
by Melanie Prague, Rui Wang and Victor De Gruttola

Abstract

Semi-parametric approaches based on generalized estimating equation (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 (CRTs), we show that other software lead to biased estimation for non-independence working correlation structure. CRTgeeDR solve this problem. We also extend the ability of existing packages to allow augmented Doubly Robust GEE estimation (DR). Simulations 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) (?) 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) (?) and for increasing precision of estimation by appropriate use of baseline covariates (AUG) (?). 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 (?). 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. ? 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 (?). ? 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 (?) and geepack (???) are computationally demanding, the package geeM allows a fast estimation through the use of sparse matrix representation (?). 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 (???) or directly plugged from a user-defined function. These approaches require the missing data process to be known or correctly specified. Some packages, such as drgee (?), 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 CausalGAM (??), but these packages are intended for analysis of observational studies, not CRTs. Finally, targeted maximum likelihood estimation (tMLE) method allows estimation of the marginal additive effect of a treatment (?). It is implemented in the packages tmle (?) and tmlenet (?) for longitudinal and correlated data. Except for ?, 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 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 (?) 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 CRTs comprised of \( n \) clusters or communities, each with \( n_i \) individuals. The cluster sample sizes are assumed fixed and non-informative. Let \( Y_{ij} = \{Y_{ij}\}_{i=1,...,n} \) 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 \( X_{ij} = \{X_{ij}\}_{i=1,...,n}, r=1,...,p \) 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 \( \mu_i = \{\mu_{ij}\}_{j=1,...,n_i} \) 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 \( i \), and consider a model for the mean as follow:
\[
g(\mu_{ij}) = X_{ij}'\beta_X + A_i\beta_A,
\]
where \( g(.) \) is a monotone differentiable link function and \( \beta = (\beta_A, \beta_X) \) is a \( (P + 1) \times 1 \) vector of regression coefficients of interest. In this article, we focus on estimation of the marginal effect of an intervention \( \beta_A \) for a binary outcome using the logit link. We assume the variance is \( \var(Y_{ij}|X_{ij}, A_i) = \phi(\mu_{ij}) \), where \( h(.) \) is the variance function and \( \phi \) is the dispersion parameter. Thus for our specific example, \( \var(Y_{ij}|X_{ij}, A_i) = \phi(\mu_{ij}) \). 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 \( \pi_{ij} \) for individual \( j \) in cluster \( i \), called the propensity score (PS), is: \( \pi_{ij}(X_{ij}, A_i, \eta_W) = P(R_{ij} = 1|X_{ij}, A_i). \) The parameters \( \eta_W \) are nuisance parameters and must be estimated.

IPW in CRTs

In presence of rMAR outcome, as in (?), we estimate \( \beta \) by using inverse probability weighted generalized estimating equation (IPW). Therefore, we must include a weight matrix \( W_i \) to the usual GEE, that is:
\[
W_i(X_{ij}, A_i, \eta_W) = \text{diag}\left(\frac{R_{ij}}{\pi_{ij}(X_{ij}, A_i, \eta_W)}\right)_{j=1,...,n_i}.
\]
This matrix \( W_i(X_{ij}, A_i, \eta_W) \), denoted simply as \( W_i \) in the following, adjusts the contribution of each individual in a given cluster by up weighting 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:
\[
0 = \sum_{i=1}^{n} D_i V_i^{-1} W_i(Y_i - \mu_i), \tag{1}
\]
where \( D_i = \partial \mu_i / \partial \beta \) is a derivative matrix and \( V_i \) is the working covariance matrix for the response \( Y_i \). In particular, \( V_i = \phi F_i^{1/2} C(\alpha) F_i^{1/2} \), where \( F_i^{1/2} = \text{diag}(h(\mu_{ij}))_{j=1,...,n_i} \) and \( C(\alpha) \) is the working
correlation structure with non-diagonal terms $\alpha$. For example, for an independence correlation structure $\alpha$ is zero; for exchangeable structure, all the elements of $\alpha$ are identical. Parameters $\alpha$ could also depend on the treatment assignment $C(A_i)$ but we do not consider this possibility in our implementation. In the package CRTgeeDR, we estimate the $\alpha$ and $\phi$ parameters using moment estimators from the Pearson residuals and the Pearson Chi-Square statistic as in geeM (?) also described in ??. In the absence of missing data, $W_i = I$ is set to identity, and the standard GEE is performed by CRTgeeDR.

In existing packages such as geepack, the Equation 1 is implemented as $0 = \sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i)$, with $V_i^{-1} = \phi \pi_i^{-1/2} W_i^{1/2} C(\alpha) W_i^{1/2} \pi_i^{-1/2}$ to ensure the fast invertibility of $V_i$. It is easy to verify that when an independence correlation structure is used $C(\alpha) = I$, 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 Web-Supplementary Material for 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 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 (??); we denote this estimator the AUG. They have also been extended to accommodate missing data using an approach based on the IPW 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 non at random, none of the methods implemented in CRTgeeDR is adequate.

**The R package CRTgeeDR**

**The main function for estimation in the package CRTgeeDR**

The call function for performing estimation is geeDREstimation:

```R
ggeeDREstimation(formula, id, data = parent.frame(), family = gaussian, +    corstr = "independence", Mv = 1, corr.mat = NULL, init.bet = NULL, +    tol=le-05, print.log = FALSE, nameTRT = "TRT", nameMISS = "MISS", +    nameY = "OUTCOME", sandwich = TRUE, sandwich.nuisance = FALSE, +    fay.adjustment = FALSE, fay.bound = 0.75, aug = NULL, pi.a = 1/2,
```

The R Journal Vol. XX/YY, AAAA 20ZZ  ISSN 2073-4859
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, geeOREstim1ation allows the implementation of standard GEE, the IPW, the AUG and the DR approaches. The algorithm is defined as follow:

1. **Determine the PS:** \( \pi_{ij}(X_{ij}, A_i, \eta_W) = P(R_{ij}|X_{ij}, A_i), \pi_{ij} \) for short. Either the \( \pi_{ij} \) are known from prior analysis or by design and the weights can be specified directly in the weights argument. Alternatively one can compute the PR by fitting a logistic regression of \( R_{ij} \) on \( (X_{ij}, A_i) \). 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 variables 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:** \( B_{ij}(X_{ij}, A_i = a) = E(Y_{ij}|A_j = a, X_{ij}) \). When the \( B_i \) 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 \( X_{ij} \) within each treatment group. In this case, the OM regression formula can be directly entered in model.augmentation.trt and model.augmentation.ctrl. A glm with logit link function 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 p1.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 (7). 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 \( \phi \) 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 \( \hat{\phi}^{(0)} \) and by setting \( \phi^{(0)} = 1 \) and \( \alpha^{(0)} = 0 \).

5. **Enter/Continue the iterative procedure:**

   (a) Use the fit from \( \hat{\beta}^{(n)} \) 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 \( \alpha^{(n+1)} \).

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

   \[
   \hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} - \left[ \frac{\partial \Phi(Y_i, R_i, A_i, X_{ij}, \hat{\beta}, \eta_W, \eta_B)}{\partial \beta} \right]^{-1} \Phi(Y_i, R_i, A_i, X_{ij}, \hat{\beta}^{(n)}, \eta_W, \eta_B).
   \]

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

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 \( \eta_W \) in the PS and \( \eta_B \) in the OM) sandwich estimators of the variance are provided, see ? 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...
We focus on a simple example to illustrate that, even in very simple cases, estimators implemented in the package `numDeriv` (?); 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 (?) setting the argument fay.adjustment to TRUE. Its implementation is derived from the function gee.var.fg in the package `geesmv` (?).

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 ? for theoretical demonstrations. User may want to check the adequacy of the selected OM model to the data by using the function getOMplot, which provide 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 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` (?) or other approaches such as described in ?. 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 presence of imbalance in baseline covariates have already been demonstrated in ?. 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 \( V^{-1} = \phi F^{1/2} W^{1/2} C(s) W^{1/2} F^{1/2} \) 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: \( X_{ij} \sim N(2, 1) \). 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 \( b_i \sim B(0.05) \). Data generation process is as follow:

\[
\begin{align*}
\logit[P(Y_{ij} = 1|A_i, X_{ij})] &= -0.5 + 0.3A_i + 0.4X_{ij} + 0.4X_{ij}A_i + b_i, \\
\logit[P(R_{ij} = 1|A_i, X_{ij})] &= 4.0 - 0.3A_i - 0.8X_{ij} - 0.8X_{ij}A_i. 
\end{align*}
\]

(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 odd-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:

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

The true OR is given by \( \frac{1}{R} \sum_{k=1}^{R} \text{OR}_k = 2.56 \) 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: DRI 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.
Table 1: Comparison of the standard GEE, the IPW, the AUG and the DR analysis with the package CRTgeeDR, geepack and geeM using independence and exchangeable working correlation structure. True value for the parameter $\beta_A$ is 0.91 (OR=2.56). The bias, the empirical and the estimated standard errors (SE) and the coverages for parameter $\hat{\beta}_A$ are computed over 10000 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 ≤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 ?, 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 (?). All the code and data associated with this study are available on dataverse, see url in ?.

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 Web-Supplementary Material. To illustrate the use of the package CRTgeeDR, we provide instructions for the DR estimator:

\begin{verbatim}
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)
\end{verbatim}

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

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

Table 2: Description of the Sanitation dataset from (?) 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.
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 analyses. 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 (\( \eta_W \) and \( \eta_B \)). Information about the PS and the OM can be obtained by using the following command lines:

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

<table>
<thead>
<tr>
<th>Method</th>
<th>( \beta_A )</th>
<th>Sandwich SE</th>
<th>Nuis-adj. SE</th>
<th>( \exp(\beta_A) ) OR</th>
<th>IC(_{min})</th>
<th>IC(_{max})</th>
<th>Unadj. p-value</th>
<th>Nuis-adj. time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEE</td>
<td>0.19</td>
<td>0.171</td>
<td>-</td>
<td>1.21</td>
<td>0.87</td>
<td>1.69</td>
<td>0.262</td>
<td>-</td>
</tr>
<tr>
<td>IPW</td>
<td>0.19</td>
<td>0.182</td>
<td>0.219</td>
<td>1.21</td>
<td>0.79</td>
<td>1.86</td>
<td>0.290</td>
<td>0.386</td>
</tr>
<tr>
<td>AUG</td>
<td>0.45</td>
<td>0.141</td>
<td>0.176</td>
<td>1.57</td>
<td>1.12</td>
<td>2.22</td>
<td>0.001</td>
<td>0.010</td>
</tr>
<tr>
<td>DR</td>
<td>0.44</td>
<td>0.143</td>
<td>0.183</td>
<td>1.55</td>
<td>1.08</td>
<td>2.21</td>
<td>0.002</td>
<td>0.016</td>
</tr>
</tbody>
</table>

**Marginal mean model:** \( \logit(\mu_{ij}) = \beta_0 + \beta_A A_i \).

**PS:** \( \logit(P(R_{ij}|A_i,X_{IND}^{ij},X_C^{ij})) = \gamma_0 + \gamma_A A_i + \sum_{k=1}^{10} \gamma_{IND}^{k} X_{IND}^{ijk} + \sum_{k=1}^{5} \gamma_{C}^{k} X_C^{ijk} \).

**OM:** \( \logit(P(Y_{ij}|A_i = a,X_{IND}^{ij},X_C^{ij}) = \xi_0 + \sum_{k=1}^{10} \xi_{IND}^{a,k} X_{IND}^{ijk} + \sum_{k=1}^{5} \xi_{C}^{a,k} X_C^{ijk} \) 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 (?) using the standard GEE, the IPW adjustment (IPW and DR), and the augmentation for imbalance (AUG and DR) assuming outcomes are rMAR.
Conclusion

We demonstrated that the IPW can be biased in CRTs if the weights are not implemented as described in ? 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 observation 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 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. 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 capability 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

Melanie Prague  
Department of biostatistics  
Harvard T.H. Chan School of Public Health  
655 Huntington Ave  
Boston, MA 02115  
(ORCiD:0000-0001-9809-7848)  
mprague@hsph.harvard.edu

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