Longitudinal analysis of ordinal data

A report on the external research project with ULg

Anne-Françoise Donneau, Murielle Mauer

June 30th 2009

Generalized Estimating Equations (Liang and Zeger, 1986)

Analysis of longitudinal data - GEE1

In GLM, score equation writes

$$S(\beta) = \sum_i \frac{\partial \mu_i}{\partial \beta} v_i^{-1}(y_i - \mu_i) = 0 \text{ , with } v_i = Var(Y_i)$$

Multivariate extension, $\boldsymbol{Y_i} = (y_{i1}, ..., y_{in_i})'$

$$S(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1}(\boldsymbol{Y_i} - \boldsymbol{\mu_i}) = 0$$

With $\boldsymbol{\mu_i} = E(\boldsymbol{Y_i})$ and $V_i = diag(v_{i1}, ..., v_{in_i})$

→ Independence estimating equation

Generalized Estimating Equations (Liang and Zeger, 1986)

Analysis of longitudinal data - GEE1

- *Problem* : Correlations among the repeated observations for a given subject
- *Solution*: Allow non-diagonal V_i, $\longrightarrow$ 'working' correlation matrix

Let $R_i(\boldsymbol{\alpha})$ be a $n_i \times n_i$ symmetric matrix which fullfills the condition to be a correlation matrix , and let $\boldsymbol{\alpha}$ be an unknown vector which fully characterizes $R_i(\boldsymbol{\alpha})$.

Generalized Estimating Equations (Liang and Zeger, 1986)

Analysis of longitudinal data - GEE1

- *Problem* : Correlations among the repeated observations for a given subject
- *Solution*: Allow non-diagonal V_i, $\rightarrow$ 'working' correlation matrix

Define,

$$V_i = A_i^{1/2}(\beta) R_i(\alpha) A_i^{1/2}(\beta)$$

in which $A_i(\beta) = diag\{v_{ij}(\mu_{ij}(\beta))\}$

$\rightarrow V_i$ is the 'working' covariance matrix of $\boldsymbol{Y_i}$ and is function of the parameter vectors β and α

Generalized Estimating Equations (Liang and Zeger, 1986)

Analysis of longitudinal data - GEE1

- *Problem* : Correlations among the repeated observations for a given subject
- *Solution*: Allow non-diagonal V_i, $\rightarrow$ 'working' correlation matrix

Using the new definition of V_i, consistent estimates of the model parameters, β, can be obtained from the solution of the estimating equations

$$\sum_{i=1}^{N} D_i' V_i^{-1} (\boldsymbol{Y_i} - \boldsymbol{\mu_i}) = 0 \rightarrow \text{GEE1}$$

GEECAT and GEEGOR (Williamson et al. 1998)

Reference:

JM Williamson, SR Lipsitz, KM Kim. GEECAT and GEEGOR: computer programs for the analysis of correlated categorical response data, Biomedicine 58: 25-34, 1999

GEECAT and GEEGOR are two user-friendly SAS macros for the analysis of clustered, correlated categorical response data

- **GEECAT: for correlated nominal or ordered categorical response data (with independent, exchangeable, banded and unstructured correlation matrices)**
- **GEEGOR: models the association of ordered categorical responses within a cluster using the global odds ratio as a measure of association**

GEECAT and GEEGOR (Williamson et al. 1998)

Issues:

- GEE valid under the MCAR assumption only
- GEECAT: issue with sparse data

Model: $Y_{i,j,k}$ i=subject, t=timepoint, k=treatment group

with $Y_{i,j,k} \in \{1, \ldots, C\}$, C categories

$\pi_{j,k,l}$=probability of category l at timepoint j for group k

$\ln(\pi_{j,k,l}/(1- \pi_{j,k,l}))=\alpha_{j,l}+\beta_{j,l} X$ X=0 for group 1

X=1 for group 2

If $\pi_{j,k,l}$=0 or 1, issue with the parametrization of the model

GEECAT and GEEGOR (Williamson et al. 1998)

Issues:

- **GEECAT: heavy assumptions of proportional odds**

Proportional odds logistic regression for Ordinal data:

$$Y_i \in \{1,2, \ldots, c\}$$

cumulative logit: $\text{logit}[P(Y_i \leq k/x_i)] = \alpha_k + \beta x_{i,}$ $k=1,\ldots,c-1$

The odds for a unit increase in an element of xi are equal to exp(β), irrespective of the cutoff

Study 26951: LDA of QoL

Phase III study
of adjuvant Procarbazine, CCNU and Vincristine chemotherapy
in patients with highly anaplastic oligodendroglioma

Dyspnoea: single item with 4 modalities

9 timepoints, 2 treatment groups

x=time1 time2 time3 time4 time5 time6 time7 time8 tt0 tt1 tt2 tt3 tt4 tt5 tt6 tt7 tt8 (dummies for time and time X treatment interaction)

- **Analysis as continuous variable, using proc nlmixed (valid under MAR)**
- **Analysis as ordinal variable, using GEEGOR (valid under MCAR)**
- **Analysis of rough data (categories with available data at each timepoint in each treatment group)**

Possible model extensions
Weighted GEE or MI

Reference:

Beunckens C, Sotto C, Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. Computational Statistics and Data Analysis 52: 1533 1548, 2008

- Robins et al. (1995) extended GEEs by using inverse probability weights, resulting in Weighted estimating equations (WGEE).

- Alternative developed by Rubin et al (1987) is Multiple imputation (MI).

Focus of the paper: to compare between WGEE and MI-GEE for incomplete data by means of a simulation study.

Possible model extensions
Weighted GEE or MI

WGEE:

GEE-based inferences are valid only under MCAR. If the working correlation structure happens to be correct, the estimates and model-based standard errors are valid under the weaker MAR. In general, the working correlation structure will not be correctly specified, and hence Robins et al. proposed a class of WGEEs to allow for MAR.

The idea is to weight each subject's contribution in the GEEs by the inverse probability that a subject drops out at the time he dropped out. Thus, anyone staying in the study is considered representative of himself as well as a number of similar subjects that did drop out from the study. The incorporation of these weights, reduces possible bias in the regression parameter estimates.

Possible model extensions
Weighted GEE or MI

Weights:

$v_{ij=}=P(D_i=j)$

$=\Pi_{k=2,\ldots,j-1}\,(1-P(R_{ik}=0/R_{i2}=\ldots=R_{i,k-1}=1))\ X\ P(R_{ij}=0/R_{i2}=\ldots=Ri,j-1=1)^{I\{j\leq J\}}$

where $j=2,3,\ldots,J+1$.

Score equations:

$S(\beta)=\sum_{i=1,\ldots,N}\sum_{d=2,\ldots,J+1}(I(D_i=d)/v_{id})\ (\partial\mu_i/\partial\beta')(d)(A_i^{1/2}R_iA_i^{1/2})^{-1}\ (d)(y(d)-\mu_i(d))=0$

Where $y_i(d)$ and $\mu_i(d)$ are the first d-1 elements of Y_i and μ_i.

Possible model extensions
Weighted GEE or MI

MI-GEE:

The key idea is to replace each missing value with a set of M plausible values drawn from the conditional distribution of the unobserved values, given the observed ones. This conditional distribution represents the uncertainty about the right value to impute. M imputed datasets are generated (imputation stage), which are then analyzed using standard complete data methods (analysis stage). Finally the results from the M analyses have to be combined into a single inference (pooling stage).

MI requires the mechanism to be MAR.

Suppose the parameter vector of the distribution of $Y_i=(Y_i^0,Y_i^m)$ is denoted by θ.
If distribution of $Y_i=(Y_i^0,Y_i^m)$ is known, Y_i^m could be imputed by drawing a value of Y_i^m from the conditional distribution $f(y_i^m/y_i^0, \theta)$. The objective is to sample from this true predictive distribution but θ unknown.

Possible model extensions
Weighted GEE or MI

MI-GEE:

- Imputation stage: Procedure MI in SAS

First estimation of θ from the data: θ*

$f(y_i^m/y_i^0, \theta^*)$ used to impute the missing data

Multiple imputation does not attempt to estimate each missing value through Simulated values. Instead, it draws a random sample of the missing values from its distribution.

- Analysis stage: Procedure MIANALYZE in SAS

With M imputations, the estimate of β is

$$\beta^* = 1/M \sum_{t=1,\ldots,M} \beta^{*,t}$$

$$(\beta - \beta^*) \sim N(0,V)$$

where $V=W+(M+1/M)B$

$W = 1/M \sum_{t=1,\ldots,M} U^t$ = average within imputation variance

and $B=1/M-1 \sum_{t=1,\ldots,M} (\beta^{*,t} - \beta^*)(\beta^{*,t} - \beta^*)^T$ = between imputation variance

Possible model extensions
Weighted GEE or MI

A simulation study:

- Asymptotic simulation study to explore the situation of large sample sizes

- Small sample sizes to give insight into the behavior of the methods in real-life Setting

1. Everything correctly specified

2. Dropout and measurement models correct, imputation model incorrect

3. Imputation and measurement models correct, droupout model incorrect

4. Imputation and dropout models correct, measurement model incorrect

Possible model extensions
Weighted GEE or MI

Conclusions (based on simulations only):

-Although asymptotically WGEE exhibits the desirable properties that it theoretically is known to possess, these are barely reproduced for small samples, even when every aspect of the analysis is correctly specified

-Moreover, the observed sensitivity of WGEE to misspecification in either the dropout or measurement model renders these asymptotic properties meaningless

-MI-GEE demonstrates a certain degree of robustness to misspecification in either the imputation or measurement model

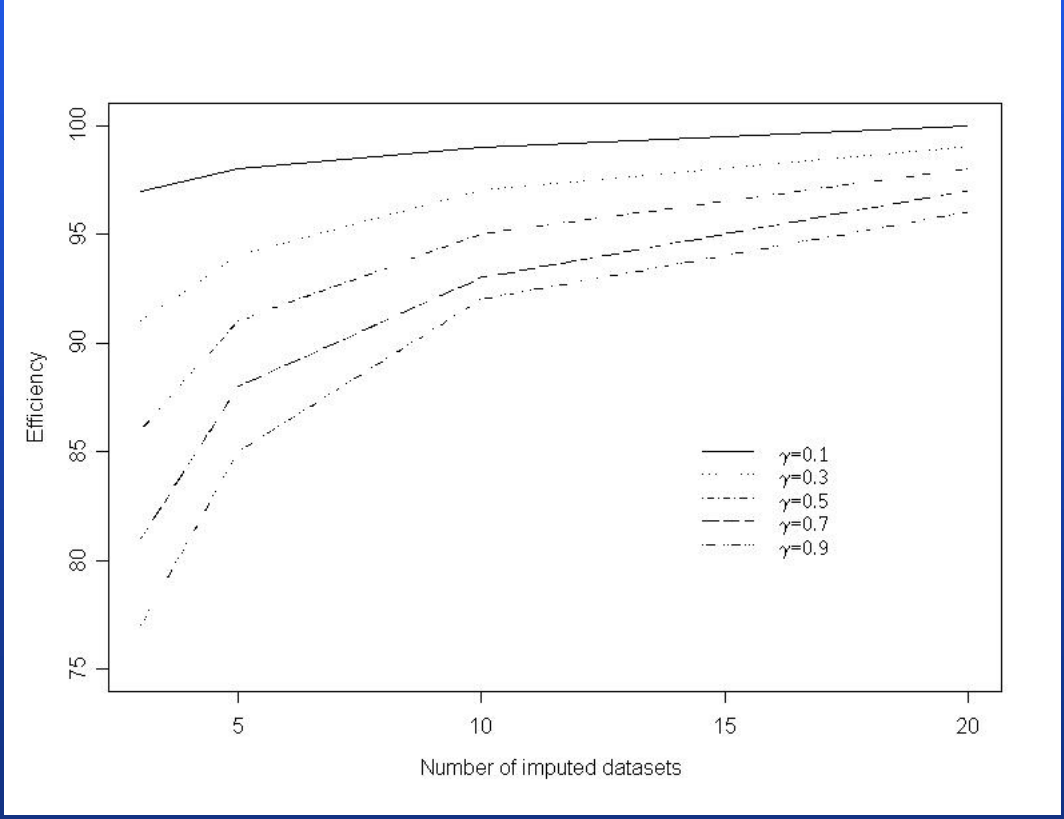
-Moreover, one can do MI under MAR with intermittent missing data

Results in alignment with previous publications

Possible model extensions
Weighted GEE or MI

Efficiency of MI (Rubin):

Good efficiency for M=5 (used in Molenberghs' simulations)



Possible model extensions
Weighted GEE or MI

Proc MI (SAS):

- Possibility to use a logistic regression to impute missing data for categorical variables but only for monotone missing patterns

- For arbitrary missing patterns, Markov Chain Monte Carlo (MCMC) method should be used to impute all missing values or just enough missing values to make the imputed data sets have monotone missing patterns

We should investigate the possibility to have different missingness mechanisms…
Otherwise MCMC with rounded values…

Possible model extensions
Partial proportional odds model

Reference:

- I Carrière and J Boyer. Random-effect models for ordinal responses: Application to self-reported disability among older persons. Rev Epidemiol Santé Publique 54: 61-72, 2006.

Carrière & Boyer: Use of proc nlmixed with the manual specification of likelihood (use of a random effect)

$\text{logit}(P(Y_{ij} \leq c / X_{ij}, u_i)) = \alpha_c + X_{ij}' \beta_c + u_i$

"The use of a random effect, u_i, independent of response category is based on the notion that a unique unknown continuous phenomenon underlies the ordinal response".

+Other models proposed: the adjacent category model, the stereotype mixed model

→ Should more depend on the model specifications than the MI-GEE

Possible model extensions
Partial proportional odds model

Reference:

- B Peterson, FE Harrell. Partial Proportional Odds Models for Ordinal Response variables. Appl. Statistics 39(2), 205-217, 1990.
- Book by Stokes, Davis & Koch (200), 533-541. Partial proportional odds model.

Peterson & Harrell: not for repeated measurements

Proposal: Just dichotomize the ordinal variable

Y_{ijk}=1 if $Y_{ij} \leq k$ and 0 otherwise for k=1,…,c-1

Analyze as multivariate variable for each subject at each timepoint

Define odds1, … oddc-1=indicators to be included as covariates

Example: patient with level 1 for an ordinal variable with 4 modalities

y	odds1	odds2	odds3
1	1	0	0
1	0	1	0
1	0	0	1

Possible model extensions
Partial proportional odds model

Example: patient with level 2 for an ordinal variable with 4 modalities

y	odds1	odds2	odds3
0	1	0	0
1	0	1	0
1	0	0	1

→ Apply MI-GEE

-Issue with sparse data can be avoided by reversing the order of the categories, as this is just a different parametrization of the model

-Can be applied even for nominal variables

Possible model extensions
Partial proportional odds model

Préliminary results for the test of the proportional odds for treatment effect at each time point (without MI):

- DYSPNOEA: TT0 TT5 TT6 TT7
- SLEEP DISORDER: TT3
- APPETITE LOSS: TT4 TT6 TT8
- CONSTIPATION: TT4 TT5 TT6 TT8
- DIHAROEA: TT4 TT5 TT7
- PAIN: TT3

- FINANCIAL PROBLEM: -
- ROLE FUNCTIONING: -
- COGNITIVE FUNCTIONING: -
- SOCIAL FUNCTIONING: -
- NAUSEA AND VOMITING: -

- PHYSICAL FUNCTIONING : Singular Matrix in the contrast
- GLOBAL HEALTH STATUS: Singular Matrix in the contrast
- FATIGUE: Singular Matrix in the contrast
- EMOTIONAL FUNCTIONING= Singular Matrix in the analysis

(difficulties with large number of categories)

TT = interaction time X treatment

Further work…

-To produce all results with MI-GEE and to validate the results

-To compare the method with proc nlmixed

-To publish…

Planned…

-Joint modeling of survival data and longitudinal data

-Competing risks