Estimation Doublement Robuste de l'Effet Marginal d'un Traitement pour les Essais Randomisés en Cluster.

Melanie Prague [1] & Rui Wang[1,2] & Victor De gruttola[1]

[1] *Harvard T.H. Chan School of Public Health, Biostatistics Department, 677 Huntington Avenue, Boston MA 02115, USA*
[2] *Division of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital, Boston MA 02115, U.S.A.*

Résumé. Les méthodes semi-paramétrique basées sur les équations d'estimation généralisées (GEE) sont largement utilisées pour analyser des données corrélées. Dans cette communication, nous présentons un estimateur doublement robust (DR) de l'effet marginal du traitement pour les essais randomisés en cluster (CRTs). Cet estimateur prend en compte la structure complexe de corrélation entre les individus, les données manquantes informatives et les phénomènes de covariable d'interférence. Les phénomènes de covariable d'interférence apparaissent lorsque la variable résultat d'un individus ne dépends pas uniquement de ses propres covariables mais aussi des covariables des autres individus dans le même cluster. L'estimateur DR combine des méthodes d'augmentation (AUG) pour ajuster sur le déséquilibre des covariables au temps de base et des méthodes de pondération par probabilités inverses (IPW) pour prendre en compte les données manquantes. Nous proposons un package R (*CRTgeeDR*) implémentant cette méthode. De manière intéressante, nous démontrons que ce package est supérieur aux autres packages disponibles sur le CRAN pour IPW pour les CRTs. Il permet une estimation non biaisés quel que soit la structure de correlation de travail choisie. Cela vient du fait que les autres packages, initialement développés pour des données longitudinales, adoptent un implémentation inadéquate des poids dans l'equation d'estimation. L'estimateur DR améliore l'efficacité de l'estimation comparé à l'IPW et simplifie les besoins de modélisation. En effet, intéractions et covariables d'interference peuvent être ignorées sous certaines conditions. Nous démontrons ces résultats par simulations et en utilisant les données d'un CRT sur l'amélioration des conditions sanitaires dans des pays en voie de développement.

Mots-clés. Augmentation, CRTgeeDR, Données corrélées, Données manquantes, Doublement robuste, Effet marginal, Equation d'estimation généralisées, Essai randomisé en cluster, Manquante au hasard (MAR), Ponderation par probabilité inverse (IPW), R.

Abstract. Semi-parametric approaches based on generalized estimating equations (GEE) are widely used to analyse correlated outcomes. In this communication, we present a doubly robust estimator (DR) of the marginal treatment effect in cluster randomized trials (CRTs), which account for complex correlation structure of individuals, informative missing data and covariate interference. Covariance interference arises when the outcome of an index subject does not only depend on this index subject's covariates but also on covariates of other subjects in the same cluster. The DR estimator combines augmentation (AUG) approaches to deal with imbalance in baseline covariates and inverse-probability weighting (IPW) to account for missing data. We present the R package *CRTgeeDR* which implements the method. Interestingly, we demonstrate that for IPW in CRTs, this package is superior to existing packages on the CRAN because it provided unbiased results whatever the correlation structure used. This is because other packages, initially developed for longitudinal data, adopt a misleading implementation of weights in the estimating equations. The DR improves efficiency compared to IPW and simplifies the modeling. Actually, treatment-covariates interactions and interfering covariates can be ignored under some conditions. We demonstrate this results in simulations and using data from a sanitation CRT in developing countries (Guiteras et al. 2015, Science).

Keywords. Augmentation, Cluster randomized trial, Correlated data, CRTgeeDR, Doubly Robust, Generalized Estimating Equation, inverse probability weighting (IPW), MAR, marginal effect, missing data, R.

1 Introduction and background

In clustered randomized clinical trials (CRTs), the unit of treatment assignment is a cluster of subjects. In such settings, outcomes are likely to be correlated among subjects within the same cluster. Often used for estimation, generalized estimating equations (GEE) based on semi-parametric methods (Zeger and Liang, 1986) target marginal effects of treatment. In CRTs, covariates may be fully observed even if the outcome is missing. If the model for the missingness mechanism represents the MAR data generating process, the IPW estimation provides Consistent and Asymptotically Normal (CAN) estimators of treatment effects by reweighing complete cases according to the probability of being observed (Robins et al., 1995). Recent methodological developments improve estimation efficiency by leveraging baseline covariates for example with augmentation approaches(Robins et al., 1994; Zhang et al., 2008). Augmentation had been extended to CRTs (Stephens et al., 2012). We have developed a method that combines the IPW and the AUG that is doubly robust (DR) (Prague et al., 2015), which implied that the resulting estimator is CAN if either the outcome model or missing data model are correctly specified – that is, they reflect the true data generation processes.

2 The DR estimator

2.1 Notations

We consider a study design in which a vector of P baseline covariates $\boldsymbol{X}_{ij} = (X_{ij}^1, \ldots, X_{ij}^P)$ and outcome Y_{ij} are recorded for each subject $j = 1, \ldots, n_i$ in cluster $i = 1, \ldots, M$. The sample size within each cluster is assumed fixed by design and non-informative. Our setting compares two arms (treated $A_i = 1$ and control $A_i = 0$); the probability of treatment assignment is known and given by $p = P(A_i = 1)$; extension to a greater number of treatments is straightforward but complicates the notation. The vector $\boldsymbol{R}_i = [R_{ij}]_{j=1,\ldots,n_i}$ is the indicator of missingness; Y_{ij} is observed when $R_{ij} = 1$. The matrix of covariates $\boldsymbol{X}_i = [\boldsymbol{X}_{ij}]_{j=1,\ldots,n_i}$ is assumed to be fully observed and consists only of pre-exposure covariates measured at baseline. Interest lies in estimating the marginal effect of the treatment given by $M_E^* = E(E(Y_{ij}|A_i = 1, \boldsymbol{X}_i) - E(Y_{ij}|A_i = 0, \boldsymbol{X}_i))$. For estimating M_E^*, we make inference about the parameters $\boldsymbol{\beta} = (\beta_0, \beta_A)^T$ indexing the marginal model $g(\mu_{ij}(\boldsymbol{\beta}, A_i)) = g(E(Y_{ij}|A_i)) = \beta_0 + \beta_A A_i$, where $\boldsymbol{\mu}_i(\boldsymbol{\beta}, A_i) = [\mu_{ij}(\boldsymbol{\beta}, A_i)]_{j=1,\ldots,n_i}$ and g is a one-to-one link function.

2.2 Assumptions

The methods is designed to analyze data collected in cluster randomized trials (CRTs) where 1) observations within a cluster may be correlated, 2) observations in separate clusters are independent, 3) a monotone transformation of expectation of the outcome is linearly related to the explanatory variables, 4) the variance is a function of the expectation, and 5) treatment is randomized at a cluster level. Regarding missing data, we make a stronger assumption than MAR that we refer to as restricted MAR (rMAR): the probability that the outcome for one individual is missing is independent of all outcomes in the cluster, conditional on baseline exposure A_i and cluster characteristics $\boldsymbol{X}_i$. The conditional probability that the outcome is observed is denoted $\pi_{ij}(\boldsymbol{X}_i, A_i) = P(R_{ij} = 1|\boldsymbol{X}_i, A_i)$ and is called the propensity score (PS). When data are rMAR, ignoring missing data leads to biased inference if missingness depends both on $\boldsymbol{X}_i$ and A_i.

2.3 Estimation

The Doubly Robust estimator is given by:

$$\begin{aligned} 0 &= \sum_{i=1}^{M} \Bigg[\boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}, A_i) \right) \\ &\qquad + \sum_{a=0,1} p^a (1-p)^{1-a} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \Big(\boldsymbol{B}_i(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B) - \boldsymbol{\mu}_i(\boldsymbol{\beta}, A_i = a) \Big) \Bigg], \end{aligned}$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta}, A_i)}{\partial \boldsymbol{\beta}^T}$ is the design matrix. The matrix $\boldsymbol{V}_i$ is the covariance matrix equal to $\phi \boldsymbol{U}_i^{1/2} \boldsymbol{C}(\boldsymbol{\alpha}) \boldsymbol{U}_i^{1/2}$ with $\boldsymbol{U}_i$ a diagonal matrix with elements $\text{var}(y_{ij})$, ϕ the dispersion parameter, and $\boldsymbol{C}(\boldsymbol{\alpha})$ is the working correlation structure with non-diagonal terms $\boldsymbol{\alpha}$. The $n_i \times n_i$ matrix of weights is $\boldsymbol{W}_i(\boldsymbol{X}_i, A_i, \boldsymbol{\eta}_W) = diag\left[R_{ij}/\pi_{ij}(\boldsymbol{X}_i, A_i, \boldsymbol{\eta}_W)\right]_{j=1,\dots,n_i}$, where the PS is obtained by fitting a binary response model that regresses the indicator R_{ij} on functions of A_i and $\boldsymbol{X}_{ij}$. The $\boldsymbol{\eta}_W$ are nuisance parameters estimated in the PS. The vector $\boldsymbol{B}_i(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B) = [B_{ij}(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B)]_{j=1,\dots,n_i}$ is an arbitrary function of $\boldsymbol{X}_i$ given for each treatment arm. The $\boldsymbol{\eta}_B$ are nuisance parameters that must be estimated. The DR estimator is most efficient if $B_{ij}(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B)$ is equal to $E(Y_{ij}|\boldsymbol{X}_i, A_i = a)$. Hence, we define $\boldsymbol{B}_i(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B)$ as the outcome model (OM).

3 Results

We propose a doubly robust method for the estimation of the marginal effect of treatment in CRTs with continuous data subject to rMAR - an assumption that arises because missingness is non-monotone in CRTs. To be CAN, the DR estimator requires that either the OM or PS model be correctly specified regardless of the choice of the working correlation matrix. In other words, if $\pi_{ij}(\boldsymbol{X}_i, A_i) = P(R_{ij} = 1|\boldsymbol{X}_i, A_i)$ or $B_{ij}(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B) = E(Y_{ij}|\boldsymbol{X}_i, A_i = a)$. When the OM is correctly specified, the DR is generally more efficient than classical IPW.

When we consider the phenomenon of covariate interference where there exists at least one individual $j' \neq j$ such that $E(Y_{ij}|\boldsymbol{X}_{ij}) \neq E(Y_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{X}_{ij'})$. Interfering covariates can be ignored if either the OM or the PS is correctly specified, which may simplify the modeling. Actually, if it is believed that covariate interference only have an effect on the outcome or the missing data generation process they do not need to be entered in the PS or the OM.

In presence of treatment-covariate interactions, if the PS is not correctly specified, covariates that interact with treatment on the outcome must be included in the OM. We accommodate these treatment-covariate interactions by modeling the OM separately for each treatment group. This may simplify the modeling because treatment-covariate interactions does not have necessarily to be entered in the PS.

We recommend using $\boldsymbol{V}_i^{-1} \boldsymbol{W}_i(\boldsymbol{X}_i, A_i, \boldsymbol{\eta}_W)$ to ensure consistency of the IPW and the DR for CRTs. Actually, the implementation, $\boldsymbol{W}_i^{1/2}(\boldsymbol{X}_i, A_i, \boldsymbol{\eta}_W) \boldsymbol{V}_i^{-1} \boldsymbol{W}_i^{1/2}(\boldsymbol{X}_i, A_i, \boldsymbol{\eta}_W)$, which is available in several software packages of the weighted GEE (such as *SAS* procedure *GENMOD* or *R* in pasckages *gee*, *geeM* and *geepack*), may lead to inconsistent results. If a working independence correlation structure is used, then the two implementations lead to the same result. When $\boldsymbol{W}_i^{1/2}(\boldsymbol{X}_i, A_i, \boldsymbol{\eta}_W) \boldsymbol{V}_i^{-1} \boldsymbol{W}_i^{1/2}(\boldsymbol{X}_i, A_i, \boldsymbol{\eta}_W)$ and an arbitrary correlation structure is used in the IPW, estimation of marginal treatment effect is not consistent. Regarding the DR, it is consistent in this case only if the OM is correctly specified but the efficiency not guaranteed.

We provide an *R* package called *CRTgeeDR* that implements the proposed DR estimator. The package can accommodate for a wide range of outcome types, link functions, and working correlation structures. The *CRTgeeDR* package is easy to use and does not require extensive programming. Moreover, it also implement standard GEE, IPW (with correct implementation of weights), AUG and DR.

Extended version of the simulations and the data analysis presented in this communication are available in the articles Prague et al. (2015) and Prague et al. (2016).

References

M Prague, R Wang, A Stephens, E Tchetgen Tchetgen, and V De gruttola. Accounting for interactions and complex inter-subject dependency for estimating treatment effect in cluster randomized trials with missing at random outcomes. *http://biostats.bepress.com/harvardbiostat/paper193/*, 2015.

M Prague, R Wang, and V De gruttola. Crtgeedr: An r package for doubly robust generalized estimating equations estimations in cluster randomized trials with missing data. *http://biostats.bepress.com/harvardbiostat/paper200/*, 2016.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89(427):846–866, 1994.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *JASA*, 90 (429):106–121, 1995.

Alisa J Stephens, Eric J Tchetgen Tchetgen, and Victor De Gruttola. Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates. *Stat. Med.*, 31(10):915–930, 2012.

Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, 1986.

Min Zhang, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.