The Effect of Missing Visits on GEE, a Simulation Study

Julia Geronimi, Gilbert Saporta

The Effect of Missing Visits on GEE, a Simulation Study

Julia Geronimi[1,2] and Gilbert Saporta[2]

[1] Institut de de Recherches Internationales SERVIER, 50 rue Carnot 92150 Suresnes (E-mail: geronimi.julia@gmail.com)
[2] Cedric-Cnam, 292 rue Saint Martin 75141 Paris Cedex 03 (E-mail: gilbert.saporta@cnam.fr)

Abstract. Clinical research is often interested in longitudinal follow-up over several visits. All scheduled visits are not carried out and it is not unusual to have a different number of visits by patient. The Generalized Estimating Equations can handle continuous or discrete autocorrelated response. The method allows a different number of visits by patients. The GEE are robust to missing completely at random data, but when the last visits are fewer, the estimator may be biased. We propose a simulation study to investigate the impact of missing visits on the estimators of the model parameters under different missing data patterns. Different types of responses are studied with an exchangeable or autoregressive of order one structure. The number of subjects affected by the missing data and the number of visits removed, vary in order to assess the impact of the missing data. Our simulations show that the estimators obtained by GEE are resistant to a certain rate of missing data. The results are homogeneous regardless to the imposed missing data structure.
Keywords: Longitudinal data, repeated correlated data, correlation, missing data, simulations, Generalized Estimating Equations.

1 Introduction

Clinical follow-up provides information on changing pattern of diseases. This allows for biological measurements and clinical criterion observation over several visits. Therefore, it is possible to study the link between several potential biological covariates and a clinical response on repeated measurements.

However, observations from the same patient cannot be handled as independent and the correlation among visits must be taken into account. Two of the most common methods which are able to deal with longitudinal data are the Generalized Linear Mixed Model, GLMM as describe by McCulloch [6] and the Generalized Estimating Equations, GEE from Liang and Zeger [5].

GLMM are a subject specific method which introduces a random effect per patient to take into account the longitudinal aspect of observations. Unfortunately, the integration over these random effects distribution may be numerically untractable. GEE are a population specific method which consider the

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

intra-subject correlations by imposing a correlation structure to the response. Advantage of the GEE method is that only correct specification of marginal means is needed for having a consistent and asymptotically normal parameter estimator. We will use this method in this paper. For a discussion on GEE, GLMM and relation between marginal and mixed effect models, reader can refer to the work of Park [9], Heagerty and Zeger[3] and Nelder and Lee[7].

Studies'design provides for a number of visits per patient which is regrettably not always complied. In the case of intermittent missing data this results in blank lines in observation matrix. No classical parametric imputation shall be performed since no information is collected at this date. Moreover the interpolation of these values is difficult because there are often few widely spaced visits which means the prediction is blurred.

Missing data, as defined by Rubin [15], are divided into three categories :

- Missing Completely at Random, like a visit randomly deleted by loss record
- Missing At Random, as a missed visit linked to the length of the study
- Missing Not At Random, such as non presence of a patient related to the latent seriousness of his condition

The GEE estimator is robust to the first case but biased in the other two as explained by Liang and Zeger[5] and Robins *et al.*[13]. In case of dropouts Robins *et al.*[13] introduced an inverse probability of censoring weighted GEE which have been studied by Preisser *et al.*[10]. They proposed a modified version of GEE in which observations or person-visits have weights inversely proportional to their probability of being observed, which is unfortunately not suitable here.

Within this context questions may arise :

- How much the GEE estimator is robust to missing visits?
- Which bias should we consider in case of MAR data?

We provide a simulation study to measure the impact of different missing data patterns on GEE estimators. Second part of this paper gives the GEE approach outline. Simulations plan and their results are shown in section 3 and 4. The paper ends by a conclusion in section 5.

2 Generalized Estimating Equation

When the population-average effect is of interest, the marginal model is commonly used to analyzing longitudinal data. Liang and Zeger[5] proposed the Generalized Estimating Equations to estimate the regression parameter, by only specifying the marginal distribution of the outcome variables in the marginal model. Both continuous and binary responses can be modeled.

Let y_{it}, of expectation μ_{it}, be the response of interest for the subject i at the visit t for $i \in \{1, .., K\}$ and $t \in \{1, .., n_i\}$. Each subject has a set of p measured covariates at each time t denoted x_{it}. For a known function $V(.)$ and a given mean-link function $g(.)$ we have :

$$\mathbb{V}ar(y_{it}) = \phi V(\mu_{it}) \tag{1}$$

$$g(\mu_{it}) = x_{it}^t \beta \tag{2}$$

β is the regression parameter to be estimated, ϕ is the dispersion parameter. We will note Y_i, the $n_i \times 1$ independent response vector and X_i, the $n_i \times p$ measured covariates matrix for subject i. Generalized Estimating Equations are defined by :

$$U(\beta) = \sum_{i=1}^{K} D_i^t V_i^{-1} (Y_i - \mu_i) = 0 \tag{3}$$

D_i is the matrix of partial derivatives with $\partial \mu_{it} / \partial \beta_k$ as its (t, k)-th element. V_i is the working covariance matrix defined by :

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2} \tag{4}$$

where $R_i(\alpha)$ is a working correlation matrix completely described by the parameter vector α of size $s \times 1$. A_i is the diagonal matrix with elements equal to the variance terms $V(\mu_{it})$. If $R_i(\alpha)$ is the true correlation matrix of Y_i then V_i is the true covariance matrix.

Liand and Zeger[5] propose an iterative estimation method. A consistent method (as the moments method) is used to estimate the couple (α, ϕ) for fixed values of $\hat{\beta}$. Then equation (3) is used to estimate $\hat{\beta}$ for fixed values of $(\hat{\alpha}, \hat{\phi})$. This leads to a consistent estimate of β.

The choice of $R_i(\alpha)$ is important. Classic structures are independent, exchangeable or auto-regressive of order 1. Selection criterion for the choice of the working correlation matrix are useful. We quote here just a few : the Quasi-log-likelihood under the independence model Information Criteria from Pan [8], the Correlation Information Criteria from Hin and Wang[4] and Rotnitzky-Jewell's criterion[14]. In order to simplify, we will suppose the working correlation known and of exchangeable or auto-regressive of order one structure.

3 Simulations plan/structure

Two types of responses are studied, a continuous and a binary outcome. Both cases introduce 4 covariates which have been simulated by a Gaussian distribution with an auto-regressive of order one with parameter $\rho = 0.3$. We denote Σ this correlation structure.

3.1 Gaussian response

The response Y_i is a multivariate normal vector with intra-subject correlation structure $R_i(\alpha)$ following the model :

$$Y_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{5}$$

where $x^l \sim \mathcal{N}(0, \Sigma)$ for $l \in \{2, ..., 5\}$. The error vector $\boldsymbol{\epsilon}_i$ is a multivariate normal vector with mean zero and variance matrix $\sigma^2 R_i(\alpha)$. The mean parameter vector is imposed equal to $\beta = (1, 0.5, -0.2, 1, -1)$, where the first component is the intercept. The variance parameter σ^2 is chosen for having a signal/noise ratio of 0.5 as described by Fu[1].

$$\frac{V(x_{it}^t\beta)}{\sigma^2} = \frac{1}{2} \Leftrightarrow \sigma^2 = 2\sum_{l=2}^{5}\beta_l^2 = 4.58 \tag{6}$$

3.2 Binary response

To simulate a binary response, the *logit* link is used and an intra-subject correlation structure equal to $R_i(\alpha)$ is imposed thanks to Qaquish[11].

$$logit(\mathbb{E}(y_{it})) = x_{it}^t\beta \tag{7}$$

where $x^l \sim \mathcal{N}(0, \Sigma)$ for $l \in \{2, ..., 5\}$. The mean parameter vector is imposed equal to $\beta = (1, 0.5, -0.2, 0.3, -0.4)$. The first component is the intercept.

For both kinds of data, the parameters vary as follows according to a full factorial design.

- K, the number of subjects on $\mathcal{K} = \{50, 100, 200, 300\}$
- n, the number of scheduled visits on $\mathcal{N} = \{4, 6, 9\}$
- $R_i(\alpha)$, the correlation structure is either exchangeable or auto-regressive of order one (both admit a scalar $\alpha \to s = 1$)
- α, the unique parameter of correlation on $\mathcal{A} = \{0.1, 0.3, 0.5, 0.6\}$

We simulated 1000 samples that we will called *completed* for each of these 96 scenarios. All of the subjects in these samples get the same number of visits. In order to evaluate the effect of missing visits on the GEE estimators we simulated 1000 other samples that we will called *uncompleted* ou *unbalanced* where we deleted some of the visits on some subjects. The percentage of concerned subjects varies according to $\mathcal{P} = \{10\%, 20\%, 30\%, 50\%\}$ and the number of deleted visits varies according to $\mathcal{V} = \{1, 2, 3\}$.

With the aim of evaluating how robust the GEE estimator is in MCAR and MAR situations, we imposed two different schemes of visits removal. First, we consider a scheme where visits follow a uniform distribution. In that case we can speak of MCAR data. In a second time we consider a probability of

deletion that will increase with the follow-up (i.e. with the number of visits). Last case imposed MAR data. We will talk about uniform unbalanced and increasing unbalanced respectively. All computations are performed using R [12] and GEE fitting performed by the package `geepack` of Halekoh *et al.*[2].

4 Results

A useful criterion for assessing the goodness of an estimator $\hat{\boldsymbol{\theta}}$ is the Absolute Relative Bias defined by $ARB(\hat{\boldsymbol{\theta}}) = \frac{||\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}||}{||\boldsymbol{\theta}||}$. We estimate this criterion by :

$$\widehat{ARB}(\hat{\boldsymbol{\theta}}) = \frac{1}{1000} \sum_{b=1}^{1000} \frac{||\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}||}{||\boldsymbol{\theta}||} \tag{8}$$

where $||.||$ is the euclidean norm which boils down to the absolute value when the parameter is a scalar. $\hat{\boldsymbol{\theta}}_b$ is the estimate of $\boldsymbol{\theta}$ on the b-th sample. The mean of the absolute relative gap between the estimator and its target is thus estimated on 1000 samples.

4.1 Continuous response results

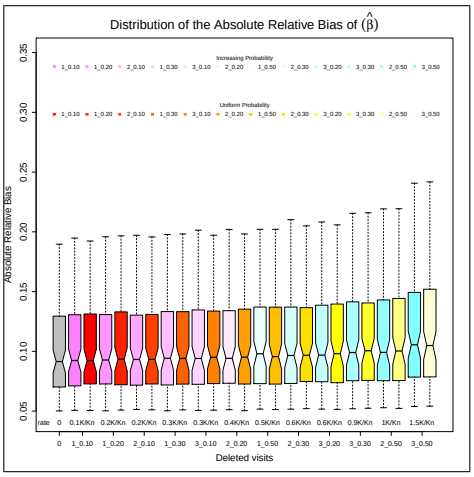

Fig. 1. $\hat{\beta}$ ARB evolution by missing rate for 96 models with a continuous response

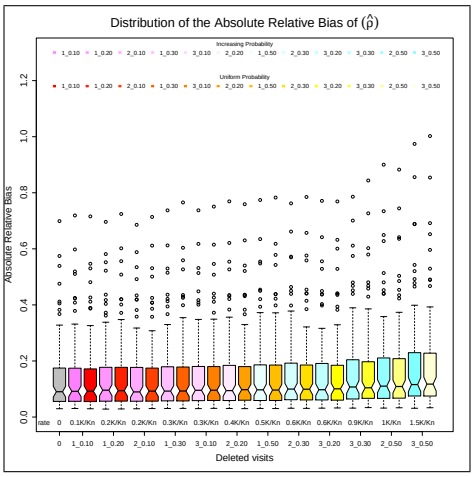

Fig. 2. $\hat{\rho}$ ARB evolution by missing rate for 96 models with a continuous response

Figures 1 and 2 show the distribution of the ARB for the GEE estimator of the parameter β and ρ on the 96 tested models for a continuous response. These graphs compare the two deletion schemes : uniform and increasing.
The boxplots show no differences between the two deletion schemes. Precisely, the difference is between $[-0.005, 0.005]$ for the Absolute Relative Bias of $\hat{\beta}$ and between $[-0.06, 0.06]$ for the ABR of $\hat{\rho}$.

The ARB slightly increases with the missing rate. The median ARB switches from 0.091 to 0.101 for $\hat{\beta}$ and from 0.09 to 0.117 for $\hat{\rho}$.
More precisely, graphics 3, 4 and 5 present the evolution of the Absolute Relative Bias for $\hat{\beta}$ in the case $K = 100$ and $n \in \{4, 6, 9\}$ with increasing unbalanced scheme.

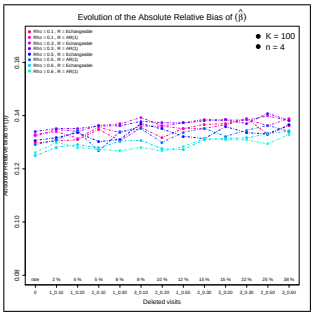

Fig. 3. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 4$ for a continuous response

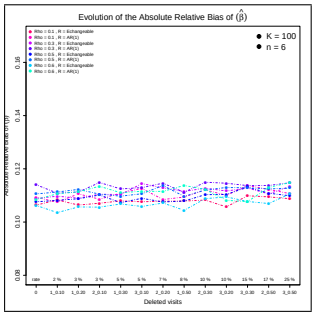

Fig. 4. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 6$ for a continuous response

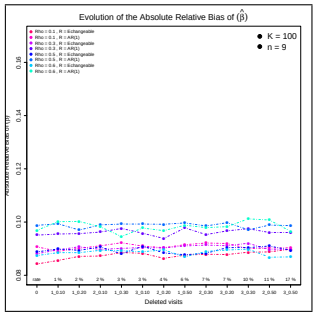

Fig. 5. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 9$ for a continuous response

4.2 Binary response results

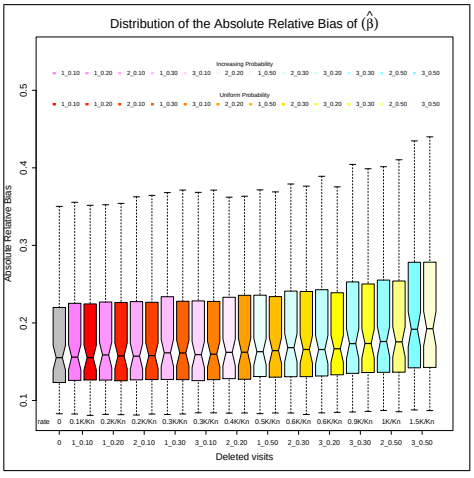

Fig. 6. $\hat{\beta}$ ARB evolution by missing rate for 96 models with a binary response

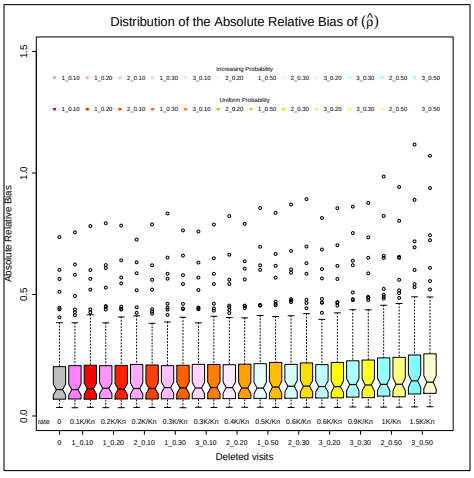

Fig. 7. $\hat{\rho}$ ARB evolution by missing rate for 96 models with a binary response

Graphs 6 and 7 show the distribution of the ARB for the GEE estimator of the parameter β and ρ on the 96 tested models for a binary response. These

graphs compare the two deletion schemes : uniform and increasing.
There are no differences between the two deletion schemes. Some differences in the range of $[-0.005, 0.005]$ and $[-0.015, 0.015]$ have been noted for the Absolute Relative Bias of $\hat{\beta}$ and $\hat{\rho}$ respectively.

The small increase of the ARB is more important for a binary response whith a median ARB switching from 0.155 to 0.193 for $\hat{\beta}$ and from 0.101 to 0.131 for $\hat{\rho}$. Graphs 8, 9 and 10 give more details about the evolution of the Absolute Relative Bias.

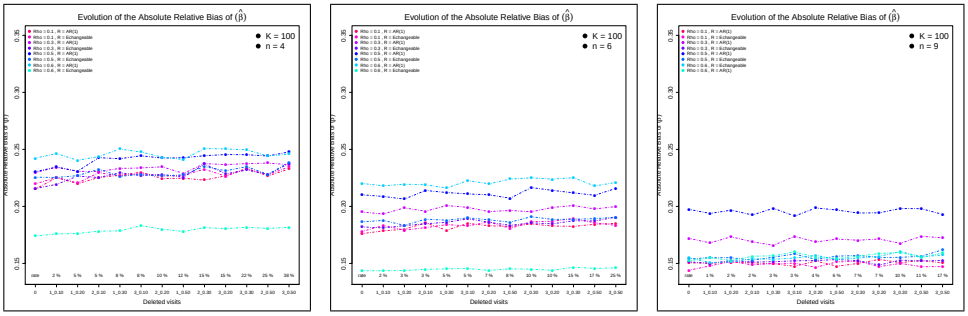

Fig. 8. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 4$ for a binary response

Fig. 9. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 6$ for a binary response

Fig. 10. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 9$ for a binary response

Results on binary response show higher Absolute Relative Bias meaning worst results. Such results were expected since it is more complicated to have an accurate estimator with a binary outcome. Nevertheless both responses, binary and continuous, show the same evolution according to the rate of missing visits. Moreover, both responses point the same lack of differences between uniform unbalanced and increasing unbalanced structure.
Figures 3, 4, 5, 8, 9 and 10 demonstrate how small the increase is with the rate of missing data. The decrease with the number of scheduled visits was expected since it means a lower rate and better estimations.

5 Conclusion

Our simulations show two important issues. First of all, the evolution of the absolute relative bias is similar regardless of the imposed missing data structure. This means that no differences have been highlighted between both schemes. Secondly, the absolute relative bias increases slowly with the missing rate, which means that our imposed missing rate does not disrupt the efficacy of GEE estimator.
We may infer that GEE estimators can be used in studies where MCAR and MAR data are present. Bias induced by MAR is negligible. However, users should pay attention to the missing data scheme and rates used here.

Since it is very complicated to prove the presence of MNAR data, this missing structure has not been studied here. Nevertheless, a complementary study with this type of missing data could bring some more information about expected bias.

References

1. W. Fu. Penalized estimating equations. *Biometrics*, 59:126–132, 2003.
2. U. Halekoh, S. Hojsgaard, and J. Yan. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006.
3. P. Heagerty and S. Zeger. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26, 2000.
4. L.-Y. Hin and Y.-G. Wang. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28(4):642–658, 2009.
5. K.-Y. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 38:13–22, 1986.
6. C. McCulloch and J. Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.
7. J. Nelder and Y. Lee. Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238, 2004.
8. W. Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125, 2001.
9. T. Park. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, 12(18):1723–1732, 1993.
10. J. Preisser, K. Lohman, and P. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21(20):3035–3054, 2002.
11. F. Qaqish. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90 (2):455–463, 2003.
12. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.
13. J. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.
14. A. Rotnitzky and N. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77 (3):485–497, 1990.
15. D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.