Electronic Journal of Statistics
Vol. 7 (2013) 2737–2761
ISSN: 1935-7524
DOI: 10.1214/13-EJS856

Estimating multiple treatment effects using two-phase semiparametric regression estimators

Cindy Yu

Department of Statistics, Iowa State University, Ames, IA 50011, USA
e-mail: cindyyu@iastate.edu

Jason Legg

Global Biostatistical Science, Amgen Inc., Newbury Park, CA 91320, USA
e-mail: jlegg@amgen.com

and

Bin Liu

Department of Statistics, Iowa State University, Ames, IA 50011, USA
e-mail: lbbb@iastate.edu

Abstract: We propose a semiparametric two-phase regression estimator with a semiparametric generalized propensity score estimator for estimating average treatment effects in the presence of the first-phase sampling. The proposed estimator can be easily extended to any number of treatments and does not rely on a prespecified form of the response or outcome functions. The proposed estimator is shown to reduce bias found in standard estimators that ignore the first-phase sample design, and can have improved efficiency compared to the inverse propensity weighted estimators. Results from simulation studies and from an empirical study of NHANES are presented.

Keywords and phrases: Propensity score, semiparametric, treatment effects, two-phase regression estimator.

Received October 2012.

Contents

1 Introduction 2738
2 Proposed two-phase semiparametric regression estimators 2740
  2.1 Basic set-up and the proposed estimator 2740
  2.2 The central limit theorem of $\widehat{\theta}_g$ 2742
3 Simulation study 2748
4 Empirical study 2753
5 Conclusion and remarks 2754
Appendix 2755

Acknowledgments 2759
References 2759

1. Introduction

Timely comparative treatment analysis is useful for physician recommendations, patient awareness, regulatory agency assessments of benefit-risk profiles, and reimbursement agency cost effectiveness assessments. The use of observational data for such purposes has grown significantly in number of studies and importance [14, 24]. In many observational databases subjects can be exposed to one or more treatment options and the treatments and study participation are self-selected. The database is often a sample from a complex survey, such as National Health and Nutrition Examination Survey (NHANES) data, a set of subjects who enroll in a particular insurance policy, a combination of clinical trials as in meta-analysis, or a set of subjects who receive care at centers that shares electronic medical records. Of particular interest due to simple interpretation and practical use in reimbursement are the average expected treatment differences for a population, termed population average treatment effects (ATE) in [8, 15].

For estimating average treatment effects in observational data analysis, literature already contains several approaches including matching estimators [1, 12, 23], inverse probability weighted (IPW) estimators ([9]; [11] for the case of two treatments; [4] in the context of multiple treatments), and doubly robust estimators [2, 17, 18, 26] that tend to be combinations of IPW estimators and outcome regression models. A recent work by [27] considers estimation of treatment effects from two-phase samples, where their observational dataset is a simple random sample from a super-population, a validation sample is drawn using stratified Poisson sampling, and observations grouped by treatment indicators are results from a self-selection process. However, there is not much work done to study the impact of a general design used to obtain the observational data, and to derive the asymptotic results from incorporation of a general first-phase design. Ignoring the sampling design for the analysis dataset can lead to biased estimators of the average treatment effects and incorrect variance estimation. In the following, we quantify the bias due to ignoring the sample design and give a motivation example to emphasize the importance of the sample design.

In general, survey data can be viewed as the outcome of two processes: in the first process the values of random variables are generated for units in a finite population according to a model called the super population model, and in the second process a sample of units is drawn from the finite population according to a sample design, termed the first-phase sample. Analytic inference is made with respect to the super population model. When the sampling probability depends on an auxiliary variable $\mathbf{z}$ or the response variable y, the observed marginal sample likelihood of the response variable y can be altered from the super population likelihood where inference is being made. Therefore sample estimators that ignore first-phase design can be biased for the super population parameters. To quantify the bias, we use the results in [21]. For a random vector $(y, \mathbf{z})$, the sample conditional probability density function (pdf) of y given $\mathbf{z}$ and the sample marginal pdf of $\mathbf{z}$ can be expressed through the super population pdf's as

$$f_s(y|\mathbf{z}) = \frac{E_\xi(\pi|y,\mathbf{z})}{E_\xi(\pi|\mathbf{z})} f_\xi(y|\mathbf{z}), \tag{1.1}$$

$$f_s(\mathbf{z}) = \frac{E_\xi(\pi|\mathbf{z})}{E_\xi(\pi)} f_\xi(\mathbf{z}), \tag{1.2}$$

where $f_s(\cdot)$ and $f_\xi(\cdot)$ are the sample and super population pdf's, $E_s(\cdot)$ and $E_\xi(\cdot)$ denote the expectations under the sample and super population distributions respectively, and π is the sampling probability. In this paper, we are interested in estimating the marginal mean of y, denoted as $\theta = \int\int y f_\xi(y|\mathbf{z}) f_\xi(\mathbf{z}) d\mathbf{z}dy$. The marginal mean estimator that disregards the sampling design is $\theta_s = \int\int y f_s(y|\mathbf{z}) f_s(\mathbf{z}) d\mathbf{z}dy$. Using equations (1.1) and (1.2), the bias in θ_s can be quantified as

$$Bias = \int\int \left(\frac{E_\xi(\pi|y,\mathbf{z})}{E_\xi(\pi)} - 1 \right) y f_\xi(y|\mathbf{z}) f_\xi(\mathbf{z}) d\mathbf{z}dy. \tag{1.3}$$

If π is independent of $(y, \mathbf{z})$, then the bias is zero. If π depends on auxiliary variable $\mathbf{z}$ only, then the bias is

$$Bias = E_\xi \left\{ \left(\frac{\pi(\mathbf{z})}{E_\xi(\pi(\mathbf{z}))} - 1 \right) \mu(\mathbf{z}) \right\}, \tag{1.4}$$

where $\mu(\mathbf{z}) = \mathbf{E}_\xi(\mathbf{y}|\mathbf{z})$. If π depends on y, which is called informative sampling, then the bias is

$$Bias = E_\xi \left\{ \left(\frac{\pi(y,\mathbf{z})}{E_\xi(\pi(y,\mathbf{z}))} - 1 \right) y \right\}. \tag{1.5}$$

In practice, π often depends on auxiliary variables and possibly design variables used for the sample selection but not included in the outcome model under consideration. The probabilities π can depend on the outcome variable in the case of self-selection. Estimators that do not account for the selection effects in the inference can be seriously biased.

As an example of a case where the first-phase sample design is important, consider a finite population generated from a super population model $y_i = \mu + \epsilon_i$, where ϵ_i is a random error variable with mean zero for subject i and y is the outcome of a treatment. Suppose subjects migrate after severe disease progression to larger hospitals with greater treatment options available. If subjects with severe disease progression are also less likely to respond to the treatment, this migration could generate clusters of subjects where subjects with homogeneous ϵ_i values are together in larger hospitals. A study designer selects a cluster sample with probability proportion to the hospital size for convenience as more data can be obtained with fewer hospitals selected. Ignoring the sample design will lead to biases in both mean and variance estimation. An analyst might include disease severity in an outcome model as an auxiliary variable, but an estimator of the marginal distribution of disease severity is needed to estimate the marginal treatment mean. Other examples with details on the importance of accounting for the sampling design can be found in [19, 21]. Due to the potential for biases, it is worthwhile to explore estimators that account for the first-phase sampling design.

In this paper, we propose a two-phase semiparametric regression estimator based on an argument in [3]. The term two-phase is used because we consider the sampling of the observational data as the first phase and subject treatment selection as the second phase. The term semiparametric is used because both the outcome model and treatment selection probabilities are estimated semiparametriclly. The key advantage of our estimator is the incorporation of the first phase sampling, similarly as in [27], thus correcting the biases in estimators that disregard the first phase design information in the ATE estimation. The paper derives asymptotic results for the proposed estimators obtained from incorporating a general first-phase design and including semiparametric estimators of the self-selection probability and outcome models. Moreover, 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 analysis dataset can be a subset selected from a larger sample of the finite population. This case covers the common situation where detailed treatment and outcome data is available for only a subsample of the data such as in a subsample with medical chart adjudication of claims records or a subsample constructed by merging multiple sources of data like claims records and electronic medical records. The proposed estimator that is designed to handle multiple treatments does not require strong model specification as in fully parametric solution and permits incorporating covariate information through regression.

The paper is organized as below. Section 2 introduces the proposed two-phase semiparametric regression estimators and their asymptotic properties. Two simulation studies are presented in Section 3 to compare the proposed estimators to other commonly used estimators. Section 4 contains two examples to illustrate the use of our approach. Section 5 discusses the caveats of using the estimator and possible extensions.

2. Proposed two-phase semiparametric regression estimators

In this section, we introduce our two-phase semiparametric regression estimators. Section 2.1 builds the framework and discusses the motivation of the estimators, and Section 2.2 contains theoretical results for asymptotic consistency and normality of the proposed estimators.

2.1. Basic set-up and the proposed estimator

Let U be a finite population containing $(\mathbf{y}_i, \mathbf{z}_i)$, where $i = 1, \ldots, N$ indexes a subject, $\mathbf{z}_i$ is a set of covariate variables, and $\mathbf{y}_i = [y_{i1}, \ldots, y_{iG}]^T$ is a vector of potential outcomes for G different treatments. Consider $(\mathbf{y}_i, \mathbf{z}_i), i = 1, \ldots, N$, to be i.i.d. realizations from a superpopulation regression model

$$y_{ig} = \mu_{zg}(\mathbf{z}_i) + \epsilon_{ig}, \tag{2.1}$$

where ϵ_{ig} are independent random variables with mean zero and variance $\nu_g(\mathbf{z}_i)$ and $\mu_{zg}(\cdot)$ is a smooth function. Let A_1 with size n index a first phase sample selected from U under a design $p_1(\cdot)$ with π_{1i} as the first order inclusion probabilities, and let A_{2g} $(g = 1, \ldots, G)$ be a collection of disjoint second-phase sample indices that partition the first-phase sample into the G treatment groups. The partitioning can be viewed as a multinomial extension of Poisson sampling with probabilities π_{2ig} (on observables) for subject i

$$\pi_{2ig} = Prob(\delta_{2ig} = 1|\mathbf{z}_i),$$

where δ_{2ig} is the indicator variable of subject i selecting treatment g, $\sum_{g=1}^{G} \delta_{2ig} = 1$, for any i, and δ_{2ig} is independent of δ_{2jh} for any subjects $i \neq j$ and any treatments g and h. The self-selection probabilities π_{2ig} can be impacted by physician/patient preferences and reimbursement guidelines, and are estimated using the sieve estimation approach of [4]. The $\mathbf{z}_i$ are assumed to be observed in A_1 and y_{ig} is observed only in A_{2g}, which is different from [27]where the observed outcome, treatment indicators and covariates are assumed to be available in the population level.

If the outcome model $\mu_{zg}(\mathbf{z}_i)$ and the selection probability model π_{2ig} were known, a two-phase regression estimator of the finite population mean $\bar{y}_{Ng} = N^{-1} \sum_{i \in U} y_{ig}$ is

$$\frac{1}{N}\left(\sum_{i \in A_1} \frac{\mu_{zg}(\mathbf{z}_i)}{\pi_{1i}} + \sum_{i \in A_{2g}} \frac{y_{ig} - \mu_{zg}(\mathbf{z}_i)}{\pi_{1i}\pi_{2ig}}\right), \text{ for any } g. \tag{2.2}$$

Estimator (2.2) is a two-phase sampling extension of the design unbiased difference estimator proposed by [3, 22], and it is usually more efficient relative to the IPW estimator $N^{-1} \sum_{i \in A_{2g}} \pi_{1i}^{-1} \pi_{2ig}^{-1} y_{ig}$ when y_{ig} is correlated with $\mathbf{z}_i$ [22]. In the following, the methods used for estimating the selection probability π_{2ig} and the outcome model $\mu_{zg}(\mathbf{z}_i)$ will be discussed.

We adopt the method in [4] to estimate π_{2ig}. Let $\{r_K(\mathbf{z}_i)\}_{k=1}^{\infty}$ be a sequence of known approximating functions, and assume that π_{2ig} can be approximated by $R_K(\mathbf{z}_i)^T \boldsymbol{\gamma}_{g,K}$ for $K = 1, 2, \ldots,$ where $R_K(\mathbf{z}_i) = [r_1(\mathbf{z}_i), r_2(\mathbf{z}_i), \ldots, r_K(\mathbf{z}_i)]$ and $\gamma_{g,K}$ is the real-valued coefficients of $R_K(\mathbf{z}_i)$ for the g-th treatment selection. Let an estimator of the $K \times G$ matrix $\boldsymbol{\gamma}_K = [\boldsymbol{\gamma}_{1,K}, \boldsymbol{\gamma}_{2,K}, \ldots, \boldsymbol{\gamma}_{G,K}]$ be

$$\hat{\boldsymbol{\gamma}}_K = \underset{\boldsymbol{\gamma}_K | \boldsymbol{\gamma}_{1,K} = \mathbf{0}_K}{argmax} \sum_{i \in A_1} \sum_{g=1}^{G} \delta_{2ig} log \left[\frac{e^{R_K(\mathbf{z}_i)' \boldsymbol{\gamma}_{g,K}}}{\sum_{g=1}^{G} e^{R_K(\mathbf{z}_i)' \boldsymbol{\gamma}_{g,K}}} \right],$$

where $\mathbf{0}_K$ represents a $K \times 1$ vector zeros used to constrain the sum to 1. The estimated probabilities are

$$\begin{aligned} \hat{\pi}_{2ig} &= \frac{e^{R_K(\mathbf{z}_i)'\widehat{\boldsymbol{\gamma}}_{g,K}}}{1+\sum\limits_{g=2}^{G} e^{R_K(\mathbf{z}_i)'\widehat{\boldsymbol{\gamma}}_{g,K}}} && \text{for g=2,3,\ldots,G} \\ &= \left(1+\sum_{g=2}^{G} e^{R_K(\mathbf{z}_i)'\widehat{\boldsymbol{\gamma}}_{g,K}}\right)^{-1} && \text{for g=1.} \end{aligned} \tag{2.3}$$

This solution is that of multinomial logistic regression. Condition B in the Appendix specifies assumptions about $R_K(\mathbf{z}_i)$, π_{2ig} and K to ensure $\hat{\pi}_{2ig}$ converges to π_{2ig} fast enough. Choices for the $r_K(\mathbf{z}_i)$ include power series, spline, and kernel expansions.

We propose estimating the g-th outcome model $\mu_{zg}(\mathbf{z}_i)$ with a semiparametric regression estimator using the base $R_K(\mathbf{z}_i)$ as in (2.3). The benefit is that the estimator has a semiparametric specification for both the probabilities and the mean functions. Let $\widehat{\mu}_{zg}(\mathbf{z}_i)$ be the predicted values for all i in A_1, and the regression is fit with elements indexed in A_{2g},

$$\widehat{\mu}_{zg}(\mathbf{z}_i) = R_K(\mathbf{z}_i)^T\widehat{\boldsymbol{\beta}}_{zg}, \tag{2.4}$$

where

$$\widehat{\boldsymbol{\beta}}_{zg} = \left(\sum_{i\in A_{2g}} \pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1} R_K(\mathbf{z}_i)R_K(\mathbf{z}_i)^T\right)^{-1} \sum_{i\in A_{2g}} \pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1} R_K(\mathbf{z}_i) y_{ig}, \tag{2.5}$$

where $R_K(\mathbf{z}_i)$ includes the intercept through the entire paper. Combining (2.2), (2.3) and (2.4), our two-phase semiparametric regression estimator for g-th marginal treatment mean is

$$\widehat{\theta}_g = \frac{1}{N}\sum_{i\in A_1} \frac{\widehat{\mu}_{zg}(\mathbf{z}_i)}{\pi_{1i}} + \frac{1}{N}\sum_{i\in A_{2g}} \frac{y_{ig}-\widehat{\mu}_{zg}(\mathbf{z}_i)}{\pi_{1i}\hat{\pi}_{2ig}}, \text{ for any } g = 1,\ldots,G. \tag{2.6}$$

2.2. The central limit theorem of $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$

The asymptotic consistency and normality of $\widehat{\theta}_g$ are established in Theorem 1 on the finite population level, and in Corollary 1 on the super-population level. For the design properties, we use the traditional finite population asymptotic framework, in which the population U and the designs are embedded into a sequence of such populations index by $\mathcal{F}_N$ with $N \to \infty$. The $o_p(\cdot)$ and $\to$ notations below are with respect to this sequence of populations and designs, see [16].

Theorem 1. *Under the regularity conditions in the Appendix,*
(i) $\widehat{\theta}_g - \bar{y}_{Ng}|\mathcal{F}_N = o_p(1)$,
(ii)

$$(V_{1g} + V_{2g})^{-\frac{1}{2}}(\widehat{\theta}_g - \bar{y}_{Ng})|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(0,1), \textit{ where}$$

$$V_{1g} = E\left\{V(\bar{\epsilon}_{2\pi,g}|A_1, \mathcal{F}_N)\right\} \tag{2.7}$$

$$V_{2g} = V(\bar{e}_{1\pi,g} + \boldsymbol{\beta}_{zg}^T \overline{R}_{z,1\pi}|\mathcal{F}_N) \tag{2.8}$$

$$\bar{y}_{Ng} = N^{-1}\sum_{i\in U} y_{ig}, \quad \overline{R}_{z,1\pi} = N^{-1}\sum_{i\in A_1} \pi_{1i}^{-1} R_K(\mathbf{z}_i) \tag{2.9}$$

$$\bar{e}_{1\pi,g} = N^{-1}\sum_{i\in A_1} \pi_{1i}^{-1} e_{ig}, \quad \bar{\epsilon}_{2\pi,g} = N^{-1}\sum_{i\in A_{2g}} \pi_{1i}^{-1}\pi_{2ig}^{-1}\epsilon_{ig} \tag{2.10}$$

$$e_{ig} = y_{ig} - R_K(\mathbf{z}_i)^T \boldsymbol{\beta}_{zg}, \quad \epsilon_{ig} = y_{ig} - \mu_g(\mathbf{z}_i) \tag{2.11}$$

and $\boldsymbol{\beta}_{zg} = \lim_{N\to\infty}\left(\sum_{i\in U} R_K(\mathbf{z}_i)R_K(\mathbf{z}_i)^T\right)^{-1}\sum_{i\in U} R_K(\mathbf{z}_i)y_{ig}$.

Two key steps in the proof (details in the Appendix) are to show

$$\widehat{\theta}_g - \bar{y}_{Ng} = (\overline{R}_{z,1\pi} - \overline{R}_{z,N})^T \boldsymbol{\beta}_{zg} + \frac{1}{N}\sum_{i\in A_{2g}} \frac{e_{ig}}{\pi_{1i}\hat{\pi}_{2ig}} - \frac{1}{N}\sum_{i\in U} e_{ig} + o_p(n^{-\frac{1}{2}}), \tag{2.12}$$

and

$$\frac{1}{N}\sum_{i\in A_{2g}} \frac{e_{ig}}{\pi_{1i}\hat{\pi}_{2ig}} = \frac{1}{N}\sum_{i\in A_1}\left\{\frac{\delta_{2ig}e_{ig}}{\pi_{1i}\pi_{2ig}} - \frac{\delta_{2ig} - \pi_{2ig}}{\pi_{1i}\pi_{2ig}} E(e_{ig}|\mathbf{z}_i)\right\} + o_p(n^{-\frac{1}{2}}). \tag{2.13}$$

Combining (2.12) and (2.13) gives

$$\widehat{\theta}_g - \bar{y}_{Ng} = (\bar{\epsilon}_{2\pi,g} - \bar{\epsilon}_{1\pi,g}) + (\bar{\epsilon}_{1\pi,g} - \bar{\epsilon}_{Ng}) + \beta_{zg}^T(\overline{R}_{z,1\pi} - \overline{R}_{z,N}) + o_p(n^{-\frac{1}{2}}), \tag{2.14}$$

where $\bar{\epsilon}_{1\pi,g} = N^{-1}\sum_{i\in A_1} \pi_{1i}^{-1}\epsilon_{ig}$ and $\bar{\epsilon}_{Ng} = N^{-1}\sum_{i\in U} \epsilon_{ig}$. This leads to the asymptotic results in Theorem 1.

Remark 1. The result in Theorem 1 holds so long as $\hat{\mu}_{zg}(\mathbf{z}_i)$ is consistent for some quantity that does not necessarily need to be $\mu_{zg}(\mathbf{z}_i)$, but the efficiency improves if $\hat{\mu}_{zg}(\mathbf{z}_i)$ approximates $\mu_{zg}(\mathbf{z}_i)$ well. Intuitively, if $\hat{\mu}_{zg}(\mathbf{z}_i)$ approximates the true $\mu_{zg}(\mathbf{z}_i)$ well, the values of $e_{ig} = y_{ig} - R_K(\mathbf{z}_i)^T\boldsymbol{\beta}_{zg}$ are small, thus $V(\bar{e}_{1\pi g}|\mathcal{F}_N)$ which is a component of V_{2g} in (2.8) becomes smaller, relative to the situation where $\hat{\mu}_{zg}(\mathbf{z}_i)$ is a poor approximation of $\mu_{zg}(\mathbf{z}_i)$. The impact can be seen under a simple random sample design (SRS), in which $V(\bar{e}_{1\pi g}|\mathcal{F}_N) = (1 - nN^{-1})n^{-1}S_{eg}^2$, where S_{eg}^2 is the variance of e_{ig}'s. However, the proof used to show the consistency in (i) of Theorem 1 does not require the consistency of $\hat{\mu}_{zg}(\mathbf{z}_i)$ to $\mu_{zg}(\mathbf{z}_i)$.

Remark 2. Our estimator performs better in terms of bias than the commonly used naive IPW estimator that ignores the first phase design, $\widehat{\theta}_g^{na-ipw} = n^{-1}\sum_{i\in A_{2g}} \hat{\pi}_{2ig}^{-1} y_{ig}$.

To quantify the bias, write

$$\begin{aligned}\widehat{\theta}_g^{na-ipw} - \bar{y}_{Ng} \quad &= \quad \left(\frac{1}{n} \sum_{i \in A_1} \epsilon_{ig} - \frac{1}{N} \sum_{i \in A_1} \frac{\epsilon_{ig}}{\pi_{1i}} \right) \\ &\quad + \left(\frac{1}{n} \sum_{i \in A_1} \mu_{zg}(\mathbf{z}_i) - \frac{1}{N} \sum_{i \in A_1} \frac{\mu_{zg}(\mathbf{z}_i)}{\pi_{1i}} \right) \\ &\quad + (\bar{y}_{1\pi g} - \bar{y}_{Ng}) + o_p(n^{-1/2}).\end{aligned}$$

Taking an expectation gives the asymptotic bias of $\widehat{\theta}_g^{na-ipw}$ as

$$Bias = E_\xi \left\{ \left(\frac{N}{n} \pi_{1i} - 1 \right) \mu_{zg}(\mathbf{z}) \right\}.$$

The magnitude of the bias depends on the correlation between the first-phase inclusion probabilities, π_{1i}, and the error in the outcome model implied by the naive IPW estimator ignoring the first-phase.

Our estimator can gain efficiency relative to the IPW estimator that incorporates the first phase sampling,

$$\widehat{\theta}_g^{ipw} = \frac{1}{N} \sum_{i \in A_{2g}} \frac{y_{ig}}{\pi_{1i} \hat{\pi}_{2ig}}. \tag{2.15}$$

To see this, we assume $R_k(\mathbf{z}) = z$ for a univariate covariate z without loss of generality. Our estimator $\widehat{\theta}_g$ can be written as

$$\begin{aligned}\widehat{\theta}_g \quad &= \quad \tilde{y}_{2\pi g} - \widehat{\beta}_{zg}(\tilde{z}_{2\pi g} - \bar{z}_{1\pi}) \\ &= \quad \tilde{y}_{2\pi g} - \beta_{zg}(\tilde{z}_{2\pi g} - \bar{z}_{1\pi}) - (\widehat{\beta}_{zg} - \beta_{zg})(\tilde{z}_{2\pi g} - \mu_z) \\ &\quad + (\widehat{\beta}_{zg} - \beta_{zg})(\bar{z}_{1\pi} - \mu_z),\end{aligned} \tag{2.16}$$

where $\tilde{y}_{2\pi g} = \widehat{\theta}_g^{ipw}$, $\tilde{z}_{2\pi g} = N^{-1} \sum_{i \in A_{2g}} z_i \pi_{1i}^{-1} \hat{\pi}_{2ig}^{-1}$, $\bar{z}_{1\pi} = N^{-1} \sum_{i \in A_1} z_i \pi_{1i}^{-1}$ and μ_z is the marginal mean of z. Because $\tilde{z}_{2\pi g} - \mu_z = O_p(n^{-1/2})$, $\bar{z}_{1\pi} - \mu_z = O_p(n^{-1/2})$ and $\widehat{\beta}_{zg} - \beta_{zg} = o_p(1)$, then

$$\widehat{\theta}_g = \widehat{\theta}_g^{ipw} - \beta_{zg}(\tilde{z}_{2\pi g} - \bar{z}_{1\pi}) + o_p(n^{-1/2}),$$

and

$$Var(\widehat{\theta}_g) \approx Var(\widehat{\theta}_g^{ipw}) + \beta_{zg}^2 Var(\tilde{z}_{2\pi g} - \bar{z}_{1\pi}) - 2 * \beta_{zg} Cov(\tilde{y}_{2\pi g}, \tilde{z}_{2\pi g} - \bar{z}_{1\pi}).$$

Our $\widehat{\theta}_g$ has a smaller variance of the linearized term than $\widehat{\theta}_g^{ipw}$ when the condition, $\beta_{zg}^2 Var(\tilde{z}_{2\pi g} - \bar{z}_{1\pi}) < 2 * \beta_{zg} Cov(\tilde{y}_{2\pi g}, \tilde{z}_{2\pi g} - \bar{z}_{1\pi})$, holds. This condition will often hold when y_{ig} and z_i are correlated and the outcome model is approximately correctly specified. Simulation studies in Section 3 illustrate cases where this efficiency gain occurs. This indicates that a combination of regression and use of estimated propensity scores can give further improvement than using estimated propensity scores alone, which is noted by several authors including [15, 25].

Remark 3. When a subset of $\mathbf{z}_i$, called $\mathbf{x}_i$, is available on the population level, estimator $\widehat{\theta}_g$ can be easily extended to incorporate this additional information.

For example, this case can occur when there are some demographic variables available in the frame. Let $\widehat{\mu}_{xg}(\mathbf{x}_i)$ for $i \in U$ denote the predicted values for the model relating y_{ig} to the $\mathbf{x}_i$. The extended three-phase estimator is

$$\widehat{\theta}_{g,p} = \frac{1}{N}\left(\sum_{i\in U}\widehat{\mu}_{xg}(\mathbf{x}_i) + \sum_{i\in A_1}\frac{\widehat{\mu}_{zg}(\mathbf{z}_i) - \widehat{\mu}_{xg}(\mathbf{x}_i)}{\pi_{1i}} + \sum_{i\in A_{2g}}\frac{y_{ig} - \widehat{\mu}_{zg}}{\pi_{1i}\hat{\pi}_{2ig}}\right), \quad (2.17)$$

where

$$\widehat{\boldsymbol{\beta}}_{xg} = \left(\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(\mathbf{x}_i)R_K(\mathbf{x}_i)^T\right)^{-1}\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(\mathbf{x}_i)y_{ig},$$

$\widehat{\mu}_{xg}(\mathbf{x}_i) = R_K(\mathbf{x}_i)^T\widehat{\boldsymbol{\beta}}_{xg}$, and the $R_K(\mathbf{x}_i)$ is the base constructed using $\mathbf{x}_i$. The asymptotic properties of $\widehat{\theta}_{g,p}$ and its variance estimation are given in Appendix C, where it is shown that the asymptotic variance of $\widehat{\theta}_{g,p}$, denoted by $AV(\widehat{\theta}_{g,p}|\mathcal{F}_N)$, is

$$AV(\widehat{\theta}_{g,p}|\mathcal{F}_N) = E\left\{V(\bar{\epsilon}_{2\pi,g})|\mathcal{F}_N\right\} + V\left\{\bar{a}_{1\pi,g}|\mathcal{F}_N\right\}, \quad (2.18)$$

where $\bar{a}_{1\pi,g} = N^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}a_{ig}$, $a_{ig} = y_{ig} - R_K(\mathbf{x}_i)^T\boldsymbol{\beta}_{xg}$ and $\boldsymbol{\beta}_{xg} = \lim_{N\to\infty}(\sum_{i\in U}R_K(\mathbf{x}_i)R_K(\mathbf{x}_i)^T)^{-1}\sum_{i\in U}R_K(\mathbf{x}_i)y_{ig}$. Comparing (2.18) to the asymptotic variance of $\widehat{\theta}_g$ which can also be expressed as

$$AV(\widehat{\theta}_g|\mathcal{F}_N) = E\left\{V(\bar{\epsilon}_{2\pi,g})|\mathcal{F}_N\right\} + V\left\{\bar{y}_{1\pi,g}|\mathcal{F}_N\right\}, \quad (2.19)$$

where $\bar{y}_{1\pi,g} = N^{-1}\sum_{i\in A_1}y_{ig}$. It can be seen that $\widehat{\theta}_{g,p}$ is usually more efficient than $\widehat{\theta}_g$ when y_{ig} is correlated with $\mathbf{x}_i$. The efficiency gain occurs because the second term in (2.18), $V\left\{\bar{a}_{1\pi,g}|\mathcal{F}_N\right\}$, is likely smaller than the second term in (2.19), $V\left\{\bar{y}_{1\pi,g}|\mathcal{F}_N\right\}$, when $R_K(\mathbf{x}_i)^T\boldsymbol{\beta}_{xg}$ can explain part of the variation in y_{ig}. In general, $V(\bar{y}_{1\pi g}|\mathcal{F}_N) = N^{-2}\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1i}^{-1}\pi_{1j}^{-1}y_{ig}y_{jg}$ and $V(\bar{a}_{1\pi g}|\mathcal{F}_N) = N^{-2}\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1i}^{-1}\pi_{1j}^{-1}a_{ig}a_{jg}$, where $\Delta_{1ij} = \pi_{1ij} - \pi_{1i}\pi_{1j}$ and π_{1ij} is the joint inclusion probability in the first phase. Assuming the SRS design is used, $V(\bar{y}_{1\pi g}|\mathcal{F}_N) = (1 - nN^{-1})n^{-1}S_{yg}^2$ and $V(\bar{a}_{1\pi g}|\mathcal{F}_N) = (1 - nN^{-1})n^{-1}S_{ag}^2$, where S_{yg}^2 and S_{ag}^2 are the variances of y_{ig} and a_{ig}. S_{yg}^2 tends to be larger than S_{ag}^2 if y_{ig} can be well approximated by $R_K(\mathbf{x}_i)^T\boldsymbol{\beta}_{xg}$. An extreme example is if $y_{ig} = R_K(\mathbf{x}_i)^T\boldsymbol{\beta}_{xg}$, then $V(\bar{a}_{1\pi g}|\mathcal{F}_N) = 0$, but $V(\bar{y}_{1\pi g}|\mathcal{F}_N) = N^{-2}\boldsymbol{\beta}_{xg}^T(\sum_{i\in U}\sum_{j\in U}\Delta_{1ij}\pi_{1i}^{-1}\pi_{1j}^{-1}R_K(\mathbf{x}_i)R_K(\mathbf{x}_i)^T)\boldsymbol{\beta}_{xg} > 0$. If only control totals are known for the population, a linear regression model can be used to estimate $\mu_{xg}(\mathbf{x}_i)$. The estimator $\widehat{\theta}_{g,p}$ in (2.17) can then be written as

$$\widehat{\theta}_{g,p} = \frac{1}{N}\left(\sum_{i\in U}\mathbf{x}_i - \sum_{i\in A_1}\frac{\mathbf{x}_i}{\pi_{1i}}\right)^T \widehat{\boldsymbol{\beta}}_{xg} + \frac{1}{N}\sum_{i\in A_1}\frac{\widehat{\mu}_{zg}(\mathbf{z}_i)}{\pi_{1i}} + \frac{1}{N}\sum_{i\in A_{2g}}\frac{y_{ig}-\widehat{\mu}_{zg}(\mathbf{z}_i)}{\pi_{1i}\hat{\pi}_{2ig}}. \tag{2.20}$$

It is worth noting that $\widehat{\theta}_g$ takes a similar form to the Simple Doubly Robust (SDR) estimator in [27], if assuming both the outcome and self-selection models are correctly specified in their set-up. The differences are that we use the known population size N and estimate π_{2ig} and $\mu_{zg}(\mathbf{z}_i)$ semiparametrically, while they use $\sum_{i\in A_1}\pi_{1i}^{-1}$ in place of N and estimate π_{2ig} and $\mu_{zg}(\mathbf{z}_i)$ parametrically. The distinction between using parametric and semiparametric estimation arises in the asymptotic results. The SDR will suffer efficiency loss if one of π_{2ig} and $\mu_{zg}(\mathbf{z}_i)$ models is wrong. The similarity between their SDR and our $\widehat{\theta}_g$ is not surprising since the SDR does not use non-validation (population level) data and we do not have population level data to use. If the first-phase is simple random sampling and the covariate is known for the whole population, then our estimator devolves into the estimator from [4], shown to be semiparametric efficient.

While Theorem 1 shows conditional convergence together for $\widehat{\theta}_g$ and $\bar{y}_{Ng}$, the goal typically is to make inference for g-th marginal treatment mean on the superpopulation level. The following corollary extends the results of $\widehat{\theta}_g$ on the finite population level to the superpopulation level with a sketch of the proof in the Appendix.

Corollary 1. *Assume $\{\mathbf{z}_i,\mathbf{y}_i\}_{i=1}^N$ are i.i.d. realizations from the super-population model (2.1), then under the conditions in the Appendix*

(i) $\widehat{\theta}_g - \theta_g^* = o_p(1)$,

(ii)

$$\left\{E_\xi(V_{1g}+V_{2g}) + \frac{\sigma_{yg}^2}{N}\right\}^{-\frac{1}{2}}(\widehat{\theta}_g - \theta_g^*) \to N(0,1) \textit{ in distribution,} \tag{2.21}$$

where $E_\xi[y_{ig}] = \theta_g^*$, $\sigma_{yg}^2 = V_\xi(y_{ig})$, V_{1g} and V_{2g} are the same as in (2.7) and (2.8), and $E_\xi(\cdot)$ and $V_\xi(\cdot)$ here are with respect to the randomness on the super-population.

In order to make inference, we next propose a variance estimator $\widehat{V}(\widehat{\theta}_g)$ and prove its consistency in Theorem 2. An estimator of V_{1g} in (2.7) is

$$\hat{V}_{1g} = \hat{V}(\bar{\hat{\epsilon}}_{2\pi g}) = \frac{1}{N^2}\sum_{i\in A_{2g}}(1-\hat{\pi}_{2ig})\frac{\hat{\epsilon}_{ig}^2\pi_{1i}^{-2}}{\hat{\pi}_{2ig}^2}, \tag{2.22}$$

where

$$\hat{\epsilon}_{ig} = y_{ig} - R_K(\mathbf{z}_i)^T\widehat{\boldsymbol{\beta}}_{gz}. \tag{2.23}$$

An estimator of V_{2g} is

$$\hat{V}_{2g} = \hat{M}_{1g} + \hat{M}_{2g} + \hat{M}_{3g}, \tag{2.24}$$

where

$$\hat{M}_{1g} = \frac{1}{N^2} \sum_{i \in A_{2g}} \sum_{j \in A_{2g}} \frac{\Delta_{1ij}}{\pi_{1ij}\hat{\pi}_{2ij,g}} \frac{\hat{e}_{ig}}{\pi_{1i}} \frac{\hat{e}_{jg}}{\pi_{1j}}, \tag{2.25}$$

$$\hat{M}_{2g} = \hat{\boldsymbol{\beta}}_{zg}^T \hat{V}(\overline{R}_{z,1\pi}) \hat{\boldsymbol{\beta}}_{zg}, \tag{2.26}$$

$$\hat{M}_{3g} = 2\hat{\boldsymbol{\beta}}_{zg}^T \frac{1}{N^2} \sum_{i \in A_{2g}} \sum_{j \in A_1} \frac{\Delta_{1ij}}{\pi_{1ij}\hat{\pi}_{2ig}} \frac{\hat{e}_{ig}}{\pi_{1i}} \frac{R_K(\mathbf{z}_j)}{\pi_{1j}}, \tag{2.27}$$

$$\hat{V}(\overline{R}_{z,1\pi}) = \frac{1}{N^2} \sum_{i \in A_1} \sum_{j \in A_1} \frac{\Delta_{1ij}}{\pi_{1ij}} \frac{R_K(\mathbf{z}_i)}{\pi_{1i}} \frac{R_K(\mathbf{z}_j)^T}{\pi_{1j}}, \tag{2.28}$$

and $\hat{e}_{ig}$ is calculated the same way as $\hat{\epsilon}_{ig}$ in (2.23) and $\hat{\pi}_{2ij,g} = \hat{\pi}_{2ig}\hat{\pi}_{2jg}$ if $i \neq j$ and $\hat{\pi}_{2ij,g} = \hat{\pi}_{2ig}$ if $i = j$. An estimator of σ_g^2 is

$$\begin{aligned} \hat{\sigma}_g^2 \quad &= \quad \frac{1}{N} \sum_{i \in A_1} \frac{\hat{\mu}_{zg}(\mathbf{z}_i)^2}{\pi_{1i}} - \left(\frac{1}{N} \sum_{i \in A_1} \frac{\hat{\mu}_{zg}(\mathbf{z}_i)}{\pi_{1i}} \right)^2 + \frac{1}{N} \sum_{i \in A_{2g}} \frac{\hat{\epsilon}_{ig}^2}{\pi_{1i}\hat{\pi}_{2ig}} \\ &\quad - \left(\frac{1}{N} \sum_{i \in A_{2g}} \frac{\hat{\epsilon}_{ig}}{\pi_{1i}\hat{\pi}_{2ig}} \right)^2. \end{aligned} \tag{2.29}$$

Combining (2.22), (2.24) and (2.29), the variance estimator for the asymptotic variance in (2.21) is

$$\widehat{V}(\widehat{\theta}_g) = \hat{V}_{1g} + \hat{V}_{2g} + \frac{\hat{\sigma}_g^2}{N}. \tag{2.30}$$

The following theorem gives the consistency of $\widehat{V}(\widehat{\theta}_g)$ and the central limit theory using $\widehat{V}(\widehat{\theta}_g)$.

Theorem 2. *Under the conditions in the Appendix,*

(i) $\widehat{V}(\widehat{\theta}_g) = E_{\xi}(V_{1g} + V_{2g}) + \frac{\sigma_g^2}{N} + o_p(n^{-1})$.

(ii) $\widehat{V}(\widehat{\theta}_g)^{-\frac{1}{2}}(\widehat{\theta}_g - \theta_g) \to N(0,1)$ *in distribution.*

A short proof is provided in Appendix B.

To obtain the inference for treatment effects and other functions of treatment means, we need to estimate $\boldsymbol{\lambda}^T\boldsymbol{\theta}^*$, where $\boldsymbol{\lambda}$ is any real-valued vector and $\boldsymbol{\theta}^* = [\theta_1^*, \ldots, \theta_g^*]^T$ is the vector of marginal treatment means from the superpopulation model. As an example, an average treatment effect $\theta_1^* - \theta_2^* = \boldsymbol{\lambda}^T\boldsymbol{\theta}^*$ where $\boldsymbol{\lambda} = [1, -1, 0, \ldots, 0]^T$. The estimator for $\boldsymbol{\theta}^*$ is $\widehat{\boldsymbol{\theta}} = [\hat{\theta}_1, \ldots, \hat{\theta}_G]^T$. The same proof in Appendix A can be directly applied to show the asymptotic consistency and normality of $\widehat{\boldsymbol{\theta}}$. If we denote the cell (g,h) of a matrix M by $[M]_{(g,h)}$ and define $\hat{\boldsymbol{\beta}}_z = [\hat{\boldsymbol{\beta}}_{z1}, \ldots, \hat{\boldsymbol{\beta}}_{zG}]$, the variance estimator for $\widehat{\boldsymbol{\theta}}$ can be expressed similarly as

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \hat{\mathbf{V}}_1 + \hat{\mathbf{M}}_1 + \hat{\mathbf{M}}_2 + \hat{\mathbf{M}}_3 + N^{-1}\hat{\boldsymbol{\Sigma}}, \tag{2.31}$$

where $\hat{\mathbf{V}}_1 = diag\{\widehat{V}(\hat{V}_{1g})\}_{g=1,\ldots,G}$,

$$[\hat{\mathbf{M}}_1]_{(g,h)} = \frac{1}{N^2}\sum_{i\in A_{2g}}\sum_{j\in A_{2h}}\frac{\Delta_{1ij}}{\pi_{1ij}\hat{\pi}_{2ig}\hat{\pi}_{2jh}}\frac{\hat{e}_{ig}}{\pi_{1i}}\frac{\hat{e}_{jh}}{\pi_{1j}} \text{ for } g,h=1,\ldots,G,$$

$$\hat{\mathbf{M}}_2 = \hat{\boldsymbol{\beta}}_z^{\;T}\hat{V}(\overline{R}_{z,1\pi})\hat{\boldsymbol{\beta}}_z,$$

$$\hat{\mathbf{M}}_3 = 2\hat{\boldsymbol{\beta}}_z^{\;T} \times [\hat{\mathbf{M}}_{31},\ldots,\hat{\mathbf{M}}_{3G}] \text{ with } \hat{\mathbf{M}}_{3g} = \frac{1}{N^2}\sum_{i\in A_{2g}}\sum_{j\in A_1}\frac{\Delta_{1ij}}{\pi_{1ij}\hat{\pi}_{2ig}}\frac{\hat{e}_{ig}}{\pi_{1i}}\frac{R_K(\mathbf{z}_j)}{\pi_{1j}},$$

$$[\hat{\mathbf{\Sigma}}]_{(g,h)} = \frac{1}{N}\sum_{i\in A_1}\frac{\hat{\mu}_{zg}(\mathbf{z}_i)\hat{\mu}_{zh}(\mathbf{z}_i)}{\pi_{1i}} - \left\{\frac{1}{N}\sum_{i\in A_1}\frac{\hat{\mu}_{zg}(\mathbf{z}_i)}{\pi_{1\pi}}\right\}\left\{\frac{1}{N}\sum_{i\in A_1}\frac{\hat{\mu}_{zh}(\mathbf{z}_i)}{\pi_{1i}}\right\}$$

for $g,h=1,\ldots,G$ and $g \neq h$, and the diagonal cells $[\hat{\mathbf{\Sigma}}]_{(g,g)}$ are the same as $\hat{\sigma}_g^2$ in (2.29). The similar arguments of Theorem 2 can be used to show the consistency of this estimator. The central limit theorem for any linear combination estimator $\boldsymbol{\lambda}^T\widehat{\boldsymbol{\theta}}$ follows immediately.

3. Simulation study

In this section, we provide two simulation examples to illustrate the performance of our two-phase semiparametric regression estimators of average treatment effects. In both examples, we consider three treatment levels and population and sample sizes $(N,n) = (12500, 250), (25000, 500)$ and $(50000, 1000)$ to illustrate convergence. These simulations are intended to demonstrate that in two-phase sampling problems, ignoring the first-phase and handling only treatment selection can lead to erroneous conclusions. The simulations will also show there are potential efficiency gains by incorporating population control data, which is often ignored in treatment comparison studies. The first phase designs chosen for the two examples are stratified and probability proportional to size sampling, which are two commonly used designs for data selection.

Example 1. We specify the simulation set-up as follows.

(1) Covariates: $\mathbf{z}_i = [z_{i1}, z_{i2}, z_{i3}]^T$ where z_{ij} is i.i.d from $Uniform[-2,2]$ for all $j = 1,2$ and 3. (2) Outcome models: the population U is stratified into two equal size strata U_t $(t = 1,2)$, in which the g-th outcome is generated as

$$\begin{aligned} y_{ig}^{(t)} &= \mu_{hg} + \beta_{g1}z_{1i} + \beta_{g2}(z_{1i}^2 - 4/3) + \beta_{g3}z_{1i}^3 \\ &\quad + \gamma_{g1}z_{2i} + \gamma_{g2}(z_{2i}^2 - 4/3) + \gamma_{g3}z_{2i}^3 + \delta_{g1}z_{i3} + \delta_{g2}z_{i3}^3 + \epsilon_{ig}, \quad (3.1)\end{aligned}$$

where $\epsilon_{ig} \sim Laplace(0,1)$, $[\beta_{11},\beta_{21},\beta_{31}] = [2,2,2]$, $[\beta_{12},\beta_{22},\beta_{32}] = [2,2,0]$, $[\beta_{13},\beta_{23},\beta_{33}] = [-2,-2,-2]$, $[\gamma_{11},\gamma_{21},\gamma_{31}] = [1,2,1]$, $[\gamma_{12},\gamma_{22},\gamma_{32}] = [-1, -2,-1]$, $[\gamma_{13},\gamma_{23},\gamma_{33}] = [2,-2,0]$, $[\delta_{11},\delta_{21},\delta_{31}] = [2,2,-2]$, $[\delta_{12},\delta_{22},\delta_{32}] = [0,0,2]$. And $[\mu_{11},\mu_{12},\mu_{13}] = [8,2/3,-8]$ for U_1, and $[\mu_{21},\mu_{22},\mu_{23}] = [-12, -20/3,12]$ for U_2. By design, all the terms in (3.1) except for the intercepts have mean zero, thus $E(y_{ig}^{(h)}) = \mu_{hg}$. The order of the means in U_1 is $Trt1 > Trt2 >$

$Trt3$ and in U_2 is $Trt1 < Trt2 < Trt3$. The overall marginal means are $E(y_{i1}) = -2, E(y_{i2}) = 0$ and $E(y_{i3}) = 2$. (3) First phase sampling: stratified random sampling with 80% of the sample coming from U_1 and 20% from U_2. For units in stratum t ($t = 1$ or 2), $\pi_{1i} = N_t^{-1} n_t$ and $\pi_{1ij} = \{N_t(N_t - 1)\}^{-1} n_t(n_t - 1)$, where n_t and N_t are the first phase sample size and stratum population size in stratum t. The joint including probability for two units in different strata is zero. (4) Second phase selection:

$$\pi_{2ig} = \frac{\exp\{\phi_{0g} + \phi_{1g} z_{1i} + \phi_{2g} z_{2i} + \phi_{3g}(z_{2i}^2 - 4/3)\}}{\sum_{g=1}^{G} \exp\{\phi_{0g} + \phi_{1g} z_{1i} + \phi_{2g} z_{2i} + \phi_{3g}(z_{2i}^2 - 4/3)\}},$$

where $(\phi_{0g}, \phi_{1g}, \phi_{2g}, \phi_{3g})$ is $(0.1, 0.1, 0.1, 0.1)$ for $g = 1$, is $(0.2, 0.2, 0.2, 0.2)$ for $g = 2$ and is $(0, 0, 0, 0)$ for $g = 3$.

Example 2. The second set-up is as follows.

(1) Covariates: $\mathbf{z}_i = [z_{i1}, z_{i2}, z_{i3}]^T$, where z_{i1} is i.i.d. from $N(0, 1)$, $z_{2i} = z_{1i} + \eta_i$ with $\eta_i \sim N(0, 0.3)$, and z_{3i} is i.i.d. from χ_1^2. (2) Outcome models:

$$y_{i1} = 5 + 10z_{1i} - 10I_{z_{1i}<-1} + 10I_{z_{1i}>-1} + 10z_{1i}I_{z_{1i}\in[-1,1]} + 3(z_{3i} - 1) + s_i e_{i1}, \quad (3.2)$$

$$y_{i2} = 5 + 10z_{1i} + s_i e_{i2}, \quad (3.3)$$

$$y_{i3} = 5 - 10z_{1i} + 10I_{z_{1i}<-1} - 10I_{z_{1i}>-1} - 10I_{z_{1i}\in[-1,1]} - 3(z_{3i} - 1) + s_i e_{i3}, \quad (3.4)$$

where $I_{(\cdot)}$ is an indicator function, $s_i = z_{1i} + 5$ and $e_{ig} \sim N(0, 1)$. Under this setup, the marginal means are $E(y_{i1}) = E(y_{i2}) = E(y_{i3}) = 5$. (3) First phase sampling: Poisson sampling with probability-proportional-to-size (PPS), where the size variable is s_i. So $\pi_{1i} = (\sum_{i\in U} s_i)^{-1} n s_i$, and n is the expected sample size. The joint inclusion probability $\pi_{1ij} = \pi_{1i}\pi_{1j}$ due to independence of the Poisson sampling. (4) Second phase selection:

$$\pi_{2ig} = \frac{\Phi\{\phi_{0g} + \phi_{1g} z_{2i} + \phi_{2g}(z_{3i} - 1) + \phi_{3g} z_{2i}^2\}}{\sum_{g=1}^{G} \Phi\{\phi_{0g} + \phi_{1g} z_{2i} + \phi_{2g}(z_{3i} - 1) + \phi_{3g} z_{2i}^2\}},$$

where $\Phi(\cdot)$ is the CDF of $N(0, 1)$, and $(\phi_{0g}, \phi_{1g}, \phi_{2g}, \phi_{3g})$ is $(0.1, 0.1, -0.1, 0.1)$ for $g = 1$, is $(0.2, 0.2, -0.2, 0.2)$ for $g = 2$ and is $(0, 0, 0, 0)$ for $g = 3$. In this example, we assumed (z_{1i}, z_{3i}) were observed in A_1 and used for estimating π_{2ig}, while the true functional form of π_{2ig} depends on (z_{2i}, z_{3i}) where z_{2i} is z_{1i} contaminated with noise η_i. The second example is of greater complexity than the first example and includes an optimal first-phase design in terms of anticipated variance (see [6] Theorem 3.1.1).

For each example and each (N, n) size combination, we simulated 2000 Monte Carlo (MC) samples. Six estimators of marginal means and average treatment effects were calculated for each Monte Carlo sample:

1. **TPR1**: Our two-phase regression estimator $\widehat{\theta}_g$ in (2.6) when there is no covariate available on the population level.
2. **TPR2**: Our three-phase regression estimator $\widehat{\theta}_{g,p}$ in (2.17) when some covariates are available in the population. We assume z_{1i} is observed for every unit in the population in both examples.

3. **IPW**: The IPW estimator $\widehat{\theta}_g^{ipw}$ in (2.15) using both π_{1i} and $\hat{\pi}_{2ig}$.
4. **NA-IPW**: The naive IPW estimator $\widehat{\theta}_g^{na-ipw} = n^{-1} \sum_{i \in A_{2g}} \hat{\pi}_{2ig}^{-1} y_{ig}$.
5. **REG**: A regression estimator using the augmented data of y_{ig}, for all g as the response variable. For example 1, the explanatory variables are $[1, Trt2_i, Trt3_i, H_i, R_K(\mathbf{z}_i)]$, where H_i is the indicator for the stratum, and $Trt2_i$ (or $Trt3_i$) is the indicator for treatment 2 (or 3). The explanatory variables in example 2 are $[1, Trt2_i, Trt3_i, R_K(\mathbf{z}_i)]$. The choices of $R_K(\mathbf{z}_i)$ for both examples will be discussed next. The estimated coefficient of $Trt2_i$ is the estimated treatment effect of $\theta_2 - \theta_1$ and the estimated coefficient of $Trt3_i$ is the estimated treatment effect of $\theta_3 - \theta_1$. Note that the covariates related to the first phase sampling, H_i in example 1 and z_{1i} in example 2, are included in the regression analysis.
6. **MT**: A one-to-one matching estimator using an approach detailed in [1]. The matching was done based on the estimated propensity scores $\hat{\pi}_{2ig}$, and the first phase sampling design weights are also included.

The NA-IPW, REG and MT are three commonly used estimators by practitioners, among which NA-IPW and REG ignore the first phase sampling design. In example 1, we used a cubic spline base of $[z_{1i}, z_{2i}, z_{3i}]$ for $R_K(\mathbf{z}_i)$ and a cubic spline base of $x_i \equiv z_{1i}$ for $R_K(\mathbf{x}_i)$ in estimation. For each variable, 10 knots were identified with locations corresponding to 10 equally spaced quantiles of the corresponding observations. In example 2, a cubic spline base of z_{1i} with 18 knots and a cubic spline base of z_{3i} with 18 knots were used to construct $R_K(\mathbf{z}_i)$, and a cubic spline bases of $x_1 \equiv z_{1i}$ with 18 knots was used to construct $R_K(\mathbf{x}_i)$. The locations of the knots were chosen such that the first one third (or the last one third) of the knots are uniformly spread between 0 and 20^{th} (or 80^{th} and 100^{th}) quantiles of the data for the corresponding variables, and the remaining one third were equally spaced between 20^{th} and 80^{th} quantiles.

Tables 1 (a) and (b) present the MC biases, variances, and mean squared errors (MSE) of the estimated treatment effects using the six estimators for each (N, n) combination and for example respectively. The NA-IPW and REG estimators as expected are highly biased in both examples due to ignoring the relationship between the first-phase design and the treatment effects. The matching estimator MT using the first phase design weights does reduce biases, compared to the NA-IPW and REG, but the IPW performs better than the MT in terms of the MSE in most of the cases. Although the IPW is consistent and has the same asymptotic efficiency as our two-phase semiparametric regression estimator (TPR1), the MC biases and variances of the IPW are greater than those of TPR1 in both examples. The MC biases and variances of the IPW though decrease when the sample size increases. The variance reduction of TPR1 over the IPW estimator indicates that gains for finite samples can be made by combining propensity and outcome regression when both models are well approximated semiparametrically. Both of our proposed estimators (TPR1 and TPR2) have similar low MC biases and much smaller MC variances and MSE relative to other estimators considered. TPR2 is more efficient than TPR1 due to the use of additional information on the population level.

Table 1
The MC biases, variances and MSEs of the estimated treatment effects, for (1): $(N, n) = (12500, 250)$*; (2):* $(N, n) = (25000, 500)$*; (3):* $(N, n) = (50000, 1000)$

(a) Example 1

		$\theta_1 - \theta_2$			$\theta_1 - \theta_3$			$\theta_2 - \theta_3$		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
(1)	TPR1	−0.008	0.309	0.309	−0.090	1.543	1.551	−0.082	1.495	1.502
	TPR2	−0.008	0.353	0.353	−0.093	0.604	0.613	−0.085	0.562	0.569
	IPW	0.395	4.900	5.056	0.681	5.321	5.785	0.286	4.766	4.848
	NA-IPW	2.138	2.312	6.883	12.259	3.104	153.395	10.122	2.554	104.999
	REG	1.579	1.424	3.917	10.930	2.814	122.276	9.351	2.905	90.342
	MT	−0.731	6.032	6.571	−0.394	4.051	4.209	0.340	3.528	3.652
(2)	TPR1	0.002	0.129	0.129	−0.036	0.673	0.674	−0.038	0.690	0.692
	TPR2	0.002	0.133	0.133	−0.023	0.241	0.241	−0.025	0.235	0.236
	IPW	0.109	2.298	2.310	0.169	2.136	2.165	0.060	1.882	1.886
	NA-IPW	2.015	0.811	4.871	12.033	1.056	145.847	10.018	0.927	101.285
	REG	1.616	0.623	3.236	10.954	1.238	121.229	9.338	1.250	88.439
	MT	−0.891	3.072	3.858	-0.543	2.041	2.342	0.348	1.773	1.891
(3)	TPR1	−0.001	0.064	0.064	−0.006	0.337	0.337	−0.004	0.332	0.332
	TPR2	0.000	0.064	0.064	−0.012	0.117	0.117	−0.012	0.115	0.115
	IPW	0.040	1.080	1.082	0.070	0.911	0.916	0.030	0.874	0.875
	NA-IPW	1.999	0.350	4.348	12.005	0.445	144.572	10.006	0.392	100.51
	REG	1.613	0.292	2.895	10.903	0.590	119.465	9.290	0.608	86.904
	MT	−0.907	1.501	2.321	−0.581	0.972	1.311	0.330	0.861	0.971

(b) Example 2

		$\theta_1 - \theta_2$			$\theta_1 - \theta_3$			$\theta_2 - \theta_3$		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
(1)	TPR1	0.081	1.349	1.356	0.193	6.629	6.667	0.112	4.539	4.551
	TPR2	0.083	1.349	1.356	0.182	1.570	1.603	0.099	1.438	1.447
	IPW	0.138	1.504	1.523	0.814	7.425	8.087	0.676	4.855	5.312
	NA-IPW	1.386	1.153	3.073	6.861	6.150	53.227	5.476	4.027	34.009
	REG	1.086	2.742	3.922	5.546	8.208	38.971	4.460	6.777	26.668
	MT	−0.310	5.331	5.423	−0.534	8.031	8.324	−0.222	5.342	5.389
(2)	TPR1	0.008	0.273	0.273	0.087	2.817	2.824	0.079	1.815	1.821
	TPR2	−0.011	0.185	0.185	-0.013	0.326	0.326	−0.002	0.232	0.232
	IPW	0.024	0.492	0.493	0.489	3.081	3.32	0.465	2.059	2.276
	NA-IPW	1.344	0.378	2.183	6.838	2.583	49.337	5.494	1.717	31.902
	REG	1.068	1.319	2.460	5.429	4.046	33.520	4.361	3.226	22.242
	MT	−0.35	2.668	2.789	−0.628	3.878	4.281	−0.276	2.800	2.881
(3)	TPR1	0.000	0.112	0.112	−0.030	1.319	1.320	−0.030	0.834	0.835
	TPR2	−0.001	0.061	0.061	−0.023	0.120	0.121	−0.022	0.084	0.084
	IPW	0.039	0.165	0.167	0.206	1.429	1.471	0.166	0.918	0.946
	NA-IPW	1.373	0.139	2.025	6.687	1.236	45.949	5.313	0.782	29.015
	REG	1.149	0.621	1.941	5.410	1.954	31.223	4.261	1.603	19.761
	MT	−0.317	1.321	1.423	−0.581	2.021	2.349	−0.255	1.321	1.389

In example 1, the order of the true marginal treatment means is $Trt1 < Trt2 < Trt3$ and our proposed two estimators, TPR1 and TPR2, and the IPW estimators estimated the treatment effect order correctly. However, if the first phase sampling is ignored, the estimates from the NA-IPW and REG reverse the order of the estimated treatment means completely. In example 2 where all treatments are marginally equivalent, the NA-IPW and REG estimate a decreasing order of treatment efficacy. These simulation results show that ignoring the first phase design can result in a serious bias in the ATE estimation.

TABLE 2

The coverage probabilities of the 95% C.I. for estimated treatment effects, for (1): $(N,n)=(12500,250)$; (2): $(N,n)=(25000,500)$; (3): $(N,n)=(50000,1000)$

(a) Example 1

		$\theta_1-\theta_2$	$\theta_1-\theta_3$	$\theta_2-\theta_3$
(1)	TPR1	0.886	0.920	0.922
	TPR2	0.876	0.879	0.893
	IPW	0.400	0.686	0.699
	NA-IPW	0.461	0.000	0.000
	REG	0.904	0.000	0.000
	MT	0.490	0.540	0.000
(2)	TPR1	0.922	0.941	0.936
	TPR2	0.916	0.917	0.924
	IPW	0.357	0.724	0.746
	NA-IPW	0.303	0.000	0.000
	REG	0.725	0.000	0.000
	MT	0.789	0.000	0.000
(3)	TPR1	0.932	0.951	0.952
	TPR2	0.935	0.932	0.932
	IPW	0.351	0.769	0.764
	NA-IPW	0.068	0.000	0.000
	REG	0.398	0.000	0.000
	MT	0.000	0.000	0.000

(b) Example 2

		$\theta_1-\theta_2$	$\theta_1-\theta_3$	$\theta_2-\theta_3$
(1)	TPR1	0.918	0.923	0.925
	TPR2	0.884	0.899	0.898
	IPW	0.786	0.882	0.878
	NA-IPW	0.420	0.139	0.136
	REG	0.980	0.394	0.538
	MT	0.510	0.340	0.000
(2)	TPR1	0.970	0.950	0.950
	TPR2	0.970	0.960	0.966
	IPW	0.891	0.920	0.914
	NA-IPW	0.207	0.010	0.008
	REG	0.960	0.138	0.254
	MT	0.461	0.000	0.000
(3)	TPR1	0.983	0.958	0.960
	TPR2	0.989	0.974	0.984
	IPW	0.940	0.940	0.942
	NA-IPW	0.023	0.000	0.000
	REG	0.898	0.010	0.046
	MT	0.000	0.000	0.000

Tables 2 (a) and (b) report the coverage probabilities of the 95% confidence interval (C.I.) for the average treatment effects. For each MC sample and each (N,n), we computed the point estimator $\widehat{\boldsymbol{\theta}}$ and the variance estimator of $\widehat{\boldsymbol{\theta}}$, and constructed the 95% C.I. for the pair differences. Variance estimation for the DE is similar to (2.30) with $\widehat{V}_{2g}$ replaced by $N^{-2}\sum_{i\in A_{2g}}\sum_{j\in A_{2g}}(\pi_{1ij}\hat{\pi}_{2ij})^{-1}\Delta_{1ij}(\pi_{1i}^{-1}y_{ig})(\pi_{1j}^{-1}y_{jg})$. Variance estimation for the NA-IPW was done by noting the NA-IPW estimator as a special case of the IPW estimator with assumed simple random sampling in the first phase. The estimated variance of the REG and the MT are provided by the regression and matching packages used in R. Note that the variance estimators for the estimators ignoring the first phase probabilities are not appropriate and can be biased under the full design. In both examples, estimators NA-IPW and REG have very poor coverage probabilities due to the large biases. Estimators IPW and MT that do not ignore the first phase sampling underestimate coverage probabilities in both examples. Our two estimators, TPR1 and TPR2, give satisfactory coverage probabilities in both examples even for a small sample size, relative to the nominal size 0.95.

4. Empirical study

In this section, we evaluate empirical performance of our proposed estimators in estimating treatment effects using a subsample from the 2005-2006 NHANES Survey. The goal of this empirical analysis is to assess the effect of nutrition label use (treatments) on body mass index (BMI) as in [5]. The NHANES is a study designed to assess the health and nutritional status of adults and children in the United States and is unique in that it combines interviews and physical examinations. A detailed description of the survey can be found at http://www.cdc.gov/nchs/nhanes.htm. The nutrition label use variable has three levels: level 1 = often, level 2 = sometimes, and level 3 = seldom. The study variable y is the BMI calculated from body weight and height. Covariates included were selected from [5]. In general, the covariates were classified into five categories: demographic, risky behavior, lifestyle, knowledge and health situation. There are totally 36 covariates and most of them are dummy variables, see detailed description in [5]. The analysis dataset contains $n = 1775$ subjects from the NHANES survey data.

NHANES uses a complex multistage probability sampling design, and the weights, i.e. π_{1i}^{-1}, are created to account for the complex survey design, survey non-response, and post-stratification. The same set of estimators (except for TPR2) evaluated in the simulation section were computed using this dataset. Since most of the covariate are dummy variables, the base used for estimating $\hat{\pi}_{2ig}$ and the outcome regression model is a vector simply containing all individual covariates. In the REG regression estimator, the explanatory variables are an intercept, treatment indicators and all the covariates. In addition, variance estimation is carried out for all estimators. Due to confidentiality issues, Mashed Variance Units (MVUs) were created and attached to the NHANES data files. The NHANES website provides an R code instruction to produce variance estimates using the MVUs. This R code was embedded into our main codes to calculate the components that are related to the first phase variance estimator in equations (2.25)–(2.28).

The estimated treatment effects, the standard errors and the 95% C.I.s are reported in Table 3. For the two estimators that incorporate the first phase design, TPR1 suggests that the estimated treatment mean of the BMI monotonically increases when the nutritional label use changes from “often” to “seldom”, while the IPW and MT estimators give an increasing trend from “often” to

Table 3

Results from the empirical study of nutrition label uses, including estimated treatment effects for three nutrition label use levels, their standard errors and their 95% C.I.s.

		Often-Sometimes	Often-Seldom	Sometimes-Seldom
TPR1	Estimate	−0.83	−1.13	−0.30
	SE	1.47	1.83	1.82
	95% C.I.	[−3.71, 2.05]	[−4.71, 2.45]	[−3.86, 3.26]
IPW	Estimate	−0.60	1.13	1.74
	SE	2.75	2.47	2.97
	95% C.I.	[−6.00, 4.80]	[−3.70, 5.97]	[−4.09, 7.56]
NA-IPW	Estimate	0.22	1.72	1.50
	SE	0.69	0.67	0.68
	95% C.I.	[−1.13, 1.57]	[0.40, 3.04]	[0.18, 2.83]
REG	Estimate	0.06	0.94	0.88
	SE	0.35	0.36	0.37
	95% C.I.	[−0.63, 0.76]	[0.24, 1.64]	[0.16, 1.60]
MT	Estimate	−0.63	0.61	1.23
	SE	0.26	0.23	0.25
	95% C.I.	[−1.31, −0.13]	[0.17, 1.05]	[0.75, 1.72]

"sometimes", but decreasing trend from "sometimes" to "seldom". A monotonic decreasing trend is present in NA-IPW and REG, leading to the strange conclusion that increasing nutritional label awareness increases BMI. Researchers generating hypotheses using results from NA-IPW and REG method could be led astray by not completely controlling for the full treatment group inclusion probabilities.

5. Conclusion and remarks

Much of the focus of observational study analysis has been on incorporating treatment selection into estimators to reduce bias due to self selection. Ignoring the first-phase sample design can have large implications for the interpretation of data. Accounting for the first-phase sample design reduces the bias and makes the target of estimation explicit. By incorporating auxiliary variables, the proposed two-phase semiparametric regression estimators are an improvement over the IPW estimators in finite sample problems. The assumptions for the two-phase regression estimators are reasonable for a large number of problems and we demonstrate that valid inference can be made with semiparametric model specificiations. However, these estimators only account for bias that can be explained by observed covariates. If the second-phase inclusion probabilities depend on unobserved variables, residual bias will exist. Further, the IPW and two-phase semiparametric regression estimators rely on a known first-phase design. In some cases, the first-order inclusion probabilities may need to be estimated and a design such as Poisson sampling is assumed. In summary, consideration of handling sample selection phases prior to treatment selection and auxiliary variables can lead to stronger and clearer evidence from observational studies. Estimating treatment effect parameters defined through a general estimation equation in observational studies is a topic for future research.

Appendix

We first discuss some technical assumptions. 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 assume

Condition A. (1) For all g, δ_{2ig} is independent of $\mathbf{y}_i$, given the variable $\mathbf{z}_i$; (2) $\mathbf{z}_i$ is distributed with density bounded away from zero on its compact support $\mathcal{Z}$; (3) For all g, $V(y_{ig}|\mathbf{z}_i)$ is uniformly bounded for all $\mathbf{z}_i \in \mathcal{Z}$; (4) For all g, π_{2ig} is bounded away from zero and one. And there exist positive constant C_1 and C_2 such that $C_1 < n^{-1}N\pi_{1i} < C_2$.

Condition B. (1) The smallest eigenvalue of $E[R_K(\mathbf{z})R_K(\mathbf{z})']$ is bounded away from zero uniformly in K; (2) There exists a sequence of constants $\xi(K)$ such that $\|R_K(\mathbf{z})\| \leq \xi(K)$ for any K; (3) For all g, $\pi_{2ig}(\mathbf{z})$ and $\mu_g(\mathbf{z}) = E[y_{ig}|\mathbf{z}]$ are s-time differentiable with $sd_z^{-1} > 2\eta + 1$, where d_z is the dimension of $\mathbf{z}$, and $\eta = log(\xi(K))[log(K)]^{-1}$; (4) $K = n^{\nu}$ with $4sd_z^{-1} - 4\eta - 2 > \nu^{-1} > 4\eta + 2$.

Condition C. (1) the limiting design covariance matrix: $nV(\bar{\mathbf{u}}_{1\pi}) \to \mathbf{\Sigma}_1$ a.s. and $nV(\bar{\mathbf{u}}_{2\pi,g}|A_1) \to \mathbf{\Sigma}_{2g}$ a.s., where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_{2g}$ are positive definite; (2) the normalized HT estimators satisfy central limit theorems: $\sqrt{n}(\bar{\mathbf{u}}_{1\pi} - \bar{\mathbf{u}}_N)|\mathcal{F}_N \to N(0, \mathbf{\Sigma}_1)$ a.s. and $\sqrt{n}(\bar{\mathbf{u}}_{2\pi,g} - \bar{\mathbf{u}}_N)|A_1, \mathcal{F}_N \to N(0, \mathbf{\Sigma}_{2g})$ a.s.; (3) consistency of variance estimators: $n(\hat{V}(\bar{\mathbf{u}}_{1\pi}) - V(\bar{\mathbf{u}}_{1\pi})) = o_p(1)$ and $n(\hat{V}(\bar{\mathbf{u}}_{2\pi}) - V(\bar{\mathbf{u}}_{2\pi})) = o_p(1)$. (4) We also assume for all g, $n(\hat{V}(\bar{\mathbf{u}}_{1\pi}) - \widetilde{V}(\bar{\mathbf{u}}_{1\pi})) = o_p(1)$, where $\widetilde{V}(\bar{\mathbf{u}}_{1\pi})) = N^{-2}\sum_{i \in A_{2g}}\sum_{i \in A_{2g}} \pi_{1ij}^{-1}\pi_{2ij,g}^{-1}\Delta_{1ij}\pi_{1i}^{-1}\mathbf{u}_i\pi_{1j}^{-1}\mathbf{u}_j^T$, and $n(\hat{V}(\bar{\mathbf{u}}_{2\pi,g}|A_1) - E[\hat{V}(\bar{\mathbf{u}}_{2\pi,g}|A_1)]) = o_p(1)$; (5) Assume $\widetilde{\boldsymbol{\beta}}_{ug} - B_{N,ug} = o_p(1)$, where $\widetilde{\boldsymbol{\beta}}_{ug} = (\sum_{i \in A_{2g}} \pi_{1i}^{-1}\pi_{2ig}^{-1}\mathbf{u}_i\mathbf{u}_i^T)^{-1}(\sum_{i \in A_{2g}} \pi_{1i}^{-1}\pi_{2ig}^{-1}\mathbf{u}_i y_{ig})$ and $B_{N,ug} = (\sum_{i \in U} \mathbf{u}_i\mathbf{u}_i^T)^{-1}(\sum_{i \in U} \mathbf{u}_i y_{ig})$.

The super-population parameter of interest is not identifiable from the data on $\{\sum_{g=1}^{G} y_{ig}\delta_{2ig}, \mathbf{z}_i\}_{i=1}^{n}$. Following the literature, we consider missing at random assumption in (A.1) to achieve identification. Condition B is general. But particularly, if $R_K(\mathbf{z}_i)$ is the power series or the spline series, (B.1) and (B.2) are satisfied automatically with $\eta = 1$ for the power series and $\eta = 0.5$ for the spline series. Condition C gives the design properties of the Horvitz and Thompson [13] estimators on both phases in the traditional finite population asymptotic framework. For any variable $\mathbf{u}$ with finite 4^{th} moment, define $\bar{\mathbf{u}}_{1\pi} = N^{-1}\sum_{i \in A_1} \pi_{1i}^{-1}\mathbf{u}_i$, and $\bar{\mathbf{u}}_{2\pi,g} = N^{-1}\sum_{i \in A_{2g}} (\pi_{1i}\pi_{2ig})^{-1}\mathbf{u}_i$, and their variance and variance estimators as $V(\bar{\mathbf{u}}_{1\pi}) = N^{-2}\sum_{i \in U}\sum_{j \in U} \Delta_{1ij}\pi_{1i}^{-1}\mathbf{u}_i\pi_{1j}^{-1}\mathbf{u}_j^T$, $\widehat{V}(\bar{\mathbf{u}}_{1\pi}) = N^{-2}\sum_{i \in A_1}\sum_{j \in A_1} \pi_{1ij}^{-1}\Delta_{1ij}\pi_{1i}^{-1}\mathbf{u}_i\pi_{1j}^{-1}\mathbf{u}_j^T$, $V(\bar{\mathbf{u}}_{2\pi,g}|A_1) = N^{-2}\sum_{i \in A_1}(\pi_{2ig}^{-1} - 1)\mathbf{u}_i\mathbf{u}_i^T$, $\widehat{V}(\bar{\mathbf{u}}_{2\pi,g}|A_1) = N^{-2}\sum_{i \in A_{2g}} \pi_{2ig}^{-1}(\pi_{2ig}^{-1} - 1)\mathbf{u}_i\mathbf{u}_i^T$. Condition C are satisfied for many commonly designs in reasonably behaved finite populations. Note that (C.3) would not hold for systematic sampling or one-per-stratum designs.

A: Proof of Theorem 1 and Corollary 1

Proof of Theorem 1. Write $\widehat{\theta}_g = N^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}R_K(\mathbf{z}_i)^T\widehat{\boldsymbol{\beta}}_{zg} = \bar{R}_{z,N}^T\boldsymbol{\beta}_{zg} + \bar{R}_{z,N}^T(\widehat{\boldsymbol{\beta}}_{zg}-\boldsymbol{\beta}_{zg})+(\bar{R}_{z,1\pi}-\bar{R}_{z,N})^T\boldsymbol{\beta}_{zg}+o_p(n^{-\frac{1}{2}})$, where $\bar{R}_{z,N} = N^{-1}\sum_{i\in U}R_K(\mathbf{z}_i)$. The first equality is true due to the inclusion of the intercept, and the second equality is from Taylor expansion and condition (C.5). Note that

$$\begin{aligned}\bar{R}_{z,N}^T(\widehat{\boldsymbol{\beta}}_{zg}-\boldsymbol{\beta}_{zg}) &= \bar{R}_{z,N}^T\left(\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(\mathbf{z}_i)R_K(\mathbf{z}_i)^T\right)^{-1}\\ &\quad\times\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}R_K(\mathbf{z}_i)e_{ig}\\ &= \left(\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}\right)^{-1}\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}e_{ig}. \qquad (A.1)\end{aligned}$$

The last equality is obtained using the Gram−Schmidt transformation. Thus,

$$\widehat{\theta}_g - \bar{y}_{Ng} = (\bar{R}_{z,1\pi}-\bar{R}_{z,N})^T\boldsymbol{\beta}_{zg} + (\widetilde{e}_{2\pi,g}-\bar{e}_{Ng}) + o_p(n^{-\frac{1}{2}}), \qquad (A.2)$$

where $\widetilde{e}_{2\pi,g} = (\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1})^{-1}\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}e_{ig}$. The key part of the proof is to show that

$$\widehat{\theta}_g - \bar{y}_{Ng} = (\bar{R}_{z,1\pi}-\bar{R}_{z,N})^T\boldsymbol{\beta}_{zg} + (\bar{e}_{1\pi,g}-\bar{e}_{Ng}) + (\bar{e}_{2\pi,g}-\bar{\epsilon}_{1\pi,g}) + o_p(n^{-\frac{1}{2}}). \quad (A.3)$$

Suppose (A.3) is true, by condition (C.1), the consistency result in Theorem 1 - (i) holds. Also under condition C, conditioning on the given finite population $\mathcal{F}_N$,

$$V_{2g}^{-\frac{1}{2}}\left((\bar{R}_{z,1\pi}-\bar{R}_{z,N})^T\boldsymbol{\beta}_{zg} + (\bar{e}_{1\pi,g}-\bar{e}_{Ng})\right)|\mathcal{F}_N \xrightarrow{d} N(0,1), a.s. \qquad (A.4)$$

where V_{2g} is defined in (2.8), and conditioning on the first phase sample A_1,

$$V_{1g}^{-\frac{1}{2}}\left(\bar{e}_{2\pi,g}-\bar{e}_{1\pi,g}\right)|A_1,\mathcal{F}_N \xrightarrow{d} N(0,1), a.s. \qquad (A.5)$$

where $V_{1g} = E\left\{V[\bar{\epsilon}_{2\pi,g}|A_1]\right\}$ is defined in (2.7). Then, using Theorem 1.3.6 of Fuller (2009), results (A.5) and (A.4) can be combined to obtain the central limit result in Theorem 1 - (ii). Next we show (A.3) holds. Define $\check{e}_{ig} = \pi_{1i}^{-1}e_{ig}$, $\mu_{eg}(\mathbf{z}_i) = E[e_{ig}|\mathbf{z}_i]$, and $\mu_{\check{e}g}(\mathbf{z}_i) = \pi_{1i}^{-1}\mu_{eg}(\mathbf{z}_i)$. In order to show (A.3), we first decompose $\sum_{i\in A_{2g}}\pi_{1i}^{-1}\hat{\pi}_{2ig}^{-1}e_{ig}$ into a sum of several terms by adding and subtracting,

$$\begin{aligned}&n^{-\frac{1}{2}}\sum_{i\in A_{2g}}\frac{e_{ig}}{\pi_{1i}\hat{\pi}_{2ig}}\\ &\quad= n^{-\frac{1}{2}}\sum_{i\in A_1}\left\{\frac{\delta_{2ig}\check{e}_{ig}}{\hat{\pi}_{2ig}} - \frac{\delta_{2ig}\check{e}_{ig}}{\pi_{2ig}} + \frac{\delta_{2ig}\check{e}_{ig}}{\pi_{2ig}^2}(\hat{\pi}_{2ig}-\pi_{2ig})\right\}\end{aligned}$$

$$
\begin{aligned}
&+ n^{-\frac{1}{2}} \sum_{i \in A_1} \left\{ -\frac{\delta_{2ig}\check{e}_{ig}}{\pi_{2ig}^2}(\hat{\pi}_{2ig} - \pi_{2ig}) + \frac{\mu_{\check{e}g}(\mathbf{z}_i)}{\pi_{2ig}}(\hat{\pi}_{2ig} - \pi_{2ig}) \right\} \\
&+ n^{-\frac{1}{2}} \sum_{i \in A_1} \left\{ -\frac{\mu_{\check{e}g}(\mathbf{z}_i)}{\pi_{2ig}}(\hat{\pi}_{2ig} - \pi_{2ig}) + \frac{\mu_{\check{e}g}(\mathbf{z}_i)}{\pi_{2ig}}(\delta_{2ig} - \pi_{2ig}) \right\} \\
&+ n^{-\frac{1}{2}} \sum_{i \in A_1} \left\{ \frac{\delta_{2ig}\check{e}_{ig}}{\pi_{2ig}} - \frac{\mu_{\check{e}g}(\mathbf{z}_i)}{\pi_{2ig}}(\delta_{2ig} - \pi_{2ig}) \right\}. \qquad \text{(A.6)}
\end{aligned}
$$

By [4] Theorem B-1,

$$
\|\hat{\pi}_{2ig} - \pi_{2ig}\| = O_p(\xi(K)K^{1/2}n^{-1/2} + \xi(K)K^{1/2}K^{-s/d_z}),
$$

so the first three terms in (A.6) can be shown to have order $o_p(1)$ asymptotically, which leads to

$$
\begin{aligned}
\widetilde{e}_{2\pi g} &= \frac{1}{N}\sum_{i \in A_1} \left\{ \frac{\delta_{2ig}\check{e}_{ig}}{\pi_{2ig}} - \frac{\mu_{\check{e}g}(\mathbf{z}_i)}{\pi_{2ig}}(\delta_{2ig} - \pi_{2ig}) \right\} + o_p(n^{-\frac{1}{2}}) \\
&= \frac{1}{N}\sum_{i \in A_1} \frac{\check{e}_{ig} - \mu_{\check{e}g}(\mathbf{z}_i)}{\pi_{2ig}} + \frac{1}{N}\sum_{i \in A_1}\sum_{i \in A_1} \mu_{\check{e}g}(\mathbf{z}_i) + o_p(n^{-\frac{1}{2}}) \\
&= \bar{\epsilon}_{2\pi,g} + \frac{1}{N}\sum_{i \in A_1}\sum_{i \in A_1} \mu_{\check{e}g}(\mathbf{z}_i) + o_p(n^{-\frac{1}{2}}). \qquad \text{(A.7)}
\end{aligned}
$$

The justification of those orders follows [4], and we refer readers to [4] for details. Therefore, by plugging (A.7) into (A.2) we have (A.3). It follows that

$$
\widehat{\theta}_g - \bar{y}_{Ng} = (\bar{\epsilon}_{2\pi,g} - \bar{\epsilon}_{1\pi,g}) + (\bar{y}_{1\pi} - \bar{y}_{Ng}) + o_p(n^{-\frac{1}{2}}). \qquad \square
$$

Proof of Corollary 1. We can decompose $\widehat{\theta}_g - \theta_g^* = \widehat{\theta}_g - \bar{y}_{Ng} + \bar{y}_{Ng} - \theta_g^*$. Then the asymptotic results are immediate by using Theorem 1.3.6 of [6] again. □

B: Proof of Theorem 2

First note that, for all g, the following results hold under condition B. $\|\hat{\pi}_{2ig} - \pi_{2ig}\| = O_p(\xi(K)K^{1/2}n^{-1/2} + \xi(K)K^{1/2}K^{-\alpha}) = o_p(1)$ (see [4]); Similarly, since $\|\hat{\mu}_{zg}(\mathbf{z}_i) - \mu_{zg}(\mathbf{z}_i)\| = o_p(1)$, then $\hat{\epsilon}_{ig} - \epsilon_{ig} = o_p(1)$; $\hat{\pi}_{2ij,g} - \pi_{2ij,g} = o_p(1)$, $\hat{\pi}_{2ij,g}^{-1} - \pi_{2ij,g}^{-1} = -\pi_{2ij,g}^{-2}o_p(1)$, $\widehat{\boldsymbol{\beta}}_{zg} - \boldsymbol{\beta}_{zg} = o_p(1)$, and $\widehat{\boldsymbol{\beta}}_{xg} - \boldsymbol{\beta}_{xg} = o_p(1)$. The term $\widehat{V}_{1g}$ in (2.22) can be written as

$$
\begin{aligned}
\widehat{V}_{1g} &= \frac{1}{N^2}\sum_{A_{2g}} \frac{1}{\pi_{2ig}}\left(\frac{1}{\pi_{2ig}} - 1\right)\epsilon_{ig}^2\pi_{1i}^{-2} + o_p(n^{-1}), \text{ by (C.4)} \\
&= V(\bar{\epsilon}_{2\pi,g}|A_1) + o_p(n^{-1}), \text{ by (C.4)} \\
&= E\left\{V(\bar{\epsilon}_{2\pi,g}|A_1)\right\} + o_p(n^{-1}) = V_{1g} + o_p(n^{-1}). \qquad \text{(B.1)}
\end{aligned}
$$

The term $\widehat{M}_{1g}$ in (2.25) is

$$
\begin{aligned}
\widehat{M}_{1g} = & \frac{1}{N^2} \sum_{i \in A_{2g}} \sum_{i \in A_{2g}} \frac{\Delta_{1ij}}{\pi_{1ij}} \left(\frac{1}{\pi_{2ij}} - \frac{1}{\pi_{2ij}^2} o_p(1) \right) \frac{e_{ig} + R_K(\mathbf{z}_i)^T o_p(1)}{\pi_{1i}} \frac{e_{jg} + R_K(\mathbf{z_j})^T o_p(1)}{\pi_{1i}} \\
& + o_p(n^{-1}), \text{ by (C.4)} \\
= & \widehat{V}(\bar{e}_{1\pi,g}) + o_p(n^{-1}) = M_{1g} + o_p(n^{-1}) \text{ by (C.3)} \qquad \text{(B.2)}
\end{aligned}
$$

The term $\widehat{M}_{2g}$ in (2.26) can be written as

$$
\begin{aligned}
\widehat{M}_{2g} &= (\boldsymbol{\beta}_{zg}^T + o_p(1))(V(\bar{R}_{z,1\pi}) + o_p(n^{-1}))(\boldsymbol{\beta}_{zg} + o_p(1)) \\
&= \boldsymbol{\beta}_{zg}^T V(\bar{R}_{z,1\pi}) \boldsymbol{\beta}_{zg} + o_p(n^{-1}) = M_{2g} + o_p(n^{-1}). \qquad \text{(B.3)}
\end{aligned}
$$

The same argument for $\widehat{M}_{1g}$ can be used to show that

$$
\widehat{M}_{3g} = M_{3g} + o_p(n^{-1}). \qquad \text{(B.4)}
$$

Following the same fashion, the four terms in $\hat{\sigma}_g^2$ of (2.29) can be shown to be consistent for terms $E[\mu_{zg}(\mathbf{z}_i)^2]$, $E^2[\mu_{zg}(\mathbf{z}_i)]$, $E[\epsilon_{ig}^2]$ and $E^2[\epsilon_{ig}]$ respectively. Thus, the $\hat{\sigma}_g^2$ in (2.29) is

$$
\begin{aligned}
\hat{\sigma}_g^2 &= E[\mu_{zg}(\mathbf{z}_i)^2] - E^2[\mu_{zg}(\mathbf{z}_i)] + E[\epsilon_{ig}^2] - E^2[\epsilon_{ig}] + o_p(1) \\
&= \sigma_g^2 + o_p(1).
\end{aligned} \qquad \text{(B.5)}
$$

Combining (B.1), (B.2), (B.3), (B.4) and (B.5), we have Theorem 2 - (i). Part (ii) in Theorem 2 can be shown using Slutsky theory.

C: Asymptotic properties of $\widehat{\boldsymbol{\theta}}_{g,p}$ and its variance estimator

Decompose $\boldsymbol{\beta}_{zg}^T$ into two parts $[\boldsymbol{\beta}_{zg,x}^T, \boldsymbol{\beta}_{zg,-x}^T]$, where $\boldsymbol{\beta}_{zg,x}$ contains the coefficients corresponding to the base $R_K(\mathbf{x}_i)$, and $\boldsymbol{\beta}_{zg,-x}$ has the remaining coefficients for bases that are not in $R_K(\mathbf{x}_i)$. Similarly, $\widehat{\boldsymbol{\beta}}_{zg}^T = [\widehat{\boldsymbol{\beta}}_{zg,x}^T, \widehat{\boldsymbol{\beta}}_{zg,-x}^T]$. Define $\boldsymbol{\alpha}_g^T = [(\boldsymbol{\beta}_{zg,x} - \boldsymbol{\beta}_{xg})^T, \boldsymbol{\beta}_{zg,-x}^T]$, where $\boldsymbol{\beta}_{xg} = \lim_{N \to \infty} (\sum_{i \in U} R_K(\mathbf{x}_i) R_K(\mathbf{x}_i)^T)^{-1} \sum_{i \in U} R_K(\mathbf{x}_i) y_{ig}$. The same asymptotic results in Theorem 1 still hold for $\widehat{\theta}_{g,p}$, after simply replacing $\boldsymbol{\beta}_{zg}$ by $\boldsymbol{\alpha}_g$ and using the similar arguments in Appendix A. The results can be easily obtained by the following expansion

$$
\begin{aligned}
\widehat{\theta}_{g,p} - \bar{y}_{Ng} &= \bar{R}_{z,N}^T \boldsymbol{\beta}_{zg} + \bar{R}_{z,N}^T (\widehat{\boldsymbol{\beta}}_{zg} - \boldsymbol{\beta}_{zg}) - (\bar{R}_{x,1\pi} - \bar{R}_{x,N})^T \boldsymbol{\beta}_{xg} \\
&\quad + (\bar{R}_{z,1\pi} - \bar{R}_{z,N})^T \boldsymbol{\beta}_{zg} + o_p(n^{-\frac{1}{2}}) \\
&= (\bar{R}_{z,1\pi} - \bar{R}_{z,N})' \boldsymbol{\alpha}_g + (\widetilde{e}_{2\pi,g} - \bar{e}_{Ng}) + o_p(n^{-\frac{1}{2}}) \text{ by (A.7)} \\
&= (\bar{R}_{z,1\pi} - \bar{R}_{z,N})' \boldsymbol{\alpha}_g + (\bar{\epsilon}_{2\pi,g} - \bar{\epsilon}_{1\pi,g}) + (\bar{e}_{1\pi,g} - \bar{e}_{Ng}) + o_p(n^{-\frac{1}{2}}),
\end{aligned}
$$

where $\bar{R}_{x,N} = N^{-1}\sum_{i\in U} R_K(\mathbf{x}_i)$ and $\bar{R}_{x,1\pi} = N^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}R_K(\mathbf{x}_i)$. It can also be shown,

$$\begin{aligned}
\widehat{\theta}_{g,p} - \bar{y}_{Ng} &= (\bar{y}_{2\pi,g} - \bar{y}_{Ng}) + (\bar{R}_{z,1\pi} - \bar{R}_{z,2\pi})'\hat{\boldsymbol{\beta}}_{zg} + (\bar{R}_{x,N} - \bar{R}_{x,1\pi})'\hat{\boldsymbol{\beta}}_{xg} \\
&= (\bar{\epsilon}_{2\pi,g} - \bar{\epsilon}_{1\pi,g}) + (\bar{y}_{1\pi,g} - \bar{y}_{Ng}) + (\bar{R}_{x,N} - \bar{R}_{x,1\pi})'\boldsymbol{\beta}_{xg} + o_p(n^{-\frac{1}{2}}) \\
&= (\bar{\epsilon}_{2\pi,g} - \bar{\epsilon}_{1\pi,g}) + (\bar{a}_{1\pi,g} - \bar{a}_{Ng}) + o_p(n^{-\frac{1}{2}}),
\end{aligned}$$

where $\bar{R}_{z,2\pi} = N^{-1}\sum_{i\in A_{2g}}\pi_{1i}^{-1}\pi_{2ig}^{-1}R_K(\mathbf{z}_i)$. The variance estimation of $\widehat{\theta}_{g,p}$ is the same as in (2.22) - (2.30), by replacing $\widehat{\boldsymbol{\beta}}_{zg}$ by the corresponding $\widehat{\boldsymbol{\alpha}}_g$. Same arguments in Appendix B can be used to show the consistency of this variance estimator.

Acknowledgments

The authors thank Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

References

[1] Abadie, A. and Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1), 235–267. MR2194325

[2] Bang, H. and Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61, 962–972. MR2216189

[3] Breidt, F. J., Claeskens, G. and Opsomer, J. D. (2005). Model-Assisted Estimation for Complex Surveys Using Penalised Splines. *Biometrika*, 92(4), 831–846. MR2234189

[4] Cattaneo, M. D. (2010). Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability. *Journal of Econometrics*, 155(2), 138–154. MR2607191

[5] Drichoutis, A. C., Nayga, R. M. and Lazaridis, P. (2009). Can Nutritional Label Use Influence Body Weight Outcomes? *Kyklos*, 62, 500–525.

[6] Fuller, W. A. (2009). *Sampling Statistics*, John Wiley & Sons.

[7] Giffin, R. B. and Woodcock, J. (2010). Comparative Effectiveness Research: Who Will Do The Studies. *Health Affairs*, 29(11), 2075–2081.

[8] Glynn, A. N. and Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, 18, 36–56.

[9] Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2), 315–331. MR1612242

[10] Hajek, J. (1971). Comment on An Essay on the Logical Foundations of Survey Sampling by Basu, D. in Godambe, V.P. and Sprott, D.A. eds. *Foundations of Statistical Inference*, Holt, Rinehart and Winston, page 236. MR0375600

[11] Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4), 1161–1189. MR1995826

[12] Hong, G. (2010). Marginal Mean Weighting Through Stratification: Adjustment for Selection Bias in Multilevel Data. *Journal of Educational and Behavioral Statistics*, 35(5), 499–531.

[13] Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47, 663–685. MR0053460

[14] Iglehart, J. K. (2009). Prioritizing Comparative-Effectiveness Research — IOM Recommendations. *The New England Journal of Medicine*, 361, 325–328.

[15] Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature, American Economic Association*, 47(1), 5–86.

[16] Isaki, C. T. and Fuller, W. A. (1982). Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77, 89–96. MR0648029

[17] Kang, J. D. Y. and Schafer J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4), 523–539. MR2420458

[18] Kim, J. K. and Haziza, D. (2010). Doubly Robust Inference with Missing Data in Survey Sampling. *Joint Statistical Meetings*, Vancouver, Canada.

[19] Korn, E. and Graubard, B. (1991). Epidemiologic Studies Utilizing Surveys: Accounting for the Sampling Design. *American Journal of Public Health*, 81(9), 1166–1173.

[20] Lorentz, G. (1986). *Approximation of Functions*, New York: Chelsea Publishing Company. MR0917270

[21] Pfeffermann, D. and Sverchkov, M. (1999). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya: The Indian Journal of Statistics, Series B*, 61(1), 166–186. MR1720710

[22] Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer. MR1140409

[23] Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software*, 42(7), 1–52.

[24] Sorenson, C. (2010). Use of Comparative Effectiveness Research in Drug Coverage and Pricing Decisions: A Six-Country Comparison. *The Commonwealth Fund*, 91, 1–14.

[25] Tan, Z. (2006). Regression and Weighting Methods for Causal Inference Using Instrumental Variables. *Journal of the American Statistical Association*, 101, 1607–1618. MR2279483

[26] Tan, Z. (2010). Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting. *Biometrika*, 94(2), 1–22. MR2672490

[27] Wang, W., Scharfstein, D., Tan, Z. and MacKenzie, E. J. (2009). Causal Inference in Outcome-dependent Two-phase Sampling Designs. *Journal of Royal Statistical Society Series B*, 71, 947–969. MR2750252