CRTgeeDR: an R Package for Doubly Robust Generalized Estimating Equations Estimations in Cluster Randomized Trials with Missing Data

by Melanie Prague, Rui Wang, and Victor De Gruttola

Abstract Semi-parametric approaches based on generalized estimating equations (GEE) are widely used to analyze correlated outcomes in longitudinal settings. In this paper, we present a package **CRTgeeDR** developed for cluster randomized trials with missing data (CRTs). For use of inverse probability weighting to adjust for missing data in cluster randomized trials, we show that other software lead to biased estimation for non-independence working correlation structure. **CRTgeeDR** solves this problem. We also extend the ability of existing packages to allow augmented Doubly Robust GEE estimation (DR). Simulation studies demonstrate the consistency of estimators implemented in **CRTgeeDR** compared to packages such as **geepack** and the gains associated with the use of the DR for analyzing a binary outcome using a logistic regression. Finally, we illustrate the method on data from a sanitation CRT in developing countries.

Introduction

We describe the R package **CRTgeeDR**, for estimating coefficients of regression in a marginal mean model. The method 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, and 4) treatment is randomized at a cluster level. The estimation approach generalizes the Generalized Estimating Equation (GEE) (Zeger and Liang, 1986) for fitting marginal generalized linear models to clustered data with possibly informative missingness of the outcome. It combines existing methods for accommodating missing data that use inverse probability weighting (IPW) (Robins et al., 1995) and for increasing precision of estimation by appropriate use of baseline covariates (AUG) (Stephens et al., 2012). We have developed a method for estimating the intervention effect in cluster randomized trials that combines the IPW and the AUG and is doubly robust (DR), meaning that the resulting estimator is consistent if either the model predicting the outcome or the model predicting the missing data is correctly specified—that is, they reflect the true data generation processes (Prague et al., 2016). Below we illustrate the use of the software on a real dataset and clarify its benefits.

The package **CRTgeeDR** not only implements the DR estimator but also the standard GEE, the IPW and the AUG. Regarding IPW, our package differs from most of those currently available in that it avoids the bias that can result from conventional implementation applied to CRTs. Lin et al. (2015) pointed out that implementation of GEE for complete longitudinal data in the current version of SAS (`GENMOD procedure`) requires use of an independence correlation structure if the observation of the outcome at one time point depends on covariates obtained at another time point; this problem had been corrected in the new `GEE procedure` in SAS/STAT 13.2 (SAS Institute Inc., 2015). Tchetgen Tchetgen et al. (2012) made a similar comment regarding the analysis of incomplete longitudinal data in which time-varying covariates and previous outcome values are needed to model the missingness process. This article clarifies this issue for CRTs and proposes an implementation in R that allows for unbiased IPW (and thus DR) estimation with non-independence working correlation structure.

GEE-based approaches for estimating the coefficients in marginal models, in particular the marginal effect of an intervention, have been implemented in only a limited number of R packages and other software for general use. Of note, most of the available software was initially developed to deal with correlated longitudinal data rather than data from CRTs. There are three R packages on CRAN, which will solve GEEs and produce standard errors: whereas **gee** (Carey et al., 2012) and **geepack** (Jun, 2002; Halekoh et al., 2006; Højsgaard and Halekoh, 2016) are computationally demanding, the package **geeM** allows a fast estimation through the use of sparse matrix representation (McDaniel et al., 2013). When interest lies in adjusting for missing outcomes using the IPW, all the packages mentioned above require specification of weights. These weights can be computed using packages such as **ipw** (van der Wal and Geskus, 2011; Geskus and van der Wal, 2015) or directly assigned from a user-defined function. These approaches require the missing data process to be known or correctly specified. Some packages, such as **drgee** (Zetterqvist and Sjölander, 2015), implement doubly robust approaches for uncorrelated data arising from observational studies. These packages provide estimates that are doubly robust in the sense that the consistency of the parameter estimator from the marginal models is guaranteed if the model linking the outcome to covariates and treatment or the model linking the treatment assignment to covariates correctly reflects the true data generation process. These methods have been extended to deal with missing data with IPW approaches in **CausalGAM** (Glynn and Quinn, 2010a,b), but these packages are intended for analysis of observational studies, not CRTs. Finally, the targeted maximum likelihood estimation (tMLE) method allows estimation of the marginal additive effect of a treatment (van der Laan, 2014a). It is implemented in the packages **tmle** (Gruber, 2014) and **tmlenet** (Sofrygin and van der Laan, 2015) for longitudinal and correlated data. Except for Porter et al. (2011), there has been little published discussion about the differences between GEE-based and tMLE estimation, and we do not delve into a comparison of the two methods. The focus of this article is only on software implementation of the doubly robust GEE for CRTs.

The paper is organized as follows. Section 2.2 introduces the theory of the doubly robust estimator and Section 2.3 describes the features of the **CRTgeeDR** and the estimating function denoted `GeedrEstimation`. Section 2.4 compares the performance of **CRTgeeDR** to **geepack** for the IPW in CRTs and illustrates that the DR is consistent and more efficient than the IPW. Section 2.5 illustrates the analysis of a dataset on sanitation in developing countries (Guiteras et al., 2015a) and illustrates the benefit of using the DR approach compared to standard GEE. Section 2.6 presents a discussion.

IPW in CRTs and doubly robust estimation

Notation

Consider a CRT comprised of n clusters or communities, each with n_i individuals. The cluster sample sizes are assumed fixed and non-informative. Let $\boldsymbol{Y}_i = [Y_{ij}]_{j=1,\ldots,n_i}$ denote the outcome vector for cluster i, some elements of which may be unobserved. Let $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ otherwise. Let $\boldsymbol{X}_{ij} = [X^r_{ij}]_{j=1,\ldots,n_i;r=1,\ldots,P}$ denote the P baseline covariates for subject j in cluster i, which is fully observed. Let A_i be the treatment assigned to cluster i; the indicator for treated condition is $A_i = 1$, and $A_i = 0$ for control condition. We assume that the probability of treatment assignment is known and fixed to $p_A = P(A_i = 1)$. The conditional mean of Y_{ij} is denoted $\mu_{ij} = E(Y_{ij}|X_{ij}, A_i)$, and we let $\boldsymbol{\mu}_i = [\mu_{ij}]_{j=1,\ldots,n_i}$ denote the full vector of means in the i^{th} cluster. We assume that the mean structure of Y_{ij} depends on the covariate vector for subject j in cluster i (Robins et al., 1999), and consider a model for the mean as follow:

$$g(\mu_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_X + A_i\beta_A,$$

where $g(.)$ is a monotone differentiable link function and $\boldsymbol{\beta} = (\beta_A, \boldsymbol{\beta}_X)$ is a $(P+1) \times 1$ is a vector of regression coefficients of interest. In this article, we focus on estimation of the marginal effect of an intervention β_A for a binary outcome using the logit link. We assume the variance is $v_{ij} = \text{var}(Y_{ij}|X_{ij}, A_i) = \phi h(\mu_{ij})$, where $h(.)$ is the variance function and ϕ is the dispersion parameter. Thus for our specific example, $v_{ij} = \phi\mu_{ij}(1-\mu_{ij})$ When data are missing at random (MAR), the observation indicator R_{ij} is a function of covariates, treatment condition, and observed outcomes. For CRTs, we assume a restricted version of MAR (rMAR), which requires that R_{ij} cannot be a function of observed outcomes. Although all the theory would hold for classical MAR assumption, it is most of the time difficult to specify the function linking the observation indicator and the observed outcomes of other individuals in the same cluster because there is no ordering. Thus, the probability of being observed π_{ij} for individual j in cluster i, called the propensity score (PS), is: $\pi_{ij}(\boldsymbol{X}_{ij}, A_i, \eta_W) = P(R_{ij} = 1|\boldsymbol{X}_{ij}, A_i)$. The parameters η_W are nuisance parameters and must be estimated.

IPW in CRTs

In presence of rMAR outcome, as in Robins et al. (1995), we estimate $\boldsymbol{\beta}$ by using inverse probability weighted generalized estimating equation (IPW). Therefore, we must include a weight matrix $\boldsymbol{W}_i$ to the usual GEE, that is:

$$\boldsymbol{W}_i(\boldsymbol{X}_{ij}, A_i, \eta_W) = \text{diag}\Big(\frac{R_{ij}}{\pi_{ij}(\boldsymbol{X}_{ij}, A_i, \eta_W)}\Big)_{j=1,\ldots,n_i}.$$

This matrix $\boldsymbol{W}_i(\boldsymbol{X}_{ij}, A_i, \eta_W)$, denoted simply as $\boldsymbol{W}_i$ in the following, adjusts the contribution of each individual in a given cluster by upweighting the contribution of individuals who are less likely to be observed according to their characteristics. Thus, if the propensity score is correctly specified, i.e., correspond to the true missingness process, the IPW equation provides consistent estimates:

$$0 = \sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{W}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i), \tag{1}$$

where $D_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ is a derivative matrix and $\boldsymbol{V}_i$ is the working covariance matrix for the response $\boldsymbol{Y}_i$. In particular, $\boldsymbol{V}_i = \phi \boldsymbol{F}_i^{1/2} \boldsymbol{C}(\boldsymbol{\alpha}) \boldsymbol{F}_i^{1/2}$, where $\boldsymbol{F}_i^{1/2} = \mathrm{diag}(h(\mu_{ij}))_{j=1,\ldots,n_i}$ and $\boldsymbol{C}(\boldsymbol{\alpha})$ is the working correlation structure with non-diagonal terms $\boldsymbol{\alpha}$. For example, for an independence correlation structure $\boldsymbol{\alpha}$ is zero; for exchangeable structure, all the elements of $\boldsymbol{\alpha}$ are identical. Parameters $\boldsymbol{\alpha}$ could also depend on the treatment assignment $\boldsymbol{C}(\boldsymbol{\alpha}(A_i))$ but we do not consider this possibility in our implementation. In the package **CRTgeeDR**, we estimate the $\boldsymbol{\alpha}$ and ϕ parameters using moment estimators from the Pearson residuals and the Pearson Chi-Square statistic as in **geeM** (McDaniel and Henderson, 2015) also described in McDaniel et al. (2013). In the absence of missing data, $\boldsymbol{W}_i = \boldsymbol{I}$ is set to identity, and the standard GEE is performed by **CRTgeeDR**.

In existing packages such as **geepack**, the Equation 1 is implemented as $0 = \sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, with $\boldsymbol{V}_i^{-1} = \phi \boldsymbol{F}_i^{1/2} \boldsymbol{W}_i^{1/2} \boldsymbol{C}(\boldsymbol{\alpha}) \boldsymbol{W}_i^{1/2} \boldsymbol{F}_i^{1/2}$ to ensure the fast invertibility of $\boldsymbol{V}_i$. It is easy to verify that when an independence correlation structure is used, $\boldsymbol{C}(\boldsymbol{\alpha}) = I$, and the two implementations are identical. Therefore, one can always use **geepack** with an independence working correlation structure. In contrast, if a non-independence working correlation structure is used, the consistency of IPW estimators do not hold. See the Web-Supplementary Material for a demonstration. Regarding other packages such as **geeM**, although the implementation was the same as in **geepack** up to version 0.8.0, it is now implemented as in Equation 1 in version 0.10.0. In the SAS `GEE` procedure, one can use the option `"type=obslevel"` (in the `missing` statement) in order to use the same implementation as in Equation 1. In general, it is necessary to check the formula used for implementation of the estimating equation in any desired software to avoid confusion.

Augmentation and doubly robust estimation

Recent advances in methods for analysis of data from CRTs have used augmented GEE to improve efficiency of inferences by incorporating baseline covariates (Stephens et al., 2012); we denote this estimator the AUG. They have also been extended to accommodate missing data using an approach based on the IPW which is doubly robust GEE (DR). The DR properties are described in Prague et al. (2016) and the estimating equation is given by :

$$\begin{aligned} 0 = \sum_{i=1}^{M} \Bigg[& \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{W}_i \left(\boldsymbol{Y}_i - \boldsymbol{B}_i(\boldsymbol{X}_{ij}, A_i, \boldsymbol{\eta}_B) \right) \\ & + \sum_{a=0,1} p_A^a (1-p_A)^{1-a} \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1} \left(\boldsymbol{B}_i(\boldsymbol{X}_{ij}, A_i = a, \boldsymbol{\eta}_B) - \boldsymbol{\mu}_i(\boldsymbol{\beta}, A_i = a) \right) \Bigg] \\ = {} & \boldsymbol{\Phi}(\boldsymbol{Y}_i, \boldsymbol{R}_i, A_i, \boldsymbol{X}_{ij}, \boldsymbol{\beta}, \boldsymbol{\eta}_W, \boldsymbol{\eta}_B). \end{aligned} \tag{2}$$

Each element of the vector $\boldsymbol{B}_i(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B) = [B_{ij}(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B)]_{j=1,\ldots,n_i}$ is an arbitrary function linking Y_{ij} with $\boldsymbol{X}_{ij}$ for each treatment arm, which we refer to as the outcome model (OM) The $\boldsymbol{\eta}_B$ are nuisance parameters. The estimator in Equation 2 is most efficient if $B_{ij}(\boldsymbol{X}_i, A_i = a, \boldsymbol{\eta}_B) = E(Y_{ij} | \boldsymbol{X}_{ij}, A_i = a)$ (Zhang et al., 2008), that is, the OM is correctly specified. If the OM is not correctly specified, i.e., does not correspond to the true data generation process, the estimation remains consistent provided that the PS model is correctly specified, but one may have a loss in efficiency. Without missing data, $\boldsymbol{W}_i = \boldsymbol{I}$ is set to identity, and the AUG is performed by **CRTgeeDR**.

Without missing data or with data missing completely at random, the use of augmentation may allow a gain in efficiency by incorporating information on baseline covariates. The PS should not be used because it will be misspecified and therefore may lead to an increase of the variance of the estimates. In presence of rMAR data, IPW alone can be used but DR should be preferred in order to increase the chances to have an unbiased estimator. Finally, as mentioned above, for data missing not at random, none of the methods implemented in **CRTgeeDR** are adequate.

The R package CRTgeeDR

The main function for estimation in the package `CRTgeeDR`

The call function for performing estimation is `geeDREstimation`:

```
R> geeDREstimation(formula, id, data = parent.frame(), family = gaussian,
+    corstr = "independence", Mv = 1, corr.mat = NULL, init.beta = NULL,
+    init.alpha = NULL, init.phi = 1, scale.fix = FALSE, maxit = 20,
+    tol=1e-05, print.log = FALSE, nameTRT = "TRT", nameMISS = "MISSING",
+    nameY = "OUTCOME", sandwich = TRUE, sandwich.nuisance = FALSE,
+    fay.adjustment = FALSE, fay.bound = 0.75, aug = NULL, pi.a = 1/2,
+    model.augmentation.trt = NULL, model.augmentation.ctrl = NULL,
+    stepwise.augmentation = FALSE, weights = NULL, typeweights = "VW",
+    model.weights = NULL, stepwise.weights = FALSE)
```

The marginal model, to be estimated on the R dataframe `data`, is given in `formula`. The link function, g, depends on the nature of the outcome, which is specified in the argument `family`. The name of the outcome `nameY`, the clustering variable `id`, the binary treatment `nameTRT` (with the convention 1 is treated and 0 is control), and the missing indicator `nameMISS` must be specified if they differ from default values. The algorithm iterates between the estimation the working correlation structure and regression parameters with a stopping rule based on stabilisation of estimates (tolerance can be set by the user; default is `tol`$=10^{-5}$ or `max.iter=20`). Depending on the specification or not of the PS and the OM, `geeDREstimation` allows the implementation of standard GEE, the IPW, the AUG and the DR approaches. The algorithm is defined as follow:

1. *Determine the PS:* $\pi_{ij}(\boldsymbol{X}_{ij}, A_i, \eta_W) = P(R_{ij}|\boldsymbol{X}_{ij}, A_i)$, π_{ij} for short. Either the π_{ij} are known from prior analysis or by design and the weights can be specified directly in the `weights` argument. Alternatively one can compute the PS by fitting a logistic regression of R_{ij} on $(\boldsymbol{X}_{ij}, A_i)$. In this case, the PS regression formula can be directly entered in `model.weights`. A `glm` with logit link function is internally processed with or without variable selection, depending on the value of the `stepwise.weights` argument. If all of the above are set to NULL or default, no IPW adjustment will be made—GEE or AUG will be used. Finally, if despite our concern about the implementation of weights, one wants to use the same implementation as in packages **geepack** or `proc GENMOD` in SAS, then one can set `typeweights="GENMOD"`.
2. *Determine group-specific OM:* $B_{ij}(X_{ij}, A_i = a) = E\left[\boldsymbol{Y}_{ij}|A_i = a, X_{ij}\right]$. When the $\boldsymbol{B}_i$ are known from prior analysis, they can be directly entered in `aug=c(ctrl=`$B_{ij}(X_{ij}, A_i = 0)$`,trt=`$B_{ij}(X_{ij}, A_i = 1)$`)`. Alternatively, we can regress Y_{ij} on $\boldsymbol{X}_{ij}$ within each treatment group. In this case, the OM regression formulas can be directly entered in `model.augmentation.trt` and `model.augmentation.ctrl`. A `glm` is then internally processed with or without variable selection depending on the value of the argument `stepwise.augmentation`. If all of the above are set to NULL or default, no augmentation adjustment will be made—GEE or IPW will be used. The probability of treatment assignment, which is known in CRTs, must be specified in the argument `pi.a`. Of note for steps 1 and 2, when using the stepwise option to compute the OM or the PS, one runs the risk of overfitting (van der Laan, 2014b). Avoiding this is possible by sparsely including only relevant variables in the selection and also by running a bootstrap diagnostic using outputs (`ps.model`, `om.model.trt` and `om.model.ctrl`). The underlying assumption is that the true OM or PS are selected at the end of the stepwise selection and then held fixed in the estimating equation in further steps.
3. *Determine the working correlation structure.* Available structures are `independence`, `exchangeable`, `M-dependent` (using `Mv`), `unstructured`, or `user-defined` (using `corr.mat`). Using the `scale.fix` argument, the dispersion parameter ϕ can be either estimated or held fixed to a specified value.
4. *Obtain initial values.* They are either specified by the user (`init.beta`, `init.alpha`, and `init.phi`) or internally defined by fitting a glm under independence to obtain initial values for $\hat{\boldsymbol{\beta}}^{(0)}$ and by setting $\phi^{(0)} = 1$ and $\boldsymbol{\alpha}^{(0)} = 0$.
5. *Enter/continue the iterative procedure* :

   (a) Use the fit from $\hat{\boldsymbol{\beta}}^{(n)}$ to compute Pearson residuals. Use Pearson residuals based formulas to compute the scale parameter ($\phi^{(n+1)}$, except if `scale.fix=`TRUE) and the parameters in the working correlation matrix ($\boldsymbol{\alpha}^{(n+1)}$).

   (b) Construct the augmented equation given in Equation 2 and solve it numerically using Newton-Raphson algorithm for $\hat{\boldsymbol{\beta}}^{(n+1)}$:

   $$\hat{\boldsymbol{\beta}}^{(n+1)} = \hat{\boldsymbol{\beta}}^{(n)} - \left[\frac{\partial \boldsymbol{\Phi}(\boldsymbol{Y}_i, \boldsymbol{R}_i, A_i, \boldsymbol{X}_{ij}, \boldsymbol{\beta}, \boldsymbol{\eta}_W, \boldsymbol{\eta}_B)}{\partial \boldsymbol{\beta}}\right]^{-1}_{\hat{\boldsymbol{\beta}}^{(n)}} \boldsymbol{\Phi}(\boldsymbol{Y}_i, \boldsymbol{R}_i, A_i, \boldsymbol{X}_{ij}, \hat{\boldsymbol{\beta}}^{(n)}, \boldsymbol{\eta}_W, \boldsymbol{\eta}_B).$$

   (c) If $\max\left|\frac{\hat{\beta}^{(n+1)} - \hat{\beta}^{(n)}}{\hat{\beta}^{(n)} + prec.machine}\right| >$ `tol` and $n+1 \leq$ `max.iter` go back to 5 else go to 6, where $prec.machine \sim 10^{-16}$.

6. Compute the requested variances of $\hat{\beta}^{(n+1)}$. If, `sandwich` and `sandwich.nuisance` are set to TRUE, classical and nuisance-adjusted (for the estimation of parameters η_W in the PS and η_B in the OM) sandwich estimators of the variance are provided, see Prague et al. (2016) for their definition. The nuisance-adjusted version is computed using numerical derivatives of score equations for PS, OM and estimating equations jointly, which are obtained by using the `jacobian` function of the package **numDeriv** (Gilbert and Varadhan, 2015); this is recommended if the AUG, the IPW or the DR estimator are considered. Finally, a small-sample-adjusted sandwich estimator of the variance can also be computed using Fay's adjustment (Fay and Graubard, 2001) setting the argument `fay.adjustment` to TRUE. Its implementation is derived from the function `gee.var.fg` in the package **geesmv** (Wang, 2015).

Adequacy of the PS and the OM to data

Consistency and efficiency of the DR estimator depend on the correct specification of the PS and the OM, see Prague et al. (2016) for theoretical demonstrations. The user may want to check the adequacy of the selected OM model to the data by using the function `getOMPlot`, which provides plots to check the `glm` model assumption. The "Residuals vs. Fitted" and the "Scale-location" graphics allow verification of the homogeneity of the variance and the adequacy of the link function. The "Normal Q-Q" checks for the normal distribution of the residuals. The "Residuals vs Leverage" plot allows detection of points that have high leverage on the regression coefficients and that should be investigated as outliers. In the same spirit, the "Cook's distance" and the "Cook's distance vs leverage" provide measures of the effect of deleting a given observation. Of note, these graphs are only interpretable for a continuous outcome. In addition, for the PS model the function `getPSPlot` provides a histogram of the weights. If weights are too large then the IPW and DR approaches are likely to be unstable. In this case, the user should compute weights externally using, for example, stabilized weights with the associated package **ipw** (van der Wal and Geskus, 2011) or other approaches such as described in Wang and Paik (2011). Finally, the user can access the `glm` objects created during the PS and OM initial steps as objects named `ps.model`, `om.model.trt`, and `om.model.ctrl` from the main function `geeDREstimation`.

Simulations

The properties of DR to accommodate complex correlation structure, rMAR outcomes, and the presence of imbalance in baseline covariates have already been demonstrated in Prague et al. (2016). In this article, we focus on the superiority of implementation of weights in the package **CRTgeeDR** compared to package **geepack**. We focus on a simple example to illustrate that, even in very simple cases, estimators implemented in broadly used R package **geepack** for IPW can be inconsistent when using an exchangeable working correlation structure. This is the case when $\boldsymbol{V}_i^{-1} = \phi \boldsymbol{F}_i^{1/2}\boldsymbol{W}_i^{1/2}\boldsymbol{C}(\boldsymbol{\alpha})\boldsymbol{W}_i^{1/2}\boldsymbol{F}_i^{1/2}$ is used in the estimating equation. We simulate data from a CRT with 100 communities of 90, 100, or 110 individuals with probability 1/3 for each. The treatment A is randomly assigned with probability $p_A = 1/2$. One covariate is of interest: $X_{ij} \sim \mathcal{N}(2,1)$. We simulate correlated outcome with exchangeable structure, and correlation between individuals is set to 0.05. This is done by using a cluster-level bridge distribution $b_i \sim \mathcal{B}(0.05)$. Data generation process is as follow:

$$\begin{array}{rcl} \text{logit}[P(Y_{ij}=1|A_i,X_{ij})] &=& -0.5+0.3A_i+0.4X_{ij}+0.4X_{ij}A_i+b_i, \\ \text{logit}[P(R_{ij}=1|A_i,X_{ij})] &=& 4.0-0.3A_i-0.8X_{ij}-0.8X_{ij}A_i. \end{array} \tag{3}$$

We simulated R=10,000 replicates. The observed average proportion of missing observations is around 25% and the observed average intraclass correlation is 0.08. Missingness is associated strongly with individual covariates and, therefore, the weights differ between individuals in the same cluster. The true value of the odds-ratio for the marginal effect of treatment is computed for each dataset k without missing data by obtaining the counterfactual values with and without treatment under this model:

$$\text{OR}_k = \frac{E(Y_{ij}=1|A_i=1)/E(Y_{ij}=0|A_i=1)}{E(Y_{ij}=1|A=0)/E(Y_{ij}=0|A_i=0)}.$$

The true OR is given by $\frac{1}{R}\sum_{k=1}^{R} OR_k$=2.56 with associated parameter for marginal intervention effect in the marginal regression $\beta_A = 0.941$. For each dataset, we first ran the analysis on the dataset without missing data for the standard GEE and the AUG using **CRTgeeDR**. Then we ran the analysis on the dataset with missing data for the IPW using **geepack** and for the standard GEE, the IPW, the AUG, and the DR using **CRTgeeDR**. Two types of DR are presented here: DR1 is the estimator using the correct models for the OM and the PS, and DR2 omits treatment-covariate interaction terms in the PS. The models for the PS and OM for analysis are described in the Table 1. Table 1 shows the bias, empirical standard error, sandwich standard error, and coverages for each analysis using independence (-I) and exchangeable (-E) working correlation structure. The code to replicate this study is available in Web-Supplementary Material.

	Independence (-I)				Exchangeable (-E)			
Method	Bias	Emp. SE	SE	Cov.	Bias	Emp. SE	SE	Cov.
No missing data:								
GEE **CRTgeeDR**	0.002	0.102	0.099	94.3	0.002	0.108	0.099	93.2
GEE **geepack**	0.003	0.102	0.101	94.6	0.003	0.102	0.101	94.6
AUG **CRTgeeDR**	0.002	0.101	0.099	94.3	0.002	0.109	0.114	95.8
With missing data:								
GEE **CRTgeeDR**	-0.257	0.103	0.177	82.0	-0.256	0.104	0.081	18.1
AUG **CRTgeeDR**	0.249	0.092	0.109	35.7	0.307	0.115	0.139	37.1
With missing data and adjustment for it:								
IPW **CRTgeeDR**	0.003	0.108	0.106	95.0	0.003	0.118	0.110	93.7
IPW **geepack**	0.008	0.107	0.104	94.8	0.582	0.577	0.357	19.4
DR1 **CRTgeeDR**	0.003	0.107	0.104	94.5	0.004	0.120	0.125	96.1
DR2 **CRTgeeDR**	0.003	0.105	0.102	94.4	0.004	0.118	0.123	96.0

Marginal mean model:
$\text{logit}(\mu_{ij}) = \beta_0 + \beta_A A_i$.
PS used for IPW and DR (true):
$\text{logit}(P(R_{ij} = 1|A_i, X_{ij})) = \gamma_0 + \gamma_A A_i + \gamma X_{ij} + \gamma_I X_{ij} A_i$.
PS used for DR2 (omitting interactions in PS):
$\text{logit}(P(R_{ij} = 1|A_i, X_{ij})) = \gamma_0 + \gamma_A A_i + \gamma X_{ij}$.
OM used for AUG, DR1 and DR2 (fitted for each group a):
$\text{logit}(P(Y_{ij} = 1|A_i = a, X_{ij})) = \xi + \xi_A A_i + \xi X_{ij}$.

Table 1: Comparison of the standard GEE, the IPW, the AUG and the DR analysis with the packages **CRTgeeDR**, **geepack**, and **geeM** using independence and exchangeable working correlation structure. True value for the parameter β_A is 0.91 (OR=2.56). The bias, the empirical and the estimated standard errors (SE), and the coverages for parameter $\widehat{\beta_A}$ are computed over 10,000 replicates. The true data generation process for outcome and missingness is provided in Equation 3. The PS and OM models for analysis are correctly specified and given in the footnote of the table.

The results for standard GEE are unbiased in the absence of missing data (<0.003 for GEE-I and GEE-E with all packages) and biased in presence of rMAR outcomes reflecting the fact that the missingness is informative. Using the IPW-I corrects for this bias (0.008 for **geepack**). All packages give a similar estimated standard error leading to acceptable coverage close to their nominal value of 95%. When using an exchangeable correlation structure, the coverage (93.7%) remains close to the nominal value for IPW-E using **CRTgeeDR**, but it drops to 19.4% using **geepack**. This is mainly driven by an increase in the bias from 0.003 for **CRTgeeDR** to 0.582 for **geepack** for IPW-E. Using the DR1 version of **CRTgeeDR** provides consistent estimates (bias $\leq$0.004 for DR1-I and DR1-E). DR-1 yields coverage that is close to or greater than 95% and gains, on average, in efficiency. For example, the empirical standard error is 0.108 for IPW-I and 0.107 for DR-I. DR2, which omits the term $X_{ij}A_i$ in the PS, yields consistent and efficient estimates even when the treatment-covariate interactions are not explicitly specified in the PS. As demonstrated in Prague et al. (2016), DR1 and DR2 have similar properties.

Illustration on the sanitation data

In this section, we present a step-by-step analysis of data from a CRT to investigate the efficacy of alternatives policies on the investment in hygienic latrines in developing countries. A total of 380 communities in rural Bangladesh were assigned to different marketing interventions—community motivation, subsidies, supply side-market, a combination of the three, and a control group. Results of this study were published in (Guiteras et al., 2015a). All the code and data associated with this study are available on dataverse, see url in Guiteras et al. (2015b).

	Side-Market supply		Control		All	
Cluster structure						
M	36 (n = 1651)		66 (n = 3186)		100 (n = 4837)	
N_i	49 (15)		48 (16)		48 (16)	
Outcome Y_{ij}	Mean	Missing %	Mean	Missing %	Mean	Missing %
Hygienic Latrine Ownership	34.8%	4.2%	30.3%	3.1%	31.8%	3.5%
Individual-level X_{ij}^{IND}	Mean	Missing %	Mean	Missing %	Mean	Missing %
Report diarrhea	4.3%	0%	4.8%	0%	4.6%	0%
Male	91.1%	<0.01%	90.0%	<0.01%	90.1%	<0.01%
Education	49.2%	0%	45.8%	0%	46.9%	0%
Muslim	83.2%	<0.01%	86.3%	<0.01%	85.2%	<0.01%
Bengali	85.6%	<0.01%	88.5%	<0.01%	87.6%	<0.01%
Agricultor	75.0%	<0.01%	70.2%	<0.01%	71.9%	<0.01%
Stoves	58.2%	<0.01%	62.9%	<0.01%	61.3%	<0.01%
Water Pipes	89.9%	<0.01%	91.3%	<0.01%	90.8%	<0.01%
Phone	64.1%	<0.01%	57.2%	<0.01%	59.5%	<0.01%
Age	39 (13)	<0.01%	39 (14)	<0.01%	39 (14)	<0.01%
Cluster-level X_{ij}^{C}	Mean	Missing %	Mean	Missing %	Mean	Missing %
Village size	230 (120)	0%	270 (190)	0%	256 (170)	0%
Nb doctors	7 (7)	0%	9 (18)	0%	8 (15)	0%
% Landless	41.6 (12)	0%	34.4 (15)	0%	36.9 (15)	0%
% Almost Landless	19.3 (11)	0%	24.0 (8)	0%	22.4 (9)	0%
% Access electricity	59.9 (26)	0%	59.1 (20)	0%	59.4 (22)	0%

Table 2: Description of the Sanitation dataset from (Guiteras et al., 2015a) considering only the Side-Market supply and the Control group. Percentages are given for qualitative covariates. Means and standard deviations in parentheses are provided for continuous covariates.

We consider only the comparison of a supply side-market versus control. The published analysis used a mixed effect model and showed that the supply side-market alone did not increase the hygienic latrine ownership (+0.3 percentage points, p-value=0.90). We reanalyze the dataset using the GEE approaches in order to get the marginal effect of intervention. Description of the outcome and variables for adjustment are available in Table 2. Because covariates were missing in less than 0.01% of the observations, we assume that covariates are missing completely at random and exclude individuals with missing covariates. The final dataset contains 4774 individuals and 380 clusters. We assume the outcomes are rMAR. As there is some evidence of imbalance in baseline covariates across arms, i.e., the descriptive distributions of covariates in Table 2 are different between treated and control groups, we use the DR approach. We assume that the correlation between any pair of individuals in the same cluster is the same and hence use an exchangeable working correlation structure. In this example, the PS and OM are fitted using a logistic regression with a linear combination of all the individual-level and cluster-level covariates described in Table 2. Variables for these models are selected using a forward stepwise regression before solving the estimating equation. Adequacy of the model has been verified. The code for analysis is available in the Web-Supplementary Material. To illustrate the use of the package **CRTgeeDR**, we provide instructions for the DR estimator:

```
R> DR <- geeDREstimation(OUTCOME ~ TRT, id = CLUSTER, data = Sanitation,
+    family = binomial("logit"), corstr = "exchangeable", typeweights = "VW",
+    model.weights = MISSING ~ TRT + DIARRHEA + ... + ELEC_ACCESS,
+    model.augmentation.trt = OUTCOME ~ DIARRHEA + ... + ELEC_ACCESS,
+    model.augmentation.ctrl = OUTCOME ~ DIARRHEA + ... + ELEC_ACCESS,
+    stepwise.weights = TRUE, stepwise.augmentation = TRUE)
R> summary(DR)
```

The output displays statistics for estimated coefficients $\boldsymbol{\beta}, \boldsymbol{\alpha}$ and ϕ, the number of Newton-Raphson iterations before convergence, and some description of the size of the clusters.

```
            Estimates Model SE Robust SE   wald        p
(Intercept)   -0.8106  0.09396    0.1088 -7.452 0.000000
TRT            0.4365  0.12890    0.1425  3.062 0.002198

 Est. Correlation:  0.07306
 Correlation Structure:  exchangeable
 Est. Scale Parameter:  0.9955

 Number of GEE iterations: 2
 Number of Clusters:  100    Maximum Cluster Size:  87
 Number of observations with nonzero weight:  4612
```

Table 3 presents the PS and OM for analysis, the estimates, the nuisance-adjusted sandwich estimates of the variance, the confidence intervals for the odd-ratios, the p-values, and the computation times for each of these analysis. For DR the computation time is 20 seconds, most of which is required for the computation of the nuisance-adjusted sandwich estimator of the variance (the estimation is $<$ 3 seconds otherwise). Whereas GEE and IPW lead to non-significant effect of supply side-market, the DR estimates are significantly different from 0 at the 0.05 level (p=0.025). Using the DR, we conclude that there is 55% [8% - 121%] greater chance of owning hygienic latrine after one year if there is a supply side-market. This effect is significant (p<0.05) even using a nuisance-adjusted SE, which is generally larger than the standard sandwich SE due to incorporation of additional variability from estimation of the nuisance parameters in the PS and the OM (η_W and η_B). Information about the PS and the OM can be obtained by using the following commands:

```
R>  summary(DR$ps.model)
R>  summary(DR$om.model.trt)
R>  summary(DR$om.model.ctrl)
R>  getPSPlot(DR)
```

		Sandwich	Nuis-adj.	$exp(\beta_A)$			p-value		time
	β_A	SE	SE	OR	IC_{min}	IC_{max}	Unadj.	Nuis-adj.	(sec.)
GEE	0.19	0.171	-	1.21	0.87	1.69	0.262	-	1
IPW	0.19	0.182	0.219	1.21	0.79	1.86	0.290	0.386	32
AUG	0.45	0.141	0.176	1.57	1.12	2.22	0.001	0.010	11
DR	0.44	0.143	0.183	1.55	1.08	2.21	0.002	0.016	20

Marginal mean model: $logit(\mu_{ij}) = \beta_0 + \beta_A A_i$.
PS: $\text{logit}(P(R_{ij}|A_i, \boldsymbol{X}_{ij}^{\text{IND}}, \boldsymbol{X}_{ij}^{\text{C}}) = \gamma_0 + \gamma_A A_i + \sum_{k=1}^{10} \gamma_k^{\text{IND}} X_{ijk}^{\text{IND}} + \sum_{k=1}^{5} \gamma_k^{\text{C}} X_{ijk}^{\text{C}}$.
OM: $\text{logit}(P(Y_{ij}|A_i = a, \boldsymbol{X}_{ij}^{\text{IND}}, \boldsymbol{X}_{ij}^{\text{C}}) = \xi_0 + \sum_{k=1}^{10} \xi_k^{a\text{IND}} X_{ijk}^{\text{IND}} + \sum_{k=1}^{5} \xi_k^{a\text{C}} X_{ijk}^{\text{C}}$, for each group a.

Table 3: Effects of the supply side-market vs. control on the probability of hygienic latrine ownership in the sanitation data analysis (Guiteras et al., 2015a) using the standard GEE, the IPW adjustment (IPW and DR), and the augmentation for imbalance (AUG and DR) assuming outcomes are rMAR.

Description of models for OM, PS and histogram of weights are given in the Web-Supplementary Material Table 1 and Figure 1. As noted in Table 3, the estimates for IPW are close to those for GEE, reflecting the fact that only 3.5% of data are missing. We also note that all of the non-null weights are close to 1 (1.035 [1.02; 1.04]) showing that no covariate of the PS explains the missingness pattern. Thus, the increased significance of the intervention in the DR analysis compared to GEE is mainly driven by the augmentation. In both groups, households with higher education and economic status (as evidenced by stoves, water pipes, phones, and other factors) are more likely to have a hygienic latrine. For cluster-level covariates the patterns differ by intervention group: a high number of doctors is positively associated with the hygienic latrine ownership only in the intervention group indicating a potential synergy between the number of doctors and the presence of side-supply markets.

Conclusion

We demonstrated that the IPW can be biased in CRTs if the weights are not implemented as described in Robins et al. (1995) and a non-independence working correlation structure is chosen. In particular, we discuss problems that arise in the package **geepack** implemented in R. These concerns apply not only for outcome data in CRTs but also to longitudinal outcome data, when the probability that an observations is missing at a given time depends on time-varying covariates measured at other times. We recommend to always check the implementation in the software that has been chosen for analysis. The **CRTgeeDR** package protects against this bias and allows for adjustment in imbalance in baseline covariates in CRTs. The package can accommodate 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. It therefore makes the augmented GEE (AUG) and the Doubly robust (DR) methodology for CRTs more accessible to applied researchers. Of note, although the **CRTgeeDR** package had been designed for CRTs, it can also be used for analysis of correlated longitudinal data from a randomized trial. The use of version 2.0 of the **CRTgeeDR** package to analyze observational clustered data (in which treatment attribution may be informative) is not straightforward, but updates with these capabilities are under development.

Acknowledgement

We thank R. Guiteras for sharing the Sanitation study on the dataverse website. This work was founded by NIH grants R37 AI 51164 and R01 MH100974. Portions of this research were conducted on the Cluster at Harvard Medical (NIH grant NCRR 1S10RR028832-01).

Bibliography

V. J. Carey, T. Lumley, and B. Ripley. *gee: Generalized Estimation Equation Solver*, 2012. URL http://CRAN.R-project.org/package=gee. R package version 4.13-19. [p105]

M. P. Fay and B. I. Graubard. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57(4):1198–1206, 2001. [p109]

R. B. Geskus and W. M. van der Wal. *ipw: Estimate Inverse Probability Weights*, 2015. URL http://CRAN.R-project.org/package=ipw. R package version 1.0-11. [p105]

P. Gilbert and R. Varadhan. *numDeriv: Accurate Numerical Derivatives*, 2015. URL http://CRAN.R-project.org/package=numDeriv. R package version 2014.2.1. [p109]

A. Glynn and K. Quinn. *CausalGAM: Estimation of Causal Effects with Generalized Additive Models*, 2010a. URL http://CRAN.R-project.org/package=CausalGAM. R package version 0.1-3. [p106]

A. N. Glynn and K. M. Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56, 2010b. [p106]

S. Gruber. *tmle: Targeted Maximum Likelihood Estimation*, 2014. URL http://CRAN.R-project.org/package=tmle. R package version 1.2.0-4. [p106]

R. Guiteras, J. Levinsohn, and A. M. Mobarak. Encouraging sanitation investment in the developing world: a cluster-randomized trial. *Science*, 348(6237):903–906, 2015a. [p106, 111, 112]

R. Guiteras, J. Levinsohn, and M. Mobarak. Encouraging sanitation investment in the developing world: A cluster-randomized trial. *Harvard Dataverse; online data*, 2015b. doi: doi/10.7910/DVN/GJDUTV. URL `http://dx.doi.org/10.7910/DVN/GJDUTV`. [p111]

U. Halekoh, S. Højsgaard, and J. Yan. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006. [p105]

S. Højsgaard and R. Halekoh. *geepack: Generalized Estimating Equations Package*, 2016. URL `http://CRAN.R-project.org/package=geepack`. R package version 1.2-0.1. [p105]

Y. Jun. geepack: Yet another R package for generalized estimating equations. *R-News*, 2(3): 12–14, 2002. [p105]

G. Lin, R. Rodriguez, and I. I. SAS. Weighted methods for analyzing missing data with the `GEE Procedure`. *Proceedings of SAS Global Forum*, Washington DC(2014 March 23th-26th): paper 166, 2015. [p105]

L. S. McDaniel and N. Henderson. *geeM: Solve Generalized Estimating Equations*, 2015. URL `http://CRAN.R-project.org/package=geeM`. R package version 0.7.4. [p107]

L. S. McDaniel, N. C. Henderson, and P. J. Rathouz. Fast pure R implementation of gee: Application of the Matrix package. *The R journal*, 5(1):181, 2013. [p105, 107]

K. E. Porter, S. Gruber, M. J. van der Laan, and J. S. Sekhon. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, 7(1):1–34, 2011. [p106]

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. *Biometrics*, 72(4):1066–1077, 2016. [p105, 107, 109, 110]

J. M. Robins, A. Rotnitzky, and L. P. 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. [p105, 106, 113]

J. M. Robins, S. Greenland, and F.-C. Hu. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700, 1999. [p106]

SAS Institute Inc. *SAS/STAT Software, Version 13.2*. Cary, NC, 2015. URL `http://www.sas.com/`. [p105]

O. Sofrygin and M. van der Laan. *tmlenet: Targeted Maximum Likelihood Estimation for Network Data*, 2015. URL `http://CRAN.R-project.org/package=tmlenet`. R package version 0.1-0. [p106]

A. J. Stephens, E. J. Tchetgen Tchetgen, and V. DeGruttola. Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates. *Statistics in medicine*, 31 (10):915–930, 2012. [p105, 107]

E. J. Tchetgen Tchetgen, M. M. Glymour, J. Weuve, and J. Robins. A cautionary note on specification of the correlation structure in inverse-probability-weighted estimation for repeated measures. *Epidemiology*, 23(4):644–646, 2012. [p105]

M. van der Laan. Causal inference for a population of causally connected units. *Journal Causal Inference*, 2(1):1374–1380, 2014a. [p106]

M. van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014b. [p108]

W. M. van der Wal and R. B. Geskus. ipw: an R package for inverse probability weighting. *Journal of Statistical Software*, 43(13):1–23, 2011. [p105, 109]

C. Wang and M. C. Paik. A weighting approach for gee analysis with missing data. *Communications in Statistics-Theory and Methods*, 40(13):2397–2411, 2011. [p109]

M. Wang. *geesmv: Modified Variance Estimators for Generalized Estimating Equations*, 2015. URL http://CRAN.R-project.org/package=geesmv. R package version 1.3. [p109]

S. L. Zeger and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, 1986. [p105]

J. Zetterqvist and A. Sjölander. *drgee: Doubly Robust Generalized Estimating Equations*, 2015. URL http://CRAN.R-project.org/package=drgee. R package version 1.1.3. [p105]

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008. [p107]

Melanie Prague
Department of Biostatistics
Harvard T.H. Chan School of Public Health
655 Huntington Ave
Boston, MA 02115
and
INRIA - INSERM U1219 - SISTM
164 rue Leo Saignat Room 23
33076 Bordeaux Cedex, France
(ORCiD:0000-0001-9809-7848)
melanie.prague@inria.fr

Rui Wang
Department of Biostatistics
Harvard T.H. Chan School of Public Health
655 Huntington Ave
Boston, MA 02115
(ORCiD:0000-0001-5007-193X)
rwang@hsph.harvard.edu

Victor De Gruttola
Department of Biostatistics
Harvard T.H. Chan School of Public Health
655 Huntington Ave
Boston, MA 02115
degrut@hsph.harvard.edu