DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 753–784, http://dx.doi.org/10.2478/JOS-2018-0035

Generalized Method of Moments Estimators for Multiple Treatment Effects Using Observational Data from Complex Surveys

Bin Liu[1], *Cindy Long Yu*[2], *Michael Joseph Price*[2], *and Yan Jiang*[3]

In this article, we consider a generalized method moments (GMM) estimator to estimate treatment effects defined through estimation equations using an observational data set from a complex survey. We demonstrate that the proposed estimator, which incorporates both sampling probabilities and semiparametrically estimated self-selection probabilities, gives consistent estimates of treatment effects. The asymptotic normality of the proposed estimator is established in the finite population framework, and its variance estimation is discussed. In simulations, we evaluate our proposed estimator and its variance estimator based on the asymptotic distribution. We also apply the method to estimate the effects of different choices of health insurance types on healthcare spending using data from the Chinese General Social Survey. The results from our simulations and the empirical study show that ignoring the sampling design weights might lead to misleading conclusions.

Key words: Observational data; propensity score; semiparametric; treatment effects; two-phase sampling design.

1. Introduction

Observational data from a complex survey has increasingly become useful for causal inference because they can provide timely results with low cost. Survey data contains information on the treatment selections, which enables us to estimate the effects of treatments that cannot feasibly be evaluated with a randomized trial. In a survey, a treatment can be broadly defined as one of the survey questions, for example whether or not an individual has quit smoking, how often an individual does a physical exam, or what types of health insurance an individual has chosen. We can use the existing survey data to estimate effects of those treatments on health care spending, even if we cannot randomize the health behavior or the health insurance enrollment of an individual. Also because a well-designed survey sample is often a good representative of the target population, the treatment effect results can be generalized to the target population level if the survey weights are appropriately incorporated. Propensity score methods are well-established statistical methods to remove treatment selection bias in observational studies if the

[1] Ant Financial, Hangzhou, China. Email: lb88701@alibaba-inc.com
[2] Iowa State University – Department of Statistics, Ames, Iowa 50011, United States. Emails: cindyyu@iastate.edu and michael.price@pioneer.com
[3] Renmin University of China - School of Statistics and The Center for Applied Statistics, Beijing, China. Email: jiangyan@ruc.edu.cn

selection probability model is correctly specified (Rosenbaum and Rubin 1983). Many observational data sets have multiple treatment options. In order to handle the complexity in multiple treatment groups, theoretical results support using the inverse of the estimated treatment selection probabilities as weights to adjust for selection bias and attain asymptotic efficiency (Hahn 1998; Hirano et al. 2003; Cattaneo 2010). This kind of estimator is called inverse probability weighted (IPW) estimator, and the estimated selection probabilities are called propensity scores. We also consider IPW estimators in this article to address the potential confounding in observational studies. However, it is very common that people ignore survey weights in observational data when using the IPW estimators yet claim that the estimated treatment effects are generalizable to the target population, causing misleading guidance in causal inference. Failure to properly account for the complex survey design may lead to biased treatment effect estimates and incorrect variance estimation.

Several authors have emphasized the importance of incorporating survey weights in their IPW estimators, for example DuGoff et al. (2014), Zanutto (2006), Ashmead (2014), and Ridgeway et al. (2015). The general idea is to multiply the inverse of the estimated propensity scores by the sampling design weights. However most of the papers, except for Ashmead (2014), do not provide theoretical justification for such survey adjusted estimators, and variance estimation is seldom discussed. Yu et al. (2013) proposes a semiparametric two-phase regression estimator to estimate marginal mean treatment effects in observational data sets from complex survey designs. This article considers a more general set up in which parameters of interest are defined through estimation equations, and uses the generalized method of moments (GMM) for parameter estimation. Similarly to Yu et al. (2013), this article draws a connection between the two-phase sampling in survey statistics and the estimation of treatment effects from an observational database. The observational data set, denoted as A_1 (with size n), is considered as a first-phase sample from a finite population, according to a known sampling probability π_{1i} for subject i. The second-phase sampling is a partitioning of the first-phase sample (observational data set) into mutually exclusive and self-selected treatment groups, $A_{21}, \ldots, A_{2G}$, where G is the number of treatments. This partitioning in the second-phase can be viewed as a multinomial sampling in survey statistics, and its self-selection probabilities π_{2ig} for subject i into group g ($g = 1, \ldots, G$) can be estimated using the semiparametric approach in Cattaneo (2010).

Our article differs from DuGoff et al. (2014), Zanutto (2006), Ashmead (2014) and Ridgeway et al. (2015) in the following ways. (i) Their papers consider two treatments, while our article deals with multi-level treatment selection. (ii) In their work, the propensity scores are estimated using a parametric linear logistic regression, while our propensity scores, that is π_{2ig} in our situation, are estimated through a semiparametric approach. Thus, our approach should be more robust to the misspecification of the selection probability model. (iii) In their work, the parameters of interest are treatment means. We are interested in estimating treatment specific parameters defined through estimation equations. In addition to providing generality, defining parameters through estimation equations can facilitate variance estimation. For example, if a parameter is a function of means, such as correlation or domain mean (see more details in Subsection 2.1), the variance estimation of GMM estimators for such parameter scan be easily calculated

through the sandwich formula associated with the asymptotic variance for a GMM estimator. Ashmead (2014) also utilizes estimation equations in their weighting estimator.

This article also differs from Yu et al. (2013) in the following aspects. We extend Yu et al. (2013), which only focuses on estimating marginal treatment means, to estimate parameters defined through estimation Equations (see $\hat{\boldsymbol{\theta}}_g^{(1)}$ in Subsection 2.3). This article also proposes the second estimator to gain efficiency by incorporating the first phase and second phase means of covariates into the estimation equations (see $\hat{\boldsymbol{\theta}}_g^{(2)}$ in Subsection 2.3). This is similar to the effect of calibrating the second phase means of covariates to their first phase means seen in the optimal two-phase regression estimator discussed in Fuller (2009). Additionally, Yu et al. (2013) assumes sample missing at random (SMAR), which is commonly used in literature, while this article considers population missing at random (PMAR), the framework proposed in Berg et al. (2016) (see more details in Subsection 2.1). It makes sense to use PMAR assumption in the context of casual inference study using observation dataset. We discuss situations when PMAR holds but SMAR fails, and argue that when it happens survey weights should be included in the estimation of π_{2ig}, that is the propensity scores.

We provide theoretical justification for our estimator in a combined framework of a finite population and a superpopulation, and also propose variance estimators. We demonstrate the validity of our estimator through simulation studies, and show that the estimator that ignores the design weights might be subject to biases. We also explore the feasibility of our method using data from the Chinese General Social Survey to estimate the effects of different choices of health insurance types on health care spending. The article is organized as below. Section 2 introduces the framework and the proposed estimators. Section 3 presents an asymptotic normality and variance estimation. Simulation studies and an empirical study are reported in Sections 4 and 5 respectively. Section 6 concludes. Appendix collects the conditions and a sketch of the proof for the main theorem in the article.

2. Proposed Estimators

In this section, we introduce our estimators. Subsection 2.1 discusses the basic set-up, Subsection 2.2 introduces the semiparametric approach for estimating the self-selection probabilities, and Subsection 2.3 proposes the estimators.

2.1. Basic Setup

Let U be a finite population with size N containing $(\mathbf{Y}_i, Z_i)$, where $i = 1, \ldots, N$ indexes a subject, Z_i is a covariate variable, and $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{iG}]^T$ is a vector of potential outcomes for G different treatments depending on covariate Z_i. Let δ_{1i} be the sampling indicator from the survey design, defined by $\delta_{1i} = 1$ if unit i is selected into A_1 and zero otherwise. Let π_{1i} and π_{1ij} be the first and second order inclusion probabilities of the sampling design, defined as,

$$[\pi_{1i}, \pi_{1ij}] = [Prob(\delta_{1i} = 1), Prob(\delta_{1i} = 1, \delta_{1j} = 1)].$$

We assume the sampling weights are appropriately adjusted for any nonresponse. If the weights are adjusted due to nonresponse, the method can be used but with awareness of

that the variation from estimating $\hat{\pi}_{1i}$ is not accounted for. Let δ_{2ig} $(g = 1, \ldots, G)$ be the self-selection indicator of subject i selecting treatment g, defined by $\delta_{2ig} = 1$ if unit i selects treatment g and zero otherwise. The self-selection process leads to the partitioning in the second phase. Assume conditioning on a covariate X_i, the self-selection indicators $\boldsymbol{\delta}_{2i} = \left[\delta_{2i1}, \ldots, \delta_{2iG}\right]$ follow a multinomial distribution with probabilities,

$$\pi_{2ig} = Prob(\delta_{2ig} = 1|X_i), \quad \text{for} \quad g = 1, \ldots, G, \tag{1}$$

that is for any subject i,

$$\boldsymbol{\delta}_{2i} = \left[\delta_{2i1}, \ldots, \delta_{2iG}\right] \sim multinomial\left(1; \pi_{2i1}, \ldots, \pi_{2iG}\right),$$

where $\sum_{g=1}^{G} \pi_{2ig} = 1$ for any i, and $\boldsymbol{\delta}_{2i}$ is independent of $\boldsymbol{\delta}_{2j}$ for any subjects $i \neq j$. Here covariates Z_i and X_i can be totally different, or can have overlap. We use separate notations in order to emphasize that the outcome response variables $\mathbf{Y}_i$ and the self-selection indicators $\boldsymbol{\delta}_{2i}$ can depend on different sets of covariates. We discuss how to identify Z_i and X_i practically in Section 4. Both Z_i and X_i have compact supports and are observed in A_1. They are written to be univariate forms in order to reduce notation burden. It is straightforward to extend to multivariate covariates, which are considered in the simulation studies and the empirical study of this article. We suppose that $\left(\mathbf{Y}_i, \delta_{1i}, \delta_{2i}, X_i, Z_i\right); i = 1, \ldots, N$ are identically independently distributed (i.i.d.) generated from a superpopulation ξ.

In the context of simple random sampling, a common missing at random (MAR) assumption is $\mathbf{Y}_i \perp \boldsymbol{\delta}_{2i}|(X_i, Z_i)$. With this MAR assumption, the selection bias can be removed by applying the propensity score method (Rosenbaum and Rubin 1983; Hirano et al. 2003). However, in the context of a complex survey, unequal probabilities of sampling can complicate the relationship between $\mathbf{Y}_i$, (X_i, Z_i), $\boldsymbol{\delta}_{2i}$ and the sample inclusion indicator δ_{1i}. Even if

$$\mathbf{Y}_i \perp \boldsymbol{\delta}_{2i}|(X_i, Z_i), \tag{2}$$

holds for a specific superpopulation model,

$$\mathbf{Y}_i \perp \boldsymbol{\delta}_{2i}|\{(X_i, Z_i), \delta_{1i} = 1\}, \tag{3}$$

may not hold. Following Berg et al. (2016), we call Assumption (2) population missing at random (PMAR), and Assumption (3) sample missing at random (SMAR) to emphasize it depends on the realized sample (that is conditional on $\delta_{1i} = 1$). The SMAR has been used previously (Pfefferman 2011 and Little 1982). However, it is natural to consider PMAR in our context because the mechanisms underlying the selection propensity are conceptualized as inherent characteristics of the subjects in the population. For example, whether or not a person decides to stop smoking heavily depends on this person's perseverance and personality type; what types of insurance a person has chosen depends on the nature of this person's work. In these examples, the self-selection probabilities depend on subjects' inherent characteristics that have nothing to do with whether or not the subjects were selected into the survey that was typically designed for other general purposes. Berg et al. (2016) also provides examples of situations in which PMAR may be considered reasonable. They argue that if both PMAR and SMAR hold, weights are not needed in their imputation model; however if PMAR holds but SMAR fails, it is necessary to include weights to produce consistent estimators. A situation in which PMAR holds

while SMAR does not can arise if a design variable omitted from the first phase sample is related to both the sampling inclusion probabilities and the response variable. An example of such a design variable is location in a situation where design strata are functions of location, the location is correlated with the response variable, but the specific location is masked from the analyst because of concerns associated with confidentiality. Using Lemma 1 of Berg et al. (2016), we identify the following two conditions of the sampling and the self-selection mechanisms for which PMAR implies SMAR: (1) $\delta_{1i} \perp \mathbf{Y}_i | (X_i, Z_i), \boldsymbol{\delta}_{2i}$; or (2) $\boldsymbol{\delta}_{2i} \perp \left(\mathbf{Y}_i, \delta_{1i}\right) | \left(X_i, Z_i\right)$. The first condition states that the sampling mechanism is noninformative given covariates (X_i, Z_i) within all the second phase self-selected groups A_{2g}. The second condition states that the self-selection mechanism is independent of either $\mathbf{Y}_i$ or sample inclusion given $\left(X_i, Z_i\right)$. Like Berg et al. (2016), we suggest to include survey weights into the estimation of the self-selection probabilities π_{2ig} when SMAR fails (see Subsection 2.2). In our simulation studies, we consider both noninformative sampling (Condition (1) above holds), and informative sampling (Condition (1) above fails).

The true parameter of interest, $\boldsymbol{\theta}_g^0$ $(g = 1, \ldots, G)$, is a d_θ-dimensional vector satisfying,

$$E\left[\mathbf{m}_g\left(Y_{ig}, Z_i; \boldsymbol{\theta}_g\right)\right] = 0, \tag{4}$$

in the superpopulation, where $\mathbf{m}_g(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$, hereafter denoted as $\mathbf{m}_{ig}(\boldsymbol{\theta}_g)$ to save space, is an r-dimensional function with $r \geq d_\theta$. Sometimes in addition to treatment marginal means, people might be interested in estimating treatment correlations or treatment domain means. For example in our empirical study, it is interesting to understand whether the correlations between annual medical expenditure and age (or household income) differ significantly across different health insurance type groups; or whether the means of annual medical expenditure for very sick people (domain means) are significantly different across health insurance type groups. The parameter defined through Equation (4) includes treatment correlations and treatment domain means as special cases. More specifically, if the parameter of interest is $\boldsymbol{\theta}_g^0 = \left[P_g, \mu_g, \sigma_g^2, R_g\right]^T$, where $P_g = Prob(Y_{ig} \leq C)$ for some C, $\mu_g = E(Y_{ig})$, $\sigma_g^2 = Var(Y_{ig})$ and $R_g = Corr(Y_{ig}, Z_i)$, then the estimation equation can be defined as,

$$\begin{aligned}\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = \Big[& 1_{Y_{ig} \leq C} - P_g, Y_{ig} - \mu_g, (Y_{ig} - \mu_g)^2 - \sigma_g^2, (Y_{ig} - \mu_g)(Z_i - \mu_z) \\ & -R_g \sqrt{\sigma_g^2} \sqrt{\sigma_z^2}, Z_i - \mu_z, (Z_i - \mu_z)^2 - \sigma_z^2 \Big]^T.\end{aligned} \tag{5}$$

If the parameter of interest is a treatment specific domain mean, $\boldsymbol{\theta}_g^0 = E(Y_{ig} | Z_i \leq C)$, then the estimation equation can be written as,

$$\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = [Y_{ig} 1_{Z_i \leq C} - \boldsymbol{\theta}_g P_z, 1_{Z_i \leq C} - P_z]^T. \tag{6}$$

Here in both examples, μ_z, σ_z^2 or P_z are all nuisance parameters.

2.2. *Semiparametric Estimation of* π_{2ig}

Because of the difficulty in specifying a parametric form for π_{2ig} and the constraint, $\sum_{g=1}^{G} \pi_{2ig} = 1$, we adopt the semiparametric method in Cattaneo (2010) to estimate π_{2ig}.

Let $\{r_k(X_i)\}_{k=1}^{\infty}$ be a sequence of known approximating functions, and assume that the generalized logit of π_{2ig} can be approximated by $R_K(X_i)^T\gamma_{g,K}$ for $K = 1, 2, \ldots$, where $R_K(X_i) = [r_1(X_i), r_2(X_i), \ldots, r_K(X_i)]^T$ and $\gamma_{g,K}$ is a vector of the real-valued coefficients of $R_K(X_i)$ for the g-th treatment selection. Let an estimator of the $K \times G$ matrix $\gamma_K = \left[\gamma_{1,K}, \gamma_{2,K}, \ldots, \gamma_{G,K}\right]$ be,

$$\hat{\gamma}_K = [\hat{\gamma}_{1,K}, \hat{\gamma}_{2,K}, \ldots, \hat{\gamma}_{G,K}] = \underset{\gamma_K | \gamma_{1,K} = \mathbf{0}_K}{argmax} \sum_{i \in A_1} b_i w_{1i} \sum_{g=1}^{G} \delta_{2ig} log\left[\frac{e^{R_K(X_i)^T\gamma_{g,K}}}{\sum_{g=1}^{G} e^{R_K(X_i)^T\gamma_{g,K}}}\right], \quad (7)$$

where $w_{1i} = \pi_{1i}^{-1}$, and $\mathbf{0}_K$ represents a $K \times 1$ zero vector used to constrain the sum $\sum_{g=1}^{G} \hat{\pi}_{2ig} = 1$. The estimated self-selection probabilities are

$$\begin{aligned} \hat{\pi}_{2ig} &= \frac{e^{R_K(X_i)^T\hat{\gamma}_{g,K}}}{1 + \sum_{g=2}^{G} e^{R_K(X_i)^T\hat{\gamma}_{g,K}}} \quad \text{for g} = 2, 3, \ldots, \text{G} \\ &= \left(1 + \sum_{g=2}^{G} e^{R_K(X_i)^T\hat{\gamma}_{g,K}}\right)^{-1} \quad \text{for g} = 1. \end{aligned} \quad (8)$$

This solution is that of multinomial logistic regression where the probability for each g is approximated using a linear combination of the series of the approximating functions $R_K(X_i)$. Condition B in the Appendix specifies assumptions about $R_K(X_i)$, π_{2ig} and K to ensure $\hat{\pi}_{2ig}$ converges to π_{2ig} fast enough. Examples of $R_K(X_i)$ include a cubic polynomial basis, $R_K(X_i) = \left[1, X_i, X_i^2, X_i^3\right]^T$, or a quadratic spline basis with q knots $R_K(X_i) = \left[1, X_i, X_i^2, (X_i - \kappa_1)_+^2, \ldots, (X_i - \kappa_q)_+^2\right]^T$ where $(t)_+ = t$ if $t > 0$ and 0 otherwise, and $\kappa_1, \ldots, \kappa_q$ are knots in the compact support of X_i.

The b_i in Equation (7) is a user-specified constant that represents the properties of the sampling and the self-selecting mechanism. As discussed in Subsection 2.1, PMAR assumption does not necessarily imply SMAR assumption. If one believes SMAR assumption holds, then one can set $b_i = w_{1i}^{-1}$, which leads to unweighted estimation of $\hat{\pi}_{2ig}$. If SMAR is not satisfied, the unweighted estimator may lead to bias, and setting $b_i = 1$ is one way to attain an approximately unbiased estimator, see Berg et al. (2016) for further discussion of the choice of b_i. If it is difficult to verify SMAR assumption, we suggest to use the conservative choice of $b_i = 1$, which leads to consistent estimators under PMAR without requiring SMAR.

2.3. *Proposed Estimators*

Since the true parameter of interest $\boldsymbol{\theta}_g^0$ is defined through an estimation equation in (4), the GMM method with propensity scores is used for estimation. It is common that people simply ignore the sampling design weights in the first-phaseand calculate a naive estimator as,

$$\hat{\boldsymbol{\theta}}_g^{nw} = \underset{\boldsymbol{\theta}_g}{\arg\min} \left[\bar{\mathbf{m}}_g^{nw}(\theta_g)\right]^T \left[\bar{\mathbf{m}}_g^{nw}(\boldsymbol{\theta}_g)\right], \quad (9)$$

where

$$\bar{\mathbf{m}}_g^{nw}(\boldsymbol{\theta}_g) = \frac{1}{n}\sum_{i\in A_{2g}} \frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\hat{\pi}_{2ig}}. \tag{10}$$

Here the superscript 'nw' means no weight. The estimator $\hat{\boldsymbol{\theta}}_g^{nw}$ ignores the sampling weights by applying equal weights to the estimation equations in (10). Although it uses the propensity score $\hat{\pi}_{2ig}$ to adjust for selection biases in the second-phase, it does not account for the survey design in the first-phase, which might lead to biases and incorrect variance estimation when estimating the treatment effect parameters on the population level. This is demonstrated in the simulation studies of Section 4. Both Ridgeway et al. (2015) and Yu et al. (2013) analytically quantify biases caused by ignoring the survey weights in complex survey.

In order to obtain a consistent estimator for $\boldsymbol{\theta}_g^0$, the first-phase survey weights need to be included into the estimation equation. We propose the following GMM estimator,

$$\hat{\boldsymbol{\theta}}_g^{(1)} = \arg\min_{\boldsymbol{\theta}_g} [\bar{\mathbf{m}}_{2\pi g}(\theta_g)]^T [\bar{\mathbf{m}}_{2\pi g}(\theta_g)], \tag{11}$$

where

$$\bar{\mathbf{m}}_{2\pi g}(\theta_g) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i} \frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\hat{\pi}_{2ig}}. \tag{12}$$

In order to improve efficiency, one can incorporate the information from covariate Z_i that is potentially correlated with the outcome responses into the estimation equations. We propose the second GMM estimator as,

$$\left(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z\right) = \arg\min_{(\boldsymbol{\theta}_g, \mu_z)} [\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)]^T \hat{\boldsymbol{\Sigma}}_{Hg}^{-1}(\boldsymbol{\theta}_g, \mu_z)[\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)], \tag{13}$$

where

$$\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z) = [\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g),\ \bar{z}_{2\pi g}(\mu_z),\ \bar{z}_{1\pi}(\mu_z)]^T, \tag{14}$$

$$\bar{z}_{2\pi g}(\mu_z) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i} \frac{Z_i - \mu_z}{\hat{\pi}_{2ig}} \quad \text{and} \quad \bar{z}_{1\pi}(\mu_z)^T = \frac{1}{N}\sum_{i\in A_1} w_{1i}(Z_i - \mu_z). \tag{15}$$

$\hat{\mu}_z$ is an estimator for the nuisance parameter $\mu_z^0 = E(Z_i)$ and $\hat{\boldsymbol{\Sigma}}_{Hg}(\boldsymbol{\theta}_g, \mu_z)$ is the variance estimator of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$, which depends on the joint inclusion probabilities and is defined in (36) of Subsection 3.2. The estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ in (13) is connected to a two phase sampling extension of the design unbiased difference estimator proposed by Särndal et al. (1992) and Breidt et al. (2005) when $\bar{\mathbf{m}}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$.

Remark 1: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$ and $X_i = Z_i$, the estimator $\hat{\boldsymbol{\theta}}_g^{(1)}$ in (11) is asymptotically equivalent to the regression estimator proposed in Yu et al. (2013).

Remark 2: The estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ in (13) is more efficient than the estimator $\hat{\boldsymbol{\theta}}_g^{(1)}$ in (11). The supplemental file provides a sketch of proof to show that $\hat{\boldsymbol{\theta}}_g^{(2)}$ is the most efficient estimator among the class of estimators $\hat{\boldsymbol{\theta}}_g^{a}$ that use any fixed positive definite matrix $\mathbf{A}$ in the quadratic form minimization, that is $\hat{\boldsymbol{\theta}}_g^{a}$ is defined as

$$\left(\hat{\boldsymbol{\theta}}_g^{a}, \hat{\mu}_z^{a}\right) = \underset{(\boldsymbol{\theta}_g, \mu_z)}{\arg\min}\, [\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)]^T \mathbf{A}^{-1} [\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)]. \tag{16}$$

If the matrix is an identity matrix, then $\hat{\boldsymbol{\theta}}_g^{a}$ obtained in (16) is equivalent to $\hat{\boldsymbol{\theta}}_g^{(1)}$. Therefore $\hat{\boldsymbol{\theta}}_g^{(1)}$ is expected to be less efficient than $\hat{\boldsymbol{\theta}}_g^{(2)}$, which has been confirmed by the simulation studies in Section 4.

Remark 3: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ corresponds to the optimal two phase regression estimator discussed in Fuller (2009) (Theory 2.2.4). The optimality in Fuller (2009) is in terms of achieving the minimum variance for the limiting distribution of design consistent estimators of the form, $\bar{Y}_{2p,reg} = \bar{Y}_{2\pi} + \left(\bar{Z}_{1\pi} - \bar{Z}_{2\pi}\right)\hat{\beta}$, where $[\bar{Y}_{2\pi}, \bar{Z}_{2\pi}] = \left(\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i}^{-1}\right)^{-1} \sum_{i\in A_2}\left(\pi_{1i}^{-1}\pi_{2i}^{-1}\right)[Y_i, Z_i]$, $\bar{Z}_{1\pi} = \left(\sum_{i\in A_a} \pi_{1i}^{-1}\right)^{-1} \sum_{i\in A_a} \pi_{1i}^{-1} Z_i$, and π_{1i} (or A_1) and π_{2i} (or A_2) are the first phase and the second phase sampling probabilities (or samples). The efficiency gain of $\bar{Y}_{2p,reg}$ over $\bar{Y}_{2\pi}$ is similar to the effect of calibrating the second phase covariate mean $\bar{Z}_{2\pi g}$ to its first phase mean $\bar{Z}_{1\pi}$.

Remark 4: It can be shown that when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$ and $Z_i \equiv 1$, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ coincides analytically with the weighting estimator discussed in Ashmead (2014) except that the propensity scores in Ashmead (2014) are estimated using a parametric logistic regression.

Remark 5: When the population mean of Z_i is available, the estimator $\hat{\boldsymbol{\theta}}_g^{(2)}$ can be easily extended to incorporate this additional information. For example, this case can occur when there are some demographic variables available on the population level. The extended estimator can be obtained by adding one more moment $\bar{z}_N(\mu_z) = N^{-1}\sum_{i\in U}(Z_i - \mu_z)$ into the $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in Equation (14). Efficiency gain should be expected since this estimator uses more information on the population level. By viewing the problem as a two-phase sampling problem, the method can be readily extended to multiple sampling phases. This extension is useful because the database A_1 can come from alarger sample within the database. This case covers the common situations where detailed treatment and outcome data is available for only a subsample of the data such as a subsample with medical chart adjudication of claims records or a subsample constructed by merging multiple sources of claims records and electronic medical records.

3. Asymptotic Normality and Variance Estimation

Since $\hat{\boldsymbol{\theta}}_g^{(1)}$ can be written as a special case of $\hat{\boldsymbol{\theta}}_g^{(2)}$, in Subsection 3.1 we derive the asymptotic normal distribution for $\hat{\boldsymbol{\theta}}_g^{(2)}$ only, and in Subsection 3.2 provide a linearized variance estimator for $\hat{\boldsymbol{\theta}}_g^{(2)}$. Subsection 3.3 gives a replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$.

3.1. Asymptotic Normality of $\hat{\boldsymbol{\theta}}_g^{(2)}$

The asymptotic normality of $\hat{\boldsymbol{\theta}}_g^{(2)}$ is established in Theorem 1 by combining two randomizations from the finite population level and the superpopulation level. For the finite population level, we consider a sequence of samples and finite populations indexed by N, where the sample size $n \to \infty$ as $N \to \infty$ (Isaki and Fuller 1982). To define the regularity conditions, we introduce the notation $\mathcal{F}_N$ to represent an element of the sequence of finite population with size N. To distinguish between the two randomizations, we use the notation "$|\mathcal{F}_N$" to indicate that the reference distribution is with respect to repeated sampling conditional on the finite population size N. For example, $E(\cdot\,|\mathcal{F}_N)$ and $V(\cdot\,|\mathcal{F}_N)$ denote the conditional mean and variance with respect to the randomization generated from repeated sampling from $\mathcal{F}_N$. And we use $E_\xi(\cdot)$, $Var_\xi(\cdot)$ and $Cov_\xi(\cdot,\cdot)$ to denote mean, variance and covariance with respect to the randomization from the superpopulation ξ. The proof of Theorem 1 uses a result given in Theorem 1.3.6 of Fuller (2009) that shows how to combine two asymptotic normalities from the finite population and the superpoulation levels. Because of the importance of this theorem to our results, we state this theorem as Fact 1:

Fact 1 (Theorem 1.3.6 of Fuller 2009): Suppose θ_0 is a true parameter on a superpopulation level, θ_N is its analogous part on a finite population level, and $\hat{\theta}$ is an estimator of θ_0 calculated from a sample. If $(\hat{\theta}-\theta_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,V_{11})$ almost surely (a.s.) and $(\theta_N-\theta_0) \xrightarrow{\mathcal{L}} N(0,V_{22})$, then, $(\hat{\theta}-\theta_0) \xrightarrow{\mathcal{L}} N\ (0,V_{11}+V_{22})$. Here $(\hat{\theta}-\theta_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,V_{11})\ a.s.$ means that $\hat{\theta}-\theta_N$ converges in a distribution to a random variable with the distribution of $N(0,V_{11})$ almost surely with respect to the process of repeated sampling from the sequence of finite populations as $N \to \infty$. V_{11} is the asymptotic variance of $\hat{\theta}$ on the finite population level, while V_{22} is the asymptotic variance of θ_N on the superpopulaton level.

The key step in our proof of Theorem 1 is to obtain an asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$,

$$\begin{aligned}\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) &= \frac{1}{N}\sum_{i\in A_{2_g}} \frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}} \\ &= \frac{1}{N}\sum_{i\in U} \frac{\delta_{1i}\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\sum_{i\in U} \frac{\delta_{1i}(\delta_{2ig}-\pi_{2ig})}{\pi_{1i}\pi_{2ig}} E_\xi(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i) + o_p(n^{-1/2}).\end{aligned} \tag{17}$$

Define

$$\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z) = [m_{ig}(\boldsymbol{\theta}_g), Z_i-\mu_z]^T, \tag{18}$$

and similary we can show an asymptotic equivalent form of $\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g,\mu_z)$ as,

$$\begin{aligned}\frac{1}{N}\sum_{i\in A_{2_g}} \frac{\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}\hat{\pi}_{2ig}} &= \frac{1}{N}\sum_{i\in U} \frac{\delta_{1i}\delta_{2ig}\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\sum_{i\in U} \frac{\delta_{1i}(\delta_{2ig}-\pi_{2ig})}{\pi_{1i}\pi_{2ig}} \\ &\quad \times E_\xi(\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)|X_i) + o_p(n^{-1/2}) \\ &= \frac{1}{N}\sum_{i\in A_1} \frac{\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)}{\pi_{1i}} + o_p(n^{-1/2}),\end{aligned} \tag{19}$$

where

$$\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z) = \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)\frac{\delta_{2ig}}{\pi_{2ig}} + \left(1 - \frac{\delta_{2ig}}{\pi_{2ig}}\right)\boldsymbol{\mu}_{Hg}(X_i; \boldsymbol{\theta}_g, \mu_z), \text{ and}$$
$$\boldsymbol{\mu}_{Hg}(X_i, \boldsymbol{\theta}_g) = E_\xi(\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)|X_i). \tag{20}$$

Thus we can write $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in (14) as,

$$\begin{aligned}\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z) &= \left[\frac{1}{N}\sum_{i\in A_{2g}}\frac{\mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\pi_{1i}\hat{\pi}_{2ig}}, \frac{1}{N}\sum_{i\in A_1}\frac{Z_i - \mu_z}{\pi_{1i}}\right]^T \\ &= \left[\frac{1}{N}\sum_{i\in A_1}\frac{\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\pi_{1i}}, \frac{1}{N}\sum_{i\in A_1}\frac{Z_i - \mu_z}{\pi_{1i}}\right]^T + o_p(n^{-1/2}).\end{aligned} \tag{21}$$

Then the large sample theory for $\hat{\boldsymbol{\theta}}_g^{(2)}$ is derived based on the asymptotic form of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g, \mu_z)$ in Equation (21). We now state Theorem 1:

Theorem 1: Under the regularity conditions in the Appendix, for any $g = 1, \ldots G$,

$$\sqrt{n}\left(\begin{bmatrix}\hat{\boldsymbol{\theta}}_g^{(2)} \\ \hat{\mu}_z\end{bmatrix} - \begin{bmatrix}\boldsymbol{\theta}_g^0 \\ \mu_z^0\end{bmatrix}\right) \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, V_g\left(\boldsymbol{\theta}_g^0, \mu_z^0\right)\right),$$

where

$$V_g(\boldsymbol{\theta}_g, \mu_z) = \left[\Gamma_g^T(\boldsymbol{\theta}_g)\Sigma_{Hg}^{-1}(\boldsymbol{\theta}_g, \mu_z)\Gamma_g^T(\boldsymbol{\theta}_g)\right]^{-1}, \tag{22}$$

$$\Gamma_g(\boldsymbol{\theta}_g) = \left[E_\xi\left[\frac{\partial \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\partial \boldsymbol{\theta}_g}\right] \quad E_\xi\left[\frac{\partial \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)}{\partial \mu_z}\right]; \quad \mathbf{0} \quad -1\right], \tag{23}$$

$$\text{and} \quad \boldsymbol{\Sigma}_{Hg}(\boldsymbol{\theta}_g, \mu_z) = \left[\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g, \mu_z) \quad \boldsymbol{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z); \quad \boldsymbol{\Sigma}_{12}^T(\boldsymbol{\theta}_g, \mu_z) \quad \boldsymbol{\Sigma}_{22}(\mu_z)\right]. \tag{24}$$

Here the notation $[\mathbf{a_{11}}, \mathbf{a_{12}}; \mathbf{a_{21}}, \mathbf{a_{22}}]$ represents a 2×2 block matrix with blocks $\mathbf{a_{ij}}$. The term $\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g, \mu_z)$ in Equation (24) is related to the asymptotic variance of the first element in Equation (21) and is defined as,

$$\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}_g, \mu_z) = \lim_{N\to\infty} V_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N} Var_\xi(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)), \tag{25}$$

$$\text{where} \quad V_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)\boldsymbol{\eta}_{jg}^T(\boldsymbol{\theta}_g, \mu_z). \tag{26}$$

The term $\boldsymbol{\Sigma}_{22}(\mu_z)$ in Equation (24) is related to the asymptotic variance of the second element in Equation (21) and is defined as,

$$\boldsymbol{\Sigma}_{22}(\mu_z) = \lim_{N\to\infty} V_{z,N}(\mu_z) + \frac{n}{N} Var_\xi(Z_i), \tag{27}$$

$$\text{where} \quad V_{z,N}(\mu_z) = nN^{-2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}(Z_i - \mu_z)(Z_j - \mu_z). \tag{28}$$

The term $\boldsymbol{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z)$ in Equation (24) is related to the asymptotic covariance between the two elements in Equation (21) and is defined as,

$$\boldsymbol{\Sigma}_{12}(\boldsymbol{\theta}_g, \mu_z) = \lim_{N\to\infty} C_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N} Cov_\xi(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i), \tag{29}$$

$$\text{where} \quad C_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_j - \mu_z). \tag{30}$$

Equation (25) is connected to Fact 1 stated above, where its first term is $nV\left(N^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g)|\mathcal{F}_N\right)$ on the finite population corresponding to V_{11} in Fact 1, and its second term is $nV_\xi\left(N^{-1}\sum_{i\in U}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g)\right)$ on the superpopulation level corresponding to V_{22} in Fact 1. The limit sign in the first term of Equation (25) indicates this is the limit with respect to the process of repeated sampling from a sequence of finite population as $N \to \infty$. Similar connections can be seen in Equations (27) and (29). The proof of Theorem 1 uses results from Pakes and Pollard (1989) (Theorems 3.2 and 3.3) which provides a general central limit theorem for estimators defined by minimization of the length of a vector valued random criterion function. The justification of Theorem 1 takes into account the finite population asymptotic framework and the semiparametric estimation of $\hat{\pi}_{2ig}$. The asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ in (17) is analytically similar to the mathematical forms of the doubly robust (DR) estimators when $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, see Kim and Haziza (2014), Haziza and Rao (2006), Tan (2006), and Robins et al. (2007). One difference is that the consistency of the DR estimators requires one of the response model and the outcome model to be correctly specified, while our estimators estimate both the self-selection probabilities π_{2ig} and the outcome model semiparametrically. The regularity conditions on the sample design and tuning parameters for the semiparametric estimation are provided in the Appendix, and an outline of the proof for Theorem 1 can be found in Appendix A.

3.2. *Variance Estimation Based on the Asymptotic Normality*

We use the asymptotic variance $V_g(\boldsymbol{\theta}_g^0, \mu_z^0)$ in (22) to estimate the variance of $\hat{\boldsymbol{\theta}}_g^{(2)}$. To estimate $\boldsymbol{\Sigma}_{Hg}(\boldsymbol{\theta}_g, \mu_z)$, an estimator of $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)$ is obtained by,

$$\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z) = \mathbf{H}_{ig}(\boldsymbol{\theta}_g, \mu_z)\frac{\delta_{2ig}}{\hat{\pi}_{2ig}} + \left(1 - \frac{\delta_{2ig}}{\hat{\pi}_{2ig}}\right)\hat{\boldsymbol{\mu}}_{Hg}\left(X_i; \boldsymbol{\theta}_g, \mu_z\right), \tag{31}$$

where $\boldsymbol{\mu}_{Hg}(X_i, \boldsymbol{\theta}_g)$ is also estimated semiparametrically using the same bases $R_K(X_i)$, that is

$$\hat{\boldsymbol{\mu}}_{Hg}(X_i; \boldsymbol{\theta}_g, \mu_z) = \hat{\beta}_g^T(\boldsymbol{\theta}_g, \mu_z)R_K(X_i), \quad \text{and} \tag{32}$$

$$\hat{\beta}_g(\boldsymbol{\theta}_g, \mu_z) = \left(\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(X_i)R_K(X_i)^T\right)^{-1}\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(X_i)\mathbf{H}_{ig}^T(\boldsymbol{\theta}_g, \mu_z). \tag{33}$$

An estimator of $V_g\left(\boldsymbol{\theta}_g^0, \mu_z^0\right)$ is calculated as follows,

$$\hat{V}_g\left(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z\right) = \left[\hat{\Gamma}_g^T\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)\hat{\boldsymbol{\Sigma}}_{Hg}^{-1}\left(\hat{\boldsymbol{\theta}}_g^{(2)}, \hat{\mu}_z\right)\hat{\Gamma}_g^T\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)\right]^{-1}, \tag{34}$$

where

$$\hat{\Gamma}_g(\boldsymbol{\theta}_g) = \frac{1}{N}\left[\sum_{i\in A_{2g}} w_{1i}\hat{\pi}_{2ig}^{-1}\frac{\partial \mathbf{H}_{ig}\left(\boldsymbol{\theta}_g, \mu_z\right)}{\partial \boldsymbol{\theta}_g}\sum_{i\in A_{2g}} w_{1i}\hat{\pi}_{2ig}^{-1}\frac{\partial \mathbf{H}_{ig}\left(\boldsymbol{\theta}_g, \mu_z\right)}{\partial \mu_z}; \quad 0 \quad -1\right], \tag{35}$$

$$\text{and} \quad \hat{\boldsymbol{\Sigma}}_{Hg}(\boldsymbol{\theta}_g, \mu_z) = \left[\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z) \quad \hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z); \quad \hat{\boldsymbol{\Sigma}}_{12}^T(\boldsymbol{\theta}_g, \mu_z) \quad \hat{\boldsymbol{\Sigma}}_{22}(\mu_z)\right]. \tag{36}$$

The term $\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{11}(\boldsymbol{\theta}_g, \mu_z) = \hat{V}_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N}\widehat{Var}_\xi\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right), \tag{37}$$

$$\text{where} \quad \hat{V}_{\eta g,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\hat{\boldsymbol{\eta}}_{jg}^T(\boldsymbol{\theta}_g, \mu_z), \tag{38}$$

and

$$\widehat{Var}_\xi\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\hat{\boldsymbol{\eta}}_{ig}^T(\boldsymbol{\theta}_g, \mu_z) - \frac{1}{N^2}\left[\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]\left[\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)\right]^T. \tag{39}$$

The term $\hat{\boldsymbol{\Sigma}}_{22}(\mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{22}(\mu_z) = \hat{V}_{z,N}(\mu_z) + \frac{n}{N}\widehat{\mathrm{Var}}_\xi\left(Z_i\right), \tag{40}$$

$$\text{where} \quad \hat{V}_{z,N}(\mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}(Z_i - \mu_z)(Z_j - \mu_z), \quad \text{and} \tag{41}$$

$$\widehat{\mathrm{Var}}_\xi\left(Z_i\right) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i - \mu_z)^2 - \frac{1}{N^2}\left[\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i - \mu_z)\right]\left[\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i - \mu_z)\right]^T, \tag{42}$$

The term $\hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z)$ is estimated using

$$\hat{\boldsymbol{\Sigma}}_{12}(\boldsymbol{\theta}_g, \mu_z) = \hat{C}_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) + \frac{n}{N}\widehat{\mathrm{Cov}}_\xi\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g, \mu_z), Z_i\right), \tag{43}$$

$$\text{where } \hat{C}_{\eta z,N}(\boldsymbol{\theta}_g, \mu_z) = nN^{-2}\sum_{i\in A_1}\sum_{j\in A_1}\frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g, \mu_z)(Z_j - \mu_z), \tag{44}$$

and

$$\widehat{\mathrm{Cov}}_{\xi}\left(\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z),Z_i\right)=\frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g,\mu_z)(Z_i-\mu_z) - \frac{1}{N^2}\left[\sum_{i\in A_1}\pi_{1i}^{-1}\hat{\boldsymbol{\eta}}_{ig}(\boldsymbol{\theta}_g,\mu_z)\right]\left[\sum_{i\in A_1}\pi_{1i}^{-1}(Z_i-\mu_z)\right]. \tag{45}$$

To construct a joint estimator for $\boldsymbol{\theta}=[\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_G]^T$, one can simply stack $\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)$ in the quadratic form of Equation (13). Define $\mathbf{H}_i(\boldsymbol{\theta},\mu_z)$ as the stacked vector of $\mathbf{H}_{ig}(\boldsymbol{\theta}_g,u_z)'s$ in Equation (18) and $\boldsymbol{\eta}_i(\boldsymbol{\theta},\mu_z)$ as the stacked vector of $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)'s$ in Equation (20). The asymptotic theory and the variance estimator for $\hat{\boldsymbol{\theta}}^{(2)}$ can be derived by simply replacing $\mathbf{H}_{ig}(\boldsymbol{\theta}_g,\mu_z)$ by $\mathbf{H}_i(\boldsymbol{\theta},\mu_z)$ and $\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g,\mu_z)$ by $\boldsymbol{\eta}_i(\boldsymbol{\theta},\mu_z)$. Then we can obtain an inference for the treatment effects or any linear combination of treatment parameters, $\boldsymbol{\lambda}^T\boldsymbol{\theta}$.

3.3. Replication Variance Estimation

In surveys conducted on land, for example surveys about natural resources (soil, forest, water, etc.), non-responses hardly occur. However, in surveys with high non-response rates, such as almost all surveys conducted on people, the joint inclusion probabilities are typically not available because sampling weights have to be appropriately adjusted for nonresponse. After such adjustments, the joint inclusion probabilities change and are hard to be derived. In practice, a set of replicate weights are often provided instead, because (1) design weights are often adjusted due to nonresponse issues and a set of replicate weights are provided to account for the weight adjustment; (2) sometimes a few design variables are masked from users to keep confidentiality. An example of such design variable is location which is used for defining design strata in a study, but the specific location is omitted from the analyst because of concerns associated with confidentiality. In this subsection, we show how to use the replicate weights to construct a Jackknife variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$. Note that $\hat{\boldsymbol{\theta}}_g^{(2)}$ depends on the joint inclusion probabilities π_{1ij} which are typically not available when replicate weights are provided. We propose to use the Jackknife (JK) variance estimator for a two-phase sampling design discussed in Fuller (2009) and Kim et al. (2006). Assume that there is a replicate variance estimator that gives a consistent estimator for the variance of the total estimator based on the first-phase sample. We write the replication variance estimator as, $\hat{V}_{JK1}(\hat{\boldsymbol{\theta}}_1)=\sum_{b=1}^{B}c_b\left(\hat{\boldsymbol{\theta}}_1^{[b]}-\hat{\boldsymbol{\theta}}_1\right)(\hat{\boldsymbol{\theta}}_1^{[b]}-\hat{\boldsymbol{\theta}}_1)^T$, where B is the number of replicates, $\hat{\boldsymbol{\theta}}_1=\sum_{i\in A_1}w_{1i}X_i$ is the total estimator of variable X using the first-phase sample, $\hat{\boldsymbol{\theta}}_1^{[b]}=\sum_{i\in A_1}w_{1i}^{[b]}X_i$ is the estimated total for the b^{th}replicate, $w_{1i}^{[b]}$ is the b^{th}replicate weights in the first-phase, and c_b is a factor associated with replicate b such that $\hat{V}_{JK1}(\hat{\boldsymbol{\theta}}_1)$ is a consistent estimator for the variance of $\hat{\boldsymbol{\theta}}_1$. Suppose the second-phase total estimator is, $\hat{\boldsymbol{\theta}}_2=\sum_{i\in A_2}w_{1i}\pi_{2i|1i}^{-1}X_i$, where $\pi_{2i|1i}$ is the conditional probability of selecting i for the phase 2 sample given that i is in the phase 1 sample, and A_2 is the phase 2 sample. Define the b^{th}replicate of $\hat{\boldsymbol{\theta}}_2$ as, $\hat{\boldsymbol{\theta}}_2^{[b]}=\sum_{i\in A_2}w_{1i}^{[b]}\pi_{2i|1i}^{-1}X_i$. A Jackknife variance estimator for $\hat{\theta}_2$

can be calculated as, $\hat{V}_{JK2}(\hat{\boldsymbol{\theta}}_2) = \sum_{b=1}^{B} c_b \left(\hat{\boldsymbol{\theta}}_2^{[b]} - \hat{\boldsymbol{\theta}}_2\right)\left(\hat{\boldsymbol{\theta}}_2^{[b]} - \hat{\boldsymbol{\theta}}_2\right)^T$. Kim et al. (2006) showed that $\hat{V}_{JK2}(\hat{\boldsymbol{\theta}}_2)$ is a consistent estimator for the variance of $\hat{\boldsymbol{\theta}}_2$.

Following the idea of Fuller (2009 Subsection 4.4), let b be the index for the deleted Jackknife groups and the corresponding replicate version of $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ be,

$$\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i\in A_{2g}} w_{1i}^{[b]} \left(\hat{\pi}_{2ig}^{[b]}\right)^{-1} \mathbf{m}_{ig}(\boldsymbol{\theta}_g), \tag{46}$$

where $\hat{\pi}_{2ig}^{[b]}$ is obtained by replacing w_{1i} by $w_{1i}^{[b]}$ in Equation (7). Then the replicate estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$ is,

$$\hat{\boldsymbol{\theta}}_g^{(1)[b]} = \arg\min_{\boldsymbol{\theta}_g} \left[\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g)\right]^T \left[\bar{\mathbf{m}}_{2\pi g}^{[b]}(\boldsymbol{\theta}_g)\right], \tag{47}$$

and the replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{(1)}$ is calcualted as,

$$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right) = \sum_{b=1}^{B} c_b \left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)\left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)^T. \tag{48}$$

Examples of $w_{1i}^{[b]}$ and c_b for a variety of designs are given in Särndal et al. (1992). For example, if the first-phase sample is drawn from a multi-stage cluster design, the Jackknife technique is usually applied at the primary sampling unit (PSU) levels. Assuming there are B PSUs and S_b is the b^{th} PSU deleted in the b^{th} replicate sample, the b^{th} replicate weight for the first-phase is defined as,

$$w_{1i}^{[b]} = \begin{cases} 0 & \text{if} \quad i \in S_b \\ \dfrac{B}{B-1} w_{1i} & \text{if} \quad i \notin S_b \end{cases}, \tag{49}$$

and $c_b = B^{-1}(B-1)$. As mentioned in Särndal et al. (1992), for stratified sampling designs, $w_{1i}^{[b]}$ and c_b need to be defined with care. We discuss this situation in Section 5 of the empirical study. If the first phase replicate weights are provided in practice, one can directly use them as $w_{1i}^{[b]}$. One thing to note is that Kim et al. (2006) assume π_{2ig} are known in their two phase replication variance estimator. The consistency theorem in Kim et al. (2006) needs to be modified to account for the variation from estimating $\hat{\pi}_{2ig}$ in our JK variance estimator, which can be our future study.

4. Simulation Study

In this section, we evaluate the performance of our estimators and variance estimators under four different simulation set-ups. We consider three treatment levels, and a population size of $N = 10{,}000$ and an expected sample size of $n = 1{,}000$. We generate i.i.d. realizations, $(\mathbf{Y}_i, \delta_{1i}, \boldsymbol{\delta}_{2i}, X_i, Z_i)$; $i = 1, \ldots, N$, according to the following super-population set-ups.

(1) Covariates: simulate covariates $\mathbf{Z}_i = [Z_{1i}, Z_{2i}]$ where $Z_{1i} \sim N(2,1)$ and $Z_{2i} \sim N(10,1)$, and $\mathbf{X}_i = [X_{1i}, X_{2i}]$ where $X_{1i} = Z_{1i}$ and $X_{2i} \sim N(0.5, 0.3^2)$.

(2) Potential response outcomes: the superpopulation model for potential outcomes is $Y_{ig} = \mu_g(\mathbf{Z}_i) + \sigma_g(\mathbf{Z}_i)\varepsilon_{ig}$, where

$$\mu_g(\mathbf{Z}_i) = \beta_{g0} + \beta_{g1}(Z_{1i} - 0.5) + \beta_{g2}(Z_{1i} - 0.5)^2 + \beta_{g3}Z_{2i},$$

$\epsilon_{ig} \sim N(0,1)$, $\sigma_g(\mathbf{Z}_i) = |\mu_g(\mathbf{Z}_i)|$, and $[\beta_{g0}, \beta_{g1}, \beta_{g2}, \beta_{g3}]$ equals to [5, 4, 2, 1] for $g = 1$, [0, 1, 0, 0] for $g = 2$, and $[-5, -4, -2, -0.5]$ for $g = 3$.

(3) First phase sampling: we consider two sampling designs, non-informative stratification sampling and informative Poisson sampling.

- Stratification (STS): population units are sorted by values of Z_{1i}, and then the population is divided into two subpopulations U_1 and U_2 with equal sizes. Simple random sampling is used to draw 80 percent of the sample from U_1 and 20 percent from U_2. For units in stratum s ($s = 1$ or 2), $\pi_{1i} = N_s^{-1} n_s$ and $\pi_{1ij} = \{N_s(N_s - 1)\}^{-1} n_s(n_s - 1)$, where n_s and N_s are the sample size and the population size in stratum s. The joint inclusion probability for two units in different strata is the product of their first order inclusion probabilities.
- Informative Poisson (Informative): the first-phase sample design is Poisson sampling with selection probability,

$$\pi_{1i} = \frac{exp\big(-1.5 - 2.5X_{2i} + 0.07\|\mathbf{Y}_i\|\big)}{1 + exp\big(-1.5 - 2.5X_{2i} + 0.07\|\mathbf{Y}_i\|\big)},$$

where $\|\mathbf{Y}_i\| = \sqrt{Y_{i1}^2 + Y_{i2}^2 + Y_{i3}^2}$. Modeling π_{1i} as a function of $\mathbf{Y}_i$ is a common way (i.e., Pfeffermann and Sverchkov 1999) to represent joint dependence of $\mathbf{Y}_i$ and π_{1i} on a design variable that is not contained in (X_i, Z_i). In this specification, we assume $\|\mathbf{Y}_i\|$ is known at the design stage of the survey, but is unavailable at the analysis stage.

(4) Second phase self-selection probability models: we consider two models for π_{2ig}.

- Logit Linear (LogitLinear):

$$\pi_{2ig} = \frac{\exp(\phi_{g0} + \phi_{g1}X_{1i} + \phi_{g2}X_{2i})}{\sum_{g=1}^{G}\exp(\phi_{g0} + \phi_{g1}X_{1i} + \phi_{g2}X_{2i})},$$

where $[\phi_{g0}, \phi_{g1}, \phi_{g2}]$ equals to $[-0.5, 0, 0]$ for $g = 1$, $[0.3, -0.3, -0.3]$ for $g = 2$, and $[0, -0.5, 0.5]$ for $g = 3$.

- Jump (JUMP):

$$\begin{aligned}[\pi_{2i1}, \pi_{2i2}, \pi_{2i3}] &= [0.90, 0.05, 0.05] \quad \text{if} \quad X_{1i} + X_{2i} \geq 3 \\ &= [1/3, 1/3, 1/3] \quad \text{if} \quad 2 \leq X_{1i} + X_{2i} < 3 \\ &= [0.05, 0.05, 0.90] \quad \text{if} \quad X_{1i} + X_{2i} < 2.\end{aligned}$$

The JUMP model violates the differentiability assumption of π_{2ig} in Condition B(2) in the Appendix. It is deliberately included in the simulation to see if our semiparametric approach can estimate a nonsmooth multiple treatment selection probabilities well.

For each $i \in U$, $\boldsymbol{\delta}_{2i}$ is simulated from $multinomial(1; \pi_{2i1}, \pi_{2i2}, \pi_{2i3})$. For $i \neq j$, $\pi_{1ij} = \pi_{1i}\pi_{1j}$. For STS design which is noninformative, SMAR holds and we set $b_i = w_{1i}^{-1}$ in Equation (7) to estimate $\hat{\pi}_{2ig}$. For Informative design, SMAR fails and we use $b_i = 1$ in Equation (7) to estimate $\hat{\pi}_{2ig}$.

We first simulate a finite population with size N from the superpopulation and then use indicators generated in (3) and (4) to obtain the first and second phase samples. We repeat the process to produce 1,000 MC samples. We are interested in estimating five parameters for each group, $\boldsymbol{\theta}_g = \left[P_g, \mu_g, \sigma_g^2, R_g, D_g\right]$, where $P_g = Prob(Y_{ig} \leq 0)$, $\mu_g = E(Y_{ig})$, $\sigma_g^2 = Var(Y_{ig})$ and $R_g = Corr(Y_{ig}, Z_{2i})$, and $D_g = E[E(Y_{ig}|Z_{1i} \leq 0.65)]$. The corresponding estimation equations $\mathbf{m}_{ig}(\boldsymbol{\theta}_g)$ can be found in Equations (5) and (6). For each MC sample, we calculate the following four estimators:

- $\hat{\boldsymbol{\theta}}_g^{(1)}$: the estimator defined in (11). When $\mathbf{m}_{ig}(\boldsymbol{\theta}_g) = Y_{ig} - \mu_g$, $\hat{\boldsymbol{\theta}}_g^{(1)}$ corresponds to the estimator in Yu et al. (2013) asymptotically.
- $\hat{\boldsymbol{\theta}}_g^{(2)}$: the estimator defined in (13).
- $\hat{\boldsymbol{\theta}}_g^{nw}$: the estimator defined in (9), and is included to see what happens when the survey weights are ignored in analyses.
- $\hat{\boldsymbol{\theta}}_g^{p}$: the estimator calculated the same way as $\hat{\boldsymbol{\theta}}_g^{(1)}$, except that $\hat{\pi}_{2ig}$ are estimated using a parametric multinomial regression. This estimator is introduced in order to have plausible comparisons in context of three treatments between our estimators and others that use parametric logistic regression to estimate propensity scores, see DuGoff et al. (2014), Zanutto (2006), Ashmead (2014), and Ridgeway et al. (2015).

We use a cubic spline base of X_{1i} for $R_K(X_{1i})$, as suggested by Breidt et al. (2005) which mentions that setting the degree of the spline equal to three is a popular choice in practice. Condition 4(B) in the Appendix gives a practical guidance for the choice of K, the number of knots in the spline. Condition 4(B) requires $K = O(n^{\nu})$, where ν has an upper bound $\nu \leq (4\eta + 2)^{-1}$ with $\eta = 1/2$ for spline bases. In our simulation studies, the sample size $n = 1{,}000$, suggesting $n^{u} = 5.6$. The choices of $K = 5, 4, 3, 2$ are tried and the corresponding $\hat{\pi}_{2ig}$ curves are plotted. It is found that there is not noticeable change in the $\hat{\pi}_{2ig}$ curves until K decreases to 2. So $K = 3$ is used and the locations of the three knots correspond to the $25th$, $50th$, and $75th$ quantiles of observed X_{1i}'s. A cubic spline base for $R_K(X_{2i})$ is constructed the same way. And the semiparametric bases are $R_K(\mathbf{X}_i) = \left[R_K^T(X_{1i}), R_K^T(X_{2i})\right]^T$.

If the dimension of $(\mathbf{X}_i, \mathbf{Z}_i)$ is big, in practice we suggest to run a multinomial regression using δ_{2i} on $(\mathbf{X}_i, \mathbf{Z}_i)$ to select covariates that are most significant, and then use them for estimation of $\hat{\pi}_{2ig}$. When using $\hat{\boldsymbol{\theta}}_g^{(2)}$, one can run a multiple linear regression of Y_{ig} on $(\mathbf{X}_i, \mathbf{Z}_i)$ in A_{2g} to identify covariates that are most useful for explaining the outcome Y_{ig}, and then add their first and second phase means in the estimation equations. It is not impossible to obtain a very small $\hat{\pi}_{2ig}$ computationally, which leads to extreme weights. A solution is to truncate such $\hat{\pi}_{2ig}$'s to a small constant L (which is set to be 0.0001 in our study), then adjust the truncated $\hat{\pi}_{2ig}$ by calibrating the second phase mean of U_i to its first phase mean, that is $\tilde{\pi}_{2ig} = F_g \hat{\pi}_{2ig}^t$ where $F_g = \left(\sum_{i \in A_1} w_{1i} U_i\right)^{-1} \sum_{i \in A_{2g}} w_{1i} \left(\hat{\pi}_{2ig}^t\right)^{-1} U_i$, and $\hat{\pi}_{2ig}^t$ is the truncated propensity score which equals to L if $\hat{\pi}_{2ig} < L$, otherwise remains

unchanged. Here the variable U_i can be an important covariate chosen by users, or a weighted mean of $(\mathbf{X}_i, \mathbf{Z}_i)$ where weights indicate importance of the covariates. We use the average of the covariate $\mathbf{X}_i$ as U_i in both of the simulation studies and the empirical study.

Figures 1–4 show side-by-side boxplots of MC estimates of the four estimators for all treatment effects. Each figure represents one of four simulation setups: (STS-LogitLinear), (STS-JUMP), (Informative-LogitLinear), and (Informative-JUMP). In each subplot, the first two boxplots are for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$, and the third and fourth boxplots are for $\hat{\boldsymbol{\theta}}_g^{p}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ respectively. When comparing our estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ with $\hat{\boldsymbol{\theta}}_g^{nw}$, $\hat{\boldsymbol{\theta}}_g^{nw}$ is highly biased in most of parameters and scenarios, due to ignoring the survey weights. The variances of $\hat{\boldsymbol{\theta}}_g^{nw}$ in general are smaller than those of $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$, which is expected especially when the survey weights are very different from each other. The coefficient of variation (CV) of the weights for the STS design is 0.75, and the CV of weights for the Informative design is 4.77. When comparing our estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ with $\hat{\boldsymbol{\theta}}_g^{p}$, biases of $\hat{\boldsymbol{\theta}}_g^{p}$ are comparable to those of $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ for the LogitLinear model because in this scenario $\hat{\theta}_g^{p}$ correctly assumes a parametric model for π_{2ig}. However, in the situation of JUMP models, $\hat{\boldsymbol{\theta}}_g^{p}$ has larger biases than $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ because π_{2ig} is misspecified parametrically. When comparing $\hat{\boldsymbol{\theta}}_g^{(1)}$ with $\hat{\boldsymbol{\theta}}_g^{(2)}$, both of their biases are comparable in all scenarios. However, the plots show that $\hat{\boldsymbol{\theta}}_g^{(2)}$ consistently has smaller variances than $\hat{\boldsymbol{\theta}}_g^{(1)}$. The variance reduction of $\hat{\boldsymbol{\theta}}_g^{(2)}$ over $\hat{\boldsymbol{\theta}}_g^{(1)}$ indicates that efficiency gain occurs after adding the first and second phase means of covariates to the estimation equations, which confirms Remark 2. Additionally, it is promising to see that both $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ have relatively small biases even if the JUMP model fails to satisfy the differentiability assumption in the theory, indicating our semiparametric approach of estimating $\hat{\pi}_{2ig}$ works well for the nonsmooth function considered. We also tabulate the MC results into four tables for readers who prefer to see numbers rather than Figures (see Supplemental file, Tables 1–4).

Tables 1–2 contain the coverage probabilities of the 95 percent confidence intervals for $\hat{\boldsymbol{\theta}}_g^{(2)}$ based on its asymptotic normality and its linearized variance estimator in Subsection 3.2, and the coverage probabilities of the 95 percent confidence intervals for $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ based on the JK approach discussed in Subsection 3.3. The replication variance estimator for $\hat{\boldsymbol{\theta}}_g^{nw}$ is calculated by replacing w_{1i} by N/n in Equation (49). This gives inappropriate variance estimation for $\hat{\boldsymbol{\theta}}_g^{nw}$ under an unequal probability sampling, but mimics what people do when they ignore survey weights. To create the JK replicates, we delete one unit at a time and set $B = 1{,}000$. The coverage probabilities for $\hat{\boldsymbol{\theta}}_g^{(2)}$ using the linearized variance estimator seem to work well, except for the marginal mean μ_g under (STS-LogitLinear) and the marginal proportion P_g under (STS-JUMP). The rest of coverage probabilities are reasonably close to the nominal size 95 percent. The JK variance estimator of $\hat{\boldsymbol{\theta}}_g^{(1)}$ gives very good coverage probabilities. However the coverage probabilities for $\hat{\boldsymbol{\theta}}_g^{nw}$ using the JK variance estimation are far away from the nominal size, especially under the Informative-JUMP model where the coverage probabilities are severely underestimated. Those under-coverages are due to the biases in $\hat{\boldsymbol{\theta}}_g^{nw}$, or inappropriate variance estimation, or both.

Our simulation studies demonstrate the validity of our estimators and variance estimators.

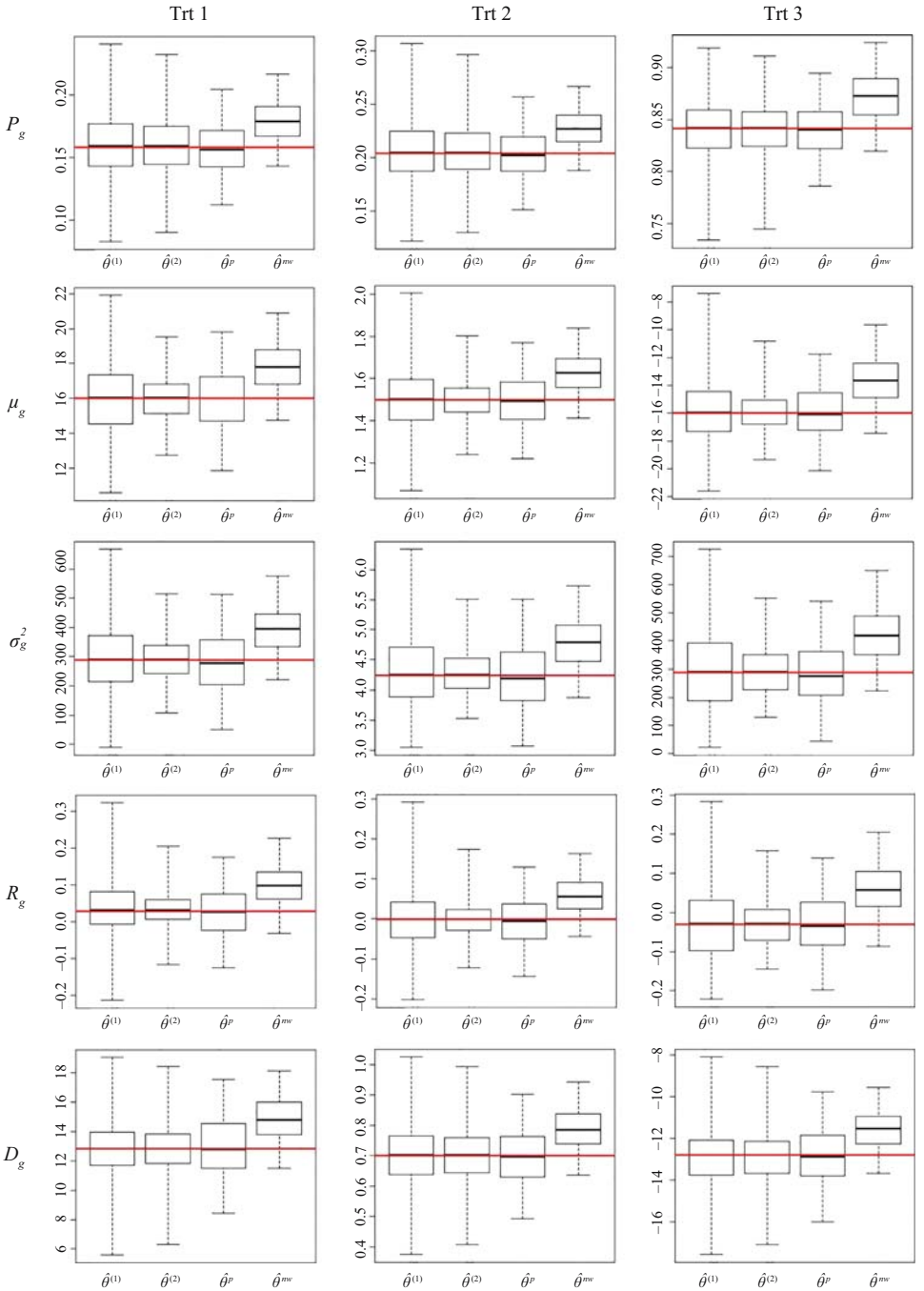

Fig. 1. **STS-LogitLinear***: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for* $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ *and* $\hat{\boldsymbol{\theta}}_g^{mw}$ *respectively in order. The horizontal line is located at the value of the true treatment effect.*

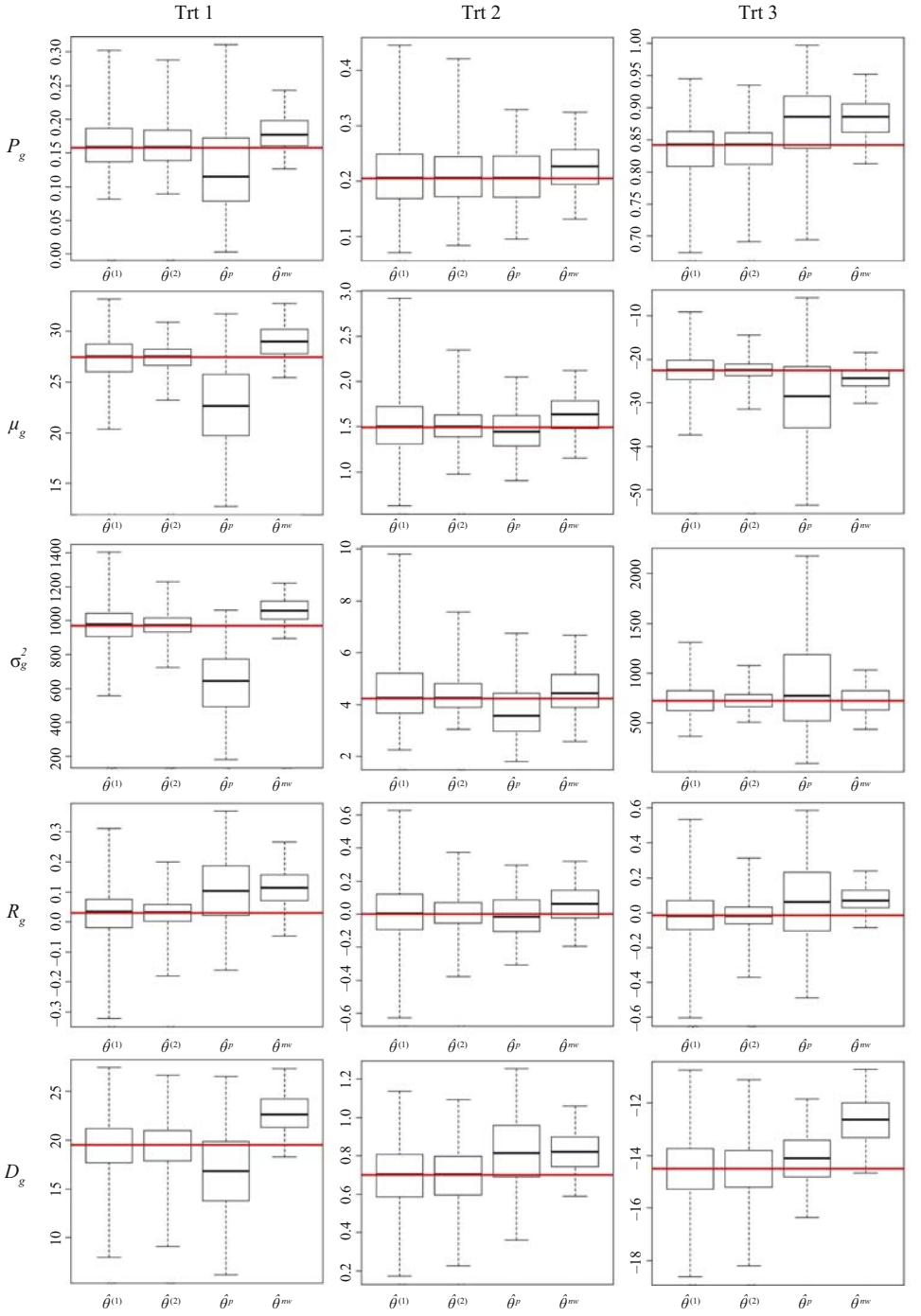

Fig. 2. **STS-JUMP***: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for* $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively in order. The horizontal line is located at the value of the true treatment effect.*

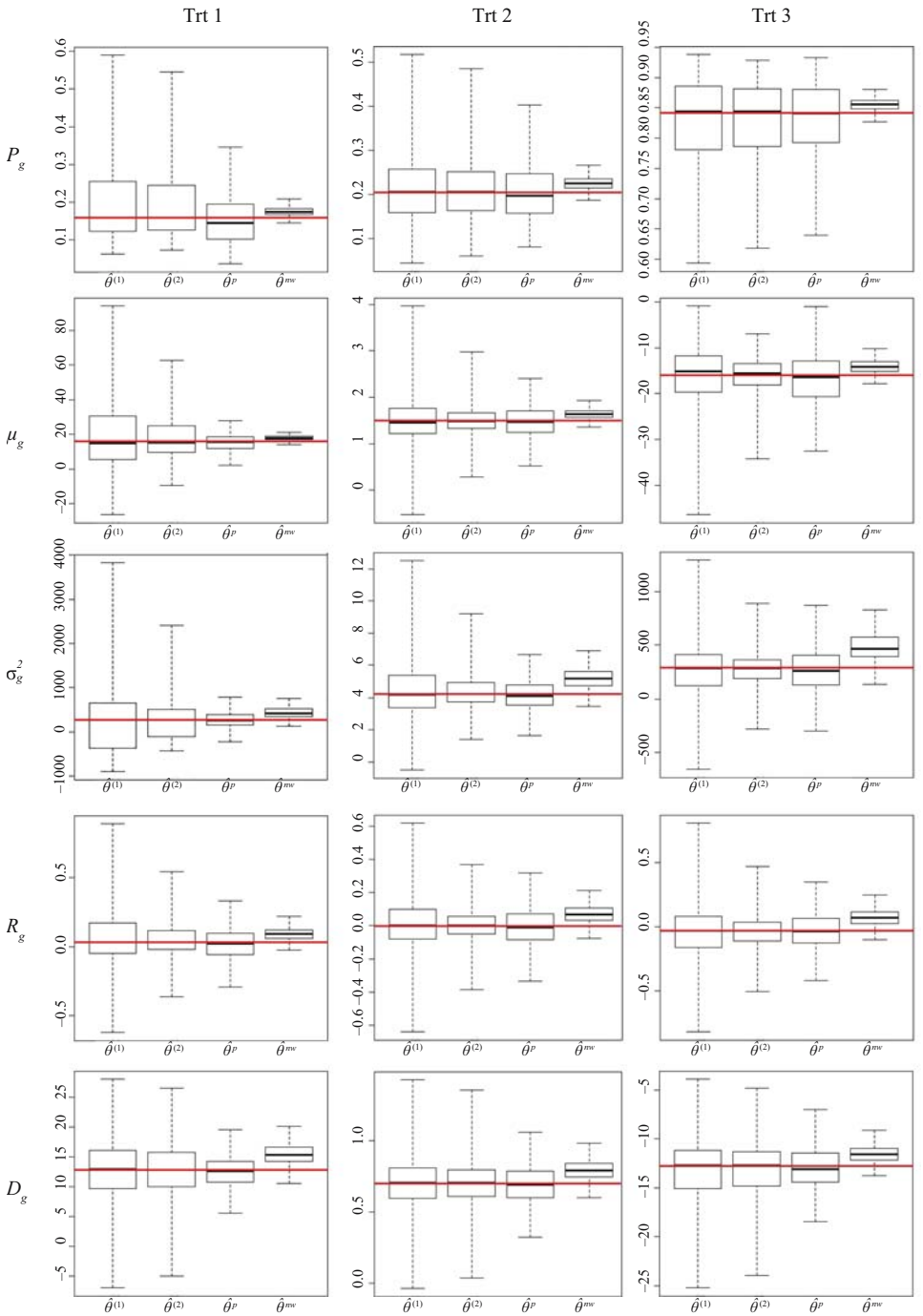

Fig. 3. ***Informative-LogitLinear****: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for* $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively in order. The horizontal line is located at the value of the true treatment effect.*

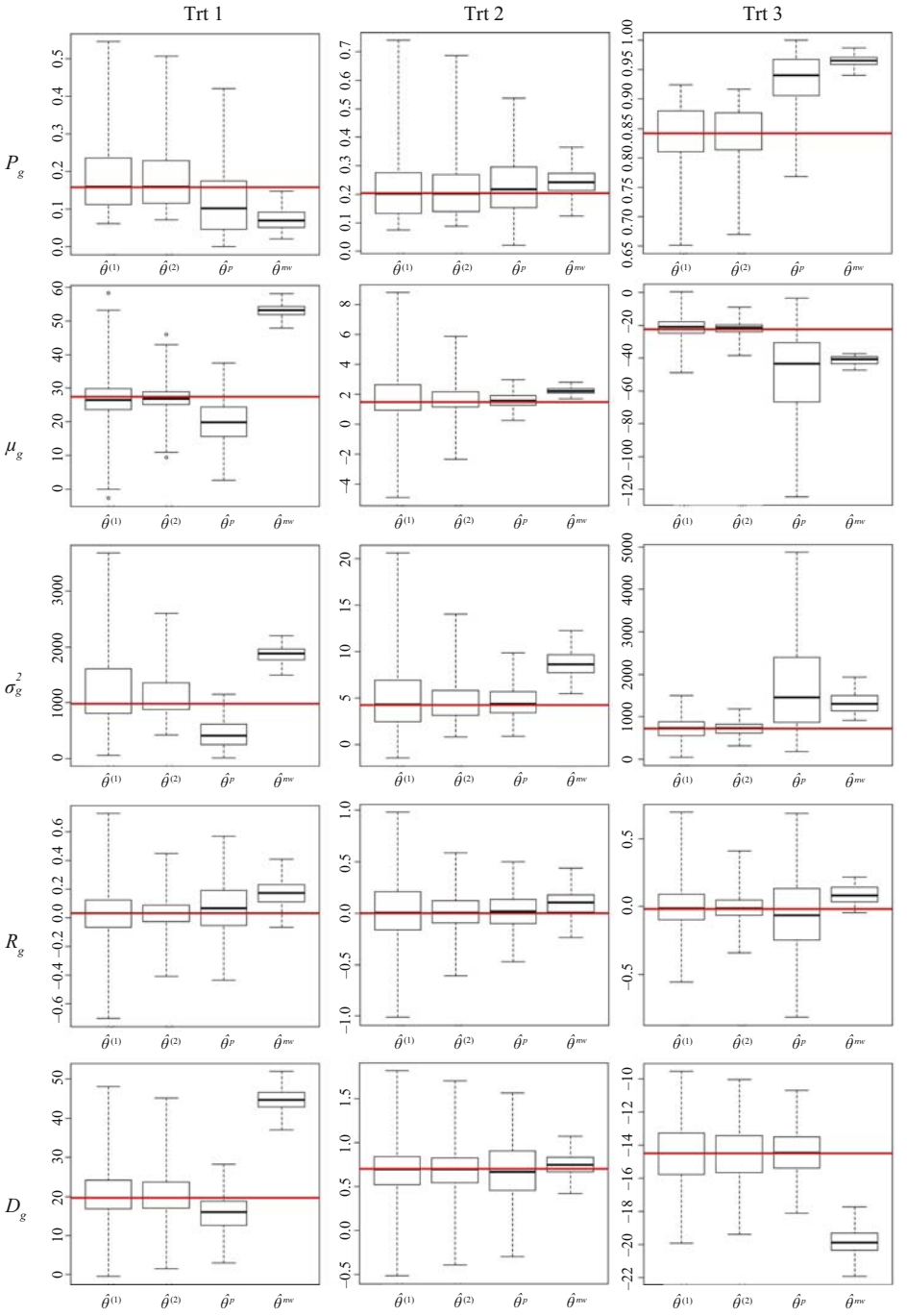

Fig. 4. ***Informative-JUMP****: Boxplots of MC estimates of the four estimators for all treatments. Each row represents a parameter, and each column represents a treatment. In each subplot, the four boxplots are for* $\hat{\boldsymbol{\theta}}_g^{(1)}$, $\hat{\boldsymbol{\theta}}_g^{(2)}$, $\hat{\boldsymbol{\theta}}_g^{p}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively in order. The horizontal line is located at the value of the true treatment effect.*

Table 1. ***Stratification:*** *The coverage probabilities of the 95 percent constructed intervals for the five estimated parameters using the linearized variance estimator* $\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(2)}$, *and the Jackknife variance estimators* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$ *and* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(1)}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively.*

(a) STS-LogitLinear

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	93.6	95.2	59.8	94.3	94.0	57.9	92.4	94.1	56.2
μ_g	95.4	95.5	58.7	95.2	95.1	57.9	88.1	94.1	62.3
σ_g^2	92.5	94.7	61.7	94.4	94.9	60.4	92.2	94.3	58.2
R_g	94.2	94.7	57.6	92.1	95.1	60.3	95.9	94.3	59.1
D_g	92.4	94.8	56.7	95.1	95.1	58.8	92.6	95.9	62.2

(b) STS-JUMP

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	89.2	92.8	71.9	95.2	92.3	80.4	92.5	95.4	50.0
μ_g	92.2	95.3	73.0	95.6	93.3	76.8	94.5	95.3	79.3
σ_g^2	94.2	93.0	56.5	93.3	96.6	83.9	95.3	96.1	86.0
R_g	95.3	94.6	60.7	93.8	95.2	81.0	92.3	95.8	61.1
D_g	92.9	95.0	50.0	96.6	93.2	61.7	94.1	96.6	28.2

Table 2. ***Informative:*** *The coverage probabilities of the 95 percent constructed intervals for the five estimated parameters using the linearized variance estimator* $\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(2)}$*, and the Jackknife variance estimators* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$ *and* $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$ *for* $\hat{\boldsymbol{\theta}}_g^{(1)}$ *and* $\hat{\boldsymbol{\theta}}_g^{nw}$ *respectively.*

(a) Informative-LogitLinear

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	94.3	95.2	58.9	94.1	96.4	49.3	95.4	94.7	49.2
μ_g	95.2	96.4	50.6	92.5	95.4	46.2	95.1	95.5	49.3
σ_g^2	92.0	94.1	46.2	95.4	95.1	50.6	93.7	96.2	44.4
R_g	94.9	96.0	45.1	90.6	94.8	42.8	93.8	95.4	38.3
D_g	93.4	96.2	59.3	93.8	95.7	48.1	93.1	94.9	45.7

(b) Informative-JUMP

	Trt1			Trt 2			Trt 3		
	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$	$\hat{V}_L\left(\hat{\boldsymbol{\theta}}_g^{(2)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right)$	$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$
P_g	94.4	96.3	6.6	97.7	94.7	72.4	92.9	96.9	0.0
μ_g	92.5	97.2	0.0	91.1	97.1	2.0	92.2	93.9	0.0
σ_g^2	94.9	97.0	0.0	95.5	92.3	2.0	94.6	96.5	20.6
R_g	92.7	93.2	41.6	95.2	95.0	71.4	94.3	95.4	48.4
D_g	92.2	95.0	0.0	91.0	94.3	81.9	95.0	96.0	0.0

Table 3. ***Empirical study with weights in estimation of*** $\hat{\pi}_{2ig}$*: The treatment effect estimates using estimators* $\hat{\boldsymbol{\theta}}_g^{nw}$ *and* $\hat{\boldsymbol{\theta}}_g^{(1)}$ *defined in Subsection 2.3. The parameter of interests are* $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ *and* $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ *where* I_{di} *is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. The standard errors are in parentheses and calculated using the Jackknife variance estimator, and the 95 percent confidence intervals are in brackets.*

(a) Treatment mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig})$

Estimators	Public – Private	Public – No insurance	Private – No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	1349.57 (215.90) [926.40 1772.74]	309.408 (28.23) [254.07 364.74]	− 1040.165 (698.47) [− 2409.17 328.84]
$\hat{\boldsymbol{\theta}}_g^{nw}$	1210.57 (353.50) [517.71 1903.44]	− 21.45 (29.17) [− 78.61779 35.71]	− 1232.03 (56.56) [− 1342.88 − 1121.18]

(b) Treatment domain mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$

Estimators	Public – Private	Public – No insurance	Private – No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	3214.18 (32.22) [3151.03 3277.34]	811.56 (38.69) [735.73 887.39]	− 2402.62 (46.48) [− 2493.73 − 2311.52]
$\hat{\boldsymbol{\theta}}_g^{nw}$	3320.93 (9.97) [3301.39 3340.47]	4.49 (2.69) [− 0.77 9.76]	− 3316.43 (240.85) [− 3788.50 − 2844.37]

5. Empirical Study

In this section, we investigate the feasibility of our method in estimating the mean annual medical expenditures under different choices of health insurance types in China. We use the data from the Chinese General Social Survey (CGSS) conducted by the National Survey Research Center at the Renming University of China in 2010. The population consisted of all Chinese adults (18+) in mainland China. A sample of 12,000 adults was drawn for the base questionnaire and a subsample of 4,000 adults was drawn for the health care questionnaire. Data were collected by in-person interviews. The sample for the CGSS survey was selected using a multi-stage cluster sampling design. In the first stage, the primary sampling units (PSUs) were districts which were divided into two strata. Stratum 1 contained 67 districts in five major cities (Shanghai, Beijing, Guangzhou, Shenzhen and Tianjin), and Stratum 2 contained 2,795 districts in the rest of the area of China. In both strata, a probability proportional to size (PPS) design with the resident population size as the size variable was used to select the PSUs (40 PSUs were selected in Stratum 1, and 100 PSUs were selected in Stratum 2). In the second stage, the secondary sampling units (SSUs) were communities. A PPS design with resident population size as the size variable was used to select 2 SSUs within each selected PSU in Stratum 1 and 4 SSUs within each selected PSU in Stratum 2. In the third stage, the ultimate sampling units (USUs) were households. In each selected SSU, 25 households were drawn by a systematic sampling method. Then a respondent was selected randomly within each household. Totally 12,000 households responded to the base questionnaire. Then every third household respondent in each SSU was selected to answer the health care questionnaire. The subsample of 4,000 was used in our investigation.

Table 4. ***Empirical study without weights in estimation of*** $\hat{\pi}_{2ig}$*: The treatment effect estimates using estimators* $\hat{\boldsymbol{\theta}}_g^{nw}$ *and* $\hat{\boldsymbol{\theta}}_g^{(1)}$ *defined in Subsection 2.3. The parameter of interests are* $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ *and* $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ *where* I_{di} *is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. The standard errors are in parentheses and calculated using the Jackknife variance estimator, and the 95 percent confidence intervals are in brackets.*

(a) Treatment mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig})$

Estimators	Public – Private	Public – No insurance	Private – No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	1301.04 (150.81) [1005.45 1596.63]	298.02 (42.79) [214.15 381.89]	−1003.02 (169.31) [−1334.87 −671.17]
$\hat{\boldsymbol{\theta}}_g^{nw}$	1205.295 (259.68) [696.32 1714.27]	−13.23 (55.84) [−122.68 96.22]	−1218.52 (260.12) [−1728.36 −708.68]

(b) Treatment domain mean effect estimates for $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$

Estimators	Public – Private	Public – No insurance	Private – No insurance
$\hat{\boldsymbol{\theta}}_g^{(1)}$	2519.35 (239.67) [2049.60 2989.10]	829.45 (87.41) [658.13 1000.77]	−1689.90 (257.46) [−2194.52 −1185.28]
$\hat{\boldsymbol{\theta}}_g^{nw}$	3207.10 (17.14) [3173.51 3240.69]	4.092 (4.30) [−4.34 12.52]	−2343.00 (180.83) [−2697.43 −1988.57]

The response variable in our study is the annual medical expenditure. The treatment variable is the health insurance type (public health insurance, private health insurance, and no health insurance). Public health insurance is sponsored by Chinese government and is the main health insurance type in China. Six relevant covariates are chosen from the health care questionnaire in our study: age, household register (urban, rural, other), annual household income, physical condition (healthy, just so-so/or a little sick, sick, very sick), chronic disease (yes, no), and treatment to illness (self-treatment, go to hospital, no treatment). Due to some nonresponse units, the final data had a sample size of 3,866. The data weights were adjusted to deal with the nonresponse issue.

We are interested in estimating the following parameters, $\boldsymbol{\theta}_g^0 = E(Y_{ig})$ and $\boldsymbol{\theta}_g^0 = E(Y_{ig}|I_{di} = 1)$ where I_{di} is the indicator for the domain of interest that contains respondents who have sick or very sick physical condition. When estimating $\hat{\pi}_{2ig}$, we use $b_i = 1$ in Equation (7) to obtain conservative estimates since it is difficult to verify SMAR assumption. For comparison, we also report the results using $b_i = w_{1i}^{-1}$ in Equation (7).

Estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$ are calculated and the Jackknife variance estimator discussed in Subsection 3.3 is used to calculate their standard errors. $\hat{\boldsymbol{\theta}}_g^{(2)}$ is not included into the empirical study because π_{1ij} are not available. Since the design is a stratified multi-stage cluster design, we use the districts (PSUs) in different strata as the deleted Jackknife groups S_b. The Jackknife variance estimator is,

$$\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{(1)}\right) = \sum_{h=1}^{2} \frac{B_h - 1}{B_h} \sum_{b=1}^{B_h} \left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)\left(\hat{\boldsymbol{\theta}}_g^{(1)[b]} - \hat{\boldsymbol{\theta}}_g^{(1)}\right)^T, \tag{50}$$

where $\hat{\boldsymbol{\theta}}_g^{(1)[b]}$ is the minimizer of Equation (47) and the replicate weight in the first-phase is defined as,

$$w_{1i}^{[b]} = \begin{cases} 0 & \text{if} \quad i \in S_b \\ \pi_{1i}^{-1} & \text{if} \quad i \notin S_b \text{ and } h(i) \neq h(b) \\ \dfrac{B_h}{B_h - 1}\pi_{1i}^{-1} & \text{if} \quad i \notin S_b \text{ and } h(i) = h(b). \end{cases} \tag{51}$$

Here $h(i)$ is the stratum where unit i belongs to, $h(b)$ is the stratum where the b^{th} deleted group S_b belongs to, and $[B_1, B_2] = [40, 100]$. The replicate estimator $\hat{\boldsymbol{\theta}}_g^{nw[b]}$ for the estimator $\hat{\boldsymbol{\theta}}_g^{nw}$ without survey weights and the variance estimator $\hat{V}_{JK}\left(\hat{\boldsymbol{\theta}}_g^{nw}\right)$ can be obtained in the same way by simply replacing π_{1i} by nN^{-1} in (51). A spline base of degree 2 with 8 equally spaced knots in the data range is constructed for the two continuous variables (age and annual household income). Dummy variables are created for the remaining categorical variables and added to the model.

Table 3 and 4 contain the estimated treatment mean effects and estimated treatment domain mean effects for physical condition, along with standard errors (in parentheses) and 95 perecnt confidence intervals (in brackets), for $b_i = 1$ and $b_i = w_{1i}^{-1}$ cases respectively. The treatment effect estimates in Table 3(a) indicate that, when the data weights are neglected, the estimated mean medical expenditure of the public health insurance group is not significantly different from that of the no health insurance group. However, when the data weights are incorporated, the public health group is found to spend significantly more on the medical expenses than the no health insurance group. This makes sense because people who have no health insurance might be reluctant to spend money to see doctors. This trend is also seen in the domain treatment effects estimates in Table 3(b). In addition, when the data weights are neglected for the treatment mean effect estimates, the estimated mean medical expenditure of the private health insurance group is significantly different from that of the no insurance group, while incorporating the data weights finds these estimated means not significantly different. Table 4 gives the same story as Table 3 when comparing the public health insurance group versus the private health insurance group, and comparing the public health insurance group versus the no health insurance group. However, when comparing the private health group with the no insurance group, Table 4 reports significant difference in the treatment mean effect for both estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{nw}$. Note that the standard errors of the unweighted estimator are not consistently smaller than those of the weighted estimator because the variation of weights in the real data is small (the CV = 0.45).

This study demonstrates that our method is feasible in real data application and suggests that ignoring the weights of an observational data might lead to a misleading conclusion.

6. Conclusions

In this article, we consider a GMM estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ to estimate treatment effects defined through an estimation equation in an observational data set that is a sample drawn by a complex survey design. The estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ include both the first-phase sampling probabilities and the estimated second-phase selection probabilities to remove

the biases due to ignoring unequal sampling design in the first-phase and the selection biases in the second-phase. The self-selection probabilities are estimated using a semiparametric approach in Cattaneo (2010) to deal with the situation with multiple treatments. Our simulation studies demonstrate that neglecting the first-phase design and handling only treatment selection could lead to erroneous treatment effect estimation. The proposed estimator is designed to handle multiple treatments and do not require strong model assumption of the selection probability as in a fully parametric solution. The estimators $\hat{\boldsymbol{\theta}}_g^{(1)}$ and $\hat{\boldsymbol{\theta}}_g^{(2)}$ can be readily extended to multiple sampling phases as well when the data set is a subsample of a larger survey sample.

Appendix

The notation of $|\cdot|$ represents the norm of a matrix, defined as $|A| = \sqrt{trace(A'A)}$ and the notation of $\|\cdot\|$ denotes the sup-norm in all arguments for functions.

We first give regular conditions on the sample designs in both phases. The following notations, I_i, π_i and π_{ij}, denote the sampling indicator, the first and second inclusion probabilities either for the first-phase design or for the second-phase design. For example, $I_i = \delta_{1i}$ or $I_i = \delta_{2ig}$ for any g, and $\pi_i = \pi_{1i}$ or $\pi_i = \pi_{2ig}$ for any g, depending on whether the design if the first-phase design or the second-phase design.

Condition A:

(1) Any variable v_i such that $E[|v_i|^{2+\delta}] < \infty$, where $\delta > 0$, satisfies $\sqrt{n}(\bar{v}_{HT} - \bar{v}_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0, V_\infty)$ *a.s.*, where $(\bar{v}_{HT}, \bar{v}_N) = N^{-1}\sum_{i=1}^N (\pi^{-1}v_i I_i, v_i)$, $V_\infty = lim_{N\to\infty} V_N$, and $V_N = nV(\bar{v}_{HT}|\mathcal{F}_N)$ is the conditional variance of the Horvitz-Thompson estimator (Horvitz and Thompson 1952), $\bar{v}_{HT}$, given $\mathcal{F}_N$.

(2) $nN^{-1} \to f_\infty \in [0, 1]$.

(3) There exist constant C_1, C_2 and C_3 such that $0 < C_1 \le nN^{-1}\pi_i^{-1} < \infty$, and $\left|n(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1}\right| \le C_3 < \infty$ *a.s.*

Condition A(1) and A(2) are regular conditions assumed for a survey design in a finite population framework. Condition A(3) is used in Fuller (2009). The part of condition A(3) related to the joint selection probabilities is used in the proofs to bound sums of covariance induced by the sample design. Condition A(3) holds for simple random sampling, where $(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1} = n^{-1}(n-1)(N-1)^{-1}N - 1$, and for Poisson sampling, where $(\pi_{ij} - \pi_i\pi_j)\pi_i^{-1}\pi_j^{-1} = 0$, and can hold for cluster sampling and stratified sampling. Fuller (2009) explains that a designer has the control to ensure condition A(3). Note that for the second-phase design in our situation, $(\pi_{2ij,g} - \pi_{2ig}\pi_{2jg})\pi_{2ig}^{-1}\pi_{2jg}^{-1} = 0$ for any g because our second-phase design is a multinomial extension of Poisson sampling.

Next we give regular conditions on the tuning parameters of the semiparametric basis. For simplicity, we consider the special case of power series and spline series.

Condition B:

(1) The smallest eigenvalue of $E[R_K(X_i)R_K(X_i)']$ is bounded away from zero uniformly in K.

(2) There exists a sequence of constant $\zeta(K)$ such that $\|R_K(X_i)\| \leq \zeta(K)$ for $K \rightarrow \infty$ and $\zeta(K)K^{1/2}n^{-1/2} \rightarrow 0$.
(3) For all g, $\pi_{2ig}(X_i)$ and $\boldsymbol{\mu}_{mg}(X_i, \boldsymbol{\theta}_g) = E[\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i]$ are s-time differentiable with $sd_x^{-1} \geq 5\eta/2 + 1/2$, where d_x is the dimension of X_i, and $\eta = 1$ or $\eta = 1/2$ depending on whether power series or spline series are used as basis function.
(4) $K = O(n^{\nu})$ with $4sd_x^{-1} - 6\eta \geq \nu^{-1} \geq 4\eta + 2$, where $\eta = 1$ or $\eta = 1/2$ depending on whether power series or spline series are used as basis function.

Condition B(1) and B(2) are standard assumptions and are automatically satisfied in the case of power series or spline series. Condition B(3) and B(4) describe the minimum smoothness required as a function of the dimension of X and the choice of basis, and the relationship between the sample size and the number of bases. Under B(3) and B(4), by Lorentz (1986), there exists a K-vector $\gamma_{g,K}^*$ for any g such that

$$\left\| log\left(\frac{\pi_{2ig}(X)}{1 - \sum_{g=2}^{G}\pi_{2ig}(X)}\right) - R_K^T(X)\gamma_{g,K}^* \right\| = O\left(K^{-\frac{s}{\nu}}\right), \tag{52}$$

where $R_K^T(X)\gamma_{g,K}^*$ is the best L_∞ approximation for the logarithm of the odds ratio of treatment g to the base treatment. The property (52) is used to derive the convergence rate of $\hat{\pi}_{2ig}$ to π_{2ig} as follows,

$$\|\hat{\pi}_{2ig} - \pi_{2ig}\| = O_p(\xi(K)K^{1/2}n^{-1/2} + \xi(K)K^{1/2}K^{-s/d_x}) = o_p(1). \tag{53}$$

For details, see Theorem B-1 of Cattaneo (2010).

Next we give regular conditions on the estimation equation function $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$.

Condition C:

(1) $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ is differentiable with respect to $\boldsymbol{\theta}_g$.
(2) Both $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ and its first derivative with respect to $\boldsymbol{\theta}_g$ have bounded $2 + \delta$ moments. More specifically, $E[|h(Y_i, Z_i; \boldsymbol{\theta})|^{2+\delta}] < M$, where $h\left(Y_i, Z_i; \boldsymbol{\theta}\right)$ denote an element of $\mathbf{m}_{ig}(Y_{ig}, Z_i; \boldsymbol{\theta}_g)$ or an element of its first derivative with respect to $\boldsymbol{\theta}_g$.
(3) $\Gamma_g\left(\boldsymbol{\theta}_g^0\right)$ is full rank.
(4) Assume that $\bar{h}_{HT}(\boldsymbol{\theta}) - \bar{h}_N(\boldsymbol{\theta})$ converges to 0 uniformly in $\boldsymbol{\theta}$, where $\bar{h}_{HT}(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}h_i(Y_i, Z_i; \boldsymbol{\theta})$, $\bar{h}_N(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^{N} h_i(Y_i, Z_i; \boldsymbol{\theta})$, and $h_i(Y_i, Z_i; \boldsymbol{\theta})$ has the same interpretation as in condition C(2) above. This condition means that for all $\epsilon > 0$, there exists a $\delta > 0$ such that $Pro(|\bar{h}_{HT}(\boldsymbol{\theta}) - \bar{h}_N(\boldsymbol{\theta})| > \epsilon) < \delta$, for all N greater than some value M, and for all $\boldsymbol{\theta}$.

A: Proof of Theorem 1

The proof of Theorem 1 proceeds in two steps. The first step is to show that the asymptotic equivalence of $\bar{\mathbf{m}}_{2\pi g}\left(\boldsymbol{\theta}_g\right)$,

$$\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i\in U}\frac{\delta_{1i}\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\frac{\delta_{1i}(\delta_{2ig} - \pi_{2ig})}{\pi_{1i}\pi_{2ig}}\boldsymbol{\mu}_{mg}(X_i; \boldsymbol{\theta}_g) + o_p(n^{-1/2}), \tag{A.1}$$

where $\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g) = E_\xi((\mathbf{m}_{ig}(\boldsymbol{\theta}_g)|X_i)$. In order to show (A.1), we first decompose $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)$ into

$$\begin{aligned}\frac{1}{N}\sum_{i\in A_{2g}}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}} &= \frac{1}{N}\sum_{i\in A_1}\left\{\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\hat{\pi}_{2ig}} - \frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} + \frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}^2}(\hat{\pi}_{2ig}-\pi_{2ig})\right\}\\ &+\frac{1}{N}\sum_{i\in A_1}\left\{-\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}^2}(\hat{\pi}_{2ig}-\pi_{2ig}) + \frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\hat{\pi}_{2ig}-\pi_{2ig})\right\}\\ &+\frac{1}{N}\sum_{i\in A_1}\left\{-\frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\hat{\pi}_{2ig}-\pi_{2ig}) + \frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\delta_{2ig}-\pi_{2ig})\right\}\\ &+\frac{1}{N}\sum_{i\in A_1}\left\{\frac{\delta_{2ig}\mathbf{m}_{ig}(\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}} - \frac{\boldsymbol{\mu}_{mg}(X_i;\boldsymbol{\theta}_g)}{\pi_{1i}\pi_{2ig}}(\delta_{2ig}-\pi_{2ig})\right\}.\end{aligned} \tag{A.2}$$

By the result in (53), the first three terms in (A.2) can be shown to have order $o_p(n^{-1/2})$ asymptotically, which leads to Equation (A.1). Similar arguments can be used to show $\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}\boldsymbol{\eta}_{ig}(\boldsymbol{\theta}_g) + o_p(n^{-1/2})$. The justification of those orders follows Cattaneo (2010), and we refer readers to Cattaneo (2010) for details.

The second step is to show the following two conditions of Pakes and Pollard (1989) hold: (1) $sup_{\boldsymbol{\theta}_g\in\Theta}|\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g))| = o_p(1)$, and (2) for every sequence of real numbers $\delta_n \to 0$, $sup_{|\boldsymbol{\theta}_g-\boldsymbol{\theta}_g^0|\le\delta_n}\left|\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)) - \bar{\mathbf{m}}_{2\pi g}\left(\boldsymbol{\theta}_g^0\right)\right| = o_p(n^{-1/2})$. By Equation (A.1), we can show that

$$\begin{aligned}E(\bar{\mathbf{m}}_{2\pi g}-E(\mathbf{m}_g(\theta_g))^2 &= E\left(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}} - \frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)(\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}} - E\left(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\right)\right)^2\\ &+o\left(n^{-1/2}\right)\\ &\le 2T_{1N}+2T_{2N}+o\left(n^{-1/2}\right),\end{aligned} \tag{A.3}$$

where $T_{1N}=E\left(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}}-E(\mathbf{m}_{ig}(\boldsymbol{\theta}_g))\right)^2$ and $T_{2N}=E\left(\frac{1}{N}\sum_{i\in U}\frac{\mathbf{m}_{ig}(\boldsymbol{\theta}_g)(\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}}\right)^2$. It is easy to show $T_{1N}=O(N^{-1})$ and $T_{2N}=O(N^{-1})$. Then we have $E(\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)-E(\mathbf{m}_g(\boldsymbol{\theta}_g)))^2=O(\frac{1}{N})\Rightarrow\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g)-E(\mathbf{m}_g(\boldsymbol{\theta}_g))=o_p(1)$. Condition (1) of Pakes and Pollard (1989) holds. Similarly, we can show that $sup_{(\boldsymbol{\theta}_g,\mu_z)}|\bar{\mathbf{H}}_{2\pi g}(\boldsymbol{\theta}_g,\mu_z)-E(\boldsymbol{H}_{ig}(\boldsymbol{\theta}_g,\boldsymbol{\mu}_z))|=o_p(1)$.

By Equation (A.1), we can also show that $\bar{\mathbf{m}}_{2\pi g}(\boldsymbol{\theta}_g) - E(\mathbf{m}_g(\boldsymbol{\theta}_g)) - \bar{\mathbf{m}}_{2\pi g}\left(\boldsymbol{\theta}_g^0\right) = T_{3N} - T_{4N} + o_p\left(n^{-1/2}\right)$, where $T_{3N} = \frac{1}{N}\sum_{i\in U}\frac{\left(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)\delta_{1i}\delta_{2ig}}{\pi_{1i}\pi_{2ig}} - E\left(\mathbf{m}_{ig}(\boldsymbol{\theta}_g) - \mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)$ and $T_{4N} = \frac{1}{N}\sum_{i\in U}\frac{E\left[\left(\mathbf{m}_{ig}(\boldsymbol{\theta}_g)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)|X\right](\delta_{2ig}-\pi_{2ig})}{\pi_{2ig}}$. When $\left|\boldsymbol{\theta}_g - \boldsymbol{\theta}_g^0\right| \le \delta_n$, we have

$$E\left(T_{3N}^2\right)=\frac{1}{N}Var\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)$$
$$+E\left[\frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\frac{\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)}{\pi_{1i}}\frac{\mathbf{m}_{jg}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{jg}\left(\boldsymbol{\theta}_g^0\right)}{\pi_{1j}}\right]$$
$$+E\left[\frac{2}{N^2}\sum_{i\in U}\left(\frac{1}{\pi_{2ig}}-1\right)\frac{\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)^2}{\pi_{1i}}\right]\leq\frac{1}{N}O\left(\delta_n^2\right)=o\left(\frac{1}{N}\right) \tag{A.4}$$

$$E\left(T_{4N}^2\right)\leq E\left[\frac{1}{N^2}\sum_{i\in U}E\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\middle|X\right)^2\right]$$
$$\leq E\frac{1}{N}E\left[\left(\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g\right)-\mathbf{m}_{ig}\left(\boldsymbol{\theta}_g^0\right)\right)^2\middle|X\right]\leq\frac{1}{N}O\left(\left|\theta_g-\theta_g^0\right|^2\right)=o\left(\frac{1}{N}\right). \tag{A.5}$$

Then we have $T_{3N}=o_p(n^{-1/2})$ and $T_{4N}=o_p(n^{-1/2})$ when $|\boldsymbol{\theta}-\boldsymbol{\theta}^0|\leq\delta_n$, thus Condition (2) of Pakes and Pollard (1989) is verified. Similarly, we can show that for every sequence of real numbers $\delta_n\rightarrow 0$,

$$sup\left|\begin{bmatrix}\boldsymbol{\theta}_g\\ \mu_z\end{bmatrix}-\begin{bmatrix}\boldsymbol{\theta}_g^0\\ \mu_z^0\end{bmatrix}\right|_{\leq\delta_n}\left|\bar{\mathbf{H}}_{2\pi g}\left(\boldsymbol{\theta}_g,\mu_z\right)-E\left(\mathbf{H}_{ig}\left(\boldsymbol{\theta}_g\right)\right)-\bar{\mathbf{H}}_{2\pi g}\left(\boldsymbol{\theta}_g^0,\mu_z^0\right)\right|=o_p(n^{-1/2}). \tag{A.6}$$

For a vector $c=[c_1,c_2]^T$, we know $|c|\leq\sqrt{2}(|c_1|+|c_2|)$. Therefore, Condition (1) and (2) of Pakes and Pollard (1989) in terms of $\mathbf{H}_{ng}(\boldsymbol{\theta}_g,\mu_z)$ can be verified. The details of the proof can be obtained upon request.

7. References

Ashmead, R. 2014. "Propensity Score Methods for Estimating Causal Effects from Complex Survey Data." Ph.D. Dissertation, Ohio State University. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=osu1417616653.

Berg, E., J.K. Kim, and C. Sinner. 2016. "Imputation under Informative Sampling." *Journal of Survey Statistics and Methodology* 4: 436–462. Doi: 10.1093/jssam/smw032.

Breidt, F.J., G. Claeskens, and J.D. Opsomer. 2005. "Model-Assisted Estimation for Complex Surveys Using Penalised Splines." *Biometrika* 92(4): 831–846. Doi: 10.1093/biomet/92.4.831.

Cattaneo, M.D. 2010. "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability." *Journal of Econometrics* 155(2): 138–154. Doi: 10.1016/j.jeconom.2009.09.023.

DuGoff, E., M. Schuler, and E. Stuart. 2014. "Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys." *Health Services Research* 49(1): 284–303. Doi: 10.1111/1475-6773.12090.

Fuller, W.A. 2009. *Sampling Statistics*, Vol. 56, John Wiley and Sons. Doi: 10.1002/9780470523551.

Hahn, J. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66(2): 315–331. Doi: 10.2307/2998560.

Haziza, D. and J.N.K. Rao. 2006. "A Nonresponse Model Approach to Inference Under Imputation for Missing Survey Data." *Survey Methodology* 32(1): 53. Doi: 12-001-X20060019257.

Hirano, K., G. Imbens, and G. Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4): 1161–1189. Doi: 10.1111/1468-0262.00442.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association* 47: 663–685. Doi: 10.1080/01621459.1952.10483446.

Isaki, C.T. and W.A. Fuller. 1982. "Survey Design under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96. Doi: 10.1080/01621459.1982.10477770.

Kim, J.K. and D. Haziza. 2014. "Doubly Robust Inference with Missing Data in Survey Sampling." *Statistica Sinica* 24: 375–394. Doi: 10.5705/ss.2012.005.

Kim, J.K. A. Navarro, and W. Fuller. 2006. "Replication Variance Estimation for Two-Phase Stratified Sampling." *Journal of the American Statistical Association* 101: 312–320. Doi: 10.1198/016214505000000763.

Little, R.J.A. 1982. "Models for Nonresponse in Sample Surveys." *Journal of the American Statistical Association* 77: 237–250. Doi: 10.1080/01621459.1982.10477792.

Lorentz, G. 1986. *Approximating of Functions*. New York: Chelsea Publishing Company. Doi: 10.1112/jlms/s1-43.1.570b.

Pakes, A. and D. Pollard. 1989. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica* 57: 1027–1057. Doi: 10.2307/1913622.

Pfeffermann, D. 2011. "Modelling of Complex Survey Data: Why Model? Why is it a Problem? How Can we Approach it?" *Survey Methodology* 37: 115–136. Retrieved from https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11602-eng.pdf?st=XWOwbI5k.

Pfeffermann, D. and M. Sverchkov. 1999. "Parametric and Semiparametric Estimation of Regression Models Fitted to Survey Data." *Sankhya B* 61: 166–186. Retrieved from http://www.jstor.org/stable/25053074.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky. 2007. "Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable." *Statistical Science* 22(4): 544–559. Doi: 10.1214/07-STS227D.

Rosenbaum, P.R. and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. Doi: 10.1093/biomet/70.1.41.

Ridgeway, G., S.A. Kovalchik, B.A. Griffin, and M.U. Kabeto. 2015. "Propensity Score Analysis with Survey Weighted Data." *Journal of Causal Inference* 3(2): 237–249. Doi: 10.1515/jci-2014-0039.

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer. Doi: 10.1007/978-1-4612-4378-6.

Tan, Z. 2006. "Regression and Weighting Methods for Causal Inference Using Instrumental Variables." *Journal of the American Statistical Association* 101: 1607–1618. Doi: 10.1198/016214505000001366.

Tan, Z. 2008. "Bounded, Efficient, and Doubly Robust Estimation with Inverse Weighting." *Biometrika* 94: 122. Doi: 10.1093/biomet/asq035.

Yu, C., J. Legg, and B. Liu. 2013. "Estimating Multiple Treatment Effects Using Two-phase Semiparametric Regression Estimators." *Electronic Journal of Statistics* 7(2013): 2737–2761. Doi: 10.1214/13-EJS856.

Zanutto, E. 2006. "A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data." *Journal of Data Science* 4: 67–91. Retrieved from http://www.jds-online.com/v4-1.

Received June 2016
Revised October 2017
Accepted November 2017