Datensatz, Fragestellung
- Modelle
- Schätzmethoden
- Ergebnisse
- Software
- Fazit

Das logistische Regressionsmodell

= das Standardmodell zur Regressionsanalyse binärer Zielgrößen

Gründe:

- Anschauliche Interpretation der Parameter
- Prognosen für das Eintreten des Zielereignisses sind möglich
- Gültig unter prospektivem und retrospektivem Sampling
- Kein Problem mit der Software

Standardannahme: Unabhängigkeit der Beobachtungen

Korrelierte Daten

fallen häufig im biometrisch-epidemiologischen Alltag an, z.B. bei

- Messungen zu verschiedenen Zeitpunkten beim gleichen Probanden
- Behandlung ein und desselben Probanden unter unterschiedlichen Bedingungen
- Beobachtung von Individuen in logisch zusammenhängenden Einheiten (Gemeinden, Familien, Kliniken etc.)
- Messung verschiedener Zielgrößen

Weniger Probleme bei stetigen Zielgrößen, mehr Probleme bei diskreten Zielgrößen

Ursprung

Titel: Meta-Analytische Verfahren zur Auswertung gemeindebezogener Interventionsstudien: Methodische Entwicklungen und Überprüfung von Anwendungsmöglichkeiten

Finanziert durch: DFG (April 2000 - März 2002)

Kooperierende Einrichtungen:

- Abteilung Klinische Sozialmedizin, Universitätsklinikum Heidelberg, Prof. Diepgen
- Abteilung Epidemiologie und Medizinische Statistik, Fakultät für Gesundheitswissenschaften, Universität Bielefeld, Prof. Blettner

- Abteilung Umweltepidemiologie, Deutsches Krebsforschungszentrum Heidelberg, Prof. Wahrendorf

Beteiligte Mitarbeiter:

- Dorothee Twardella, MPH
- Oliver Kuß
- Thomas Bruckner
- Dr. med. Wolfgang Scheuermann

Die Deutsche Herz-Kreislauf-Präventionsstudie (DHP)

- Größte gemeindebezogene Interventionsstudie in Deutschland

- **Hypothese:** Durch gemeindebezogene Präventionsmaßnahmen kann eine Reduktion der kardiovaskulären Risikofaktoren, kardiovaskulärer Morbidität und Mortalität erreicht werden

- **Design:** 6 Interventionsregionen (RUS) und eine nationale Stichprobe (NUS) als Kontrolle, Querschnittstudie mit 3 Erhebungszeitpunkten

- **Datenerhebung:** mittels Fragebogen und medizinischer Untersuchung

Vorbereitung der Daten der DHP

- Definition von 7 Interventions- und 7 Kontrollgemeinden

- Nur Berücksichtigung des 1. und 3. Survey (1984, 1991)

Hier: Beschränkung auf die einzige binäre Zielgröße ‘Rauchverhalten’

Situation

Gegeben: Rauchprävalenz in 14 Regionen (7 Interventions-, 7 Kontrolle) zu zwei verschiedenen Zeitpunkten

SAS:

```
data rauchen;
     input cluster time interven inttime n smoke;
     cards;
     1   0  0  0   209   79
     1   1  0  0   307  119
     2   0  0  0   354  123
     2   1  0  0   361  144
             ...
     7   0  0  0   796  261
     7   1  0  0   864  298
     8   0  1  0  1797  743
     8   1  1  1  1280  499
             ...
     14  0  1  0   690  197
     14  1  1  1   503  148
;run;
```

Statistisches Modell: Logistisches Regressionsmodell mit korrelierten Daten

Cluster-Randomized Trials

Aspekte der Planung, Durchführung und Analyse von Cluster-Randomized Trials sind ein aktuelles Thema in der medizinschen Statistik. → Sonderheft von *Statistics in Medicine* (SiM, 20(3), 2001)

Randomization *by cluster* accompanied by an analysis appropriate to randomization *by individual* is an exercise in self-deception, however, and should be discouraged. (Cornfield, 1978)

Marginale vs. Random Effects Modelle

Es existieren zwei Klassen von statistischen Modellen mit unterschiedlicher Motivation und unterschiedlicher Interpretation der Parameter

- **Marginale Modelle:** Getrennte Modellierung von Kovariablen und Intra-Cluster-Korrelation, Korrelation als Störgröße, *Population-averaged*

- **Random Effects Modelle:** Aufnahme eines zufälligen Effektes in die Modellgleichung, der *gleichzeitig* Heterogenität zwischen den Clustern und Korrelation innerhalb der Cluster modelliert, *Subject-specific*

Marginales Modell für die DHP

Modellgleichung:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \beta_0 + \beta_1 I_{ij} + \beta_2 T_{ij} + \beta_3 IT_{ij} \\ \text{Var}(Y_{ij}) &= \pi_{ij}(1-\pi_{ij}) \\ \text{Corr}(Y_{ij}, Y_{ik}) &= \alpha \end{aligned}$$

mit

$$\begin{aligned} i &= 1, \ldots, 14, \\ j &= 1, \ldots, n_i, \\ \pi_{ij} &= \text{p}(Y_{ij} = 1), \\ Y_{ij} &= \text{Zielgröße (Rauchen ja/nein)}, \\ I &= \text{Interventionsgruppe}, \\ T &= \text{Zeit}, \\ IT &= \text{Interventionseffekt}, \end{aligned}$$

$$Y_{ij} \sim \text{Binominal}(1, \pi_{ij})$$

Random Effects Modell für die DHP

(= Generalized Linear Mixed Model)

Verallgemeinerung des GLM für zufällige Effekte (oder: Verallgemeinerung des gemischten linearen Modells auf nicht-stetige Zielgrößen)

Modellgleichung:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 I_{ij} + \beta_2 T_{ij} + \beta_3 IT_{ij} + u_i$$

mit

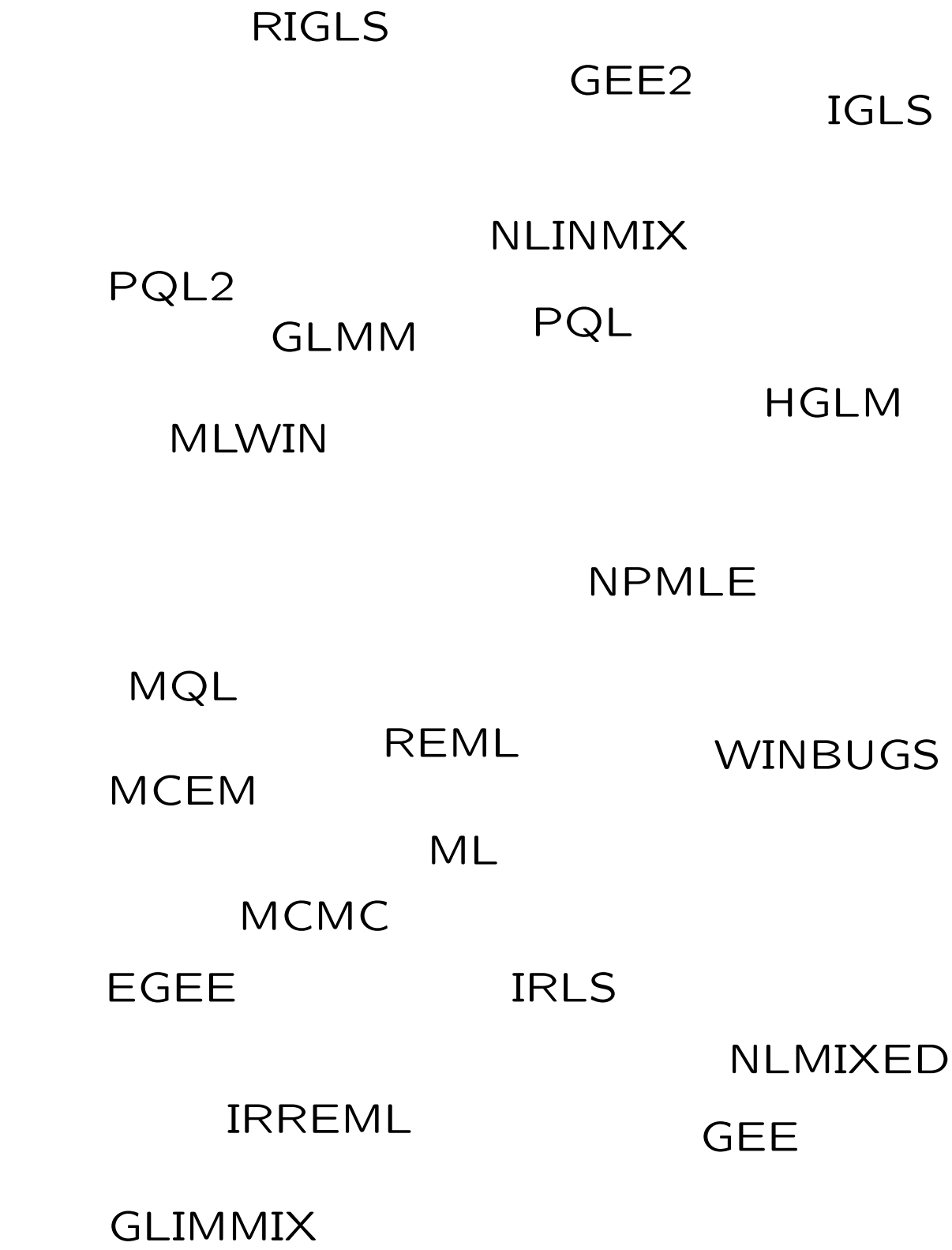
$$\begin{array}{rcl} i & = & 1, \ldots, 14, \\ j & = & 1, \ldots, n_i, \\ \pi_{ij} & = & \text{p}(Y_{ij} = 1 | u_i), \\ Y_{ij} & = & \text{Zielgröße (Rauchen ja/nein)}, \\ I & = & \text{Interventionsgruppe}, \\ T & = & \text{Zeit}, \\ IT & = & \text{Interventionseffekt}, \\ u_i & = & \text{Gemeinde} \end{array}$$

$$\begin{array}{rcl} Y_{ij}|u_i & \sim & \text{Binominal}(1, \pi_{ij}) \\ u_i & \sim & N(0, \sigma^2) \end{array}$$

Daraus resultiert die Likelihood-Funktion

$$L(\beta, \sigma^2; \mathbf{y}) = \prod_{i=1}^{14} \int \left[\prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \right] \times \frac{1}{2\pi\sigma^2} exp \left[-\frac{u_i^2}{2\sigma^2} \right] du_i$$

als Produkt von 14 nicht geschlossen lösbaren Integralen!



(Idee von M. Davidian)

Meta-Regression

- Interpretation der Cluster als einzelne Studien
- Anwendung der Methoden der Meta-Analyse
- hier: Meta-Regression
- Fixed-Effect und Random-Effect-Modelle sind möglich
- Problem (?): ignoriert die Information über die Originalbeobachtungen

GEE

= Generalized Estimation Equation (Liang/Zeger)

- Standardschätzmethode für marginales Modell
- Iterierte Schätzung zwischen Kovariablen und Korrelationsparamter α
- Parameter konsistent auch unter Missspezifikation der Korrelationsmatrix
- Benötigte Fallzahl???
- Erweiterungen möglich (Verbesserte Schätzung der Korrelationsparameter, GEE2, EGEE)

Penalized/Pseudo Quasi-Likelihood

- Maximiert (Taylor-)Approximation an die Likelihoodfunktion (Quasi-Likelihood)
- Ergibt nur approximative ML-Schätzer
- Schätzalgorithmus ist iterierte Schätzung eines gewichteten gemischten Modelles mit Pseudozielgröße
- Variante: MQL (Marginal Quasi-Likelihood), benutzt ungenauere Taylor-Approximation

Numerische Integration

- Approximation der RE-Verteilung durch Gauss-Hermite-Quadratur, wobei der Integrand durch eine gewichtete Summe der Funktionswerte an definierten Stellen approximiert wird.

- Liefert ‘exakte’ ML-Schätzer

NPMLE

- Nichtparametrische Schätzung der RE-Verteilung. Verteilung ist dann diskret und das Modell kann iterativ als Mixture von logistischen Modellen geschätzt werden.

- Keine Annahmen über die Verteilung des zufälligen Effekts

- Es werden keine SE geschätzt

- Der Schätzer für die Verteilung ist nicht-konsistent

MCMC

- Simulation der gemeinsamen Verteilung aller Parameter gegeben die Daten durch Konstruktion einer Markov-Kette.

- Das ist erstaunlich einfach!!

- Stochastische Integration

- Imitation der ML-Schätzung durch nicht-informative Prior-Verteilungen

Conditional Maximum Likelihood

- Maximierung der bedingten Likelihood-Funktion des RE-Modells, gegeben die suffizienten Schätzer der Parameter

- Vorteil: Zufälliger Term wird komplett eliminiert, d.h. auch keine Anforderungen an dessen Verteilung

- Praktisch: Likelihood-Funktion beim Modell mit zufälligem Intercept ist gleich der in einer gematchten Fall-Kontroll-Studie

Ergebnisse

Methode	$\widehat{\beta}_3$	SE($\widehat{\beta}_3$)	$\widehat{\sigma}^2$
Meta-Regression	-0.1246	0.05219	0
GEE	-0.0573*	0.1615	-
PQL	-0.1246	0.05216	0.04769
MQL	-0.1232	0.05180	0.04702
Num. Integration	-0.1246	0.05217	0.03993
NPMLE	-0.1273	0.04964	0.01124**
MCMC	-0.1249	0.05208	0.05756
CML	—	—	—

Software

Meta-Regression	SAS PROC MIXED
GEE	SAS PROC GENMOD
PQL	SAS %GLIMMIX
MQL	SAS %GLIMMIX
Num. Integration	SAS PROC NLMIXED
NPMLE	SAS
MCMC	WinBUGS
CML	SAS PROC PHREG/ PROC LOGISTIC

Siehe auch: How to Use SAS for Logistic Regression with Correlated Data, SUGI27, 2002, P261-27

Fazit

- Es existieren zwei verschiedene Modellklassen für die logistische Regression mit korrelierten Daten

- Unterschiedliche Ergebnisse zwischen den Modellklassen, keine relevanten Unterschiede bei den Schätzungen für das RE Modell

- Wer hätte recht gehabt, wenn es Unterschiede gegeben hätte???

- Software existiert