Analysing missing data in longitudinal binary outcomes using generalized linear mixed-effects model with Markov correlation structure

M. Salomé Cabral & M. Helena Gonçalves

CEAUL, DEIO, Faculdade de Ciências da Universidade de Lisboa, Portugal

CEAUL, FCT, Universidade do Algarve, Portugal

Introduction

Longitudinal binary data are routinely collected in medical studies in which repeated observations of response variable are taken over time on each individual in one or more groups of treatments. A common problem in these studies is the presence of missing data since it is difficult to have complete records of all subjects for a variety of reasons. The generalized linear mixed-effects model (GLMM) approach is widely used to analyse longitudinal binary data. However in GLMM it is assumed that the observations of the same subject are independent conditional to the random effects and covariates which may be not true. In the `R` package `bild` [2] the methodology implemented overcomes this problem using a generalized linear mixed effects model with binary Markov chain (GLM3C) as the basic stochastic mechanism to accommodate serial dependence and odds-ratio to measure dependence between successive observations. In GLM3C as well as in GLMM approaches missing values on the response are allowed provided they are MAR. A simulation study was performed to give a statistical assessment of GLM3C when compared with the GLMM approach.

Model

Let y_{it} the binary response value at time t ($t = 1, \ldots, T_i$) from subject i ($i = 1, \ldots, n$), and Y_{it} its generating random variable whose mean value is $\Pr(Y_{it} = 1) = \theta_{it}$. The equation of the GLMM for binary responses with random intercept assumes the form

$$\text{logit}[E(Y_{it}|b_i)] = \text{logit}(\theta^b_{it}) = x^\top_{it}\beta + b_i, \qquad (i = 1, \ldots, n)$$

where x_{it} is a set of p covariates associate to each observation and each subject, β is the $p \times 1$ vector of unknown parameters and $b_i \sim N(0, \sigma^2)$ are assumed to be sampled independently from each other. The serial dependence between successive observations of the he same subject is of Markovian type.

- First order dependence structure (MC1)

$$\psi_1 = OR(Y_t, Y_{t-1}) = \frac{\Pr(Y_{t-1} = Y_t = 1)\ \Pr(Y_{t-1} = Y_t = 0)}{\Pr(Y_{t-1} = 0,\ Y_t = 1)\ \Pr(Y_{t-1} = 1,\ Y_t = 0)} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

  where $p_j = \Pr(Y_t = 1|Y_{t-1} = j), \quad j = 0, 1$ are the transition probabilities.

- Second order dependence structure (MC2)

$$OR(Y_{t-1}, Y_{t-2}) = \psi_1 = OR(Y_{t-1}, Y_t)$$
$$OR(Y_{t-2}, Y_t|Y_{t-1} = 0) = \psi_2 = OR(Y_{t-2}, Y_t|Y_{t-1} = 1)$$

  with transition probabilities given by

$$p_{hj} = \Pr(Y_t = 1|Y_{t-2} = h, Y_{t-1} = j), \quad h, j = 0, 1.$$

Serial dependence is regulated by $\lambda = (\lambda_1, \lambda_2) = (\log\psi_1, \log\psi_2)$ for MC2 models and by λ_1 ($\lambda_2 = 0$) for MC1 models. Missing values are allowed on the response, provided they are MAR with some restrictions:

- If MC1 dependence model is considered and if there is a missing value at time point $t-1$, it is required that there are observations at time points $t-2$ and t.
- If MC2 dependence model is considered and if there is a missing value at time point $t-2$, it is required that there are observations at time points $t-4, t-3, t-1$ and t.

Simulation

A simulation study was carried out to GLM3C and GLMM approaches.

$$\Pr(Y_{it} = 1|t) = \frac{\exp(\beta_0 + b_i + \beta_1 t + \beta_2 x_i + \beta_3(x_i \times t))}{1 + \exp(\beta_0 + b_i + \beta_1 t + \beta_2 x_i + \beta_3(x_i \times t))}$$

where $x_i = 0$ for half the population and 1 for the remainder, $\beta_0 = -1$, $\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = 1$ and $b_i \sim N(0, \sigma^2)$, with $\sigma^2 = 0.5$.

Several designs were considered:

- Length of the profile $T = 13$. Number of subjects $n = 20$ (small) or $n = 50$ (large).
  1. MC1 serial dependence, $\lambda_1 = 1, 2$.
  2. MC2 serial dependence $(\lambda_1, \lambda_2) = (1, 1), (2, 2)$.

Let $\mathbf{R}_i = 1$ be a $T \times 1$ vector of indicator variables for the ith subject, where $R_{it} = 1$ if Y_{it} is observed, and $R_{it} = 0$ if Y_{it} is missing. An intermittent missing-data mechanism MAR was generated assuming that the binary response on the first occasion is always observed, $R_{i1} = 1$ and for $t > 1$, R_{it} is generated with probability of success given by $(1-\phi)^{1-y_{it-1}}$, where ϕ is the nonresponse parameter dependence with $\phi = 0, 0.25, 0.5$ ($\phi = 0$, complete data).

The whole estimation procedure was repeated for 1000 runs and several characteristics were computed such as CI (coverage probabilities of nominal 95% confidence intervals) and RE (relative efficiency of the estimators). RE> 1 means GLM3C estimator is preferred.

The `bild` function in the `R` package `bild` [2] was used when GLM3C approach was considered. The `glmer` function in the `R` package `lme4` [1] was used when the GLMM approach was considered.

The results of our simulation are summarized to MC1 on Figures 1 and 3 when $\lambda_1 = 1$ and to MC2 on Figures 2 and 3 when $(\lambda_1, \lambda_2) = (2, 2)$ for the time effect (β_1) and group-time interaction effect (β_3), the most interest effects in a longitudinal study.

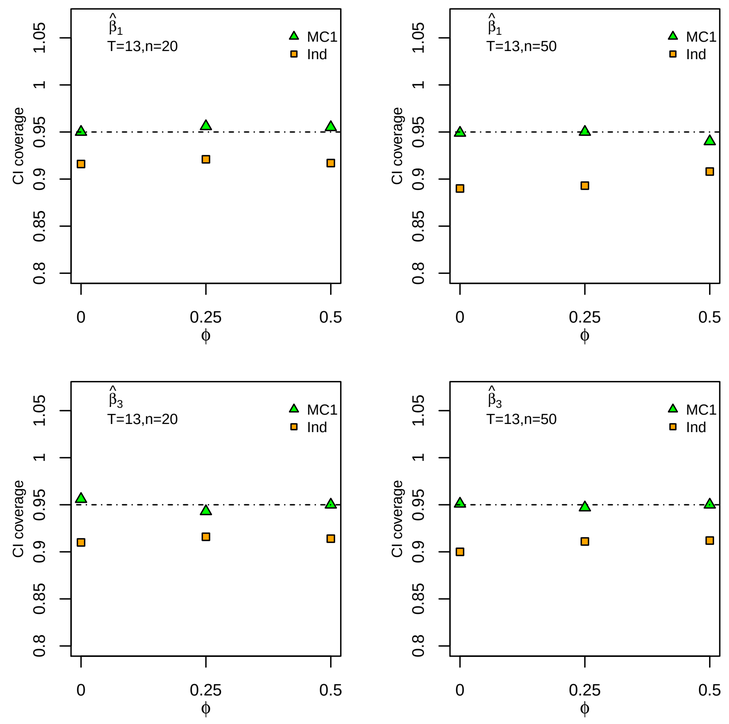

Figure 1: Coverage probabilities of nominal 95% confidence intervals (CI coverage) for β_1 ($\hat{\beta}_1$) and β_3 ($\hat{\beta}_3$) when $\lambda_1 = 1$. Coding for estimation procedures: MC1 (GLM3C) and Ind (GLMM).

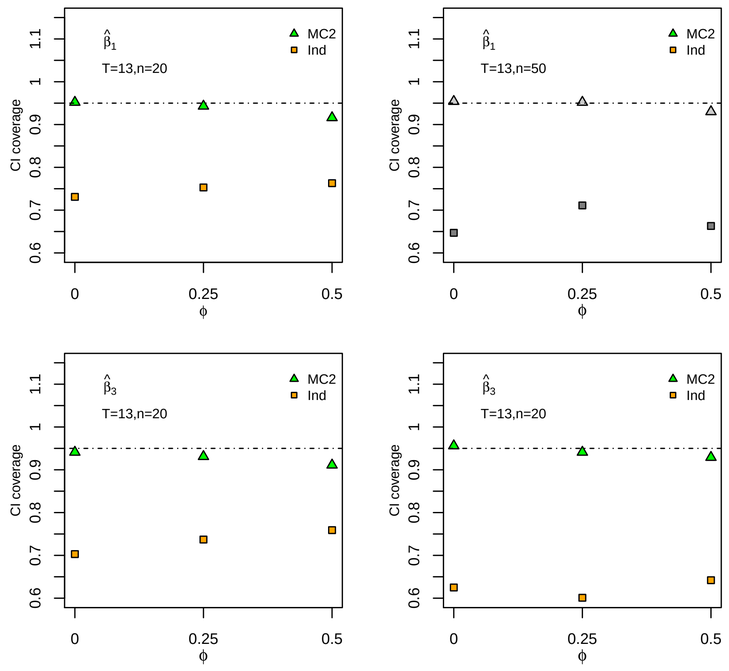

Figure 2: Coverage probabilities of nominal 95% confidence intervals (CI coverage) for β_1 ($\hat{\beta}_1$) and β_3 ($\hat{\beta}_3$) when $(\lambda_1, \lambda_2) = (2, 2)$. Coding for estimation procedures: MC2 (GLM3C) and Ind (GLMM).

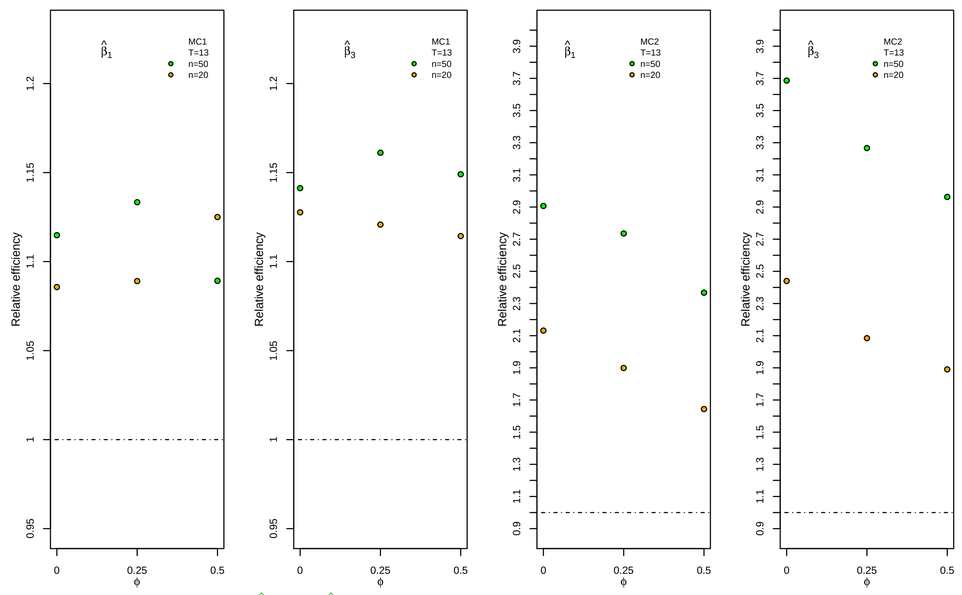

Figure 3: Relative efficiency of $\hat{\beta}_1$ and $\hat{\beta}_3$. Serial dependence MC1 regulated by $\lambda_1 = 1$. Serial dependence MC2 regulated by $(\lambda_1, \lambda_2) = (2, 2)$.

Conclusion

For both MC1 and MC2 serial dependences the impact of intermittent missingness status on the estimation of β_1 and β_3 is greater to GLMM than to GLM3C approach. To all the situations considered the GLM3C approach seems to be preferable to the GLMM since that gives coverage probabilities closer to nominal and more efficient estimators.

References

[1] D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2013. R package version 1.1-3.

[2] M. H. Gonçalves, M. S. Cabral, and A. Azzalini. *bild: A package for BInary Longitudinal Data*, 2015. version 1.1-5.

Acknowledgements

This work was partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal-FCT, through the project UID/MAT/00006/2013).