Mixed Effects Logistic Regression Models in Cluster-Randomized Trials: Relaxing the Assumption of Normality for the Random Effects Distribution

Oliver Kuss

Institute of Medical Epidemiology,
Biostatistics, and Informatics,
University of Halle-Wittenberg

GMDS 2002, Berlin

Contents

Introduction

Many study designs in biostatistics give rise to correlated data: Subjects are followed over time, are repeatedly treated under different experimental conditions, or are observed in clustered units (e.g. communities, clinics, families)

An adequate statistical analysis has to account for this correlation

One possible statistical model in this situations is the mixed model which is available for all kinds of responses

Mixed Effect Logistic Regression

Here: Restriction to random intercept

Model equation:

$$\text{logit}(\pi_{ij}) = X\beta + u_i$$

with

$$\begin{array}{rcl} i & = & 1, \ldots, I \\ j & = & 1, \ldots, n_i, \\ \pi_{ij} & = & \text{p}(Y_{ij} = 1 | u_i), \\ Y_{ij} & = & \text{response}, \\ X & = & \text{Matrix of fixed covariates}, \\ \beta & = & \text{Vector of fixed parameters}, \\ u_i & = & \text{random intercept} \end{array}$$

$$Y_{ij}|u_i \sim \text{Binomial}(1, \pi_{ij})$$

Distribution of the Random Intercept

Standard assumption: $u_i \sim N(0, \sigma^2)$

However, as the u_i are unobservable, this assumption is difficult to check

It can be shown that if the assumption fails to hold, parameter estimates in nonlinear and generalised linear mixed models are biased. (Spiessens et al., Hartford/Davidian)

Several ways to circumvent the problem have been given, we concentrate on two of them, mainly due to computational reasons

Mixture of Normals (Spiessens et al., 2001)

New Assumption:

$$u_i \sim \sum_{g=1}^{G} \pi_g N(\mu_g, \sigma^2)$$

where G is the number of mixture components, π_g is the probability of belonging to component g, $\sum_{g=1}^{G} \pi_g = 1$

Choose optimal G by comparing information criteria (AIC, BIC) of the models

Parameter estimation can be realized with the EM-algorithm

Spiessens et al. supply SAS macro %HETNLMIXED

The EM algorithm supplies estimates $\widehat{\pi}_{ig}$ (probability of the i-th subject to belong to the g-th component). These can be used for clustering the clusters

Nonparametric Estimation of the Distribution (Anderson/Hinde, 1988)

New Assumption:

u_i has a discrete distribution with K mass-points z_k and masses π_k, π_k is the probability of belonging to component k, $\sum_{k=1}^{K} \pi_k = 1$

Choose optimal K by comparing deviances of the models with different K

Parameter estimation can be realized (for fixed K) with the EM-algorithm

It turns out that the EM-algorithm in this case is a iteration of weighted logistic regression models

The estimate of the distribution is not consistent, the distribution is considered a nuisance parameter

Analogously to the MoN case, we get estimates $\widehat{\pi}_{ik}$

The German Cardiovascular Prevention Study

Largest community-based study in Germany (funded 1979-1994)

Main research question: Can cardiovascular risk factors (cholesterol, blood pressure, body mass index, smoking) be reduced by population-based prevention activities?

Included 7 intervention regions and as control a sample from the whole German population. 3 independent cross-sectional surveys

For our analysis the national sample was divided in 7 "virtual" communities (region, community size)

Here: Restriction on 1985 and 1991 survey, Response: smoking

Results I: Intervention Effect

Standard methods (with $u_i \sim N(0, \sigma^2)$

Method	$\widehat{\beta}_3$	SE($\widehat{\beta}_3$)	$\widehat{\sigma}^2$
PQL	-0.1246	0.05216	0.04769
Num. Integration	-0.1246	0.05217	0.03993
MCMC	-0.1249	0.05208	0.05756

New Methods

Method	$\widehat{\beta}_3$	SE($\widehat{\beta}_3$)	$\widehat{\sigma}^2$
MoN (G=2)	-0.1256	0.0907	—-
NPMLE (K=3)	-0.1270	0.0459	—-

Results II: Posterior Probabilities

Mixture of Normals

Cluster	$\widehat{\pi}_1$	$\widehat{\pi}_2$
1	0.99588	0.00412
2	0.99476	0.00524
3	0.02405	0.97595
4	0.99888	0.00112
⋮	⋮	⋮

NPMLE

Cluster	$\widehat{\pi}_1$	$\widehat{\pi}_2$	$\widehat{\pi}_3$
1	0.77746	0.22100	0.00154
2	0.57813	0.41863	0.00324
3	0.00000	0.00463	0.99537
4	0.95598	0.04402	0.00000
⋮	⋮	⋮	⋮

Results III: Meta-Clusters

Mixture of Normals

Meta-Cluster	Cluster
1	1,2,4,5,6,7,8,9,11,12
2	3,10,13,14

NPMLE

Meta-Cluster	Cluster
1	1,2,4,8,9
2	5,6,7,11,12
3	3,10,13,14

Interpretation:

Cluster 1,2,8,9: Urban Areas in Northern Germany (Berlin, Bremen)

Cluster 3,10: Rural Areas in Southern Germany (Traunstein)

Conclusions

- No relevant differences between fixed effects estimates were found
- New methods (MoN, NPMLE) yield data-dependent meta-clusters
- Software (for MoN and NPMLE) exists or can easily programmed

Literature

Anderson, D.A., Hinde, J.P. (1988): Random effects in generalized linear models and the EM algorithm. *Communications in Statistics - Theory and Methods* 17, 3847-56.

Forschungsverbund DHP (editors) (1998): Die Deutsche Herz-Kreislauf-Präventionsstudie: Design und Ergebnisse. Hans Huber, Bern.

Hartford, A., Davidian, M. (2000): Consequences of misspecifying assumptions in nonlinear mixed effects models. *Computational Statistics and Data Analysis* 34, 139-164.

Spiessens, B., Verbeke, G., Komárek, A. (2001): A SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalised linear mixed models. Preprint, http://www.med.kuleuven.ac.be/biostat/research/hetnlmixed.zip.