L'effet de visites manquantes sur l'estimateur des GEE, une étude par simulation

Julia Geronimi, Gilbert Saporta

L'effet de visites manquantes sur l'estimateur des GEE, une étude par simulation

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

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

Résumé. La recherche clinique s'intéresse régulièrement au suivi longitudinal du patient au cours de plusieurs visites. Toutes les visites prévues ne sont pas effectuées et il n'est pas rare d'avoir un nombre de visites différent selon les individus. Les Generalized Estimating Equations permettent d'étudier une réponse continue ou discrète autocorrélée. Cette méthode permet un nombre de visites qui diffère selon les patients. Les GEE sont robustes aux données manquantes complètement aléatoires. Cependant dans le cas où les visites de fin d'étude sont moins nombreuses, l'estimateur peut être biaisé. Nous proposons une étude par simulation pour étudier l'impact de visites non effectuées sur les estimateurs obtenus par GEE sous divers schéma de données manquantes. Deux types de réponses sont étudiées avec une structure échangeable ou auto-régressive d'ordre un. Le nombre de sujets touchés et le nombre de visites supprimées varient afin d'évaluer leur impact. Nos simulations montrent que les estimateurs calculés par GEE sont résistants jusqu'à un certain taux de données manquantes. Les résultats sont homogènes quelle que soit la structure de données manquantes imposée.

Mots-clés. Données longitudinales, données répétées corrélées, autocorrélation, données manquantes, simulations, Generalized Estimating Equations

Abstract. Clinical research is regularly 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. However 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 GEE estimators under different missing data pattern. 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 their impact. 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

Le suivi clinique de patients permet de récolter des information sur l'évolution des pathologies et donne ainsi la possibilité de mettre en relation un critère clinique avec certains paramètres biologiques. Dans ce contexte, les observations d'un même patient ne peuvent être considérées comme indépendantes et la corrélation entre les observations d'un même sujet doit être prise en compte. Les Generalized Estimating Equations de Liang and Zeger (1986) sont une méthode marginale, spécifique à la population. Les GEE prennent en compte la corrélation intra-sujet en imposant la même structure de corrélation à l'ensemble des patients. Nous utiliserons cette méthode par la suite.

Le design des études prévoit un certain nombre de visites par patient qui n'est malheureusement pas toujours respecté. Il est possible que des échantillons ne soient pas récoltés de façon aléatoire ou qu'un patient soit trop malade pour venir à une visite. Ce dernier schéma implique que la donnée manquante est informative. Ces absences ne peuvent être imputées par un modèle paramétrique puisqu'aucune des informations du patient ne sera récoltée à cette date. Une interpolation de la valeur à la date fixée est envisageable mais le design implique souvent peu de visites très espacées dans le temps.

Les données manquantes, comme définies par Rubin (1976), sont divisées en 3 catégories. Les données Missing Completely at Random, comme une visite supprimée aléatoirement par perte de dossier, les données Missing At Random comme une visite non effectuée car l'étude est trop longue, et les données Missing Not At Random comme la non présence d'un patient en raison de la gravité de son état. L'estimateur par GEE est robuste au premier cas et biaisé dans les deux autres Liang and Zeger (1986); Robins et al. (1995); Robins and Rotnitzky (1995). Dans le cas de perte de suivi Robins et al. (1995); Robins and Rotnitzky (1995) ont mis en place une version pondérée des GEE.

Deux questions se posent alors, *à quel point l'estimateur des GEE est-il robuste aux visites manquantes? Quel biais doit on envisager en cas de données MAR?* Nous proposons une étude par simulation afin d'évaluer l'effet de certains types de données manquantes sur les estimateurs obtenus par GEE.

La deuxième partie présente quelques rappels sur la méthode des GEE. Les plans de simulations et les résultats sont détaillés en partie 3 et 4.

2 GEE

Considérons une étude longitudinale dont la variable d'intérêt notée y_{it} représente la variable réponse, discrète ou continue, pour l'individu i à la visite t pour $i \in \{1, .., K\}$ et $t \in \{1, .., n_i\}$. Pour chaque individu un ensemble de p covariables est mesuré à chaque temps t noté x_{it}. Nous noterons alors Y_i, de taille $n_i \times 1$, le vecteur de réponses pour l'individu i et X_i, de taille $n_i \times p$, la matrice des covariables mesurées pour l'individu i. Nous noterons μ_{it} l'espérance de y_{it} conditionnellement à x_{it} et $v(y_{it}) = V(\mu_{it})$, la variance de y_{it}, pour une fonction $V(.)$ donnée. Pour une fonction de lien $g(.)$ choisie l'espérance

s'écrit $\mu_{it} = \mathbb{E}(y_{it}|x_{it}) = g(x_{it}^t\beta)$. β représente le vecteur de paramètres à estimer. Les GEE utilisent une matrice de corrélation de travail $R_i(\alpha)$ ce qui induit une matrice de variance covariance de travail définie par :

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

où α est un vecteur de paramètre qui définit la structure de corrélation commune aux individus et A_i est une matrice diagonale composée des variances $V(\mu_{it})$. Pour $R_i(\alpha)$ donnée l'estimateur des GEE est solution de :

$$U(\beta) = \sum_{i=1}^{K} D_i^t V_i^{-1}(Y_i - \mu_i) = 0 \tag{2}$$

D_i est la matrice des dérivées partielles dont le (t,k)-ème élément est $\partial\mu_{it}/\partial\beta_k$. Il est alors possible d'estimer, par une méthode consistante, le vecteur de paramètres α en utilisant l'estimateur $\hat{\beta}$. Liang and Zeger (1986) proposent ainsi une méthode d'estimation itérative jusqu'à convergence où $\hat{\alpha}$ est obtenu par la méthode des moments. Le choix de la structure de $R_i(\alpha)$ est important. Les structures classiques sont de type indépendante, échangeable ou autorégressive d'ordre 1. Il existe des critères similaires à l'AIC Pan (2001); Hin and Wang (2009) permettant de sélectionner une matrice de corrélation de travail. Pour plus de clarté, nous supposerons la structure de corrélation connue, en imposant soit une structure échangeable, soit une structure autorégressive d'ordre 1.

3 Plan des simulations

Deux types de variables réponses ont été étudiés, une continue gaussienne et une discrète binaire. Dans les deux cas, 4 covariables ont été simulées selon une loi normale centrée réduite admettant pour structure de corrélation une autorégressive d'ordre 1 de coefficient $\rho = 0.3$ notée Σ.

Nous avons simulé une variable réponse Y_i continue, gaussienne, admettant pour structure de corrélation $R_i(\alpha)$ selon le modèle $Y_i = X_i\beta + \epsilon_i$, où la variable $x^l \sim \mathcal{N}(0,\Sigma)$ pour $l \in \{2,...,5\}$. Le vecteur ϵ_i est simulé selon une loi normale centrée de variance σ^2 et de matrice de corrélation $R_i(\alpha)$ grâce à la décomposition de Choleski. Le vecteur de paramètre est imposé égal à $\beta = (1, 0.5, -0.2, 1, -1)$, la première composante correspondant à l'ordonnée à l'origine. Le paramètre de variance σ^2 est choisi pour avoir un rapport signal/bruit $\frac{V(x_{it}^t)}{\sigma^2}$, égal à 0.5 comme utilisé par Fu (2003). Des rapports égaux à 0.7 et 1.4 ont été testés.

Le deuxième jeu de données utilise le lien *logit* pour simuler une variable réponse binaire tout en imposant la structure de corrélation $R_i(\alpha)$ à l'aide la méthode de Qaqish (2003). La réponse y_{it} est modélisée par le modèle $logit(\mathbb{E}(y_{it})) = x_{it}^t\beta$ où $x^l \sim \mathcal{N}(0,\Sigma)$ pour $l \in \{2,...,5\}$. Le vecteur de paramètre est donné par $\beta = (1, 0.5, -0.2, 0.3, -0.4)$.

La première composante correspond à l'ordonnée à l'origine.

Pour ces deux types de jeux de données nous avons fait varier plusieurs paramètres :

- K, le nombre de sujets sur $\mathcal{K} = \{50, 100, 200, 300\}$
- n, le nombre de visites sur $\mathcal{N} = \{4, 6, 9\}$
- $R_i(\alpha)$, la structure de corrélation, soit échangeable, soit autorégressive d'ordre 1
- α, l'unique paramètre de corrélation sur $\mathcal{A} = \{0.1, 0.3, 0.5, 0.6\}$

Pour chacun de ces 96 scénarios, 288 pour une réponse continue, nous avons simulé 1000 jeux de données que nous dirons *complets*. Pour tester l'effet de visites manquantes sur les estimateurs des paramètres, nous avons simulé 1000 autres jeux de données que nous dirons *incomplets* ou *déséquilibrés* en supprimant chez quelques individus certaines visites. Nous avons fait varier le pourcentage, 10%, 20%, 30% ou 50%, d'individus qui manqueraient 1, 2 ou 3 visites.

Afin de tester la résistance des estimateurs aux données MCAR et MAR nous avons imposé deux types de schémas de suppression de visites. Dans un premier temps, les visites sont choisies selon une loi uniforme sur l'ensemble des visites possibles ce qui implique des données MCAR. Dans un second temps, la probabilité est croissante en fonction du temps imposant ainsi des données MAR. Nous parlerons alors de déséquilibre uniforme et croissant. Tous les calculs ont été réalisés à l'aide du logiciel R Development Core Team (2008) et du package geepack de Hojsgaard et al. (2006).

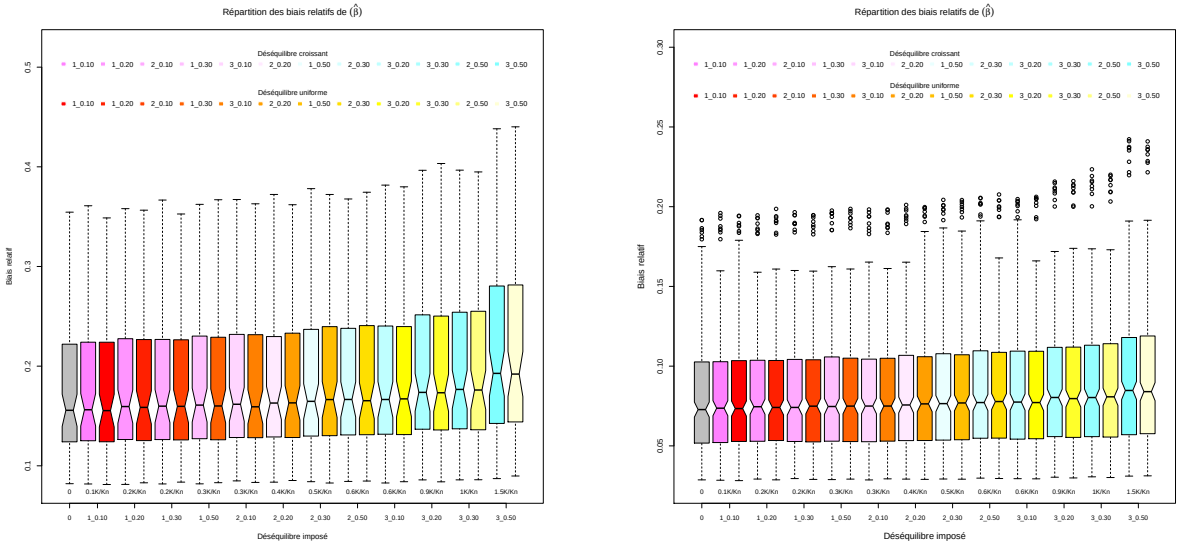
4 Résultats

Un critère utile pour mesurer la précision d'un estimateur $\hat{\boldsymbol{\theta}}$ est le biais relatif absolu définit par $\frac{||\mathbb{E}(\hat{\boldsymbol{\theta}})-\boldsymbol{\theta}||}{||\boldsymbol{\theta}||}$ que nous pouvons estimer sur 1000 échantillons indépendant par :

$$BR(\hat{\boldsymbol{\theta}}) = \frac{1}{1000}\sum_{b=1}^{1000}\frac{||\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}||}{||\boldsymbol{\theta}||} \tag{3}$$

$||.||$ représente la norme euclidienne et $\hat{\boldsymbol{\theta}}$ est le paramètre estimé sur le b-ème échantillon. Ce critère mesure la moyenne de l'écart relatif absolu entre l'estimateur et sa cible sur 1000 échantillons.

Le graphique (1) représente un boxplot des biais relatifs absolus de l'estimateur $\hat{\beta}$ en fonction du déséquilibre imposé. Chaque colonne représente la répartition du biais relatif absolu sur les différents modèles testés, 96 pour une réponse binaire, 288 pour une



(a) Variable réponse binaire

(b) Variable réponse continue

Figure 1: Comparaison de l'évolution du biais relatif de $\hat{\beta}$ en fonction du taux de données manquantes pour deux types données manquantes et deux types de variables réponse.

réponse continue. Ces deux graphiques mettent en parallèle les résultats dans le cas d'un déséquilibre uniforme et croissant.

Les résultats montrent que le biais relatif augmente faiblement avec le taux de données manquantes passant d'un biais relatif absolu médian de 15.6% à 19.2% dans le cas d'une réponse binaire et de 7.3% à 8.4% dans le cas d'une réponse continue. On remarque un *décrochement* pour une réponse binaire lorsque l'on supprime 3 visites chez 50% des patients. Dans l'ensemble les résultats sont similaires pour les deux types de déséquilibre avec de très faibles différences. L'estimateur obtenu par GEE est assez robuste aux taux de données manquantes que nous avons imposé. Les biais relatifs étant comparables entre les deux types de donnés manquantes, l'estimateur est robuste à notre schéma de données MAR.

5 Conclusion

Nos études par simulation montrent que l'estimateur obtenu par GEE admet un biais relatif constant jusqu'à un certain taux de données manquantes. De plus, cet estimateur est robuste à notre schéma de suppression de visites. Cet estimateur peut donc être utilisé pour des études où le taux de données manquantes reste raisonnable. Le cas où la donnée manquante n'est pas aléatoire n'est pas ici étudié. Une étude complémentaire avec ce type de schéma pourrait être envisagée.

Bibliographie

Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, 59:126–132.

Hin, L.-Y. and Wang, Y.-G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in medicine*, 28(4):642–658.

Hojsgaard, U. H. S., , and Yan, J. (2006). The r package geepack for generalized estiamtin equations. *Journal of Statistical Software*, 15(2).

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 38:13–22.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125.

Qaqish, F. B. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2):455–463.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.