Modelling Strategies for Longitudinal Data with Missingness.

Ivy Jansen[1] and Geert Molenberghs[1]

[1] Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, B-3590 Diepenbeek, Belgium , ivy.jansen@luc.ac.be

Abstract: A lot of research has been devoted to modelling strategies for longitudinal data with missingness, especially within the MNAR context. In this paper, an overview will be given of several existing methods, and the relatively unexplored domain of non-monotone missingness with multivariate ordinal responses will be broached. In this context, the Multivariate Dale model (Molenberghs and Lesaffre 1994) will be used.

Keywords: Longitudinal data; Missing data; Multivariate Dale model.

1 Introduction

In applied sciences, one is often confronted with the collection of *correlated data* or otherwise hierarchical data. This generic term embraces a multitude of data structures. In particular, studies are often designed to investigate changes in a specific parameter which is measured repeatedly over time in the participating persons. Longitudinal studies are conceived for the investigation of such changes, together with the evolution of relevant covariates.
In longitudinal settings, each unit (respondent, cluster, patient,...) typically has a *vector* $\boldsymbol{Y}$ of responses. This leads to several, generally non-equivalent, extensions of univariate models. In a *marginal model*, marginal distributions are used to describe the outcome vector $\boldsymbol{Y}$, given a set $\boldsymbol{X}$ of predictor variables. The correlation among the components of $\boldsymbol{Y}$ can then be captured either by adopting a fully parametric approach or by means of working assumptions, such as in the semiparametric approach of Liang and Zeger (1986). Alternatively, in a *random-effects model*, the predictor variables $\boldsymbol{X}$ are supplemented with a vector $\boldsymbol{\theta}$ of random effects, conditional upon which the components of $\boldsymbol{Y}$ are usually assumed to be independent. This does not preclude that more elaborate models are possible if residual dependence is detected (Longford 1993). Finally, a *conditional model* describes the distribution of the components of $\boldsymbol{Y}$, conditional on $\boldsymbol{X}$ but also conditional on (a subset of) the other components of $\boldsymbol{Y}$. Well-known members of this class of models are log-linear models (Gilula and Haberman, 1994).

2 Current Practice

The analysis of longitudinal clinical trials is almost invariably hampered by dropout. In current practice methods such as *last observation carried forward* (LOCF) or *complete case* analysis (CC) are very prominent. Such less than optimal methods fall within the *missing completely at random* category (MCAR), where dropout is independent of the measurement process, and part of the literature, supported by the biopharmaceutical industry and the regulatory authorities (FDA in the United States, EMEA in Europe, and their Japanese and other national counterparts), maintains that these methods are to be preferred for reasons of simplicity and validity.
The academic research community, on the other hand, focuses to a large extent on methods for *missing not at random* (MNAR) where dropout is allowed to depend on unobserved measurements. Some researchers believe that ever more complicated MNAR methods will eventually be sufficiently general to encompass the true data generating mechanism.

3 Overview of MNAR Models for Categorical Data

In the MNAR setting, we will make a distinction between models for monotone and non-monotone missingness. The model proposed by Molenberghs, Kenward and Lessafre (1997), which combines a Dale model for the measurements with a logistic regression for dropout (as in the Diggle and Kenward (1994) philosophy), can handle monotone ordinal data. For non-monotone patterns, Baker, Rosenberger, and DerSimonian (1992) proposed a model for bivariate binary data subject to non-random nonresponse, which is reformulated by Jansen et al. (2003) using 2 loglinear models, such that its membership of the selection model family is unambiguously clear, to accommodate for, possibly continuous, covariates, turning the model into a regression tool for several categorical outcomes, and to avoid the risk of invalid solutions. A disadvantage of those BRD models, is that the parameters cannot be interpreted marginally, which is actually what clinicians want.
As we can see, until now there does not exist a model that allows for non-monotone missingness with more than 2 possible outcomes. A solution will be presented in the next section.

4 A Method for Non-monotone Categorical Outcomes

Since the multivariate Dale model (Molenberghs and Lesaffre 1999), which extends the bivariate global cross-ratio model described by Dale (1986), accounts for the dependence between multiple ordinal responses, as well

as their dependence on covariate vector(s), which may be time-varying, continuous and/or discrete, this model is very useful for our purpose.
The model arises from a decomposition of the joint probabilities into main effects (described by marginal probabilities) and interactions (described by cross-ratios of second and higher orders).
This model will be used for the measurements $\boldsymbol{Y}$ and for the missingness given the measurements $\boldsymbol{R}|\boldsymbol{Y}$, such that again a selection model is obtained and both discrete and continuous covariates can be included in the measurement model as well as in the missingness model.
Results will be presented for simulated data, and for a data set from a multicenter, postmarketing study involving 315 patients that were treated by fluvoxamine for psychiatric symptoms described as possibly resulting from a dysregulation of serotonine in the brain.

5 Need for a Sensitivity Analysis

The route of a sensitivity analysis has been explored many times in the context of categorical data. For the model by Baker, Rosenberger and DerSimonian (1992), Molenberghs, Kenward and Goetghebeur (2001) developed the intervals of ignorance and uncertainty. To the reformulated model by Jansen et al. (2003) a local influence is applied by the same authors. This local influence is also applied to the Dale model with dropout (Molenberghs, Kenward and Lessafre, 1997) by Van Steen et al. (2001). Future work will be devoted to a sensitivity analysis on the model for non-monotone categorical outcomes, that was introduced in the previous section.

References

Baker, S.G. (1995). Marginal regression for repeated binary data with outcomes subject to non-ignorable non-response. *Biometrics*, **51**, 1042–1052.

Baker, S.G., Rosenberger, W.F., and DerSimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**, 643–657.

Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Diggle, P.D. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Gilula, Z. and Haberman, S. (1994). Conditional log-liear models for analyzing categorical panel data. *Journal of the American Statistical Association*, **89**, 645–656.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 409–418.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Longford, N.T. (1993). *Random Coefficient Models*. London: Oxford University Press.

Molenberghs, G., Kenward, M.G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics*, **50**, 15–29.

Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, **84**, 33–44.

Molenberghs, G., and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18**, 2237–2255.

Van Steen, K., Molenberghs, G., Verbeke, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling:* , **1**, 125–142.