Research Article

Akanksha Negi*

Doubly weighted M-estimation for nonrandom assignment and missing outcomes

https://doi.org/10.1515/jci-2023-0016
received March 30, 2023; accepted October 21, 2023

Abstract: This article proposes a class of M-estimators that double weight for the joint problems of nonrandom treatment assignment and missing outcomes. Identification of the main parameter of interest is achieved under unconfoundedness and missing at random assumptions with respect to the treatment and sample selection problems, respectively. Given the parametric framework, the asymptotic theory of the proposed estimator is outlined in two parts: first, when the parameter solves an unconditional problem, and second, when it solves a stronger conditional problem. The two parts help to summarize the misspecification scenarios permissible under the given framework and the role played by double weighting in each. As illustrative examples, the article also discusses the estimation of causal parameters like average and quantile treatment effects. With respect to the average treatment effect, this article shows that the proposed estimator is *doubly robust*. Finally, a detailed application to Calónico and Smith's (The women of the national supported work demonstration. J Labor Econom. 2017;35(S1):S65–S97.) reconstructed sample from the National Supported Work training program is used to demonstrate the estimator's performance in empirical settings.

Keywords: unconfoundedness, missing at random, double weighting, M-estimation, treatment effects

MSC 2020: 62D20

1 Introduction

When interest lies in causal inference, the prevalence of missing data poses a major identification challenge. In observational studies, causal effects estimation is complicated due to a nonrandom selection of individuals into programs or interventions (nonrandom treatment assignment). If, in addition, the observed outcome of interest is missing due to attrition or nonresponse, then this creates a double identification challenge for the estimation of treatment effects. This article proposes a class of doubly weighted M-estimators that are consistent and asymptotically normal for the joint problems of nonrandom treatment assignment and missing outcomes.

Despite the ubiquity of missing data problems in observational and experimental studies, the traditional inverse probability weighted (IPW) literature has only considered treatment and sample selection problems in isolation. In this article, I extend the literature to incorporate both issues simultaneously in a general framework. The main parameter of interest is defined to solve a population objective function. To correct for the two selection issues at hand, the article proposes weighting by both the propensity score and the missing outcome probability to identify the true parameter. The two key assumptions are unconfoundedness and missing at random, which represent ignorable assignment and missing data mechanisms, respectively [1]. Estimation

* **Corresponding author: Akanksha Negi,** Department of Econometrics and Business Statistics, Monash University, Wellington Road, Clayton, Victoria 3800, Australia; Webpage: www.anegi.net, e-mail: akanksha.negi@monash.edu

follows in two steps: first the probabilities are estimated using binary response maximum likelihood, and second, the estimated probabilities are plugged in as weights to solve some objective function.

In the missing data IPW literature, Ding and Li [2] and Cao et al. [3] considered estimation of the population mean in the presence of missing outcomes. Robins and Rotnitzky [4] considered the estimation of regression parameters when the outcome is censored. Chen et al. [5] and Graham et al. [6] considered estimation of parameters indexing moment conditions, whereas Wooldridge [7] focused on the parameter solving an optimization problem. See the study by Seaman and White [8] for a review of IPW in the missing data context. While most of this literature draws parallels between missing data and missing potential outcomes, none have considered the problem of treatment selection. On the other hand, the causal inference literature uses propensity score weighting for identifying treatment effects but does not address any traditional missing data issues [9–13].

A few articles have looked at the double selection problem. For instance, Huber [14] presented a systematic treatment of the forms of attrition (i.e., whether selection is on observables or unobservables) that yield true average treatment effects (ATEs) in an experiment. His results under the "selection-on-observables" assumption are nested within this article as a special case. In another article, Huber [15] used an instrument for sample selection along with a conditionally exogeneous treatment. He then proposes a weighted estimator that nests the sample selection probability as an additional covariate in the propensity score. This is different from the weighting scheme employed in this article which neither involves any instruments nor nests probabilities. Rather, it exploits the sequential relationship between the two selection mechanisms; individuals may be more or less likely to attrit after selecting into a particular treatment group beyond what is dictated by their covariates. Other articles that take a selection-on-unobservables view of one or both problems include [16,17]. Typically, the discussion in these articles is limited to identification results for the ATE along with an incomplete characterization of the estimator in question. One exception is Huber [15] who also considered the unconditional quantile treatment effect (QTE) for the selected subpopulation.

This article makes the following contributions. First, it attempts to provide a comprehensive treatment of the two problems under a selection-on-observables framework where the parameter of interest minimizes an objective function. Consequently, the double-weighting solution applies to a wide range of procedures that can be framed as an M-estimation problem. The asymptotic theory of the proposed estimator is characterized in two parts. The first part describes cases where a conditional feature of interest is potentially misspecified, which is formalized in terms of a weak identification assumption. This part requires the weights to be correctly specified in order to achieve identification. In contrast, the second part assumes a conditional feature of interest to be correctly specified and allows the weights to be wrong. Together, they summarize two important cases of misspecification that can arise in this double-weighting setup and help us determine what can be consistently estimated under each setting. For instance, with respect to the estimation of QTEs, this article shows how one can estimate conditional QTE (CQTE) or a *linear approximation* to CQTE depending on whether the conditional quantile function is assumed to be correctly specified or not.

This article also shows that certain quasi-log-likelihood and mean function combinations deliver a "doubly robust" (DR) estimator of ATE under the given framework. This is different from the augmented-IPW style of estimators that are well known for being DR in the missing data literature ([18–20], see [21] for a review). DR estimators involve models for the conditional mean and propensity score and are consistent if at least one of two models is correctly specified. The extant literature on DR estimators have either focused on sample selection [2,6] or treatment selection [7,13], but not both at the same time [22]. A related contribution is to propose a DR estimator for ATE, which is distinct from the augmented-IPW class, when both problems are present. An advantage of this estimator is that it is less sensitive to extreme values of the weights and also ensures that the range of the estimated mean function aligns with the nature of the outcomes being studied [13]. Simulations show that the doubly weighted ATE and QTE estimates have the lowest finite sample bias compared to alternatives that ignore one or both problems. More recently, Bia et al. [23] have adapted the double machine learning framework of Chernozhukov et al. [24] to accommodate both treatment and sample selection problems in the presence of high-dimensional covariates.

This article also adds to the existing IPW literature on "efficiency puzzle" which finds that estimated nuisance parameters often provide a more efficient estimator for the main parameter of interest. This result appears while characterizing the variance of the proposed estimator under the first half of the asymptotic theory. The estimation of the weights helps to exploit the correlation between the first- and second-step moment conditions obtained from the binary response and M-estimation problems, respectively, which the known-weights estimator fails to do. However, this ceases to be true in the second half where there are no efficiency improvements from using estimated weights. This puzzle has been well studied in [7,25–27], and more recently in the studies by Lok [28] and Su et al. [29]. This article also discusses conditions when weighting may be inefficient compared to not weighting at all.

Finally, the proposed method is applied to estimate the average and distributional impacts of the National Supported Work (NSW) training program on earnings for the Aid to Families with Dependent Children (AFDC) target group. The sample was obtained from Calónico and Smith [30] who recreated Lalonde's within-study analysis for this women's sample. The presence of experimental and non-experimental comparison groups in the data help to evaluate whether the doubly weighted estimator brings us close to the experimental benchmark relative to other alternatives. I find that the empirical bias for the doubly weighted estimate is much smaller than that for the unweighted estimate.

The rest of this article proceeds as follows: Section 2 describes the basic potential outcomes framework and provides a short description of the population models with an introduction to the naive unweighted estimator; Section 3 discusses the treatment assignment and missing outcome mechanisms, which leads us directly to the identification lemma; Section 4 develops the first half of the asymptotic theory for the doubly weighted estimator with a focus on misspecification of a conditional feature of interest and correct weights; in contrast, Section 5 considers the second half where a conditional model of interest is correctly specified but the weights may be misspecified; Section 6 studies the estimation of average and QTEs within the proposed framework; Section 7 provides supporting Monte Carlo evidence under two cases of misspecification: the correct conditional model with misspecified weights and misspecified conditional model with correct weights; Section 8 applies the proposed method to the NSW job training program; and Section 9 concludes.

2 Potential outcomes and the population models

Let $Y(1)$ and $Y(0)$ denote potential outcomes for the treatment and control states, respectively, and let W be an indicator for the binary treatment. Then,

$$Y = Y(0)\cdot(1 - W) + Y(1)\cdot W. \tag{1}$$

Also, let $\mathbf{X}$ be a fixed vector of covariates, which includes an intercept. Some feature of the distribution of $(Y(g), \mathbf{X})$ is assumed to depend on a finite vector $\boldsymbol{\theta}_g$. Let $q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)$ be an objective function where some examples include the smooth least squares function, $q(\cdot) = (Y(g) - \mathbf{X}\boldsymbol{\theta}_g)^2$, or the non-smooth quantile regression, $q(\cdot) = c_\tau(Y(g) - \mathbf{X}\boldsymbol{\theta}_g)$, where $c_\tau(u) = (\tau - \mathbf{1}\{u < 0\})u$ is the asymmetric loss function for a random variable, u.

Assumption 1. (Identification of $\boldsymbol{\theta}_g^{\mathbf{0}}$) The parameter vector $\boldsymbol{\theta}_g^{\mathbf{0}} \in \boldsymbol{\Theta}_g$ is a unique solution to the population minimization problem, $\min_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g} \mathbb{E}[q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)]$, for each $g = 0, 1$.

Assumption 1 defines the parameter of interest as the one that *uniquely* solves the population optimization problem. If $q(\cdot)$ involves a misspecified conditional feature like a conditional mean, variance, or even the full conditional distribution, Assumption 1 guarantees a unique pseudo-true solution [31]. In that case, determining whether the pseudo-truth, $\boldsymbol{\theta}_g^{\mathbf{0}}$, is a meaningful parameter will depend on the conditional feature being studied and the estimation method used. For example, least squares can provide us with the best linear approximation to the true conditional mean, even if one mis-specifies the true function. Similarly, Angrist et al. [32] established the approximation properties of quantile regression where one can still estimate the best linear approximation to the true conditional quantile function under misspecification. In both cases, $\boldsymbol{\theta}_g^{\mathbf{0}}$ indexes linear projections (LPs) to different conditional models. On the other hand, when $q(\cdot)$ involves a correctly specified model, $\boldsymbol{\theta}_g^{\mathbf{0}}$ has a straightforward interpretation.

2.1 Nonrandom treatment assignment and missing outcomes

Let S be a binary indicator that denotes whether the outcome is observed or missing. Then,

$$Y = \begin{cases} Y(0)\cdot(1 - W) + Y(1)\cdot W, & \text{if } S = 1 \\ \text{missing}, & \text{if } S = 0. \end{cases} \tag{2}$$

Given that the main objective of this article is to consistently estimate $\boldsymbol{\theta}_g^{\mathbf{0}}$, a common empirical strategy is to only use complete cases for estimation. This means solving treatment and control group problems,

$$\min_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \sum_{i=1}^{N} S_i \cdot W_i \cdot q(Y_i, \mathbf{X}_i, \boldsymbol{\theta}_1) \quad \text{and} \quad \min_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0} \sum_{i=1}^{N} S_i \cdot (1 - W_i)\cdot q(Y_i, \mathbf{X}_i, \boldsymbol{\theta}_0), \tag{3}$$

where $q(Y, \mathbf{X}, \boldsymbol{\theta}_g) = W \cdot q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1) + (1 - W)\cdot q(Y(0), \mathbf{X}, \boldsymbol{\theta}_0)$. The estimator that solves (3) is called the unweighted estimator and is denoted by $\hat{\boldsymbol{\theta}}_g^{\boldsymbol{u}}$. This will be consistent if it identifies $\boldsymbol{\theta}_g^{\mathbf{0}}$ in the population. For example, consider

$$Y(g) = \mathbf{X}\boldsymbol{\theta}_g + U(g), \quad g = 0, 1, \quad \text{where } \mathbb{E}[\mathbf{X}'U(g)] = \mathbf{0}.$$

In this case, even if the treatment is randomly assigned, missingness in the outcome may still be correlated with treatment, observable factors, or both. Hence, the population first-order condition for the selected sample, $\mathbb{E}[S \cdot W \cdot \mathbf{X}'U(g)]$, is not zero even though $\mathbb{E}[\mathbf{X}'U(g)] = \mathbf{0}$. Therefore, $\boldsymbol{\theta}_g^{\mathbf{0}}$ cannot be identified.

3 Identification of the parameter of interest

For identification of the main parameter, I make the following assumption.

Assumption 2. (Strong ignorability) Assume, $Y(0), Y(1) \perp\!\!\!\perp W|\mathbf{X}$.
(i) The vector of covariates, $\mathbf{X}$, is always observed for the entire sample.
(ii) For all $\mathbf{x} \in \mathbf{X}$, define $p(\mathbf{x}) = \mathbb{P}(W = 1|\mathbf{X} = \mathbf{x})$ such that $\kappa < p(\mathbf{x}) < 1$ for a constant $\kappa > 0$.

This indicates that conditioning on covariates is enough to parse out any systematic differences between the treatment and control groups, also known as unconfoundedness. Part (i) requires that we observe these covariates for all individuals. Part (ii) is an overlap condition that ensures that we observe units in both the treatment and control groups for each value of $\mathbf{x}$ in the population. Previous literature has found several situations where unconfoundedness is a tenable assumption. This is especially true when pre-treatment values of the outcome variable are available. For example, Lalonde [33] and Hotz et al. [34] have shown that controlling for pre-training earnings alone reduces significant bias between non-experimental and experimental estimates. The literature assessing teacher impact on student achievement has reported similar findings with pre-test scores [35–37].

Assumption 3. (Missing at Random) Assume, $Y(0), Y(1) \perp\!\!\!\perp S|\mathbf{X}, W$. (i) In addition to $\mathbf{X}$, W is always observed for the entire sample. (ii) For each $(\mathbf{x}, w) \in (\mathbf{X}, W)$, define $r(\mathbf{x}, w) = \mathbb{P}(S = 1|\mathbf{X} = \mathbf{x}, W = w)$ such that $r(\mathbf{x}, w) > \eta$ for a constant $\eta > 0$ and $w = 0, 1$.

Assumption 4. (Random Sampling) $\{(Y_i, \mathbf{X}_i, W_i, S_i);\ i = 1, 2, \ldots, N\}$ are *i.i.d* draws from an infinite population.

Assumption 3 is known as missing at random or MAR and represents an ignorable missing data mechanism. It implies that missingness depends only on observables and not on the missing values of the variable itself [38]. This includes missing completely at random as a special case [1]. The need to condition on W in addition to X helps to deal with the possibility that treatment itself can alter the probability of observing the outcome. This is especially useful in explaining cases of differential nonresponse. Parts (i) and (ii) have similar interpretations as before. Finally, Assumption 4 is a standard random sampling assumption.

Lemma 1. (Identification) *Given Assumptions* 1–3, *assume* (*i*) $q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)$ *is a real-valued function for all* $(Y(g), \mathbf{X})$ *and* (*ii*) $\mathbb{E}[|q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)|] < \infty$ *for all* $\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g$, $g = 0, 1$, *then*

$$\mathbb{E}\left[\frac{S \cdot W}{r(\mathbf{X}, W)\cdot\, p(\mathbf{X})} \cdot q(Y, \mathbf{X}, \boldsymbol{\theta}_1)\right] = \mathbb{E}[q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1)]$$

and

$$\mathbb{E}\left[\frac{S \cdot(1 - W)}{r(\mathbf{X}, W)\cdot(1 - p(\mathbf{X}))} \cdot q(Y, \mathbf{X}, \boldsymbol{\theta}_0)\right] = \mathbb{E}[q(Y(0), \mathbf{X}, \boldsymbol{\theta}_0)].$$

Define $\omega_1 = \frac{S \cdot W}{r(\mathbf{X}, W) \cdot p(\mathbf{X})}$, and $\omega_0 = \frac{S \cdot (1 - W)}{r(\mathbf{X}, W) \cdot (1 - p(\mathbf{X}))}$ for notational simplicity. Lemma 1 implies that solving the population problem with weights, ω_g, is equivalent to solving the original M-estimation problem given in Assumption 1. The proof uses two applications of the law of iterated expectations (LIEs) with unconfoundedness and MAR to arrive at the above result.

Remark. Given that treatment selection is widely viewed as a form of a missing data problem, one argument is to simply combine the two selection problems into one. Imagine a single binary indicator, $D = S \cdot W$. Arguably, existing IPW results could then be applied directly using the single indicator, D. While this may be convenient, such an approach fails to acknowledge the fact that (i) S and W often represent two different selection problems, and (ii) combining them into one may lead to a loss in efficiency. Hence, a more rigorous treatment necessitates considering each issue separately.

4 Asymptotic theory when the conditional feature of interest is misspecified

Given that $r(\mathbf{X}, W)$ and $p(\mathbf{X})$ are unknown, the following assumptions posit that we have a correctly specified model for the propensity score and missing outcome probability. Since both W and S are binary responses, estimation of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\delta}_0$ using maximum likelihood (MLE) will be asymptotically efficient under correct specification of these functions. Consistency and asymptotic normality for $\boldsymbol{\gamma}_0$ and $\boldsymbol{\delta}_0$ follow from Theorems 2.5 and 3.3 of [39].

Assumption 5. (Correct parametric specification of probability models) Assume that

(i) there exists a known parametric function $G(\mathbf{X}, \boldsymbol{\gamma})$ for $p(\mathbf{X})$, where $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ and $0 < G(\mathbf{X}, \boldsymbol{\gamma}) < 1$. Similarly, there exists a known parametric function $R(\mathbf{X}, W, \boldsymbol{\delta})$ for $r(\mathbf{X}, W)$, where $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ and $R(\mathbf{X}, W, \boldsymbol{\delta}) > 0$;

(ii) there exists $\boldsymbol{\gamma}_0 \in \boldsymbol{\Gamma}$ and $\boldsymbol{\delta}_0 \in \boldsymbol{\Delta}$ s.t. $p(\mathbf{X}) = G(\mathbf{X}, \boldsymbol{\gamma}_0)$ and $r(\mathbf{X}, W) = R(\mathbf{X}, W, \boldsymbol{\delta}_0)$.

The “doubly weighted” estimator is then defined as follows:

$$\hat{\boldsymbol{\theta}}_g = \underset{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g}{\operatorname{argmin}} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot q(Y_i, \mathbf{X}_i, \boldsymbol{\theta}_g), \tag{4}$$

where

$$\widehat{\omega}_{i1} \equiv \frac{S_i \cdot W_i}{R(\mathbf{X}_i, W_i, \hat{\boldsymbol{\delta}}) \cdot G(\mathbf{X}_i, \hat{\boldsymbol{\gamma}})} \quad \text{and} \quad \widehat{\omega}_{i0} \equiv \frac{S_i \cdot (1 - W_i)}{R(\mathbf{X}_i, W_i, \hat{\boldsymbol{\delta}}) \cdot (1 - G(\mathbf{X}_i, \hat{\boldsymbol{\gamma}}))}$$

are the estimated weights for solving the treatment and control group problems, respectively. Let $\mathbf{d}_i$ and $\mathbf{b}_i$ denote scores of the binary response log-likelihood problems for estimating the propensity score and missing outcome probability models evaluated at probability limits $\boldsymbol{\gamma}_0$ and $\boldsymbol{\delta}_0$, respectively. Also, let $\mathbf{h}(\boldsymbol{\theta}_g^0) \equiv \mathbf{h}_g$ denote the score of $q(\cdot)$ at the true parameter value and assume that it exists with probability one.

Theorem 1. (Asymptotic normality) *Under Assumptions* 1–5 *and conditions* (1)–(13) *in the Appendix*, $\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^0) \overset{d}{\rightarrow} N(\boldsymbol{0}, \mathbf{H}_g^{-1}\boldsymbol{\Omega}_g\mathbf{H}_g^{-1})$, *where* $\boldsymbol{\Omega}_g = \mathbb{E}(\mathbf{l}_{ig}\mathbf{l}_{ig}') - \mathbb{E}(\mathbf{l}_{ig}\mathbf{b}_i')\mathbb{E}(\mathbf{b}_i\mathbf{b}_i')^{-1}\mathbb{E}(\mathbf{b}_i\mathbf{l}_{ig}') - \mathbb{E}(\mathbf{l}_{ig}\mathbf{d}_i')\mathbb{E}(\mathbf{d}_i\mathbf{d}_i')^{-1}\mathbb{E}(\mathbf{d}_i\mathbf{l}_{ig}')$ *for each* $g = 0, 1$ *and* $\mathbf{l}_{ig} \equiv \omega_{ig}\mathbf{h}_{ig}$ *is score of the weighted objective function evaluated at* $\boldsymbol{\theta}_g^0$.

The asymptotic variance expression derived above offers some interesting insights. First, the middle term, $\boldsymbol{\Omega}_g$, represents the variance of the residual from the population regression of the weighted score, $\mathbf{l}_{ig}$, on the two binary response scores, $\mathbf{b}_i$ and $\mathbf{d}_i$. Note that the covariance term between the two MLE scores is zero since they are conditionally independent.

Second, the expression for $\boldsymbol{\Omega}_g$ has an efficiency implication for $\hat{\boldsymbol{\theta}}_g$. When one is only willing to assume identification of $\boldsymbol{\theta}_g^0$ in the unconditional sense of Assumption 1, it is potentially more efficient to estimate the two weights even when they are known. To show this formally, let us assume that $p(\mathbf{X})$ and $r(\mathbf{X}, W)$ are known and $\tilde{\boldsymbol{\theta}}_g$ is the doubly weighted estimator that uses known weights, ω_g. Then,

Corollary 1. (Efficiency gain with estimated weights) *Under the assumptions of Theorem* 1,

$$\text{Avar}[\sqrt{N}(\tilde{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^0)] - \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^0)] = \mathbf{H}_g^{-1}\boldsymbol{\Sigma}_g\mathbf{H}_\mathbf{g}^{-1} - \mathbf{H}_\mathbf{g}^{-1}\boldsymbol{\Omega}_g\mathbf{H}_g^{-1} = \mathbf{H}_g^{-1}(\boldsymbol{\Sigma}_g - \boldsymbol{\Omega}_g)\mathbf{H}_g^{-1}$$

is positive semi-definite and where $\boldsymbol{\Sigma}_g = \mathbb{E}(\mathbf{l}_{ig}\mathbf{l}_{ig}')$.

In other words, we do no worse, asymptotically, by estimating the weights even when we actually know them. This result has also been called the "efficiency puzzle" and has been studied in Wooldridge [7] and others [26–29]. It is understood to arise from the suboptimal use of moment conditions in two-step procedures.

5 The conditional feature of interest is correctly specified

This section discusses the second half of the asymptotic theory for the doubly weighted estimator where identification of the parameter in question is formalized using a strong identification assumption.

Assumption 6. (Strong conditional identification of $\boldsymbol{\theta}_g^0$) The parameter vector $\boldsymbol{\theta}_g^0 \in \boldsymbol{\Theta}_g$ is the unique solution to the population minimization problem, $\min_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g} \mathbb{E}[q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)|\mathbf{X}]$, for each $g = 0, 1$.

Assumption 6 describes situations where a conditional feature of interest is correctly specified. This can be seen as strengthening the identification assumption in Section 4 since LIE implies that $\boldsymbol{\theta}_g^0$ will also be a solution to the unconditional M-estimation problem.

An implication of this identification argument is that $\boldsymbol{\theta}_g^0$ solves the conditional score of the objective function, i.e., $\mathbb{E}[\mathbf{h}_g|\mathbf{X}] = \mathbf{0}$. For instance, the conditional score will be zero when estimating a correctly specified conditional mean function using either least squares or quasi maximum likelihood in the linear exponential family. It would also hold for a correctly specified conditional quantile function estimated either using quantile regression or quasi maximum likelihood in the tick exponential family [40].

Under this conditional identification assumption, correct specification of the probability weights is not required for the doubly weighted estimator to be consistent for $\boldsymbol{\theta}_g^{\mathbf{0}}$. In other words, $R(\cdot,\cdot,\boldsymbol{\delta})$ and $G(\cdot,\boldsymbol{\gamma})$ are allowed to be misspecified. Formally,

Assumption 7. (Parametric specification of probability models) Assume that
(i) First part of Assumption 5 holds.
(ii) There exists $\boldsymbol{\gamma}^* \in \boldsymbol{\Gamma}$ and $\boldsymbol{\delta}^* \in \boldsymbol{\Delta}$ such that plim($\hat{\boldsymbol{\gamma}}$) = $\boldsymbol{\gamma}^*$ and plim($\hat{\boldsymbol{\delta}}$) = $\boldsymbol{\delta}^*$, respectively.

Under this setting, the weights are given by

$$\omega_1^* = \frac{S \cdot W}{R(\mathbf{X}, W, \boldsymbol{\delta}^*)\cdot G(\mathbf{X}, \boldsymbol{\gamma}^*)} \quad \text{and} \quad \omega_0^* = \frac{S \cdot (1 - W)}{R(\mathbf{X}, W, \boldsymbol{\delta}^*)\cdot(1 - G(\mathbf{X}, \boldsymbol{\gamma}^*))},$$

where $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$ solve the same binary response problems as before but converge to probability limits given by pseudo-true values $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$, respectively [31]. The identification argument in this case can be briefly explained as follows:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_g = \underset{\boldsymbol{\theta}_g}{\text{argmin}} \sum_{i=1}^{N} \omega_{ig}^* \cdot q(Y_i, \mathbf{X}_i, \boldsymbol{\theta}_g) &\xrightarrow{p} \underset{\boldsymbol{\theta}_g}{\text{argmin}}\, \mathbb{E}[\omega_g^* \cdot q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)] \\ &= \underset{\boldsymbol{\theta}_g}{\text{argmin}}\, \mathbb{E}[\xi_g(\mathbf{X})\cdot q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)] \\ &= \underset{\boldsymbol{\theta}_g}{\text{argmin}}\, \mathbb{E}[\xi_g(\mathbf{X})\cdot \mathbb{E}(q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)|\mathbf{X})] \\ &= \boldsymbol{\theta}_g^{\mathbf{0}},\end{aligned} \tag{5}$$

where $\xi_g(\mathbf{X}) > 0$ is a function of weights. If $\boldsymbol{\theta}_g^{\mathbf{0}}$ is a solution to the conditional problem $\mathbb{E}[q(\cdot)|\mathbf{X}]$, it will also solve equation (5) (multiplication by $\xi_g(\mathbf{X})$ will not affect the conditional minimization problem). Therefore, solving the doubly weighted objective function identifies the parameter even if the weights are misspecified. Theorem 2 establishes asymptotic results under this case.

Theorem 2. (Asymptotic normality under strong identification) *Under Assumptions* 2–4, 6, *and* 7, *and conditions* (1)–(13) *in the Appendix,* $\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}}) \xrightarrow{d} N(\boldsymbol{0}, \mathbf{H}_g^{-1}\boldsymbol{\Omega}_g\mathbf{H}_g^{-1})$, *where* $\boldsymbol{\Omega}_g = \mathbb{E}(\mathbf{l}_{ig}\mathbf{l}_{ig}')$ *with* $\mathbf{H}_g$ *and* $\mathbf{l}_{ig}$ *defined in Theorem* 1 *with asymptotic weights given by* ω_{ig}^*.

Unlike the previous section, $\boldsymbol{\Omega}_g$ now is simply the variance of the weighted score of $q(\cdot)$ without the first-stage adjustment of the estimated probabilities. This is because under Assumption 6, the correlation between the score functions of the first- and second-step estimating equations is zero i.e., $\mathbb{E}(\mathbf{l}_{ig}\mathbf{b}_i') = \mathbb{E}(\mathbf{l}_{ig}\mathbf{d}_i') = \mathbf{0}$. In other words, when $\boldsymbol{\theta}_g^{\mathbf{0}}$ is correctly specified for a conditional feature of interest and an appropriate estimation method is used, there is optimal usage of moment conditions.

A simpler expression for $\boldsymbol{\Omega}_g$ also means that we can no longer exploit the correlation between scores to obtain an efficient estimator of $\boldsymbol{\theta}_g^{\mathbf{0}}$. Again, let $\tilde{\boldsymbol{\theta}}_g$ uses true weights, ω_g. Then, we have the following result.

Corollary 2. (No gain with estimated weights under strong identification) *Under the assumptions of Theorem* 2, Avar[$\sqrt{N}(\tilde{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}})$] = Avar[$\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}})$] = $\mathbf{H}_g^{-1}\boldsymbol{\Omega}_g\mathbf{H}_g^{-1}$.

A special case is when ω_g^* is just a constant since $R(\cdot,\cdot)$ and $G(\cdot,\cdot)$ are allowed to be any positive functions of $\mathbf{X}$ and W. This implies the unweighted estimator, $\hat{\boldsymbol{\theta}}_g^{\boldsymbol{u}}$, which does not weight at all, is also consistent for $\boldsymbol{\theta}_g^{\mathbf{0}}$ under the results of Theorem A.2. In this case, one may turn to asymptotic efficiency to guide our choice between weighting or not weighting at all. The following result says that if the objective function satisfies the generalized conditional information matrix equality (GCIME), the unweighted estimator is asymptotically more efficient than any weighted counterpart (correct weights or not).

Corollary 3. (Efficiency gain with unweighted estimator under GCIME) *Under assumptions of Theorem* 2, *if we additionally suppose that the objective function satisfies GCIME in the population, which is defined as*:

$$\mathbb{E}[\mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})\mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})'|\mathbf{X}] = \sigma_{0g}^2 \cdot \nabla_{\boldsymbol{\theta}_g}\mathbb{E}[\mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})|\mathbf{X}] = \sigma_{0g}^2 \cdot \mathbf{A}(\mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}}), \tag{6}$$

then, $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}})] = \mathbf{H}_g^{-1}\boldsymbol{\Omega}_g\mathbf{H}_g^{-1}$ *and* $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_g^{\boldsymbol{u}} - \boldsymbol{\theta}_g^{\mathbf{0}})] = (\mathbf{H}_g^{\mathbf{u}})^{-1}\boldsymbol{\Omega}_g^{\boldsymbol{u}}(\mathbf{H}_g^{\mathbf{u}})^{-1}$, *and*

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}})] - \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_g^{\boldsymbol{u}} - \boldsymbol{\theta}_g^{\mathbf{0}})]$$

is positive semi-definite.

The proof of this theorem follows from noting that we can express the difference in the two asymptotic variances as the expected outer product of population residuals from the regression of $\mathbf{B}_i$ on $\mathbf{D}_i$, which are weighted versions of square root of matrix $\mathbf{A}_i$ (see Appendix B for details). Hence, the difference is positive semi-definite.

We know GCIME in a variety of estimation contexts. In the case of full maximum likelihood, GCIME holds for $q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g) = -\ln f_g(Y|\mathbf{X}, \boldsymbol{\theta}_g)$, where $f_g(\cdot|\cdot)$ is the true conditional density with $\sigma_{0g}^2 = 1$. For estimating conditional mean parameters using quasi maximum likelihood estimation in the linear exponential family, GCIME holds if $\text{Var}(Y(g)|\mathbf{X}) = \sigma_{0g}^2 \cdot v[m(\mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})]$. In other words, GCIME will be satisfied if $\text{Var}(Y(g)|\mathbf{X})$ satisfies the generalized linear model assumption, irrespective of whether the higher-order moments of the conditional distribution correspond with the chosen quasi-log likelihood or not. For estimation using nonlinear least squares, GCIME will hold for $q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g) = [Y(g) - m(\mathbf{X}, \boldsymbol{\theta}_g)]^2$ with the homoskedasticity assumption. In all these cases, the unweighted estimator will be more efficient than any weighted counterpart. Otherwise, the two may not be easy to rank.

6 Estimation of treatment effects

I now use asymptotic results discussed in Sections 4 and 5 for estimation of ATE and QTEs, which can be expressed as functions of the doubly weighted estimator, $\hat{\boldsymbol{\theta}}_g$.

6.1 Double robust estimation of ATE

Let $m(\mathbf{X}, \boldsymbol{\theta}_g)$ be a parametric model for the conditional mean, $\mathbb{E}[Y(g)|\mathbf{X}]$. Define

$$\Delta_{\text{ate}} = \mathbb{E}[m(\mathbf{X}, \boldsymbol{\theta}_1) - m(\mathbf{X}, \boldsymbol{\theta}_0)]. \tag{7}$$

6.1.1 First half: Correct conditional mean

If the mean model is correctly specified, one could consistently estimate $\boldsymbol{\theta}_g^{\mathbf{0}}$ using nonlinear least squares where the estimator solves a weighted nonlinear least squares problem, i.e.,

$$\hat{\boldsymbol{\theta}}_g = \underset{\boldsymbol{\theta}_g}{\text{argmin}} \sum_{i=1}^{N} \widehat{\omega}_{ig} \cdot (Y_i - m(\mathbf{X}_i, \boldsymbol{\theta}_g))^2.$$

Since we are operating under Assumption 6, results from Section 5 dictate that the doubly weighted estimator would be consistent for $\boldsymbol{\theta}_g^{\mathbf{0}}$ irrespective of correct or incorrect weights. This is what forms the “first part” of the

DR result for ATE estimation. One could also replace $q(\cdot)$ with quasi maximum likelihood and still consistently estimate conditional mean parameter, $\boldsymbol{\theta}_g^{\mathbf{0}}$ for $g = 0, 1$.

6.1.2 Second half: Correct weights

A consistent estimator for the ATE can generally not be obtained using equation (7) if the conditional mean function is misspecified. In the generalized linear model literature, certain combinations of quasi log likelihood and link functions lead to first-order conditions such that

$$\mathbb{E}[Y(g)] = \mathbb{E}[h(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}})]$$

even though $h(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}})$ is misspecified for the true conditional mean. By allowing misspecification in the mean model, we are operating under Assumption 1, which implies that weighting is crucial for identification of the pseudo-true parameter, $\boldsymbol{\theta}_g^{\mathbf{0}}$ [31]. This forms the "second half" of the DR result for ATE estimation and draws on the theory outlined in the first half (Section 4) under the weak identification assumption.

In particular, the estimation strategy is to choose the mean model to be the *canonical* link, $h^{-1}(\cdot)$ (where $h(\cdot)$ is a strictly increasing function on the real line), and the quasi-log likelihood associated with a linear exponential family. This choice will depend on the range and nature of Y. Then, $\boldsymbol{\theta}_g^{\mathbf{0}}$ solves the following population first-order conditions:

$$\mathbb{E}\left[\omega_g \cdot \frac{\nabla_{\theta_g} h(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}}) \cdot \mathbf{X}' \cdot (Y - h(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}}))}{v[h(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}})]}\right] = \mathbf{0}, \tag{8}$$

where $v[h(\cdot)]$ is variance of the mean function. With the chosen canonical link, equation (8) simplifies to

$$\mathbb{E}[\omega_g \cdot \mathbf{X}' \cdot (Y - h(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}}))] = \mathbf{0}. \tag{9}$$

Since $\mathbf{X}$ includes an intercept, the first-order conditions give us

$$\mathbb{E}[Y(g)] = \mathbb{E}[h(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}})]. \tag{10}$$

The doubly weighted estimator, $\hat{\boldsymbol{\theta}}_g$, then solves the sample analogue of equation (9), i.e.,

$$\sum_{i=1}^{N} \hat{\omega}_{ig} \cdot \mathbf{X}_i' \cdot (Y_i - h(\mathbf{X}_i \hat{\boldsymbol{\theta}}_g)) = \mathbf{0}.$$

For Y with unrestricted support, normal quasi-log likelihood and identity link function ($h(\mathbf{X}\boldsymbol{\theta}_g) = \mathbf{X}\boldsymbol{\theta}_g$) deliver the mean fitting property. Other combinations of quasi-log likelihood and canonical link functions can be found in Table 2 of the study by Negi and Wooldridge [41].

Summary. DR estimation of ATE with double weighting

Case 1: Correct conditional mean and misspecified (or correct) weights: In this case, we are operating under Assumption 6 along with Assumption 7 (or Assumption 5 in the case of correct weights). Results in Section 5 will apply. The first half of this section discusses estimation of ATE under this scenario.

Case 2: Correct weights and misspecified mean: In this case, we are operating under Assumption 1. Results in Section 4 apply. The second half of this section discusses estimation of ATE under this scenario.

Combining the two halves, $\hat{\Delta}_{\text{ate}} = \frac{1}{N}\sum_{i=1}^{N}\{m(\mathbf{X}_i, \hat{\boldsymbol{\theta}}_1) - m(\mathbf{X}_i, \hat{\boldsymbol{\theta}}_0)\}$ gives us a DR estimator of Δ_{ate}.

A similar result is discussed in Theorem 14.2 of the study by Ding [42], which shows that a weighted least squares fit of Y_i on $(1, W_i, \mathbf{X}_i, W_i \cdot \mathbf{X}_i)$ with inverse propensity score weights produces a DR estimator for the ATE. This is identical to the inverse propensity weighted linear regression adjustment estimator discussed in the study by Imbens and Wooldridge [43], which is well known as being DR under the treatment selection problem.

6.2 Estimation of QTEs

In this section, I use double weighting to illustrate the estimation of three different quantile parameters, namely, unconditional quantile treatment effect (UQTE), CQTE, and *a weighted linear approximation* to the CQTE, each of which may be of interest to the researcher depending on whether quantiles of the conditional or unconditional outcomes distribution are of interest.

CQTE$_\tau$: Let $\mathrm{q}_\tau(\mathbf{X}, \boldsymbol{\theta}_g(\tau))$ be a correctly specified parametric model for the conditional quantile function. Then,

$$\mathrm{CQTE}_\tau(\mathbf{X}) = \mathrm{q}_\tau(\mathbf{X}, \boldsymbol{\theta}_1(\tau)) - \mathrm{q}_\tau(\mathbf{X}, \boldsymbol{\theta}_0(\tau)).$$

In this case, there are two methods that will ensure consistent estimation of $\boldsymbol{\theta}_g^{\mathbf{0}}(\tau)$. The first is quantile regression [44], where

$$\hat{\boldsymbol{\theta}}_g(\tau) = \arg\min_{\boldsymbol{\theta}_g(\tau)\in\boldsymbol{\Theta}_g} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot c_\tau(Y_i - \mathrm{q}_\tau(\mathbf{X}_i, \boldsymbol{\theta}_g(\tau))). \tag{11}$$

Since we are operating under Assumption 6, the results in Section 5 dictate that weighting the objective functions, irrespective of whether the weights are correctly specified or not, will yield a consistent estimator of $\boldsymbol{\theta}_g(\tau)$. One could also use quasi maximum likelihood in the special "*tick-exponential*" family of distributions to consistently estimate the conditional quantile parameters. As shown by Komunjer [40], quantile regression proposed in Koenker and Bassett [44] is a special case of this quasi-maximum likelihood class of estimators.

LP to CQTE$_\tau$: In the event that $\mathrm{q}_\tau(\mathbf{X}, \boldsymbol{\theta}_g(\tau))$ is misspecified as being linear, one can still interpret it as providing the best linear approximation to the true conditional quantile function. This property of quantile regression is analogous to the approximation property of linear regression under conditional mean misspecification and was established in the study by Angrist et al. [32]. Given such misspecification in the conditional quantile function, the first half requires the weights to be correct in order to consistently estimate the LP parameters, $\boldsymbol{\theta}_g(\tau)$, i.e.,

$$\hat{\boldsymbol{\theta}}_g(\tau) = \operatorname{argmin}_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot c_\tau(Y_i - \mathbf{X}_i\boldsymbol{\theta}_g(\tau)) \tag{12}$$

will be consistent for the pseudo-true parameter, $\boldsymbol{\theta}_g^{\mathbf{0}}(\tau)$, which indexes the population LP to the true conditional quantile function. Then,

$$\widehat{\mathrm{LP}}[\mathrm{CQTE}_\tau(\mathbf{X})] = \mathbf{X}[\hat{\boldsymbol{\theta}}_1(\tau) - \hat{\boldsymbol{\theta}}_0(\tau)]$$

is interpreted as providing the best LP to the true CQTE.

Unconditional QTE$_\tau$ (UQTE$_\tau$): Let $\theta_g(\tau)$ be the τth unconditional quantile of the potential outcome, $Y(g)$. Then,

$$\widehat{\mathrm{UQTE}}_\tau = \hat{\theta}_1(\tau) - \hat{\theta}_0(\tau),$$

where

$$\hat{\theta}_g(\tau) = \arg\min_{\theta_g(\tau)\in\Theta_g} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot c_\tau(Y_i - \theta_g(\tau)).$$

In this case, weighting is crucial for consistent estimation of $\theta_g(\tau)$ since no quantile model is being specified, and we are operating under Assumption 1.

7 Simulations

This section compares the empirical distributions of average and QTE estimators using unweighted, propensity score weighted (ps-weighted), and doubly weighted (d-weighted) estimators. For estimating ATE, data are simulated using a probit as, $Y(g) = \mathbf{1}(\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}} + U(g) > 0)$, where $\mathbf{X}$ includes an intercept and two covariates.

For estimating the QTE parameters, I use an exponential data generating process where the potential outcomes are generated as $Y(g) = \exp[\mathbf{X}\boldsymbol{\theta}_g^{\mathbf{0}} + U(g)]$, for $g = 0, 1$. For each setting, the covariates and latent errors have been drawn from two independent bivariate normal distributions. The treatment assignment and missing outcome mechanisms satisfy the assumptions of unconfoundedness and MAR with a 41% probability of being treated and 38% of being observed in the population, respectively. The empirical distributions of the unweighted, ps-weighted, and d-weighted estimators are then obtained for a sample of size 5,000 using 1,000 replication draws from a population of 1 million observations. Additional details can be found in Section S.1 of the Supplementary Material.

The discussion of the results is centered around two main misspecification scenarios: (1) when some conditional model (conditional mean function or conditional quantile function) is misspecified and (ii) when the weights are misspecified (enumerated in Tables S.1.1 and S.1.2). These correspond to scenarios outlined in the first and second half of the asymptotic theory.

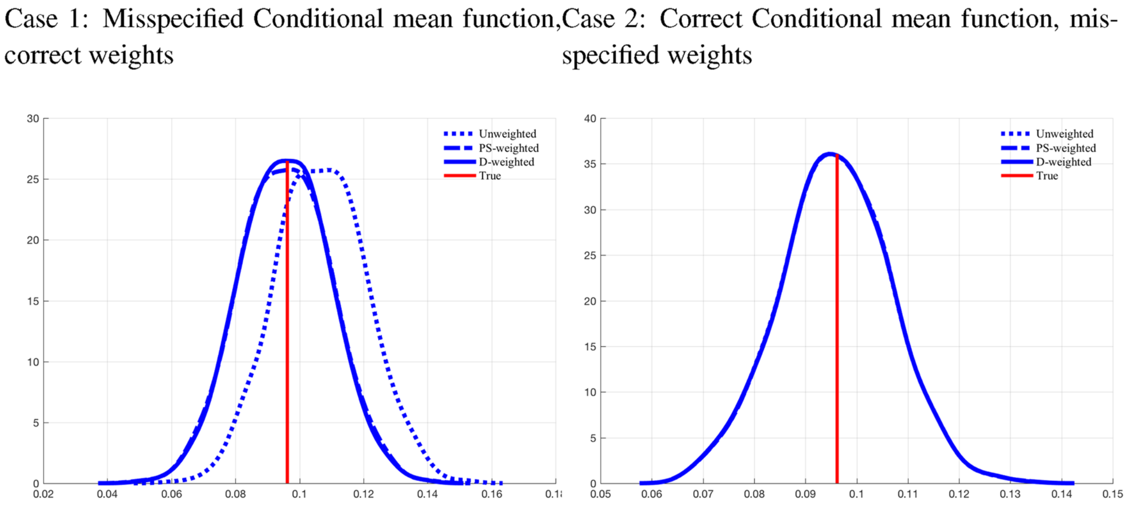
7.1 ATE: Results

Case 1 in Figure 1 considers a misspecified mean function but correct probability weights. This is the principal case covered in Section 4 where weighting is crucial. As one can see, the empirical distribution of the doubly weighted estimator is centered on the true ATE, whereas the distribution for the unweighted estimator is shifted to the right.

Case (2) of the same figure depicts the scenario of a correctly specified conditional mean function but misspecified weights. Here, weighting does not matter for consistent estimation of ATE since the mean function is correctly specified. All three empirical distributions, namely, unweighted, ps-weighted, and d-weighted, coincide and are centered on the true ATE.

7.2 QTE: Results

Figure 2 discusses the case when conditional quantile function is misspecified but the weights are correct. Using results obtained in the study by Angrist et al. [32], I interpret the solution to the double-weighted


Case 1: Misspecified Conditional mean function, correct weights

Case 2: Correct Conditional mean function, misspecified weights

Figure 1: Empirical distribution of estimated ATE for $N = 5{,}000$. *Notes*: This figure plots the empirical distributions of the unweighted, ps-weighted, and d-weighted ATE estimates using 1,000 Monte Carlo simulation draws of sample size 5,000. The average treated sample size is $N_1 = 5{,}000 \times 0.41 \times 0.38 = 779$, and the average control sample size is $N_0 = 5{,}000 \times (1 \;\; 0.41) \times 0.38 = 1,121$. The true ATE = 0.096, and the population is generated using a million observations. The unweighted estimator does not weight the observed data. The ps-weighted estimator weights to correct only for nonrandom assignment, and the d-weighted estimator weights by both the treatment and missing outcomes probabilities.

problem given in equation (12) as providing a consistent weighted linear approximation to the true conditional quantile function, which is then used to estimate an LP to the true CQTE. Figure 2 plots the bias in estimated LP using the three estimators relative to the true LP as a function of X_1. We see that the double-weighted estimator performs the best.

Next, I consider the case when the conditional quantile function is correctly specified but the weights are wrong. Since in this case one can consistently estimate the CQTE irrespective of the weights, Figure 3 plots the CQTE curve as a function of X_1. In this case, as theory suggests, all three estimators of the CQTE function, i.e., unweighted, ps-weighted, and d-weighted, coincide with the true CQTE. Section S.1.2 of the Supplementary Material provides details about plotting the CQTE curve.

Finally, Figure 4 plots the empirical distribution of UQTE for the three estimators. We find that both unweighted and d-weighted estimators have a comparable finite sample bias. Propensity score weighting performs the worst in both cases. All results correspond to the twenty-fifth quantile of the outcome distribution. The results for other quantiles can be found in Section S.5 of the Supplementary Material.

8 Returns to job training

In this section, I apply the proposed estimator to the AFDC sample of women from the NSW training program compiled by Calónico and Smith [30] (CS, thereafter). NSW was a transitional and subsidized work experience program, which was implemented as a randomized experiment in the United States between 1975–1979. CS replicate Lalonde's [33] within-study analysis for the AFDC women in the program, where the purpose of such an analysis is to evaluate how training estimates obtained from using non-experimental identification strategies (assuming unconfoundedness) compare to experimental estimates. To compute the non-experimental estimates, CS combine the NSW experimental sample with two non-experimental comparison groups drawn from panel study of income dynamics (PSID), called PSID-1 and PSID-2. I utilize this within-study feature to estimate how close the d-weighted estimates are to the experimental benchmark compared with ps-weighted and unweighted estimates.

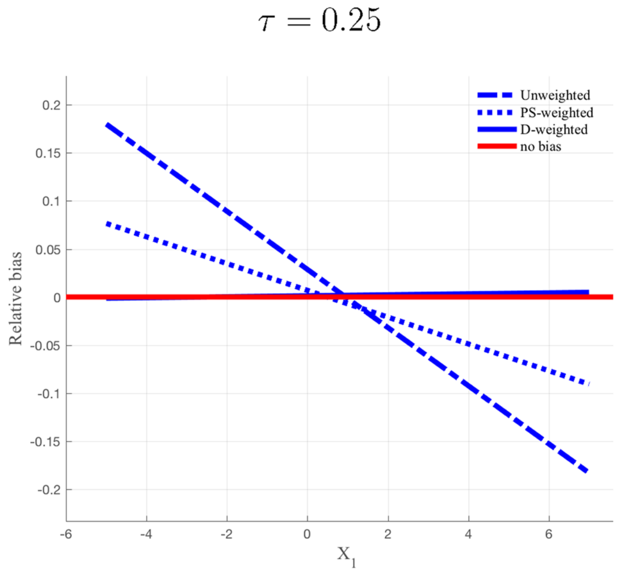


Figure 2: Relative bias of the estimated LP (CQTE) as a function of X_1 when the conditional quantile function is misspecified but weights are correct. *Notes:* This figure plots the bias in the unweighted, ps-weighted, and d-weighted LPs to CQTE relative to the true population LP for $N = 5{,}000$. The average treated sample size is $N_1 = 5{,}000 \times 0.41 \times 0.38 = 779$, and the average control sample size is $N_0 = 5{,}000 \times (1 \;\; 0.41) \times 0.38 = 1{,}121$. The unweighted estimator does not weight the observed data. The ps-weighted estimator weights to correct only for nonrandom assignment, and the d-weighted estimator weights by the treatment and missing outcomes probabilities.

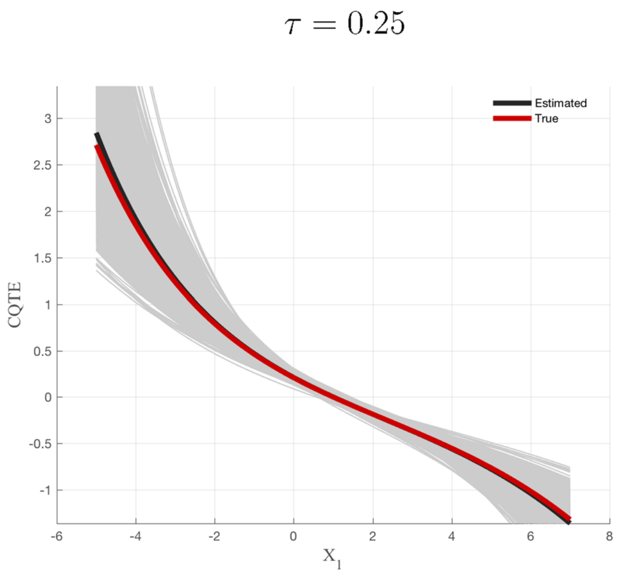


Figure 3: Estimated CQTE with true CQTE as a function of X_1 when conditional quantile function is correct but weights are misspecified. *Notes:* This figure plots the average d-weighted CQTE function with the true CQTE along X_1 for 1,000 Monte Carlo simulation draws of sample size $N = 5{,}000$. Along with these two graphs, the figure also plots the individual function across the 1,000 simulation draws. The average treated sample is $N_1 = 5{,}000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5{,}000 \times (1 \;\; 0.41) \times 0.38 = 1{,}121$.

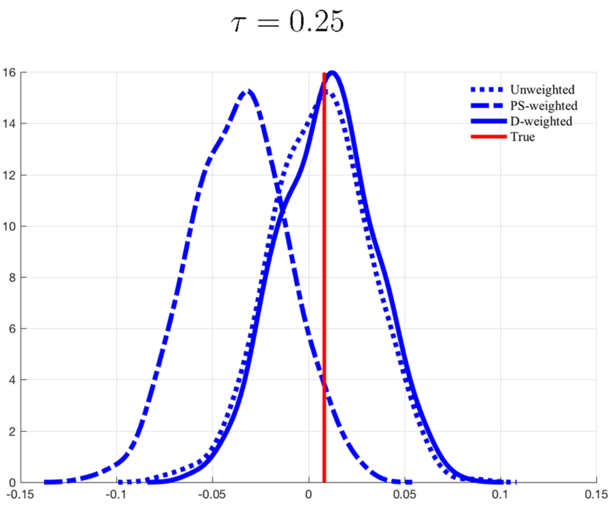


Figure 4: Empirical distribution of estimated UQTE. *Notes:* This figure plots the empirical distributions of the unweighted, ps-weighted, and d-weighted UQTE estimates using 1,000 Monte Carlo simulation draws of sample size 5,000. The average treated sample is $N_1 = 5{,}000 \times 0.41 \times 0.38 = 779$, and the average control sample is $N_0 = 5{,}000 \times (1 \;\; 0.41) \times 0.38 = 1{,}121$. The unweighted estimator does not weight the observed data. The ps-weighted estimator weights to correct only for nonrandom assignment, and the d-weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with nonrandom assignment and missing outcome problems.

To construct these empirical bias measures, I first augment the CS sample to allow for women who had missing earnings information in 1979. This renders 26% of the experimental and 11% of the PSID samples missing. I then combine the experimental treatment group of NSW with three distinct comparison groups present in the CS dataset, namely, the experimental control group, and the two PSID samples, to compute the unweighted, ps-weighted, and d-weighted training estimates, respectively. The difference between the non-experimental estimate, obtained from using the d-weighted estimator, and the experimental estimate provides

Table 1: Covariate means and p-values from the test of equality of two means for the observed and missing samples

Covariates	Experimental						Non-experimental					
	Control			Treatment			PSID-1			PSID-2		
	Missing	Observed	$P(\lvert T\rvert > \lvert t\rvert)$	Missing	Observed	$P(\lvert T\rvert > \lvert t\rvert)$	Missing	Observed	$P(\lvert T\rvert > \lvert t\rvert)$	Missing	Observed	$P(\lvert T\rvert > \lvert t\rvert)$
Age, years	33.36	33.74	0.51	32.15	33.77	0.01	34.00	37.07	0.01	33.32	34.54	0.62
	(7.30)	(7.15)		(7.39)	(7.40)		(10.50)	(10.57)		(10.81)	(9.34)	
Years of education	10.29	10.26	0.85	10.29	10.31	0.89	11.44	11.30	0.60	11.05	10.49	0.18
	(1.93)	(2.03)		(2.05)	(1.88)		(2.17)	(2.77)		(1.73)	(2.13)	
Proportion of high school dropouts	0.70	0.68	0.57	0.69	0.70	0.77	0.43	0.45	0.73	0.55	0.59	0.68
	(0.46)	(0.47)		(0.46)	(0.46)		(0.50)	(0.50)		(0.51)	(0.49)	
Proportion married	0.05	0.04	0.61	0.03	0.02	0.75	0.00	0.02	0.00	0.00	0.01	0.16
	(0.21)	(0.19)		(0.16)	(0.15)		(0.00)	(0.14)		(0.00)	(0.10)	
Proportion Black	0.81	0.82	0.81	0.83	0.84	0.87	0.74	0.65	0.10	0.91	0.86	0.50
	(0.39)	(0.39)		(0.38)	(0.37)		(0.44)	(0.48)		(0.29)	(0.35)	
Proportion Hispanic	0.12	0.13	0.87	0.13	0.12	0.64	0.01	0.02	0.82	0.05	0.02	0.62
	(0.33)	(0.33)		(0.33)	(0.32)		(0.11)	(0.12)		(0.21)	(0.15)	
Number of children in 1975	2.33	2.23	0.34	2.14	2.19	0.69	1.54	1.71	0.33	2.41	2.97	0.05
	(1.29)	(1.34)		(1.32)	(1.29)		(1.45)	(1.78)		(1.14)	(1.79)	
Real earnings in 1975	621.54	879.28	0.12	610.77	861.65	0.11	6927.95	7510.92	0.50	896.56	2211.45	0.02
	(1,523.00)	(2,194.93)		(1,677.36)	(2,005.53)		(7,330.74)	(7,541.41)		(2,315.12)	(3,567.50)	
Observations	795	795		796	796		729	729		204	204	

Notes: Along with the covariate means and standard deviation (in parentheses), the table also reports p-values from the test of equality for two means between the observed and missing samples. Real earnings in 1975 are expressed in terms of 1982 dollars.

the first measure of estimated bias associated with the proposed strategy. Combining the experimental control group with the non-experimental comparison group gives a second measure of estimated bias [45]. I report both these bias measures for the average returns to training estimates.

Given the growing importance of estimating distributional impacts of job training programs, I also estimate returns to training at every tenth quantile of the 1979 earnings distribution. The role of double weighting is strong for estimating marginal quantiles since it serves to remove biases arising from the two selection problems.

8.1 Results

First, to evaluate whether women with missing earnings in 1979 were significantly different than those who were observed, Table 1 reports the mean and standard deviation of the woman's age, years of schooling, pre-training earnings, and other characteristics across the observed and missing samples. In terms of age, the women who were observed in the experimentally treated group of NSW and the PSID-1 sample were, on average, older than those who were missing. The observed women in PSID-1 were also more likely to be married. For the PSID-2 sample, women who were observed had, on average, more kids with higher pre-training earnings. All these differences are statistically significant indicating that covariates were statistically different among the missing and observed PSID women (see the non-experimental columns in Table 1). For the experimental group, we do not find the covariates to be systematically different between those who were observed vs. those who were missing (see the experimental columns in Table 1).

The presence of non-experimental control groups implies that assignment was nonrandom and therefore an issue in the sample. This is because the comparison groups were drawn from PSID after imposing only a partial version of the full NSW eligibility criteria. Table 2 provides descriptive statistics for the covariates by the treatment status. As can be expected, the treatment and control groups of NSW are not observably different. In contrast, the women in PSID-1 and PSID-2 groups are statistically different from the treatment group.

Table 3 reports the d-weighted, ps-weighted, and unweighted average returns to training estimates using three different comparison groups: NSW control, PSID-1, and PSID-2. The unweighted (unadjusted and

Table 2: Covariate means and p-values from the test of equality of two means, by treatment status

Covariates	Treatment	Control	$P(\lvert T\rvert > \lvert t\rvert)$	PSID-1	$P(\lvert T\rvert > \lvert t\rvert)$	PSID-2	$P(\lvert T\rvert > \lvert t\rvert)$
Age, years	33.37	33.64	0.46	36.73	0.00	34.41	0.11
	(7.42)	(7.19)		(10.60)		(9.48)	
Years of education	10.30	10.27	0.72	11.32	0.00	10.55	0.07
	(1.92)	(2.00)		(2.71)		(2.09)	
Proportion of high school dropouts	0.70	0.69	0.73	0.45	0.00	0.59	0.00
	(0.46)	(0.46)		(0.50)		(0.49)	
Proportion married	0.02	0.04	0.03	0.02	0.05	0.01	0.08
	(0.15)	(0.20)		(0.13)		(0.10)	
Proportion Black	0.84	0.82	0.29	0.66	0.00	0.87	0.13
	(0.37)	(0.39)		(0.47)		(0.34)	
Proportion Hispanic	0.12	0.13	0.59	0.02	0.00	0.02	0.00
	(0.32)	(0.33)		(0.12)		(0.16)	
Number of children in 1975	2.17	2.26	0.21	1.70	0.00	2.91	0.00
	(1.30)	(1.32)		(1.75)		(1.73)	
Real earnings in 1975	799.88	811.19	0.91	7446.15	0.00	2069.65	0.00
	(1931.92)	(2041.32)		(7515.59)		(3474.10)	
Observations	796	795		729		204	

Notes: Along with the covariate means and standard deviation (in parentheses), the table also reports p-values from the test of equality for two means. Column 4 tests for differences between the NSW treatment and control groups, and columns 6 and 8 report the same using PSID-1 and PSID-2 comparison groups, respectively. Real earnings in 1975 are expressed in terms of 1982 dollars.

Table 3: Unweighted and weighted earnings comparisons and estimated training effects using NSW and PSID comparison groups

Comparison group	Unadjusted			Adjusted			Adjusted		
	Unweighted	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
				Post-training earnings estimates					
NSW	821	848	824	845	852	828	864	850	826
N = 1,185	(307.22)	(304.04)	(304.61)	(303.60)	(302.94)	(303.53)	(303.47)	(302.96)	(303.58)
PSID-1	799	827	803	298	909	907	335	905	904
N = 1,016	(444.84)	(503.00)	(503.26)	(428.60)	(497.76)	(501.54)	(440.18)	(518.54)	(522.97)
PSID-2	31	569	566	492	1,040	996	698	1,082	1,049
N = 720	(713.88)	(1041.81)	(1027.12)	(664.46)	(961.74)	(953.80)	(784.28)	(1264.18)	(1217.46)
				Bias estimates using NSW control					
PSID-1	-1,620	169	156	493	40	21	568	38	21
N = 1,001	(431.75)	(561.74)	(553.07)	(427.93)	(499.91)	(501.44)	(434.59)	(504.19)	(507.02)
PSID-2	853	228	212	109	207	200	378	17	24
N = 705	(707.87)	(1041.44)	(1025.87)	(663.80)	(962.85)	(954.61)	(759.75)	(1195.47)	(1156.39)
Adjusted covariates									
Pre-training earnings (1975)				✓	✓	✓	✓	✓	✓
Age				✓	✓	✓	✓	✓	✓
Age^2				✓	✓	✓	✓	✓	✓
Education				✓	✓	✓	✓	✓	✓
High school dropout				✓	✓	✓	✓	✓	✓
Black				✓	✓	✓	✓	✓	✓
Hispanic				✓	✓	✓	✓	✓	✓
Marital status				✓	✓	✓	✓	✓	✓
Number of Children (1975)							✓	✓	✓

Notes: This table reports unadjusted and adjusted post-training earnings differences between the NSW treatment and three different comparison groups, namely, NSW control, PSID-1 and PSID-2. The first row reports experimental training estimates that combine the NSW treatment and control group, whereas the second and third rows report non-experimental estimates computed from using the PSID-1 and PSID-2 groups, respectively. Each of the non-experimental estimates should be compared to the experimental benchmark. The second panel of the table reports bias estimates computed from combining the NSW control with PSID-1 and PSID-2 comparison groups, respectively. These represent a second measure of bias which should be compared to zero. Bootstrapped standard errors are given in parentheses and have been constructed using 10,000 replications. All values are in 1982 dollars. The samples used for estimating the training and bias estimates have been trimmed to ensure common support in the distribution of weights for the treatment and comparison groups. For more detail, see Section S.3 of the Supplementary Material.

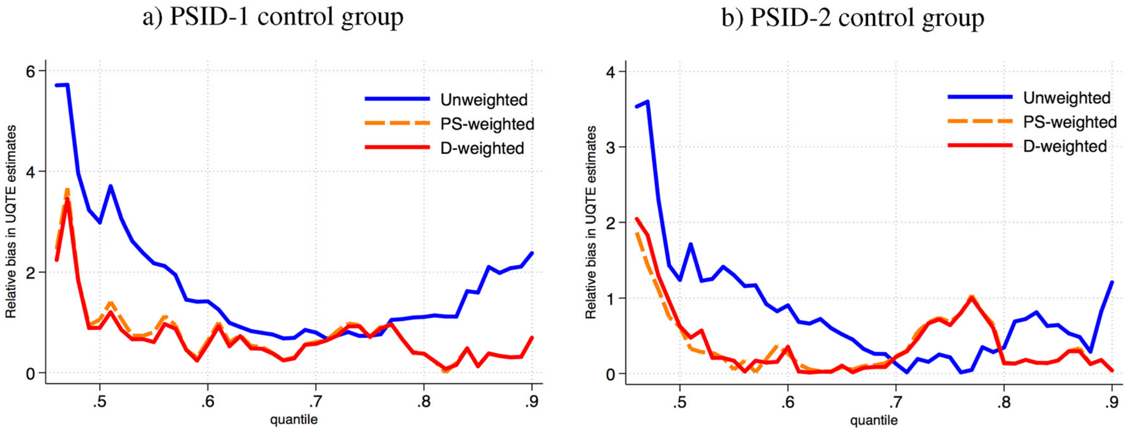


Figure 5: Relative estimated bias in UQTE estimates at different quantiles of the 1979 earnings distribution. (a) PSID-1 control group and (b) PSID-2 control group. *Notes:* This graph plots the bias in the unweighted, ps-weighted, and d-weighted UQTE estimates relative to the true experimental estimates across different quantiles of the 1979 earnings distribution. Panel (a) plots the relative bias estimates using the PSID-1 comparison group and panel (b) plots the same using the PSID-2 comparison group. The treatment and missing outcome propensity score models have been estimated as flexible logits, and the samples used for constructing these estimates have been trimmed to ensure common support across the two groups. The treatment propensity score has been estimated using the full experimental sample along with either PSID-1 or PSID-2 comparison group. The UQTE estimates for $\tau < 0.46$ are omitted from the graph since these are zero.

adjusted) experimental estimates given in row 1 are the same as the estimates reported by CS in Table 3 of their article. Overall, one can see that the double-weighted experimental estimates are more stable than the single-weighted or unweighted estimates across the different regression specifications, with a range between $824 and $828. Moreover, the d-weighted estimator is DR for the ATE, which makes it more reliable as opposed to the other two.

For computing the ps-weighted and d-weighted non-experimental estimates, I first trim the sample to ensure common support between the treatment and comparison groups. Appendix S.3 describes estimation of the two probability weights along with the sample trimming criteria. This reduces the sample size from 1,248 to 1,016 observations for the PSID-1 estimates and from 782 to 720 observations for the PSID-2 estimates. A pattern that is consistent across the two sets of non-experimental estimates is that weighting gets us much closer to the benchmark relative to not weighting at all. For instance, the unweighted simple difference in means estimate of training, which uses the PSID-1 comparison group, is −$799, whereas the weighted estimates are $827 and $803. For the PSID-2 comparison group, the unweighted estimate that controls for all covariates is $335, whereas the weighted estimates are $905 and $904.

The second panel of Table 3 reports the bias in training estimates from combining the experimental control group with the PSID comparison groups. A similar pattern is seen here with weighted bias estimates being much closer to zero than the unweighted estimates. These results suggest that the argument for weighting is strong when using a non-experimental comparison group where nonrandom assignment and missing outcomes are significant problems. Note that the large standard errors for the non-experimental estimates can be attributed to the small sample sizes and to the large residual variance of earnings in the PSID-1 and PSID-2 populations.

Figure 5 plots the relative bias in UQTE estimates at every tenth quantile of the 1979 earnings distribution. Much like the average training estimates, we see that the weighted estimates consistently lie below the unweighted estimates for most quantiles, irrespective of whether we use the PSID-1 or PSID-2 nonexperimental group.

9 Conclusion

In empirical research, the problems of nonrandom assignment and missing outcomes threaten identification of treatment effects. This article proposes a class of M-estimators that double weight by the propensity score and missing outcome probability to correct for the two problems within a selection-on-observables framework. The asymptotic theory of the proposed estimator is characterized in two halves where the first half allows misspecification in some conditional feature of interest and the second allows for misspecified weights. Together, the two parts completely characterize the kinds of misspecification scenarios permissible under the given framework.

As illustrative examples, the article utilizes results from the first and second half to discuss estimation of causal parameters like the average and QTEs. In the case of ATE, the proposed estimator is shown to be DR irrespective of whether the mean function or the weights are misspecified (not both). For the case of QTEs, one may either obtain the CQTE if the conditional quantile function is assumed correct or a linear approximation to it. This is demonstrated in the simulations where we find that the double-weighted ATE and QTE estimates have the lowest bias when compared to naive alternatives (unweighted and ps-weighted estimators). Finally, an application of the procedure to Calónico and Smith's (2017) reconstructed NSW sample helps to quantify the degree of distortion created by the two problems on the returns to training estimates through a comparsion with the experimental benchmark.

Even though missing outcomes are a common concern in empirical analysis, it is equally common to encounter missing data on the covariates. A particularly important future extension can be to allow for missing data on both. In this case, using a generalized method of moments framework, which incorporates information on complete and incomplete cases, could provide efficiency gains over just using the observed data. A different possibility would be to relax the identifying restrictions to allow for selection on unobservables and possibly explore estimation of local ATE.

Acknowledgements: I am grateful to Jeffrey M. Wooldridge, Steven Haider, Ben Zou, and Kenneth Frank. Special thanks to Tim Vogelsang, Wendun Wang, Alyssa Carlson, Christian Cox, Tymon Słoczyński, and seminar & conference participants, for insightful comments and suggestions on this article.

Funding information: No funding to declare.

Conflict of interest: The author states that there is no conflict of interest.

Ethical approval: The conducted research is not related to either human or animals use.

Data availability statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on request.

References

[1] Ding P, Li F. Causal inference: a missing data perspective. Stat Sci. 2018;33(2):214–37.

[2] Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika. 2009;96(3):723–34.

[3] Vansteelandt S, Carpenter J, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology. 2010.

[4] Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. J Amer Stat Assoc. 1995;90(429):122–9.

[5] Chen X, Hong H, Tarozzi A. Semiparametric efficiency in GMM models with auxiliary data. Ann Stat. 2008;36(2):808–43.

[6] Graham BS, de Xavier Pinto CC, Egel D. Inverse probability tilting for moment condition models with missing data. Rev Econom Stud. 2012;79(3):1053–79.

[7] Wooldridge JM. Inverse probability weighted estimation for general missing data problems. J Econom. 2007;141(2):1281–301.

[8] Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Meth Med Res. 2013;22(3):278–95.

[9] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[10] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66:315–31.
[11] Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services Outcomes Res Methodology. 2001;2(3–4):259–78.
[12] Firpo S. Efficient semiparametric estimation of quantile treatment effects. Econometrica. 2007;75(1):259–76.
[13] Słoczyński T, Wooldridge JM. A general double robustness result for estimating average treatment effects. Econom Theory. 2018;34(1):112–33.
[14] Huber M. Identification of average treatment effects in social experiments under alternative forms of attrition. J Educat Behav Stat. 2012;37(3):443–74.
[15] Huber M. Treatment evaluation in the presence of sample selection. Econom Rev. 2014;33(8):869–905.
[16] Frölich M, Huber M. Treatment evaluation with multiple outcome periods under endogeneity and attrition. J Amer Stat Assoc. 2014;109(508):1697–711.
[17] Fricke H, Frölich M, Huber M, Lechner M. Endogeneity and non-response bias in treatment evaluation-nonparametric identification of causal effects by instruments. J Appl Econ. 2020;35(5):481–504.
[18] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Amer Stat Assoc. 1994;89(427):846–66.
[19] Scharfstein D, Rotnitzky A, Robins J. Comments and rejoinder. J Amer Stat Assoc. 1999;94(448):1121–46.
[20] Robins JM, Rotnitzky A, van der Laan M. On profile likelihood: comment. J Amer Stat Assoc. 2000;95(450):477–82.
[21] Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22(4):523–39.
[22] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–73.
[23] Bia M, Huber M, Lafférs L. Double machine learning for sample selection models. 2020. arXiv: http://arXiv.org/abs/arXiv:201200745.
[24] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018;21(1):C1–C68.
[25] Henmi M, Eguchi S. A paradox concerning nuisance parameters and projected estimating functions. Biometrika. 2004;91(4):929–41.
[26] Hitomi K, Nishiyama Y, Okui R. A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. Econom Theory. 2008;24(6):1717–28.
[27] Prokhorov A, Schmidt P. GMM redundancy results for general missing data problems. J Econom. 2009;151(1):47–55.
[28] Lok JJ. How estimating nuisance parameters can reduce the variance (with consistent variance estimation). 2021. arXiv: http://arXiv.org/abs/arXiv:210902690.
[29] Su F, Mou W, Ding P, Wainwright M. When is the estimated propensity score better? High-dimensional analysis and bias correction. 2023. arXiv: http://arXiv.org/abs/arXiv:230317102.
[30] Calónico S, Smith J. The women of the National Supported Work demonstration. J Labor Econom. 2017;35(S1):S65–97.
[31] White H. Maximum likelihood estimation of misspecified models. Econometr J Econometric Soc. 1982;50:1–25.
[32] Angrist J, Chernozhukov V, Fernández-Val I. Quantile regression under misspecification, with an application to the U.S. wage structure. Econometrica. 2006;74(2):539–63.
[33] Lalonde RJ. Evaluating the econometric evaluations of training programs with experimental data. Amer Econom Rev. 1986;76:604–20.
[34] Hotz VJ, Imbens GW, Klerman JA. Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program. J Labor Econom. 2006;24(3):521–66.
[35] Chetty R, Friedman JN, Rockoff JE. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. Amer Econom Rev. 2014;104(9):2593–632.
[36] Kane TJ, Staiger DO. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607. *National Bureau of Economic Research* (2008).
[37] Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. J Amer Stat Assoc. 2008;103(484):1334–44.
[38] Little RJ, Rubin DB. Statistical analysis with missing data. vol. 793. Hoboken, New Jersey: John Wiley & Sons; 2019.
[39] Newey WK, McFadden D. Large sample estimation and hypothesis testing. Handbook Econometrics. Vol. 4. North Holland, Amsterdam: Elsevier; 1994. pp. 2111–245.
[40] Komunjer I. Quasi-maximum likelihood estimation for conditional quantiles. J Econometrics. 2005;128(1):137–64.
[41] Negi A, Wooldridge JM. Revisiting regression adjustment in experiments with heterogeneous treatment effects. Econometric Rev. 2021;40(5):504–34.
[42] Ding P. A first course in causal inference. 2023. arXiv: http://arXiv.org/abs/arXiv:230518793.
[43] Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. J Econom Literature. 2009;47(1):5–86.
[44] Koenker R, Bassett G. Regression quantiles. Econometrica. 1978;46(1):33–50.
[45] Heckman J, Ichimura H, Smith J, Todd P. Characterizing selection bias using experimental data. Econometrica. 1998;66(5):1017–98.

Appendix

A Regularity conditions for asymptotic theory

Let the population problem be denoted as, $Q_0(\boldsymbol{\theta}_g) \equiv \mathbb{E}[\omega_g \cdot q(Y, \mathbf{X}, \boldsymbol{\theta}_g)]$, and its sample analog be given as, $Q_N(\boldsymbol{\theta}_g) \equiv \frac{1}{N\hat{\rho}_g}\sum_{i=1}^{N}\widehat{\omega}_{ig} \cdot q(Y_i, \mathbf{X}_i, \boldsymbol{\theta}_g)$, where $\hat{\rho}_g = N_g/N$ and $N\hat{\rho}_g \to \infty$ as $\hat{\rho}_g \to \rho_g$.

(1) $\boldsymbol{\Theta}_g$ is compact for $g = 0, 1$.

(2) $G(\mathbf{X}, \boldsymbol{\gamma})$ and $R(\mathbf{X}, W, \boldsymbol{\delta})$ satisfy Assumption 5 and is continuous for each $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ on the support of $\mathbf{X}$ and $(\mathbf{X}, W)$, respectively.

(3) $q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)$ is continuous at each $\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g$ with probability one.

(4) $\mathbb{E}\left[\sup_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g} |q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)|\right] < \infty$.

(5) $\boldsymbol{\theta}_g^{\mathbf{0}} \in \operatorname{int}(\boldsymbol{\Theta}_g)$.

(6) $q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)$ is continuously differentiable on $\operatorname{int}(\boldsymbol{\Theta}_g)$ with probability one.

(7) $\frac{1}{N}\sum_{i=1}^{N}\widehat{\omega}_{ig} \cdot \mathbf{h}(Y_i(g), \mathbf{X}_i, \hat{\boldsymbol{\theta}}_g) = o_{\mathbb{P}}(N^{-1/2})$.

(8) $\mathbb{E}[\sup_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g} ||\mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)||^2] < \infty$.

(9) $G(\cdot, \boldsymbol{\gamma})$ and $R(\cdot, \boldsymbol{\delta})$ are both twice continuously differentiable on $\operatorname{int}(\Gamma)$ and $\operatorname{int}(\Delta)$, respectively.

(10) $\mathbb{E}\left[\sup_{\boldsymbol{\delta} \in \Delta} ||\mathbf{b}(\mathbf{X}, W, S, \boldsymbol{\delta})||^2\right] < \infty$, $\mathbb{E}\left[\sup_{\boldsymbol{\gamma} \in \Gamma} ||\mathbf{d}(\mathbf{X}, W, \boldsymbol{\gamma})||^2\right] < \infty$.

(11) $\mathbb{E}[\omega_g \cdot \mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)]$ is continuously differentiable on $\operatorname{int}(\boldsymbol{\Theta}_g)$.

(12) $\mathbf{H}_g \equiv \nabla_{\boldsymbol{\theta}_g}\mathbb{E}[\omega_g \cdot \mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})]$ is non-singular.

(13) $\{\boldsymbol{v}_N(\boldsymbol{\theta}_g) : N \geq 1\}$ is stochastically equicontinuous, where

$$\boldsymbol{v}_N(\boldsymbol{\theta}_g) \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{\widehat{\omega}_{ig}\mathbf{h}_{ig}(\boldsymbol{\theta}_g) - \mathbb{E}[\widehat{\omega}_{ig}\mathbf{h}_{ig}(\boldsymbol{\theta}_g)]\}. \tag{A1}$$

A.1 Consistency of the doubly weighted estimator

Given the two-step nature of the estimation problem, wherein the first step uses binary response MLE for estimating the probability weights and the second step solves an objective function using the first-step weights, the asymptotic theory utilizes results for two-step estimators with a non-smooth objective function to establish the large sample properties of $\hat{\boldsymbol{\theta}}_g$. The following theorem fills in the primitive regularity conditions for applying the uniform law of large numbers.

Theorem A.1. (Consistency under weak identification) *Under Assumptions* 1–4 *and conditions* (1)–(4), $\hat{\boldsymbol{\theta}}_g \xrightarrow{p} \boldsymbol{\theta}_g^{\mathbf{0}}$ *for each* $g = 0, 1$.

The proof follows from verifying the conditions in Lemma 2.4 of the study by Newey and McFadden [39].

Proof. It has already been established that

$$\mathbb{E}[\omega_g \cdot q(Y, \mathbf{X}, \boldsymbol{\theta})] \equiv \mathbb{E}[\omega_g \cdot q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)] = \mathbb{E}[q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)]$$

for both $g = 0, 1$. By (iii), $\omega_g(\boldsymbol{\gamma}, \boldsymbol{\delta})$ is continuous in $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ and is bounded in absolute value by Assumption 5. Moreover, $\omega_g(\cdot, \boldsymbol{\gamma}, \boldsymbol{\delta})q(\cdot, \boldsymbol{\theta})$ is continuous with probability one. Then, along with (v), dominated convergence theorem (DCT), and boundedness of $\omega_g(\cdot, \cdot)$, we obtain

$$\sup_{(\boldsymbol{\theta}_g,\boldsymbol{\gamma},\boldsymbol{\delta})\in(\boldsymbol{\Theta}_g,\tilde{\boldsymbol{\Gamma}},\tilde{\boldsymbol{\Delta}})} \left| \frac{1}{N}\sum_{i=1}^{N} \omega_{ig}(\boldsymbol{\gamma},\boldsymbol{\delta})\cdot q(Y_i(g),\mathbf{X}_i,\boldsymbol{\theta}_g) - \mathbb{E}[\omega_g(\boldsymbol{\gamma},\boldsymbol{\delta})\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)]\right| \xrightarrow{p} 0 \tag{A2}$$

using Lemma 2.4 in the study by Newey and McFadden [39] since $\tilde{\boldsymbol{\Gamma}}$ and $\tilde{\boldsymbol{\Delta}}$ are compact neighborhoods around $\boldsymbol{\gamma}_0$ and $\boldsymbol{\delta}_0$. By triangle inequality,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g}\left|\frac{1}{N}\sum_{i=1}^{N}\hat{\omega}_{ig}\cdot q(Y_i(g),\mathbf{X}_i,\boldsymbol{\theta}_g) - \mathbb{E}[\omega_g\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)]\right|\\
&\le \sup_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g}\left|\frac{1}{N}\sum_{i=1}^{N_g}\hat{\omega}_{ig}\cdot q(Y_i(g),\mathbf{X}_i,\boldsymbol{\theta}_g) - \mathbb{E}[\hat{\omega}_g\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)]\right|\\
&\quad + \sup_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g}|\mathbb{E}[\hat{\omega}_g\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)] - \mathbb{E}[\omega_g\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)]|.
\end{aligned} \tag{A3}$$

(A.3) is $o_{\mathbb{P}}(1)$ due to $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_0$, $\hat{\boldsymbol{\delta}} \xrightarrow{p} \boldsymbol{\delta}_0$, and uniform continuity of $\mathbb{E}[\omega_g\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)]$ on $\boldsymbol{\Theta}_g\times\tilde{\boldsymbol{\Gamma}}\times\tilde{\boldsymbol{\Delta}}$. Then, consistency of $\hat{\boldsymbol{\theta}}_g$ for $\boldsymbol{\theta}_g^{\mathbf{0}}$ follows from Theorem 2.1 of the study by Newey and McFadden [39]. □

Theorem A.2. (Consistency under strong identification) *Under Assumptions* 2–4, 6, *and* 7, *and regularity conditions* (1)–(4), $\hat{\boldsymbol{\theta}}_g \xrightarrow{p} \boldsymbol{\theta}_g^{\mathbf{0}}$ *as* $N\to\infty$.

Proof. We first establish that $\boldsymbol{\theta}_g^{\mathbf{0}}$ solves

$$\mathbb{E}[\omega_g^*\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)].$$

The proof of uniform convergence follows similar to the proof of Theorem A.1 where we replace ω_g by ω_g^*. Then, consistency of $\hat{\boldsymbol{\theta}}_g$ for $\boldsymbol{\theta}_g^{\mathbf{0}}$ follows from Theorem 2.1 in [39]. To show that $\boldsymbol{\theta}_g^{\mathbf{0}}$ is still a solution to the double-weighted population problem with misspecified weights, consider the following argument:

$$\hat{\boldsymbol{\theta}}_g = \operatorname*{argmin}_{\boldsymbol{\theta}_g}\sum_{i=1}^{N}\hat{\omega}_{ig}\cdot q(Y_i,\mathbf{X}_i,\boldsymbol{\theta}_g) \xrightarrow{p} \operatorname*{argmin}_{\boldsymbol{\theta}_g}\mathbb{E}[\omega_g^*\cdot q(Y,\mathbf{X},\boldsymbol{\theta}_g)] \equiv \operatorname*{argmin}_{\boldsymbol{\theta}_g}\mathbb{E}[\omega_g^*\cdot q(Y(g),\mathbf{X},\boldsymbol{\theta}_g)]. \tag{A4}$$

Consider, $\mathbb{E}[\omega_1^*\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)]$. Then, using law of LIEs,

$$\begin{aligned}
\mathbb{E}[\omega_1^*\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)] &= \mathbb{E}[\mathbb{E}(\omega_1^*\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)|\mathbf{X},W,Y(1))]\\
&= \mathbb{E}\left[\frac{1}{R(\mathbf{X},W,\boldsymbol{\delta}^*)}\cdot\frac{W}{G(\mathbf{X},\boldsymbol{\gamma}^*)}\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)\cdot\mathbb{P}(S=1|\mathbf{X},W,Y(1))\right]\\
&= \mathbb{E}\left[\frac{r(\mathbf{X},W)}{R(\mathbf{X},W,\boldsymbol{\delta}^*)}\cdot\frac{W}{G(\mathbf{X},\boldsymbol{\gamma}^*)}\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)\right],
\end{aligned}$$

where the second equality applies the inner expectation to S and the third equality uses the fact that $\mathbb{P}(S=1|\mathbf{X},W,Y(1)) = \mathbb{P}(S=1|\mathbf{X},W)$ because of MAR. Applying LIE again,

$$\begin{aligned}
&\mathbb{E}\left[\frac{r(\mathbf{X},W)}{R(\mathbf{X},W,\boldsymbol{\delta}^*)}\cdot\frac{W}{G(\mathbf{X},\boldsymbol{\gamma}^*)}\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)\right]\\
&= \mathbb{E}\left[\mathbb{E}\left(\frac{r(\mathbf{X},W)}{R(\mathbf{X},W,\boldsymbol{\delta}^*)}\cdot\frac{W}{G(\mathbf{X},\boldsymbol{\gamma}^*)}\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)|\mathbf{X},Y(1)\right)\right]\\
&= \mathbb{E}\left[\frac{1}{G(\mathbf{X},\boldsymbol{\gamma}^*)}\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)\cdot\mathbb{E}\left\{\frac{r(\mathbf{X},W)}{R(\mathbf{X},W,\boldsymbol{\delta}^*)}\cdot W|\mathbf{X}\right\}\right]\\
&= \mathbb{E}[\xi_1(\mathbf{X})\cdot q(Y(1),\mathbf{X},\boldsymbol{\theta}_1)].
\end{aligned} \tag{A5}$$

Here, the second equality uses unconfoundedness and the third equality recognizes that $\frac{1}{G(\mathbf{X},\boldsymbol{\gamma}^*)}\cdot\mathbb{E}\{\cdot|\mathbf{X}\}$ is a function of $\mathbf{X}$, which I denote by $\xi_1(\mathbf{X})$. One can show $\mathbb{E}[\omega_0^*\cdot q(Y,\mathbf{X},\boldsymbol{\theta}_0)] = \mathbb{E}[\xi_0(\mathbf{X})\cdot q(Y(0),\mathbf{X},\boldsymbol{\theta}_0)]$ analogously. Then,

$$\underset{\boldsymbol{\theta}_g}{\operatorname{argmin}}\mathbb{E}[\omega_g^* \cdot q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)] = \underset{\boldsymbol{\theta}_g}{\operatorname{argmin}}\mathbb{E}[\xi_g(\mathbf{X})\cdot q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)] = \underset{\boldsymbol{\theta}_g}{\operatorname{argmin}}\mathbb{E}[\xi_g(\mathbf{X})\cdot \mathbb{E}(q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g)|\mathbf{X})], \tag{A6}$$

where the second equality holds due to LIE. □

B Proofs

Proof of Lemma 1. Let us first consider the argument for $\boldsymbol{\theta}_1^{\mathbf{0}}$. By LIE and using the fact that $q(Y, \mathbf{X}, \boldsymbol{\theta}_g) = W \cdot q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1) + (1 - W)\cdot q(Y(0), \mathbf{X}, \boldsymbol{\theta}_0)$, we can write

$$\begin{aligned}\mathbb{E}[\omega_1 \cdot q(Y, \mathbf{X}, \boldsymbol{\theta}_1)] &= \mathbb{E}\left[\mathbb{E}\left(\frac{S}{r(\mathbf{X}, W)} \cdot \frac{W}{p(\mathbf{X})} \cdot q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1)|Y(1), \mathbf{X}, W\right)\right]\\ &= \mathbb{E}\left[\frac{W}{r(\mathbf{X}, W)\cdot p(\mathbf{X})} \cdot q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1)\cdot \mathbb{P}(S = 1|Y(1), \mathbf{X}, W)\right]\\ &= \mathbb{E}\left[\frac{W}{r(\mathbf{X}, W)\cdot p(\mathbf{X})} \cdot q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1)\cdot \mathbb{P}(S = 1|\mathbf{X}, W)\right]\\ &= \mathbb{E}\left[\frac{W}{p(\mathbf{X})} \cdot q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1)\right].\end{aligned}$$

Using another application of LIE along with unconfoundedness, we obtain

$$\mathbb{E}\left[\frac{W}{p(\mathbf{X})} \cdot q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1)\right] = \mathbb{E}[q(Y(1), \mathbf{X}, \boldsymbol{\theta}_1)],$$

where the third equality follows from MAR and fourth follows from part (ii) of Assumption 3. The proof for $\boldsymbol{\theta}_0^{\mathbf{0}}$ follows analogously. □

Given the two-step nature of the estimation problem, wherein the first step uses binary response MLE for estimating the probability weights and the second step solves an objective function using the first-step weights, the asymptotic theory utilizes results for two-step estimators with a non-smooth objective function to establish the large sample properties of $\hat{\boldsymbol{\theta}}_g$. The following theorem fills in the primitive regularity conditions for applying the uniform law of large numbers.

Proof of Theorem 2. Explicit dependence on data is suppressed for notational simplicity. Then, expanding $\widehat{\omega}_{ig}$ around ω_{ig},

$$\begin{aligned}N^{-1/2}\sum_{i=1}^{N}\widehat{\omega}_{ig} \cdot \mathbf{h}_{ig} &= N^{-1/2}\sum_{i=1}^{N}\{\omega_{ig}\mathbf{h}_{ig} - \widetilde{\omega}_{ig}\mathbf{h}_{ig} \cdot \mathbf{b}_i'(\tilde{\boldsymbol{\delta}})\cdot(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_{\mathbf{0}}) - \widetilde{\omega}_{ig}\mathbf{h}_{ig} \cdot \mathbf{d}_i'(\tilde{\boldsymbol{\gamma}})\cdot(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\mathbf{0}})\} + o_{\mathbb{P}}(1)\\ &= N^{-1/2}\sum_{i=1}^{N}\omega_{ig}\mathbf{h}_{ig} - N^{-1}\sum_{i=1}^{N}\widetilde{\omega}_{ig}\mathbf{h}_{ig}\mathbf{b}_i'(\tilde{\boldsymbol{\delta}})\cdot \sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_{\mathbf{0}}) - N^{-1}\sum_{i=1}^{N}\widetilde{\omega}_{ig}\mathbf{h}_{ig}\mathbf{d}_i'(\tilde{\boldsymbol{\gamma}})\cdot \sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\mathbf{0}}) + o_{\mathbb{P}}(1),\end{aligned}$$

where $\tilde{\boldsymbol{\delta}}$ lies between $\hat{\boldsymbol{\delta}}$ and $\boldsymbol{\delta}_{\mathbf{0}}$ and $\tilde{\boldsymbol{\gamma}}$ lies between $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}_{\mathbf{0}}$. Now let, $(\boldsymbol{\theta}_g^*, \boldsymbol{\delta}^*) = \arg\sup_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g, \boldsymbol{\delta} \in \Delta}||\mathbf{h}(\boldsymbol{\theta}_g)\cdot \mathbf{b}'(\boldsymbol{\delta})||$. Then,

$$(\mathbb{E}[||\mathbf{h}(\boldsymbol{\theta}_g^*)\mathbf{b}'(\boldsymbol{\delta}^*)||])^2 \le \mathbb{E}[||\mathbf{h}(\boldsymbol{\theta}_g^*)||^2]\mathbb{E}[||\mathbf{b}'(\boldsymbol{\delta}^*)||^2] \le \mathbb{E}\left[\sup_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g}||\mathbf{h}(\boldsymbol{\theta}_g)||^2\right]\mathbb{E}\left[\sup_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g}||\mathbf{b}'(\boldsymbol{\delta})||^2\right] < \infty, \tag{A7}$$

where the first inequality holds by Cauchy–Schwartz, the second inequality holds due to the definition of supremums, and the third inequality holds by conditions (iv) and (vi). Then,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g,\boldsymbol{\delta}\in\boldsymbol{\Delta}}||\mathbf{h}(\boldsymbol{\theta}_g)\mathbf{b}'(\boldsymbol{\delta})||\right] \le \left(\mathbb{E}\left[\sup_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g,\boldsymbol{\delta}\in\boldsymbol{\Delta}}||\mathbf{h}(\boldsymbol{\theta}_g)\mathbf{b}'(\boldsymbol{\delta})||\right]\right)^2 < \infty,$$

where the first inequality holds trivially and the second inequality holds because of (A7). An analogous argument may be made for showing $\mathbb{E}[\sup_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g,\boldsymbol{\gamma}\in\boldsymbol{\Gamma}}||\mathbf{h}(\boldsymbol{\theta}_g)\mathbf{d}'(\boldsymbol{\gamma})||] < \infty$. Using the fact that $\omega_g(\boldsymbol{\gamma}, \boldsymbol{\delta})$ is continuous and bounded along with continuity of $\mathbf{l}(\boldsymbol{\theta}_g)$ (condition (ii)), $\mathbf{b}(\boldsymbol{\delta})$, $\mathbf{d}(\boldsymbol{\gamma})$ (condition (iii) of Theorem A.1), we obtain

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\widetilde{\omega}_{ig}\mathbf{h}_{ig}\mathbf{b}_i'(\tilde{\boldsymbol{\delta}}) &= \mathbb{E}[\omega_{ig}\mathbf{h}_{ig}\mathbf{b}_i'] + o_{\mathbb{P}}(1),\\ \frac{1}{N}\sum_{i=1}^{N}\widetilde{\omega}_{ig}\mathbf{h}_{ig}\mathbf{d}_i'(\tilde{\boldsymbol{\gamma}}) &= \mathbb{E}[\omega_{ig}\mathbf{h}_{ig}\mathbf{d}_i'] + o_{\mathbb{P}}(1)\end{aligned} \tag{A8}$$

using Lemma 4.3 in [39] as $\tilde{\boldsymbol{\gamma}} \to_p \boldsymbol{\gamma}_0$ and $\tilde{\boldsymbol{\delta}} \to_p \boldsymbol{\delta}_0$. Rewriting equation (7) using influence function representations for $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$ along with (A8), we obtain

$$\begin{aligned}N^{-1/2}\sum_{i=1}^{N}\widehat{\omega}_{ig}\mathbf{h}_{ig} &= N^{-1/2}\sum_{i=1}^{N}\{\mathbf{l}_{ig} - \mathbb{E}[\mathbf{l}_{ig}\mathbf{b}_i']\cdot \mathbb{E}[\mathbf{b}_i\mathbf{b}_i']^{-1}\mathbf{b}_i - \mathbb{E}[\mathbf{l}_{ig}\mathbf{d}_i']\cdot \mathbb{E}[\mathbf{d}_i\mathbf{d}_i']^{-1}\mathbf{d}_i\} + o_{\mathbb{P}}(1)\\ &\equiv N^{-1/2}\sum_{i=1}^{N}\mathbf{u}_{ig} + o_{\mathbb{P}}(1)\\ &\xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_g),\end{aligned} \tag{A9}$$

where $\mathbf{u}_{ig} \equiv \mathbf{l}_{ig} - \mathbb{E}[\mathbf{l}_{ig}\mathbf{b}_i']\cdot \mathbb{E}[\mathbf{b}_i\mathbf{b}_i']^{-1}\mathbf{b}_i - \mathbb{E}[\mathbf{l}_{ig}\mathbf{d}_i']\cdot \mathbb{E}[\mathbf{d}_i\mathbf{d}_i']^{-1}\mathbf{d}_i$. Since $\mathbb{E}(\mathbf{u}_{ig}) = \mathbf{0}$,

$$\boldsymbol{\Omega}_g = \mathbb{E}(\mathbf{l}_{ig}\mathbf{l}_{ig}') - \mathbb{E}(\mathbf{l}_{ig}\mathbf{b}_i')\mathbb{E}(\mathbf{b}_i\mathbf{b}_i')^{-1}\mathbb{E}(\mathbf{b}_i\mathbf{l}_{ig}') - \mathbb{E}(\mathbf{l}_{ig}\mathbf{d}_i')\mathbb{E}(\mathbf{d}_i\mathbf{d}_i')^{-1}\mathbb{E}(\mathbf{d}_i\mathbf{l}_{ig}').$$

The next part of the proof uses the theory of empirical processes for obtaining asymptotic normality of the doubly weighted estimator. Using the definition in (A1) along with the fact that $\mathbb{E}[\widehat{\omega}_{ig}\mathbf{h}_i(\boldsymbol{\theta}_g)] \xrightarrow{p} \mathbb{E}[\omega_{ig}\mathbf{h}_i(\boldsymbol{\theta}_g)]$ (by continuity of $\omega(\boldsymbol{\gamma}, \boldsymbol{\delta})\mathbf{h}(\boldsymbol{\theta}_g)$, condition iv) and DCT as $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}) \xrightarrow{p} (\boldsymbol{\gamma}_0, \boldsymbol{\delta}_0)$, rewrite

$$\boldsymbol{v}_N(\boldsymbol{\theta}_g) = \boldsymbol{v}_N^*(\boldsymbol{\theta}_g) + o_{\mathbb{P}}(1), \tag{A10}$$

where $\boldsymbol{v}_N^*(\boldsymbol{\theta}_g) \equiv \frac{1}{N}\sum_{i=1}^{N}\{\widehat{\omega}_{ig}\mathbf{h}_i(\boldsymbol{\theta}_g) - \mathbb{E}[\omega_{ig}\mathbf{h}_i(\boldsymbol{\theta}_g)]\}$. Let

$$\bar{\boldsymbol{m}}_N(\boldsymbol{\theta}_g) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\omega}_{ig}\mathbf{h}_i(\boldsymbol{\theta}_g) \quad \text{and} \quad \boldsymbol{m}_N^*(\boldsymbol{\theta}_g) = \mathbb{E}[\omega_{ig}\mathbf{h}_i(\boldsymbol{\theta}_g)].$$

Then, performing element by element mean value expansions of $\boldsymbol{m}_N^*(\hat{\boldsymbol{\theta}}_g)$ around $\boldsymbol{\theta}_g^{\mathbf{0}}$, we obtain

$$\mathbf{0} = \sqrt{N}\boldsymbol{m}_N^*(\boldsymbol{\theta}_g^{\mathbf{0}}) = \sqrt{N}\boldsymbol{m}_N^*(\hat{\boldsymbol{\theta}}_g) - \nabla_{\boldsymbol{\theta}_g}\boldsymbol{m}_N^*(\tilde{\boldsymbol{\theta}}_g)'\cdot \sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}}),$$

where $\tilde{\boldsymbol{\theta}}_g$ lies between $\hat{\boldsymbol{\theta}}_g$ and $\boldsymbol{\theta}_g^{\mathbf{0}}$. Since the population first-order condition is zero at the truth,

$$\mathbf{0} = \nabla_{\boldsymbol{\theta}_g}\mathbb{E}[\omega_g \cdot q(Y(g), \mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})] = \mathbb{E}[\omega_g \cdot \mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})] \equiv \boldsymbol{m}_N^*(\boldsymbol{\theta}_g^{\mathbf{0}}).$$

The second equality follows from dominance condition (iv) and application of Lemma 3.6 in the study by Newey and McFadden [39]. Then, by the continuity of $\nabla_{\boldsymbol{\theta}_g}\mathbb{E}[\omega_{ig}\mathbf{h}_i(\boldsymbol{\theta}_g)]$ (condition (vi)),

$$\nabla_{\boldsymbol{\theta}_g}\boldsymbol{m}_N^*(\tilde{\boldsymbol{\theta}}_g) \xrightarrow{p} \mathbf{H}_g.$$

By continuous mapping theorem and condition (viii),

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}}) = (\mathbf{H}_g^{-1} + o_{\mathbb{P}}(1))\cdot \sqrt{N}\boldsymbol{m}_N^*(\hat{\boldsymbol{\theta}}_g). \tag{A11}$$

Consider,

$$\begin{aligned}-\sqrt{N}\boldsymbol{m}_N^*(\hat{\boldsymbol{\theta}}_g) &= \boldsymbol{v}_N^*(\hat{\boldsymbol{\theta}}_g) - \sqrt{N}\bar{\boldsymbol{m}}_N(\hat{\boldsymbol{\theta}}_g)\\ &= \boldsymbol{v}_N^*(\hat{\boldsymbol{\theta}}_g) - \boldsymbol{v}_N^*(\boldsymbol{\theta}_g^{\mathbf{0}}) + \boldsymbol{v}_N^*(\boldsymbol{\theta}_g^{\mathbf{0}}) - \sqrt{N}\bar{\boldsymbol{m}}_N(\hat{\boldsymbol{\theta}}_g)\\ &= \boldsymbol{v}_N^*(\boldsymbol{\theta}_g^{\mathbf{0}}) + o_{\mathbb{P}}(1),\end{aligned}$$

where $\boldsymbol{v}_N^*(\hat{\boldsymbol{\theta}}_g) - \boldsymbol{v}_N^*(\boldsymbol{\theta}_g^{\mathbf{0}}) = o_{\mathbb{P}}(1)$ by asymptotic equivalence in (A10) and stochastic equicontinuity by condition (ix). Moreover, $\sqrt{N}\bar{\boldsymbol{m}}_N(\hat{\boldsymbol{\theta}}_g) = o_{\mathbb{P}}(1)$ by condition (iii). Therefore,

$$\boldsymbol{v}_N^*(\boldsymbol{\theta}_g^{\mathbf{0}}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\omega}_{ig}\mathbf{h}_{ig} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_g)$$

by (A9). Then, using (A11) along with slutsky's theorem, $\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}}) \xrightarrow{d} N(\boldsymbol{0}, \mathbf{H}_g^{-1}\boldsymbol{\Omega}_g\mathbf{H}_g^{-1})$. □

Proof of Corollary 1. Consider,

$$\begin{aligned}\boldsymbol{\Sigma}_g - \boldsymbol{\Omega}_g &= \mathbb{E}(\mathbf{l}_{ig}\mathbf{l}_{ig}') - \{\mathbb{E}(\mathbf{l}_{ig}\mathbf{l}_{ig}') - \mathbb{E}(\mathbf{l}_{ig}\mathbf{b}_i')\mathbb{E}(\mathbf{b}_i\mathbf{b}_i')^{-1}\mathbb{E}(\mathbf{b}_i\mathbf{l}_{ig}') - \mathbb{E}(\mathbf{l}_{ig}\mathbf{d}_i')\mathbb{E}(\mathbf{d}_i\mathbf{d}_i')^{-1}\mathbb{E}(\mathbf{d}_i\mathbf{l}_{ig}')\}\\ &= \mathbb{E}(\mathbf{l}_{ig}\mathbf{b}_i')\mathbb{E}(\mathbf{b}_i\mathbf{b}_i')^{-1}\mathbb{E}(\mathbf{b}_i\mathbf{l}_{ig}') + \mathbb{E}(\mathbf{l}_{ig}\mathbf{d}_i')\mathbb{E}(\mathbf{d}_i\mathbf{d}_i')^{-1}\mathbb{E}(\mathbf{d}_i\mathbf{l}_{ig}').\end{aligned}$$

Since each component matrix in the above expression is positive semi-definite, therefore the sum of the two matrices is also positive semi-definite. □

Proof of Theorem 4. The proof follows in the manner of Theorem 1 where we replace ω_g by ω_g^*. Also, $\boldsymbol{\Omega}_g$ now denotes the variance of the score of the objective function, $\mathbf{l}_{ig}$, without the first stage adjustment for the estimated weights. This is because, $\mathbb{E}(\mathbf{l}_{ig}\mathbf{b}_i') = \mathbb{E}(\mathbf{l}_{ig}\mathbf{d}_i') = \mathbf{0}$ because the conditional score of $\mathbf{l}_{ig}$ is zero under strong identification of $\boldsymbol{\theta}_g^{\mathbf{0}}$ i.e. $\mathbb{E}[\mathbf{h}(Y(g), \mathbf{X}, \boldsymbol{\theta}_g^{\mathbf{0}})|\mathbf{X}] = \mathbf{0}$. □

Proof of Corollary 2. The proof follows from Theorem 2, Note that the asymptotic variance of the estimator that uses known weights is

$$\text{Avar}[\sqrt{N}(\tilde{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^{\mathbf{0}})] = \mathbf{H}_g^{-1}\boldsymbol{\Omega}_g\mathbf{H}_g^{-1},$$

where $\boldsymbol{\Omega}_g = \mathbb{E}(\mathbf{l}_{ig}\mathbf{l}_{ig}')$. The result follows immediately. □

Proof of Corollary 3 (Efficiency gain with unweighted estimator under GCIME). Using two applications of LIE and invoking MAR and unconfoundedness, I can rewrite

$$\mathbb{E}\left[\frac{S_i\cdot W_i}{R(\mathbf{X}_i, W_i, \boldsymbol{\delta}^*)\cdot G(\mathbf{X}_i, \boldsymbol{\gamma}^*)}\cdot q(Y_i(1), \mathbf{X}_i, \boldsymbol{\theta}_1^{\mathbf{0}})\right] = \mathbb{E}\left[\frac{r(\mathbf{X}_i, 1)}{R(\mathbf{X}_i, 1, \boldsymbol{\delta}^*)}\cdot\frac{p(\mathbf{X}_i)}{G(\mathbf{X}_i, \boldsymbol{\gamma}^*)}\cdot q(Y_i(1), \mathbf{X}_i, \boldsymbol{\theta}_1^{\mathbf{0}})\right].$$

Using another application of LIE, I can rewrite the above as follows:

$$\mathbb{E}\left[\frac{r(\mathbf{X}_i, 1)}{R(\mathbf{X}_i, 1, \boldsymbol{\delta}^*)}\cdot\frac{p(\mathbf{X}_i)}{G(\mathbf{X}_i, \boldsymbol{\gamma}^*)}\cdot\mathbb{E}\{q(Y_i(1), \mathbf{X}_i, \boldsymbol{\theta}_1^{\mathbf{0}})|\mathbf{X}_i\}\right].$$

Then,

$$\begin{aligned}\mathbf{H}_1 &= \mathbb{E}\left[\frac{r(\mathbf{X}_i, 1)}{R(\mathbf{X}_i, 1, \boldsymbol{\delta}^*)}\cdot\frac{p(\mathbf{X}_i)}{G(\mathbf{X}_i, \boldsymbol{\gamma}^*)}\cdot\nabla_{\boldsymbol{\theta}_1}\mathbb{E}\{\mathbf{h}(Y_i(1), \mathbf{X}_i, \boldsymbol{\theta}_1^{\mathbf{0}})|\mathbf{X}_i\}\right]\\ &= \mathbb{E}\left[\frac{r(\mathbf{X}_i, 1)}{R(\mathbf{X}_i, 1, \boldsymbol{\delta}^*)}\cdot\frac{p(\mathbf{X}_i)}{G(\mathbf{X}_i, \boldsymbol{\gamma}^*)}\cdot\mathbf{A}(\mathbf{X}_i, \boldsymbol{\theta}_1^{\mathbf{0}})\right].\end{aligned}$$

Similarly, use LIE to express $\boldsymbol{\Omega}_1$ as

$$\begin{aligned}\boldsymbol{\Omega}_1 &= \mathbb{E}\left[\frac{r(\mathbf{X}_i,1)}{R^2(\mathbf{X}_i,1,\boldsymbol{\delta}^*)}\cdot\frac{p(\mathbf{X}_i)}{G^2(\mathbf{X}_i,\boldsymbol{\gamma}^*)}\cdot\mathbb{E}\{\mathbf{h}(Y_i(1),\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})\mathbf{h}(Y_i(1),\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})'|\mathbf{X}_i\}\right]\\ &= \sigma_{01}^2\cdot\mathbb{E}\left[\frac{r(\mathbf{X}_i,1)}{R^2(\mathbf{X}_i,1,\boldsymbol{\delta}^*)}\cdot\frac{p(\mathbf{X}_i)}{G^2(\mathbf{X}_i,\boldsymbol{\gamma}^*)}\cdot\mathbf{A}(\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})\right].\end{aligned}$$

For the unweighted estimator, the variance simplifies, and this happens precisely due to the GCIME. To see this, consider $\mathbf{H}_1^{\mathbf{u}}$. Then, using LIE, we can rewrite

$$\begin{aligned}\mathbf{H}_1^{\mathbf{u}} &= \mathbb{E}[r(\mathbf{X}_i,1)\cdot p(\mathbf{X}_i)\cdot\nabla_{\boldsymbol{\theta}_1}\mathbb{E}\{\mathbf{h}(Y_i(1),\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})|\mathbf{X}_i\}]\\ &= \mathbb{E}[r(\mathbf{X}_i,1)\cdot p(\mathbf{X}_i)\cdot\mathbf{A}(\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})],\end{aligned}$$

and similarly, we can rewrite $\boldsymbol{\Omega}_1^{\boldsymbol{u}}$ using LIE as

$$\begin{aligned}\boldsymbol{\Omega}_1^{\boldsymbol{u}} &= \mathbb{E}[r(\mathbf{X}_i,1)\cdot p(\mathbf{X}_i)\cdot\mathbb{E}\{\mathbf{h}(Y_i(1),\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})\mathbf{h}(Y_i(1),\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})'|\mathbf{X}_i\}]\\ &= \sigma_{01}^2\cdot\mathbb{E}[r(\mathbf{X}_i,1)\cdot p(\mathbf{X}_i)\cdot\mathbf{A}(\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})].\end{aligned}$$

Therefore, the asymptotic variance simplifies to simply

$$\mathrm{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_1^{\boldsymbol{u}}-\boldsymbol{\theta}_1^{\mathbf{0}})] = \sigma_{01}^2\cdot(\mathbb{E}[r(\mathbf{X}_i,1)\cdot p(\mathbf{X}_i)\cdot\mathbf{A}(\mathbf{X}_i,\boldsymbol{\theta}_1^{\mathbf{0}})])^{-1}.$$

For showing that the two variances are positive semi-definite, consider the following:

$$\begin{aligned}&[\mathrm{Avar}\{\sqrt{N}(\hat{\boldsymbol{\theta}}_1^{\boldsymbol{u}}-\boldsymbol{\theta}_1^{\mathbf{0}})\}]^{-1}-[\mathrm{Avar}\{\sqrt{N}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^{\mathbf{0}})\}]^{-1}\\ &= \frac{1}{\sigma_{01}^2}\cdot\left\{\mathbb{E}(r_{i1}\cdot p_i\cdot\mathbf{A}_i)-\mathbb{E}\left(\frac{r_{i1}\cdot p_i}{R_{i1}\cdot G_i}\cdot\mathbf{A}_i\right)\cdot\mathbb{E}\left(\frac{r_{i1}\cdot p_i}{R_{i1}^2\cdot G_i^2}\cdot\mathbf{A}_i\right)^{-1}\cdot\mathbb{E}\left(\frac{r_{i1}\cdot p_i}{R_{i1}\cdot G_i}\cdot\mathbf{A}_i\right)\right\}.\end{aligned}$$

Let $\mathbf{B}_i = r_{i1}^{1/2}\cdot p_i^{1/2}\cdot\mathbf{A}_i^{1/2}$ and $\mathbf{D}_i = (r_{i1}^{1/2}/R_{i1})\cdot(p_i^{1/2}/G_i)\cdot\mathbf{A}_i^{1/2}$, then $= \frac{1}{\sigma_{01}^2}\{\mathbb{E}(\mathbf{B}_i'\mathbf{B}_i)-\mathbb{E}(\mathbf{B}_i'\mathbf{D}_i)\cdot\mathbb{E}(\mathbf{D}_i'\mathbf{D}_i)^{-1}\cdot\mathbb{E}(\mathbf{D}_i'\mathbf{B}_i)\}$.

The quantity inside the brackets is nothing but the variance of the residuals from the population regression of $\mathbf{B}_i$ on $\mathbf{D}_i$. Hence, the difference is positive semi-definite. The results for $g = 0$ can be proven analogously. □