Generalized Estimating Equations Notes on the Choice of the Working Correlation Matrix

Editorial of "Generalized Estimating Equations: Notes on the Choice of the Working Correlation Matrix (A. Ziegler, M. Vens)"

Geert Molenberghs[1]

[1] Universiteit Hasselt, Diepenbeek, Belgium and Katholieke Universiteit Leuven, Leuven, Belgium

Formatted: Dutch (Belgium)

Formatted: Dutch (Belgium)

Editorial

It is a privilege to be able to peruse the fine article by Ziegler and Vens, as well as the contributions made by nine discussants.

Over the last 25 years, generalized estimating equations (GEE) have seen an ever further spreading use. Nonetheless, it is an technique confronted with confusion and, at times, misunderstanding. The user must carefully read the technique's manual. Let us highlight a few important principles.

First, GEE is a method of estimation rather than a model or a modeling family. That said, it is virtually always applied in the context of marginal models, even though Zeger, Liang, and Albert (1988) applied it to generalized linear mixed models.

Second, comparisons between marginal and hierarchical models (usually termed random-effects models) need to be done with caution. Each family serves its own purpose and precisely which of the two is chosen should depend predominantly on the research question, although pragmatic, computational considerations are perfectly legitimate, too. In this sense, GEE should not be seen as a "downward biased version" of GLMM. Such a view would violate this and the previous principle.

Third, and related, there is the delicate relationship between GEE and fully specified marginal models. The latter allow for full likelihood or fully Bayesian estimation methodology, a choice one may wish to make should higher-order moments (including variances and correlations) be of interest, perhaps next to marginal mean parameters. When GEE is chosen, this is often because only the first moment is of scientific relevance, upon which then the second moment (correlation) is considered a nuisance. That said, the correlation structure does deserve attention, which is where the strong contribution of Ziegler and Vens lies. While one can misspecify the correlation structure and still reach consistency, important issues remain. Indeed, one can question, along with Chaganty and Sabo, whether asymptotic theory applies when the correlation is misspecified. Results by Molenberghs and Kenward (2010) are encouraging in this respect; they show that a full distribution can always be constructed, as soon as marginal mean and correlation are compatible. Such compatibility is not straightforward, as restated by Chaganty and Sabo, and the resulting model may be contrived. But the crucial fact is

that there does exist such a model, thence alleviating somewhat the concern, raised by Chaganty and Sabo, about the lack of a solid asymptotic theory. In this respect, the fact that the marginal mean structure has implications for the correlation parameter space, beyond positive definiteness, is a crucial observation pointed out by Chaganty and Sabo. One should not loose sight of the fact that GEE builds on generalized linear models (GLM), with their well-known mean-variance relationship. GLM are, after all, non-linear models, in spite of the linear predictors they carry, the implications of which are at the same time non-trivial and profound.

But there are instances where more than one moment would be of interest. It is then natural to model two or even more moments, such as in GEE2 (second-order generalized estimating equations) or pseudo-likelihood. These methods are reviewed in Molenberghs and Verbeke (2005). This point is touched upon by Lechner. It is only when higher moments are explicitly modeled that they can be made part of formal inferences.

Fourth, matters become even more complicated when data are incomplete, a situation eloquently described by Daniel and Kenward. It is important to note in this respect that, while missingness completely at random (MCAR) is a sufficient condition to ensure validity of conventional GEE, that this condition is not necessary. Daniel and Kenward describe key situations where GEE is still valid under missingness at random (MAR). Additionally, there are a number of extensions of and modifications to GEE, such as weighted GEE (Molenberghs *et al* 2010) and multiple-imputation GEE (Beunckens, Sotto, and Molenberghs 2008) that further broaden the scope of the methodology. For reviews, see Molenberghs and Kenward (2007).

Fifth, GEE is related to other developments in statistics, as brought forward by Breitung, Lechner, and Zorn. These discussants bring out the connection with the GMM and CMM techniques. The latter techniques emerged at the econometrics side, whereas GEE finds its roots in biostatistics, medical statistics, and epidemiology. Bringing out such cross-field links is doubly beneficial: (a) Perceiving links between a method and other ones allows for a better understanding of a technique's underpinnings. Here, the context is estimating functions based on moments. One gets to see the broader link with semi-parametric theory as well. Also, pitfalls are seen more clearly, such as sometimes distressingly poor small-sample properties; (b) Drawing out links across fields provides opportunities to further the spread of a method from one area of application to the next.

Let us return to the main focus of the paper, the working correlation structure. Martus and Wang, inspired by the Ziegler and Vens, point to similarities, but also to differences with model selection. Martus and Wang also refer to the need to keep an eye on the principal inferential focus, which is often in terms of the marginal mean parameters, such as a treatment effect. Thus indeed, the correlation structure's choice should be in view of stability, efficiency, and unbiasedness of such estimators. Together with the main authors and Zorn, I should like to add that biological and/or substantive considerations, jointly with aspects of the statistical design, ought to be brought into the picture when selecting a working correlation structure. The sensitivity analysis recommendation of Ziegler and Vens nicely fits in with this.

A very noteworthy feature is the special status of the independence working correlation structure, leading to independence estimating equations (IEE). Ziegler and Vens, as well as various discussants, do a fine job in bringing out this message clearly and convincingly. The statements are in line with early claims made in the literature, as alluded to by the discussants. Also in the context of missing data, does the independence

choice prove to be beneficial (Fitzmaurice, Molenberghs, and Lipsitz 1995). At the same time, problems regarding the parameter space are less dramatic with IEE. After all, the independence case reduces GEE for binary data to a simple but important "correlation-corrected logistic regression." Ziegler and Vens do a fine job in elucidating when IEE are as efficient as GEE. Generally, their messages are wise, providing a scientifically sound yet pragmatically useful road map. For example, all users will benefit from their recommendation: "In general, a reasonable non-independence working correlation structure should be chosen. Intensive modeling of the working correlation structure may provide only negligible additional gains in efficiency."

I should like to conclude by seconding Zorn's words: "Ziegler and Vens have done a substantial service for applied researchers who use generalized estimating equations (GEE) models in their work. The question of the optimal specification of the working correlation matrix is one that has vexed analysts for more than two decades. By outlining the relevant theoretical findings on this question, and providing a clear set of guidelines and criteria, the authors have contributed significantly to the usability of GEE models in clinical and epidemiological (as well as other scientific) settings."

References

Beunckens, C., Sotto, C., and Molenberghs, G. (2008) 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.

Fitzmaurice, G.M., Molenberghs, G., and Lipsitz, S.R. (1995) Regression models for longitudinal binary responses with informative dropouts. *Journal of the Royal Statistical Society, Series B*, **57**, 691-704.

Molenberghs, G. and Kenward, M.G. (2007) *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons.

Molenberghs, G. and Kenward, M.G. (2010) Semi-parametric marginal models for hierarchical data and their corresponding full models. *Computational Statistics and Data Analysis*, **54**, 585-597.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G., Kenward, M.G., Verbeke, G., and Teshome Ayele, B. (2010). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, **00**, 000-000.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049-1060.