Document information

- *Title:* Generalized estimating equations in longitudinal data analysis: a review and recent developments
- *Author(s):* Ming Wang
- *DOI:* `http://dx.doi.org/10.1155/2014/303728`
- *File name:* `wang_2014.pdf`

Introduction

This sections discusses the basic development of generalized estimating equations (GEES) then mentions some limitations to GEEs and proposed methodological advances. The paper will not focus on the new methods, but on model selection, power analysis, and the issue of informative cluster size, and recent developments for GEEs.

Method

Notation and GEE

The notation for GEEs is defined here along with the solutions to common working correlation matrices (Table 1), scale parameter that is based on Person residuals, and $cov(\mathbf{Y}_i) = \hat{r}_i \hat{r}_i^\top$, where $\hat{r}_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$. Also, remember if $\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$ is properly specified, then the covariance matrix of $\hat{\boldsymbol{\beta}}$ based on the sandwhich estimator reduces to model-based variance estimator.

Model selection of GEE

There are several reasons why model selection of GEE models is important and necessary:

1. GEE has gained increasing attention in biomedical studies which may include a large group of predictors. Therefore, variable selection is necessary for determining which are included in the final regression model.
2. One feature of GEE is that the consistency of parameter estimates can still hold even when the "working" correlation structure can definitely enhance the efficiency of the parameter estimates in particular when the samplize is not large enough. Therefore, how to select intrasubject correlation matrix plays a vital role in GEE with improved finite-sample performance.
3. Additionally, the variance function $v(\mu)$ is another potiental factor affecting the goodness-of-fit of GEE. Correctly specified variance function can assist in the selection of covariates and an appropriate correlation structure.

Selection techniques for GEEs

Rotnizky and Jewll proposed

$$RJ(R) = \sqrt{(1 - RJ_1)^2 + (1 - RJ_2)^2},$$

based on $\mathbf{\Gamma} = \left(\sum_{i=1}^K \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \hat{\mathbf{M}}_{LZ}$, where - $\hat{\mathbf{M}}_{LZ} = \sum_{i=1}^K \mathbf{D}_i^\top \mathbf{V}_i^\top cov(\mathbf{Y}_i) \mathbf{V}_i^\top \mathbf{D}_i$,

- $RJ_1 = tr(\mathbf{\Gamma})/p$, and
- $RJ_2 = tr(\mathbf{\Gamma}^2)/p$.

If the "working" correlation structure $\mathbf{R}$ is correctly specified then RJ_1 and RJ_2 will be close to 1, leading to $RJ(R)$ approaching 0.

Shults and Chaganty proposed a criterion based on the minimization of the generalized error sum of squares (ESS), where

$$\begin{aligned} ESS(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{i=1}^{K} (\mathbf{Y}_i - \mathbf{u}_i)^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{u}_i) \\ &= \sum_{i=1}^{K} Z_i^\top (\boldsymbol{\beta}) \mathbf{R}_i^{-1} (\boldsymbol{\alpha}) Z_i(\boldsymbol{\beta}), \end{aligned}$$

where $Z_i(\boldsymbol{\beta}) = \mathbf{A}^{1/2}(\mathbf{Y}_i - \mathbf{u}_i)$. The criterion is defined by

$$SC = \frac{ESS(\boldsymbol{\alpha}, \boldsymbol{\beta})}{N - p - q},$$

where N is the total number of observations, p is the number of regression parameters, and q is the number of correlation coefficients.

- An extension of SC was proposed by Carey and Wang, where the Gaussian psedudolikelihood (GP) was adpoted and a better "working" correlation structure yields a larger GP. They showed that GP criterion performed better than RJ via simulation.

Pan proposed a modified Akaike information criterion (AIC) in adaption to GEE based on the quasi-likelihood under independence called QIC, which is defined as

$$QIC(R) = -2\Psi\left(\widehat{\boldsymbol{\beta}}; I\right) + 2tr\left(\widehat{\Omega}_I \widehat{V}_{LZ}\right),$$

where Ψ is the quasi-likelihood function. For more information on the properties of QIC, which allows for simultaneous selection of the mean and "working" correlation structures, see `pan_2001a.pdf`. There are several modifications to QIC that are presented below.

- Hardin and Hilbe (2003) made a slight modification on QIC by using $\hat{\boldsymbol{\beta}}$ and ϕ assuming independence "working" correlation structure for more stability. This criterion does not perform well distinguishing the independence and exchangeable "working" correlation structures because, in certain cases, the same regression parameters estimates can be obtained under these two structures.
- Using the penalty term, $tr\left(\widehat{\Omega}_I \widehat{V}_{LZ}\right)$ from QIC, Hin and Wang proposed the correlation informaiton criterion (CIC). They showed that CIC outperforms QIC when the outcomes were binary. However, one limitation of this criterion is that it cannot penalized the overparameterization; thus the performance isnot well in comparison with two correlation structures having quite different numbers of correlation parameters.
- Another criterion is the extended QIC (EQIC) proposed by Wang and Hin by using the extended quasi-likelihood (EQL) defined by Nelder and Pregibon based on the deviance function. However, the authors indicate that the mean structure were first selected with QIC, and the varianace function could be identified as the one minimizing EQIC given the selected covariates; then "working" correlation structure selection could be achieved based on CIC.

Besides those criteria mentioned above, there are more proposed criteion for GEEs, but no "best" selection criterion has been identified. Wang et al can be followed up as the rule of thumb.

Simulation

Results from the simulations can be seen in Tables 2, 3, and 4, which show that no model selection method is the "best" overall.

Future direction and discussion

Potiential future directions related to my dissertation listed are

- A robust and optimal model selection criterion of GEE under missing at random (MAR) or missing not at random (MNAR).