Doubly weighted M-estimation for nonrandom assignment and missing outcomes

Akanksha Negi†

November 23, 2020

Abstract

This paper proposes a new class of M-estimators that double weight for the twin problems of nonrandom treatment assignment and missing outcomes, both of which are common issues in the treatment effects literature. The proposed class is characterized by a ‘robustness’ property, which makes it resilient to parametric misspecification in either a conditional model of interest (for example, mean or quantile function) or the two weighting functions. As leading applications, the paper discusses estimation of two specific causal parameters; average and quantile treatment effects (ATE, QTEs), which can be expressed as functions of the doubly weighted estimator, under misspecification of the framework’s parametric components. With respect to the ATE, this paper shows that the proposed estimator is doubly robust even in the presence of missing outcomes. Finally, to demonstrate the estimator’s viability in empirical settings, it is applied to Calónico and Smith (2017)’s reconstructed sample from the National Supported Work training program.

Keywords: Unconfoundedness, Missing at random, Double weighting, M-estimation, Treatment effects

JEL Classification: C13, C18, C31

*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 earlier drafts of this paper.

†Department of Econometrics and Business Statistics, Monash University. Email: akanksha.negi@monash.edu, Webpage: www.anegi.net
1 Introduction

When interest lies in causal inference, the prevalence of missing data poses a major identification challenge. A common issue is that the outcome of interest is missing for some proportion of the sample. In this case, the complete data method that drops observations with missing outcomes is widely used. While dropping is practically convenient, it not only leads to substantial loss of information but more importantly creates a nonrandom sample for estimation. In turn, dropping can generally lead to inconsistent treatment effect estimates. This paper proposes an estimator that double weights for the twin problems of nonrandom treatment assignment and missing outcomes by using information on covariates.

Weighting has been used extensively in both the missing data [Horvitz and Thompson (1952), Robins et al. (1994), Robins and Rotnitzky (1995), Wooldridge (2007)] and treatment effect [Rosenbaum and Rubin (1983), Hahn (1998), Hirano and Imbens (2001), Firpo (2007), Słoczyński and Wooldridge (2018)] literatures. However, a weighting approach that corrects for general missingness in the outcome to estimate treatment effects using observational data is yet to be proposed. Previous studies have considered weighting to deal with specific missing data issues such as attrition and non-response in the presence of endogenous treatment selection [Frölich and Huber (2014), Huber (2014), Fricke et al. (2020)]. Typically, the identification argument in these papers is based on one or more instruments with discussion centered around estimation of average treatment effects.

This paper introduces inverse probability weighting alongside propensity score (PS) weighting in a general M-estimation framework to address two prevalent problems in the causal inference literature. Moreover, the objective function being solved is permitted to be non-smooth in the underlying parameters thereby covering both average and quantile treatment effects. A key feature of the proposed estimator is its robustness to parametric misspecification in either a conditional model of interest (such as mean or quantile) or the two weighting functions. In addition, the ATE estimator which uses the proposed strategy is shown to be ‘doubly robust’ [Słoczyński and Wooldridge (2018)] even in the presence of missing outcomes.

The key identifying assumptions for consistency of the doubly weighted estimator of a population level parameter are unconfoundedness\(^1\) and missing at random. Put differently, the two restrictions imply that the treatment assignment and missing outcomes mechanisms are as good as randomly assigned after conditioning on covariates. With respect to missingness, the mechanism also allows sample observability to depend on the treatment status. As such it allows for differential non-response, attrition, and even non-compliance to the extent that conditioning variables predict it.

For many observational studies, unconfoundedness may be a reasonable assumption. Previous literature has found several situations where such an assumption is tenable, especially when pre-treatment values of the outcome variable are available. For example, LaLonde (1986) and Hotz et al. (2006) 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 [Chetty et al. (2014), Kane and Staiger (2008), and Shadish et al. (2008)], indicating the plausibility of unconfoundedness in

\(^1\)This is a widely used assumption in the treatment effects literature and is known by a variety of names such as exogeneity, ignorability, selection on observables, and conditional independence assumption (CIA).
such settings.

Estimation then follows in two steps. The first step estimates the treatment and missing outcome probabilities using binary response maximum likelihood\(^2\) and second step plugs in the estimated probabilities as weights to solve a general objective function. Given the parametric nature of the first and second steps, this paper highlights a \textit{robustness} property which allows the estimator to remain consistent for a parameter of interest under misspecification of either a conditional model or the two probability weights. Consequently, the asymptotic theory in this paper distinguishes between these two halves. The first half focuses on misspecification of either a conditional expectation function (CEF) or a conditional quantile function (CQF), whereas the second half considers misspecification in the weighting functions.

As illustrative examples, the paper discusses robust estimation of two specific causal parameters, namely, the ATE and QTEs, expressed as functions of the doubly weighted estimator. Consistent estimation of the ATE is achievable under both misspecification scenarios. Of particular interest is the case when the conditional mean function is misspecified. For estimation of quantile treatment effects, the paper considers three different parameters, namely, conditional quantile treatment effect (CQTE), a linear approximation to CQTE, and unconditional quantile treatment effect (UQTE), each of which may be of interest to the researcher depending on whether features of the conditional or unconditional outcomes distribution are of interest. Simulations show that the doubly weighted ATE and QTE estimates have the lowest finite sample bias compared to alternatives which ignore one or both problems.\(^3\)

Finally, the proposed method is applied to estimate 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 is obtained from Calónico and Smith (2017) who recreate Lalonde’s within-study analysis for the AFDC women. The idea behind choosing this empirical application is to utilize the presence of experimental and non-experimental comparison groups for evaluating whether the strategy of double weighting brings us close the experimental benchmark relative to other alternatives. The paper finds that the empirical bias for the doubly weighted estimate is much smaller than that for the unweighted estimate.

The rest of this paper is structured 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. This half also requires the weights to be correct for delivering parameter identification. In contrast, section 5 considers the other half where a conditional model of interest is correctly specified but the weights may be misspecified. Identification here relies on the parameter solving a conditional problem. Section 6 studies the specifics of robustness for estimating ATE and QTEs in rigorous detail. Section 7 provides supporting Monte Carlo evidence under three interesting cases of misspecification; correct conditional model with misspecified weights, misspecified conditional model with correct weights, and misspecified model and weights. Section 8 applies the proposed method to job training data from Calónico and Smith (2017) and section 9

\(^2\)As a practical matter, researchers typically follow the convention of estimating these probabilities as flexible logit functions.

\(^3\)Such as the unweighted estimator which drops missing outcomes and does not weight or the ps-weighted estimator which drops the missing data and weights by the propensity score to correct for nonrandom assignment.
concludes with directions for future research.

## 2 Potential outcomes and the population models

Consider the standard Neyman-Rubin causal model. Let \( Y(1) \) and \( Y(0) \) denote potential outcomes corresponding to the treatment and control states and let \( W \) be an indicator for whether an individual received the treatment. Then observed outcome is

\[
Y = Y(0) \cdot (1 - W) + Y(1) \cdot W
\]  

Also, let \( X \) be a vector of pre-treatment characteristics which includes an intercept.\(^4\) Some feature of the distribution of \( (Y(g), X) \subseteq \mathbb{R}^M \) is assumed to depend on a finite \( P_g \times 1 \) vector \( \theta_g \), contained in a parameter space \( \Theta_g \subseteq \mathbb{R}^{P_g} \).\(^5\) Let \( q(Y(g), X, \theta_g) \) be an objective function that depends on outcomes, covariates, and the parameter vector, \( \theta_g \). Then, the parameter of interest is defined to be a solution to the following M-estimation problem.

**Assumption 1.** (Identification of \( \theta^0_g \)) The parameter vector \( \theta^0_g \in \Theta_g \) is a unique solution to the population minimization problem

\[
\min_{\theta_g \in \Theta_g} \mathbb{E} [q(Y(g), X, \theta_g)]
\]  

for each \( g = 0, 1 \).

Examples include the smooth ordinary least squares (OLS) function, \( q(Y(g), X, \theta_g) = (Y(g) - X \theta_g)^2 \) or the non-smooth conditional quantile regression (CQR) of Koenker and Bassett (1978), \( q(Y(g), X, \theta_g) = c_\tau(Y(g) - X \theta_g) \). Other examples of \( q(\cdot) \) can be log-likelihood and quasi-log-likelihood (QLL) functions.

An implicit point in assumption 1 is that \( \theta^0_g \) is not assumed to be correctly specified for a conditional feature like a conditional mean, variance, or even the full conditional distribution. It simply requires \( \theta^0_g \) to uniquely minimize the population problem in (2). If \( \theta^0_g \) is correctly specified for any of the above mentioned quantities, then the parameter is of direct interest to researchers. However, if \( \theta^0_g \) is misspecified for any of these distributional features, assumption 1 guarantees a unique pseudo true solution, \( \theta^*_g \) [White (1982)]. In the case of misspecification, determining whether \( \theta^*_g \) is meaningful will depend on the conditional feature being studied and the estimation method used. For example, in the case of OLS, \( \theta^0_g \) will index a linear projection if one is agnostic about linearity of the CEF. Angrist et al. (2006) establish analogous approximation properties for quantiles, where a misspecified CQF can still provide the best weighted mean square approximation to the true CQF.

\(^4\)As mentioned in Negi and Wooldridge (2020), \( X \) may include functions of covariates such as levels, squares, and interactions which will be chosen by the researcher. The dimension of the covariate vector is assumed fixed and does not grow with the sample size.

\(^5\)For generality, the dimension of \( \theta_g \) is allowed to be different for the treatment and control group problems and is also different than the dimension of \( X \), where \( X \in \mathcal{X} \subseteq \mathbb{R}^{\dim(X)} \).

\(^6\)For a random variable \( u, c_\tau(u) = (\tau - 1\{u < 0\}) u \) is the asymmetric loss function for estimating quantiles and \( 1\{\cdot\} \) is an indicator function.
Let \( S \) be a binary indicator such that \( S = 1 \) if the outcome is observed and \( S = 0 \) otherwise. The objective of this paper is to consistently estimate \( \theta^0_g \). In the presence of missing outcomes, a common empirical strategy is to solve the following M-estimation problems for the treatment and control groups, respectively.

\[
\begin{align*}
\min_{\theta_1 \in \Theta_1} & \sum_{i=1}^{N} S_i \cdot W_i \cdot q(Y_i(1), X_i, \theta_1) \\
\min_{\theta_0 \in \Theta_0} & \sum_{i=1}^{N} S_i \cdot (1 - W_i) \cdot q(Y_i(0), X_i, \theta_0)
\end{align*}
\] (3)

Let us refer to the estimator that solves (3) as the unweighted M-estimator and denote it as \( \hat{\theta}^u_g \). This estimator uses the available sample after dropping the missing data to estimate \( \theta^0_g \). Using the reverse analogy principle, \( \hat{\theta}^u_g \) will be consistent for \( \theta^0_g \) if it solves the population analogue of (3), which may not be true. As an example, consider

\[
Y(g) = X\theta_g + U(g), \ g = 0, 1
\]

\[
\mathbb{E}[X'U(g)] = 0
\]

In this case, even if the treatment is randomly assigned, missingness may still be correlated with the treatment, observable factors, or both. Hence, the population first order condition for the selected sample, \( \mathbb{E}[S \cdot W \cdot X'U(g)] \), is not zero even though \( \mathbb{E}[X'U(g)] = 0 \). So identification of \( \theta^0_g \) is now confounded on two grounds; nonrandom assignment which renders the treatment and control groups incomparable and missing outcomes which violates the ‘random sampling’ assumption. The next section discusses the identification approach taken in this paper.

### 3 Identification of parameter of interest

Without imposing any structure on the assignment and missingness mechanisms in the population, estimating \( \theta^0_g \) remains difficult. To proceed with identification, I assume that the treatment is unconfounded on covariates.\(^7\) Formally,

**Assumption 2. (Strong ignorability)** Assume,

\[
\{Y(0), Y(1) \perp \perp W\} \mid X
\] (4)

i) The vector of pre-treatment covariates, \( X \), is always observed for the entire sample.

ii) For all \( x \in \mathcal{X} \subset \mathbb{R}^{\text{dim}(X)} \), define \( p(x) = \mathbb{P}(W = 1|X = x) \) such that \( p(x) > \kappa \) for a constant \( \kappa > 0 \).

Equation (4) indicates that conditioning on covariates is enough to parse out any systematic differences that may exist between the treatment and control groups. One advantage of unconfoundedness is that, intuitively, it has a better chance of holding once we control for a rich set of

\(^7\)Like most other assumptions, unconfoundedness is non-refutable. For methods that indirectly test for its validity, see Huber and Melly (2015), de Luna and Johansson (2014), and Heckman and Hotz (1989).
variables in $X$. Note that unconfoundedness not only includes cases where the treatment is a deterministic function of the covariates, for example stratified (or block) experiments, but also cases where the treatment is a stochastic function of covariates. Part i) requires that we observe these covariates for all individuals. Part ii) is an overlap condition which ensures that for all values of $x$ in $X$, we observe units in both the treatment and control groups.

With respect to the missing outcomes mechanism, I assume selection on observables Assumption 3. (Missing at Random (MAR)) Assume,

$$\{Y(0), Y(1) \perp S| X, W\}$$

i) In addition to $X$, $W$ is always observed for the entire sample.

ii) For each $(x, w) \in (X, W) \subset \mathbb{R}^{\dim(X)+1}$, define, $r(x, w) \equiv P(S = 1|X = x, W = w)$ such that $\eta < r(x, w) < 1$ for a constant $\eta > 0$ and $w = 0, 1$.

Equation (5) states that conditional on covariates and the treatment status, the individuals whose outcomes are missing do not differ systematically from those who are observed. This implies that adjusting for $X$ and $W$ renders the outcomes as good as randomly missing. In the statistics literature, this assumption is known as MAR and represents a mechanism wherein missingness only depends on observables and not on the missing values of the variable itself [Little and Rubin (2019)]. Special cases covered under this mechanism are patterns such as missing completely at random (MCAR) and exogenous missingness considered in Wooldridge (2007). Allowing the missingness probability to be a function of the treatment indicator is particularly useful in cases of differential nonresponse. For instance, in NSW, people assigned to the treatment group were less likely to drop out of the program compared to the control group. In such cases, covariates alone may not be sufficient for predicting missingness. To the extent that being observed in the sample is predicted by $X$ and $W$, assumption 3 can accommodate non-observability due to sampling design, item non-response, and attrition in a two period panel.

Part i) of the above assumption ensures that $X$ and $W$ are fully observed and part ii) again imposes an overlap condition. It states that there is a positive probability of observing people in the sample for a given $X$ and $W$.

Then solving the doubly weighted population problem given below is the same as solving the original M-estimation problem in (2). The following lemma establishes this equality

Lemma 1. (Identification) Given assumptions 1, 2, 3, assume i) $q(Y(g), X, \theta_g)$ is a real valued function for all $(Y(g), X) \subset \mathbb{R}^M$ ii) $E \left[ |q(Y(g), X, \theta_g)| \right] < \infty$ for all $\theta_g \in \Theta_g, g = 0, 1$, then

$$E[\omega_1 \cdot q(Y(g), X, \theta_g)] = E[q(Y(g), X, \theta_g)]$$

where $\omega_1 = \frac{S \cdot W}{r(X, W) \cdot p(X)}$, $\omega_0 = \frac{S \cdot (1 - W)}{r(X, W) \cdot (1 - p(X))}$.

---

8For example, Hirano and Imbens (2001) control for a rich set of prognostic factors to justify unconfoundedness while estimating the effects of right heart catheterization (RHC) on survival rates of patients.

9Methods for checking overlap involve calculating normalized sample average differences for each covariate and checking the empirical distribution of propensity scores.

10For the case of attrition, one must assume that second period missingness is ignorable conditional on initial period covariates and the treatment status.
The proof uses two applications of the law of iterated expectations (LIEs) with unconfoundedness and MAR to arrive at the above result. It implies that one can now address the identification issue due to nonrandom assignment and missing outcomes by solving the doubly weighted population problem.\(^\text{11}\)

### 4 Asymptotic theory under weak identification

Lemma 1 is important for us as it helps to illustrate the role of double weighting in dealing with the two issues at hand. However, to operationalize this argument, we first need to estimate \(r(X, W)\) and \(p(X)\) before introducing the estimator and studying its asymptotic properties.

The following assumptions posit that we have a correctly specified model for the two probabilities and that we estimate them using binary response maximum likelihood. Since both \(W\) and \(S\) are binary responses, estimation of \(\gamma_0\) and \(\delta_0\) using MLE will be asymptotically efficient under correct specification of these functions. Consistency and asymptotic normality for \(\gamma_0\) and \(\delta_0\) follow from theorems 2.5 and 3.3 of Newey and McFadden (1994).

**Assumption 4.** (Correct parametric specification of propensity score) Assume that i) There exists a known parametric function \(G(X, \gamma)\) for \(p(X)\) where \(\gamma \in \Gamma \subset \mathbb{R}^I\) and \(0 < G(X, \gamma) < 1\) for all \(X \in \mathcal{X}, \gamma \in \Gamma\); ii) There exists \(\gamma_0 \in \Gamma\) s.t. \(p(X) = G(X, \gamma_0)\); iii) \(\hat{\gamma}\) is the binary response maximum likelihood estimator that solves

\[
\max_{\gamma \in \Gamma} \sum_{i=1}^{N} \{W_i \log G(X_i, \gamma) + (1 - W_i) \log (1 - G(X_i, \gamma))\} \tag{7}
\]

**Assumption 5.** (Correct parametric specification of missing outcomes probability) Assume that i) There exists a known parametric function \(R(X, W, \delta)\) for \(r(X, W)\) where \(\delta \in \Delta \subset \mathbb{R}^K\) and \(R(X, W, \delta) > 0\) for all \(X \in \mathcal{X}, \delta \in \Delta\); ii) There exists \(\delta_0 \in \Delta\) s.t. \(r(X, W) = R(X, W, \delta_0)\); iii) \(\hat{\delta}\) is the binary response maximum likelihood estimator that solves

\[
\max_{\delta \in \Delta} \sum_{i=1}^{N} \{S_i \log R(X_i, W_i, \delta) + (1 - S_i) \log (1 - R(X_i, W_i, \delta))\} \tag{8}
\]

The influence function representations for \(\hat{\gamma}\) and \(\hat{\delta}\) can then be written as

\[
\sqrt{N} (\hat{\gamma} - \gamma_0) = \mathbb{E} (d_i d_i')^{-1} N^{-1/2} \sum_{i=1}^{N} d_i + o_p(1)
\]

\[
\sqrt{N} (\hat{\delta} - \delta_0) = \mathbb{E} (b_i b_i')^{-1} N^{-1/2} \sum_{i=1}^{N} b_i + o_p(1) \tag{9}
\]

\(^{11}\)Define \(q(Y, X, \theta) = q(Y(1), X, \theta_1)\) for \(W = 1\) and \(q(Y(0), X, \theta_0)\) for \(W = 0\), then \(E[\omega_g \cdot q(Y, X, \theta_g)] = E[\omega_g \cdot q(Y(\theta), X, \theta_g)]\) which makes it function of the observed random vector \(\{Y_i, X_i, W_i, S_i\} : i = 1, 2, \ldots, N\).
where \( d_i \) and \( b_i \) are scores of the binary response log-likelihood problems in (7) and (8) evaluated at the probability limits \( \gamma_0 \) and \( \delta_0 \), respectively. The \textit{doubly weighted} estimator is then defined as:

\[
\hat{\theta}_g = \arg\min_{\theta_g \in \Theta_g} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot q(Y_i(g), X_i, \theta_g)
\]

where \( \hat{\omega}_{i1} = \frac{S_i \cdot W_i}{R(X_i, W_i, \delta) \cdot G(X_i, \gamma)} \) and \( \hat{\omega}_{i0} = \frac{S_i (1 - W_i)}{R(X_i, W_i, \delta) \cdot (1 - G(X_i, \gamma))} \) are the estimated weights for solving the treatment and control group problems, respectively.\( ^{12} \)

Given the two-step nature of the estimation problem; first step uses binary response MLE for estimating the probability weights and 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 \( \hat{\theta}_g \). The following theorem fills in the primitive regularity conditions for applying the uniform law of large numbers.

**Theorem 1.** (Consistency) Suppose assumption 1 holds and that i) \( \{(Y_i, X_i, W_i, S_i); i = 1, 2, \ldots, N\} \) are i.i.d draws satisfying assumptions 2 and 3; ii) \( \Theta_g \) is compact for \( g = 0, 1 \); iii) \( G(X, \gamma) \) satisfies assumption 4 and is continuous for each \( \gamma \) on the support of \( X \). Similarly, \( R(X, W, \delta) \) satisfies assumption 5 and is continuous for each \( \delta \) on the support of \( (X, W) \); iv) \( q(Y(g), X, \theta_g) \) is continuous at each \( \theta_g \in \Theta_g \) with probability one; v) \( \mathbb{E} \left[ \sup_{\theta_g \in \Theta_g} |q(Y(g), X, \theta_g)| \right] < \infty \). Then, \( \hat{\theta}_g \xrightarrow{p} \theta_g^0 \).

The proof follows from verifying the conditions in Lemma 2.4 of Newey and McFadden (1994). Under the dominance condition given in v), uniform convergence of sample averages holds quite generally.

For establishing asymptotic normality, I provide primitive conditions for the general case of non-smooth objective functions. Let the score of \( q(Y(g), X, \theta_g) \) at the true parameter, \( \theta_g^0 \), be denoted as \( h(Y(g), X, \theta_g^0) \equiv h_g \) and suppose it exists with probability one. Let the population problem be denoted as

\[
Q_0(\theta_g) \equiv \mathbb{E} \left[ \omega_g \cdot q(Y(g), X, \theta_g) \right]
\]

and the sample analogue be given as

\[
Q_N(\theta_g) \equiv \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot q(Y_i(g), X_i, \theta_g)
\]

where \( \hat{\rho}_g = N_g / N \) and \( N \rho_g \to \infty \) as \( \hat{\rho}_g \to \rho_g \).\( ^{13} \) For the sake of asymptotics, we may ignore the division by \( \hat{\rho}_g \). The main condition needed for establishing asymptotic normality is stochastic equicontinuity of the empirical process

\[
v_N(\theta_g) \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \hat{\omega}_{ig} h_{ig}(\theta_g) - \mathbb{E} \left[ \hat{\omega}_{ig} h_{ig}(\theta_g) \right] \right\}
\]

\( ^{12} \)When necessary, the estimated weights will also be denoted as \( \omega_g(\hat{\delta}, \hat{\gamma}) \equiv \hat{\omega}_g \).

\( ^{13} \)The sampling fractions \( N_0 = \sum_{i=1}^{N_0} S_i \cdot W_i \) and \( N_1 = \sum_{i=1}^{N_1} S_i \cdot (1 - W_i) \) are random which implies that \( N = N_0 + N_1 \) is also random as opposed to being fixed ahead of time.
Corollary 1. (Efficiency gain with estimated weights) Under the assumptions of theorem 2, \( \omega \) potentially more efficient to estimate the two weights even when they are known. To show this, \( \theta \) when a researcher is only willing to assume identification of conditionally independent. The covariance between the two MLE scores, that term is zero on account of the two scores being probability one; iii) \( \frac{1}{N} \sum_{i=1}^{N} h(Y_i(g), X_i, \hat{\theta}_g) = o_p(N^{-1/2}); \) iv) \( \sup_{\theta \in \Theta} \|h(Y(g), X, \theta_g)\|^2 < \infty; \) v) \( G(\cdot, \gamma) \) and \( R(\cdot, \delta) \) are both twice continuously differentiable on \( \text{int}(\Gamma) \) and \( \text{int}((\Delta), \) respectively; vi) \( \sup_{\delta \in \Delta} \|b(X, W, S, \delta)\|^2 < \infty, \) \( \sup_{\gamma \in \Gamma} \|d(X, W, \gamma)\|^2 < \infty; \) vii) \( E[\omega_g \cdot h(Y(g), X, \theta_g)] \) is continuously differentiable on \( \text{int}(\Theta_g); \) viii) \( H_g \equiv \nabla_{\theta_g} E[\omega_g \cdot h(Y(g), X, \theta_g^0)] \) is nonsingular; ix) \( \{v_N(\theta_g) : N \geq 1\} \) is stochastically equi-continuous. Then,

\[
\sqrt{N}(\hat{\theta}_g - \theta_g^0) \overset{d}{\to} N \left( 0, H_g^{-1} \Omega_g H_g^{-1} \right)
\]

where \( \Omega_g = E(l_{ig} l_{ig}') - E(l_{ig} b_{ig}') E(b_{ig} b_{ig}')^{-1} E(b_{ig} l_{ig}') - E(l_{ig} d_{ig}') E(d_{ig} d_{ig}')^{-1} E(d_{ig} l_{ig}') \) for each \( g = 0, 1 \) and \( l_{ig} \equiv \omega_{ig} h_{ig} \) is score of the weighted objective function evaluated at \( \theta_g^0. \)

Sufficient primitive conditions for stochastic equicontinuity may be found in Andrews (1994). The asymptotic variance expression derived above offers some interesting insights. First, the middle term, \( \Omega_g, \) represents the variance of the residual from the population regression of the weighted score, \( l_{ig}, \) on the two binary response scores, \( b_i \) and \( d_i. \) Note that even though \( \Omega_g \) would involve covariance between the two MLE scores, that term is zero on account of the two scores being conditionally independent.

Second, the expression for \( \Omega_g \) has an efficiency implication for the second step estimate, \( \hat{\theta}_g. \) When a researcher is only willing to assume identification of \( \theta_g^0 \) in the unconditional sense, it is potentially more efficient to estimate the two weights even when they are known. To show this formally, let us assume that \( p(X) \) and \( r(X, W) \) are known and \( \hat{\theta}_g \) is the doubly weighted estimator that uses known weights, \( \omega_g. \) Then,

Corollary 1. (Efficiency gain with estimated weights) Under the assumptions of theorem 2,

\[
\text{Avar} \left[ \sqrt{N}(\hat{\theta}_g - \theta_g^0) \right] - \text{Avar} \left[ \sqrt{N}(\hat{\theta}_g - \theta_g^0) \right] = H_g^{-1} \Sigma_g H_g^{-1} - H_g^{-1} \Omega_g H_g^{-1}
\]

\[
= H_g^{-1} (\Sigma_g - \Omega_g) H_g^{-1}
\]

is positive semi-definite and where \( \Sigma_g = E(l_{ig} l_{ig}'). \)

In other words, we do no worse, asymptotically, by estimating the weights even when we actually know them. This result can be seen an extension of Wooldridge (2007) to the case when one has two sets of probability weights being estimated in the first stage.\(^{14}\)

\(^{14}\)In the missing data literature, this result has also been called the “efficiency puzzle”. Prokhorov and Schmidt (2009) study this puzzle in a GMM framework using an augmented set of moment conditions, where the first set of moments correspond to the weighted objective function and the second set belongs to the missing outcomes (or selection in their case) problem.
5 A conditional feature of interest is correctly specified

The asymptotic results in the previous section were derived under the assumption that some feature of the conditional distribution of outcomes may be misspecified. This was implicit in defining \( \theta^0_g \) as a solution to the unconditional M-estimation problem. Examples include estimating a misspecified linear conditional mean or quantile function. In contrast, this section highlights the other half of the asymptotic theory which is formalized using a strong version of the identification assumption and allowing the weights to be misspecified.

**Assumption 6.** (Strong identification of \( \theta^0_g \)) The parameter vector \( \theta^0_g \in \Theta_g \) is the unique solution to the population minimization problem

\[
\min_{\theta_g \in \Theta_g} \mathbb{E} \left[ q(Y(g), X, \theta_g) | X \right]; \quad g = 0, 1
\]

under unconfoundedness (defined in 2) and MAR (defined in 3) for each \( X \in \mathcal{X} \subset \mathbb{X}^{\text{dim}(X)} \).

The above can be seen as a strengthening of the identification assumption in section 4 since LIE implies that \( \theta^0_g \) is also a solution to the unconditional M-estimation problem. By requiring \( \theta^0_g \) to solve (12), assumption 5 is intended for situations where a conditional feature of interest is correctly specified. An implication of this strengthened identification is that \( \theta^0_g \) now solves the conditional score of the objective function i.e. \( \mathbb{E} \left[ h(Y(g), X, \theta^0_g) | X \right] = 0 \).

For instance, the conditional score will be zero in the case of estimating a correctly specified CEF with either OLS or quasi maximum likelihood estimation (QMLE) in the linear exponential family (LEF). This would also hold for a correctly specified CQF estimated either using quantile regression or QMLE in the tick exponential family [Komunjer (2005)].

Delineating these two identification scenarios is important for determining which causal parameter can be estimated consistently under each setting. As we will see in the next section, it is possible to estimate the ATE under both cases of misspecification. However the same cannot be said for QTE parameters. In addition to assumption 6, the asymptotic results in this half do not rely on correct specification of weights. In other words, assuming \( R(\cdot, \cdot, \delta) \) and \( G(\cdot, \gamma) \) to be correctly specified is rather restrictive and not required for the doubly weighted estimator to be consistent for \( \theta^0_g \).

**Assumption 7.** (Parametric specification of propensity score) Assume that conditions i) and iii) of assumption 4 hold where condition ii) is defined for some \( \gamma^* \in \Gamma \) such that \( \text{plim}(\hat{\gamma}) = \gamma^* \).

**Assumption 8.** (Parametric specification of missingness probability) Assume that conditions i) and iii) of assumption 5 hold where condition ii) is defined for some \( \delta^* \in \Delta \) such that \( \text{plim}(\hat{\delta}) = \delta^* \).

Note that assumptions 7 and 8 do not require the parametric models for the two probabilities to be correctly specified. Nevertheless, we continue to assume that \( \hat{\gamma} \) and \( \hat{\delta} \) solve the same binary response problem as in Assumptions 4 and 5 with probability limits given by pseudo true values \( \gamma^* \) and \( \delta^* \), respectively [White (1982)]. To show that \( \theta^0_g \) is still a solution to the doubly weighted population problem with misspecified weights, a sketch of the argument is given below. Consider,

\[
\mathbb{E} \left[ \omega_g^* \cdot q(Y(g), X, \theta_g) \right]
\]

(13)
where $\omega_g^*$ are asymptotic weights which use $G(X, \gamma^*)$ and $R(X, W, \delta^*)$. Using LIE along with unconfoundedness and MAR, I can rewrite the above expectation as

$$\mathbb{E} \left[ \xi_g(X) \cdot \mathbb{E}\{q(Y(g), X, \theta_g)|X\} \right]$$

where $\xi_g(X)$ is a function of weights for $g = 0, 1$. The strong identification assumption implies

$$\mathbb{E}[q(Y(g), X, \theta_g^0)|X] \leq \mathbb{E}[q(Y(g), X, \theta_g)|X], \forall \theta_g \in \Theta_g$$

Further, since $\xi_g(X) > 0$,

$$\mathbb{E}\left[ \omega_g^* \cdot q(Y(g), X, \theta_g^0) \right] \leq \mathbb{E}\left[ \omega_g^* \cdot q(Y(g), X, \theta_g) \right], \theta_g \in \Theta_g$$

where the inequality is strict when $\theta_g \neq \theta_g^0$. Therefore, solving the doubly weighted problem identifies the parameter even if the weights are wrong. In general, the parameter that solves (13) will be different from the one that solves the same problem with correct weights. But as long as $\theta_g^0$ is a unique solution, solving (13) will identify it.

The following two theorems establish consistency and asymptotic normality of the doubly weighted estimator.

**Theorem 3.** (Consistency under strong identification) Under assumptions 2, 3, 6, 7, and 8 with regularity conditions (1), (2) and (3) of Theorem 1, $\hat{\theta}_g \xrightarrow{p} \theta_g^0$ as $N \to \infty$ where $\hat{\theta}_g$ is the doubly-weighted estimator that solves (13).

**Theorem 4.** (Asymptotic Normality under strong identification) Under the assumptions of theorem 3 and the regularity conditions of theorem 2 where MLE estimators $\hat{\gamma}$ and $\hat{\delta}$ have probability limits given by $\gamma^*$ and $\delta^*$, then $\sqrt{N}(\hat{\theta}_g - \theta_g^0) \xrightarrow{d} N(0, H_g^{-1} \Omega_g H_g^{-1})$ where $\Omega_g = \mathbb{E}(l_{ig} l_{ig}')$ with $H_g$ and $l_{ig}$ defined in Theorem 2 except with asymptotic weights given by $\omega_{ig}^*$.

Substantially, there is no real difference in the proof of the above theorem compared to those derived in section 4 except that now $\hat{\gamma}$ and $\hat{\delta}$ are converging to probability limits that could be potentially different from those indexing the true treatment and missing outcome probabilities. A consequence of the objective function solving the conditional problem is reflected in the asymptotic variance expression. Compared to the previous section, $\Omega_g$ now is simply the variance of weighted score of the objective function without first stage adjustment of the estimated probabilities. This is because under assumption 6, $\mathbb{E}(l_{ig} b_i') = \mathbb{E}(l_{ig} d_i') = 0$. A sketch of the proof for $\mathbb{E}(l_{ig} b_i') = 0$ is provided below. The argument for $\mathbb{E}(l_{ig} d_i')$ follows analogously.

$$\mathbb{E}(l_{ig} b_i') = \mathbb{E}(\omega_{ig} h_{ig} b_i') = \mathbb{E}\left[ \xi_g(X_i) \cdot \mathbb{E}(h(Y_i(g), X_i, \theta_g^0)|X_i) \right] = 0$$

where $\xi_g(X)$ is a function of weights. The first equality uses the definition of $l_{ig}$ with misspecified weights and second equality applies LIEs with unconfoundedness and MAR. In other words, the reason for obtaining a simpler expression for $\Omega_g$ is because the correlation between weighted score of the objective function and the two binary response scores is zero when $\theta_g^0$ is correctly specified for a conditional feature of interest and we use an appropriate method to estimate it.

---

15When $R(X, W, \delta^*) = r(X, W)$ and $G(X, \gamma^*) = p(X)$, then solving (13) will be the same as solving the problems in section 3.
A simpler expression for $\Omega_q$ also means that we can no longer exploit these correlations between scores to obtain asymptotic efficiency for estimating $\theta_g^0$. Again, let $\hat{\theta}_g$ be the doubly weighted estimator that uses true weights, $\omega_g$, then

**Corollary 2.** *(No gain with estimated weights under strong identification)* Under the assumptions of theorem 4

$$ \text{Avar}[\sqrt{N}(\hat{\theta}_g - \theta_g^0)] = \text{Avar}[\sqrt{N}(\hat{\theta}_g - \theta_g^0)] = H_g^{-1} \Omega_g H_g^{-1} $$

Hence knowledge of the weights does little when for instance we have a correctly specified CEF or CQF and we use either OLS or QR to estimate the parameters indexing these conditional models of interest.

A special case of weights misspecification is when $\omega_g^*$ is a constant. This is plausible since $R(X, W, \delta^*)$ and $G(X, \gamma^*)$ are allowed to be any bounded positive functions of $X$ and $W$. In other words, the unweighted estimator, $\hat{\theta}_g^u$, which does not weight to correct for either problem is also consistent for $\theta_g^0$ under the results of theorem 3. In fact, assumptions 7 and 8 suggest that any weighted estimator will suffice for estimating $\theta_g^0$. 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 of its weighted counterpart (correctly specified weights or not).

**Corollary 3.** *(Efficiency gain with unweighted estimator under GCIME)* Under assumptions of theorem 4 if we additionally suppose that the objective function satisfies GCIME in the population which is defined as:

$$ \mathbb{E}[h(Y(g), X, \theta_g^0)h(Y(g), X, \theta_g^0)'|X] = \sigma_{0g}^2 \cdot \nabla_{\theta_g} \mathbb{E}[h(Y(g), X, \theta_g^0)|X] = \sigma_{0g}^2 \cdot A(X, \theta_g^0) \quad (14) $$

Then, $\text{Avar}[\sqrt{N}(\hat{\theta}_g - \theta_g^0)] = H_g^{-1} \Omega_g H_g^{-1}$ and $\text{Avar}[\sqrt{N}(\hat{\theta}_g^u - \theta_g^0)] = (H_g^u)^{-1} \Omega_g^u (H_g^u)^{-1}$

and,

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

is positive semi-definite. \qed

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 $B_i$ on $D_i$, which are weighted versions of square root of matrix $A_i$ (See appendix F for details). Hence the difference is positive semi-definite.

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

6 Estimation of treatment effects

The asymptotic theory can now be used to discuss estimation of specific causal estimands like ATE and QTEs which can be expressed as functions of the doubly weighted estimator, \( \theta_g^0 \).

6.1 Average treatment effect

As discussed in Słoczyński and Wooldridge (2018), DR estimators remain consistent for the population ATE despite misspecification in either the conditional mean function or the propensity score, but not both. The current doubly weighted framework along with results developed in sections 4 and 5 allow us to extend this result to the case with missing outcomes.

Let \( m(X, \theta_g) \) be a parametric model for the conditional mean which is said to be correctly specified for the CEF if for some \( \theta_g^0 \in \Theta_g \)

\[
\mathbb{E}[Y(g)|X] = m(X, \theta_g^0)
\]
or equivalently, \( Y(g) = m(X, \theta_g^0) + U(g) \) such that \( \mathbb{E}[U(g)|X] = 0 \). Then, let us consider the following two scenarios in turn.

6.2 Double robustness

First half: Correct conditional mean  When the conditional mean function is correct, there is more than one estimation method that can be used to consistently estimate \( \theta_g^0 \), namely, nonlinear least squares (NLS) and QMLE with LEF. For both these examples, results from section 5 dictate that weighting is not needed for consistency. The fact that one could weight by the misspecified weights and still consistently estimate \( \theta_g^0 \) is what forms the ‘first part’ of the DR result with double weighting.

Once \( \theta_g^0 \) has been estimated by solving the sample version of the NLS or QMLE problem, ATE can be estimated as follows,

\[
\hat{\Delta}_{ate} = \frac{1}{N} \sum_{i=1}^{N} m(X_i, \hat{\theta}_1) - \frac{1}{N} \sum_{i=1}^{N} m(X_i, \hat{\theta}_0)
\]

If in addition to having a correct conditional mean, I also assume the error variance of the outcomes to be homoskedastic \( (\mathbb{E}[U^2(g)|X] = \text{Var}[U(g)|X] = \sigma_{0g}^2) \), then the NLS estimator that does not weight at all is the preferred alternative from an efficiency perspective. This is due to GCIME being satisfied with NLS under homoskedasticity.

Second half: Correct weights  If one acknowledges misspecification in the conditional mean model, there is no general way of consistently estimating the ATE. However, a useful mean fitting property of QMLEs in LEF along with double weighting can be used here to obtain consistent
estimates of the unconditional means, \( \mathbb{E}[Y(g)] \), despite misspecification in the conditional means, \( \mathbb{E}[Y(g)|X] \).

In the generalized linear model (GLM) literature, the link function, \( h^{-1}(\cdot) \), relates the mean of the distribution to a linear index as follows

\[
h^{-1}(\mathbb{E}[Y(g)|X]) = X\theta_g
\]  

The estimation strategy then is to choose \( m(X, \theta_g) \) to be the function, \( h(\cdot) \), with the QLL corresponding to a choice of LEF density. Then the population first order conditions from solving this QMLE problem give us

\[
\mathbb{E}\left[ \nabla_{\theta_g} h(X\theta_g^*)' \cdot (Y(g) - h(X\theta_g^*)) \right] = 0
\]  

where \( v[h(\cdot)] \) is variance of the mean function and \( \theta_g^* \) denotes the pseudo true parameter indexing the misspecified conditional mean model [White (1982)]. In particular, by choosing \( h^{-1}(\cdot) \) to be the canonical link for the QLL associated with the density, the gradient in numerator of (16) cancels with the variance term in the denominator. Note that this occurs only when one uses the canonical link function.

Such cancellation of terms ensures that if one includes an intercept in \( X \), the misspecified mean model fits the overall mean of the distribution (see Wooldridge (2010) chapter 13 for more detail) so that,

\[
\mathbb{E}[Y(g)] = \mathbb{E}[h(X\theta_g^*)]
\]

With nonrandom assignment and missing outcomes, solving the sample GLM FOC in (16) would still not be sufficient for consistently estimating \( \theta_g^* \). Therefore, one would instead solve the doubly weighted FOC given below.

\[
\sum_{i=1}^{N} \omega_{i1} \cdot X_i' \cdot [Y_i - h(X_i\hat{\theta}_1)] = 0
\]

\[
\sum_{i=1}^{N} \omega_{i0} \cdot X_i' \cdot [Y_i - h(X_i\hat{\theta}_0)] = 0
\]

The role played by weighting is crucial here for \( \hat{\theta}_g \) to be consistent for the pseudo true parameter \( \theta_g^* \). This forms the ‘second half’ of the DR result with double weighting.

If \( h(\cdot) \) is the identity function, the first order conditions above can be recognized as those belonging to OLS with the line of best fit passing through the mean of \( Y \). This is because OLS is a QMLE with normal QLL and identity link function, typically used for outcomes with unrestricted support. Other combinations of QLLs and canonical link functions can be found in Table 2 of Negi.
and Wooldridge (2020) and have to be chosen depending on the range and nature of $Y$.

**Summary.** DR estimation of ATE with double weighting

**Case 1: Correct mean, misspecified weights**

1. Consistent estimates for the conditional mean parameters, $\theta^0_g$, can be obtained by either using NLS or QMLE in LEF.

2. A consistent estimator of ATE is obtained as

$$\hat{\Delta}_{\text{ate}} = \frac{1}{N} \sum_{i=1}^{N} m(X_i, \hat{\theta}_1) - \frac{1}{N} \sum_{i=1}^{N} m(X_i, \hat{\theta}_0)$$

**Case 2: Misspecified mean, correct weights**

1. Depending upon the range and nature of the outcome, $Y$, choose an appropriate QLL associated with an LEF density. Choose the mean function, $m(X, \theta_g) = h(X_{\theta_g})$, where $h(\cdot)$ is the inverse canonical link function associated with the chosen density. Using this combination of mean function and QLL, use the moment conditions in (17) to obtain consistent estimates, $\hat{\theta}_g$.

2. Consistent estimates of ATE can then be obtained as follows

$$\hat{\Delta}_{\text{ate}} = \frac{1}{N} \sum_{i=1}^{N} h(X_i, \hat{\theta}_1) - \frac{1}{N} \sum_{i=1}^{N} h(X_i, \hat{\theta}_0)$$

where $X$ includes an intercept and $\hat{\theta}_g$ solves the GLM first order conditions.

### 6.3 Quantile treatment effects

Unlike the case of ATE, it is generally not possible to obtain UQTE by averaging CQTE over the distribution of $X$. In this section, I use double weighting to illustrate estimation of three different quantile estimands, namely, UQTE, CQTE, and a weighted linear approximation (LP) to the true CQTE, each of which may be of interest to the researcher depending on whether features of the conditional or unconditional outcomes distribution are of interest. Whether $\theta^0_g$ indexes the true CQF or an approximation depends on what is being assumed about the conditional quantile model and the estimation method used.

Let’s assume that the two potential outcomes are continuous in $\mathbb{R}$. It is typical to define the $\tau^{th}$ quantile of $Y(g)$ as

$$Q_{\tau,g} = \inf\{y : F_g(y) \geq \tau\}, \; 0 < \tau < 1$$

Then the UQTE for the $\tau^{th}$ quantile is defined as the difference in the marginal quantiles of the outcomes distributions,

$$\text{UQTE}_\tau = Q_{\tau,1} - Q_{\tau,0}$$
Similarly, one may define the $\tau^{th}$ conditional quantile of $Y(g)$ for $X = x$ as,

$$Q_{\tau,g}(x) = \inf\{y : F_{g}(y|x) \geq \tau\}, \ 0 < \tau < 1$$

where $F_{g}(\cdot|x)$ denotes the conditional distribution function of $Y(g)$ given $X = x$. Then, CQTE for the $\tau^{th}$ quantile for some subgroup defined by $X$ is

$$\text{CQTE}_{\tau}(X) = Q_{\tau,1}(X) - Q_{\tau,0}(X)$$

Let $q_r(X, \theta_g(\tau))$ be a parametric model for the $\tau^{th}$ conditional quantile of $Y(g)$ which is said to be correctly specified if for some $\theta_g(\tau) \in \Theta_g$

$$Q_{\tau,g}(x) = q_r(X, \theta_g(\tau))$$  \hspace{1cm} (18)

**Estimation of \text{CQTE}_{\tau}**: Incidentally, much like conditional mean, if CQF $\tau$ is correctly specified, there are two methods that will ensure consistent estimation of the CQF parameters, $\theta_g(\tau)$. The first is CQR of Koenker and Bassett (1978) and the second is a class of QML estimators that use a special ‘tick-exponential’ family of distributions to suggest consistent estimators of conditional quantile parameters. This QMLE class has been proposed by Komunjer (2005). The method is analogous to estimating a correctly specified conditional mean function using QMLE in the linear exponential family.

For estimation that uses CQR, $\theta_g(\tau)$ will actually solve the stronger conditional problem,

$$\theta_g(\tau) = \arg\min_{\theta_g \in \Theta_g} \mathbb{E} \left[ c_{r}(Y(g) - q_r(X, \theta_g(\tau)))|X \right]$$  \hspace{1cm} (19)

For estimation via QMLE, as long as the CQF is correct and we choose an appropriate QLL then,

$$\theta_g(\tau) = \arg\min_{\theta_g \in \Theta_g} \mathbb{E} \left[ -\ln \left\{ \phi^\tau(Y(g), q_r(X, \theta_g(\tau))) \right\} |X \right]$$  \hspace{1cm} (20)

where $\phi^\tau(\cdot, \cdot)$ is the density that belongs to the tick-exponential family.\(^{18}\) As dictated by results in section 5, weighting the QR or QML objective functions, irrespective of whether the weights are correctly specified or not will also deliver a consistent estimator of $\theta_g(\tau)$.

Once we have obtained $\hat{\theta}_g$ either by solving the QR or QML problem, the $\tau^{th}$ conditional quantile treatment effect for subgroup X can be estimated as $\text{CQTE}_{\tau}(X) = q_r(X, \hat{\theta}_1(\tau)) - q_r(X, \hat{\theta}_0(\tau))$.

**Estimation of LP to CQTE$_{\tau}$**: The traditional literature on conditional quantile estimation has focused on correct specification. However, Angrist et al. (2006) establish an approximation property of CQR that is analogous to the approximation property of linear regression. The main implication of such a result is that solving CQR with $q_r(X, \theta_g(\tau)) = X\theta_g(\tau)$ would still identify a

\(^{18}\phi^\tau(y, \eta) = \phi^\tau(\eta) = \exp \left[ -(1 - \tau)[a(\eta) - b(\eta)]1\{y \leq \eta\} + \tau[a(\eta) - c(\eta)]1\{y > \eta\}\right] \) is a probability density and $\eta$ is the $\tau$-quantile of $\phi^\tau$ such that $\int_{-\infty}^{\eta} \phi^\tau(y, \eta)dy = \tau$. Komunjer (2005) shows that CQR of Koenker and Bassett (1978) is a special case of this QMLE class.
weighted linear approximation to CQF. Therefore, the difference in LPs of \( \tau \)-quantile CQFs is interpretable as identifying an LP to the CQTE. 

As before, weighting becomes crucial in the presence of nonrandom assignment and missing outcomes for identifying the LP parameters.

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

(21)

In other words, one would need to weight the CQR problem with correct weighting functions for \( \hat{\theta}_g(\tau) \xrightarrow{P} \theta^*_g(\tau) \), which indexes the true LP to CQF for group \( g \). Then,

\[
\text{LP}[\text{CQTE}_\tau(X)] = X[\hat{\theta}_1(\tau) - \hat{\theta}_0(\tau)]
\]  

(22)

**Direct estimation of UQTE**:

As mentioned in the beginning of this section, estimating UQTE from CQTE is generally not possible even if we assume a correct model for the conditional quantiles of \( Y(\cdot) \). In other words, one cannot obtain unconditional quantiles from averaging conditional quantiles over the distribution of \( X \). In this case, we can directly estimate \( Q_{\tau,g} \) by running a quantile regression of the outcome on an intercept (similar to Firpo (2007)).

In the present case, the solution to the doubly weighted objective function gives us,

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

such that \( \hat{\theta}_g(\tau) \xrightarrow{P} Q_{\tau,g} \). Weighting by \( G(\cdot) \) and \( R(\cdot) \) is crucial here since these functions serve to remove biases arising due to nonrandom assignment and missing outcomes. One can then obtain the unconditional quantile treatment effect as,

\[
\hat{\text{UQTE}}_{\tau} = \hat{\theta}_1(\tau) - \hat{\theta}_0(\tau)
\]

An alternative method of estimating UQTE is to use recentered influence functions suggested by Firpo et al. (2009) (see appendix B).

The next section discusses results from a Monte Carlo study which evaluates the finite sample behavior of doubly weighted ATE and QTE estimators under three different misspecification scenarios.

### 7 Simulations

This section compares the empirical distributions of ATE and QTEs using unweighted, ps-weighted, and d-weighted estimators. The discussion is centered around three common misspecification scenarios that are interesting from an empirical standpoint. These cases are enumerated in tables
A.1 and A.2 for estimating ATE and QTEs, respectively. Two of them describe situations implicit in the first and second half of the asymptotic theory, whereas the third case considers all three parametric components of the framework to be misspecified. Even though the theory developed in this paper is silent for the third case, simulation results appear to be promising.

7.1 Average treatment effect: Results

Case (1) in Table A.1 considers a misspecified mean function but correct probability weights. This is the principal case covered in section 4 wherein weighting is crucial. As one can see, the empirical distribution of the doubly weighted estimator is centered on the true ATE whereas that for the unweighted estimator is shifted to the right (see figure A.1, Case 1).

Case (2) looks at what happens when everything, conditional mean and the two weights, is misspecified. The theory in this paper does not address this particular case. However, this characterizes an interesting possibility given that misspecification of all components is a valid concern. The simulation results do offer some insight here. The doubly weighted estimator seems to be the only choice that delivers the true ATE on average whereas the others distributions are shifted away from the truth (see figure A.1, Case 2).

Finally, case (3) depicts the possibility of a correctly specified conditional mean function but misspecified weights. Here weighting does not have any bite in resolving the identification issue, beyond what is already achieved from having a correct mean function. In figure A.1, case 3, the empirical distributions of the estimated ATE for the unweighted, ps-weighted, and d-weighted estimators all coincide and are centered on the true ATE.

[Figure A.1 here]

7.2 Quantile treatment effects: Results

As discussed earlier, there are really three parameters worth discussing when one talks about QTEs; CQTE, LP to CQTE, and UQTE. Misspecification in the CQF shifts attention to consistently estimating a linear projection to the true CQTE. First case in Table A.2 considers exactly such a scenario. Using the results in Angrist et al. (2006), I interpret the solution to the doubly weighted problem given in (21) as providing a consistent weighted linear projection to the true CQTE which is then used to estimate an LP to the true CQTE. Case 1 of Figure A.2 plots the bias in estimated LP relative to the true LP as a function of $X_1$ for the three estimators. Note that weighting here is crucial for consistently estimating the LP. The relative bias of the doubly weighted estimator is the lowest amongst all and coincides with the line of no bias. Case 2 considers the situation when along with a misspecified CQF, the weights are also wrong. We still find the proposed estimator performing the best in terms of bias.

Finally figure A.3 considers a correctly specified CQF in which case we can estimate the CQTE.21 One can observe in the figure that the estimated function using double weighting coincides with the true CQTE irrespective of how we weight. All three estimators; unweighted, ps-weighted, and doubly weighted will be consistent for the true CQTE. Misspecification in the weights will not affect this result.

---

21 See section A of the online appendix for details regarding plotting the CQTE curve.
I also consider direct estimation of UQTE which does not require parametric specification of the CQF since it is simply a difference of the marginal quantiles. So the two weights are the only relevant components of the framework which will affect consistency of UQTE. In figure A.4, case 1, when both weights are correct, not weighting and double weighting both bring us close to the true parameter. For the second case where both probability models are misspecified, double weighting does a little worse than not weighting at all. However, the results at other quantiles reflect more favorably upon double weighting (see section H of the online appendix for results at 50th and 75th quantiles). Propensity score weighting performs the worst in both cases suggesting instances where weighting for nonrandom assignment after dropping data that is missing may not be the preferred alternative.

[Figure A.2 here] [Figure A.3 here] [Figure A.4 here]

8 Returns to job training

In this section, I apply the proposed estimator to the Aid to Families with Dependent Children (AFDC) sample of women from the National Supported Work program compiled by Calónico and Smith (2017) (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 (1986)’s 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 (for example, CIA) compare to experimental estimates. To compute the non-experimental estimates, CS combine the NSW experimental sample with two non-experimental comparison groups drawn from PSID, called PSID-1 and PSID-2.\textsuperscript{22} In this paper, I utilize the within-study feature of this empirical application to estimate how close the doubly weighted estimates get to the experimental estimate compared with ps-weighting and unweighted estimates.

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.\textsuperscript{23} The difference in the non-experimental estimate, obtained from using the doubly weighted estimator, and the experimental estimate provides 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 [Heckman et al. (1998)]. Much like CS, I report both these estimates across a range of regression specifications 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 10th quantile of the 1979 earnings distribution. The role of double weighting is highlighted for the case of estimating marginal quantiles since covariates,

\textsuperscript{22}The PSID-1 sample constructed by CS involves keeping all female household heads continuously from 1975-1979 who were between 20 and 55 years of age in 1975 and were not retired in 1975. The sample labeled PSID-2 further restricts PSID-1 to include only those women who received AFDC welfare in 1975.

\textsuperscript{23}For details regarding sample construction and estimation of weights, see section E of the online appendix.
which primarily serve to remove biases arising from nonrandom assignment and missing outcomes, enter the estimating equation only through the two weights.

### 8.1 Results

First, to evaluate whether women with missing earnings in 1979 were significantly different than those who were observed, Table A.2 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. Apart from these minor differences, the observed women did not appear to be systematically different that those who were missing, as measured through observable characteristics.

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 A.1 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, indicating the strong role that randomization plays in producing comparable groups. In contrast, the women in PSID-1 and PSID-2 groups are statistically different than the treatment group members implying substantial scope for nonrandom assignment.

Table A.3 reports the d-weighted, ps-weighted and unweighted average returns to training estimates which using three different comparison groups; NSW control, PSID-1 and PSID-2. The unweighted (unadjusted and adjusted) experimental estimates given in row 1, are same as the estimates reported by CS in Table 3 of their paper. Overall, one can see that the doubly weighted experimental estimates are more stable than the single weighted or unweighted estimates across the different regression specifications, with a range between $824-$828.

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. 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 which controls for all covariates is $335 whereas the weighted estimates are $905 and $904.

The second panel of Table A.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. For instance, the doubly weighted estimate that adjusts for all covariates using the PSID-1 comparison group is -$21 whereas the unweighted estimates is -$568. These results suggest that the argument for weighting is strong when using a non-experimental comparison group where nonrandom assignment and

---

24 Appendix E describes estimation of the two probability weights along with the sample trimming criteria.
missing outcomes are significant problems.\textsuperscript{25}

Figure A.5 plots the relative bias in UQTE estimates at every 10th 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 non-experimental group. Note that I do not plot UQTE estimates for quantiles less than 0.46, since these are all zero.\textsuperscript{26}

This empirical application illustrates the role of proposed estimator in both experimental and observational data contexts. The comparison involving the treatment and control group of NSW demonstrates its use in an experiment with missing outcomes, whereas the non-experimental sample demonstrates its use in the more realistic observational data setting.

[Table A.1 here] [Table A.2 here] [Table A.3 here] [Figure A.5 here]

9 Conclusion

In empirical research, the problems of nonrandom assignment and missing outcomes threaten identification of causal parameters. This paper proposes a new class of consistent and asymptotically-normal M-estimators that address these two issues using a double weighting procedure. The method combines propensity score weighting with weighting for missing outcomes in a general M-estimation framework, which can be applied to a range of estimation methods, such as ordinary least squares, quasi maximum likelihood, and quantile regression. In addition, the proposed class has a robustness property which allows us to estimate meaningful causal quantities of interest despite misspecification in either a conditional model of interest or the two weighting functions.

As leading applications, the paper discusses estimation of ATE and QTEs. A Monte Carlo study indicates that the doubly weighted estimates of average and quantile treatment effects have the lowest bias compared to naive alternatives (unweighted or propensity score weighted estimators) under three realistic cases of misspecification. Finally, the estimator is applied to the data on AFDC women from the NSW program compiled by Calónico and Smith (2017). The presence of experimental and non-experimental comparison groups in this application help to quantify the estimated bias in the doubly weighted returns to training estimates as well as the other two estimators.

Since the severity and magnitude of bias introduced from ignoring either problem cannot be assessed ex-ante, a safe bet from the practitioner’s perspective is to report both doubly weighted and unweighted causal effect estimates. Practically, the doubly weighted estimator for the ATE is easy to implement. Appendix D provides an example code that uses Stata’s \texttt{gmm} command for implementing it. Computation of analytically correct standard errors, however, requires additional coding and is still a work in progress. Alternatively, one can use bootstrapped standard errors which will provide asymptotically correct inference.

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.

\textsuperscript{25}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.

\textsuperscript{26}There are a lot of women in the experimental and PSID samples with zero real earnings in 1979.
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 average treatment effect (LATE).

References

ANDREWS, D. W. (1994): “Empirical process methods in econometrics,” *Handbook of econometrics*, 4, 2247–2294.

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): “Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure,” *Econometrica*, 74, 539–563.

CALÓNICO, S. AND J. SMITH (2017): “The women of the national supported work demonstration,” *Journal of Labor Economics*, 35, S65–S97.

CHETTY, R., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014): “Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates,” *American Economic Review*, 104, 2593–2632.

DE LUNA, X. AND P. JOHANSSON (2014): “Testing for the unconfoundedness assumption using an instrumental assumption,” *Journal of Causal Inference*, 2, 187–199.

FIRPO, S. (2007): “Efficient semiparametric estimation of quantile treatment effects,” *Econometrica*, 75, 259–276.

FIRPO, S., N. M. FORTIN, AND T. LEMIEUX (2009): “Unconditional quantile regressions,” *Econometrica*, 77, 953–973.

FIRPO, S. AND C. PINTO (2016): “Identification and estimation of distributional impacts of interventions using changes in inequality measures,” *Journal of Applied Econometrics*, 31, 457–486.

FRICKE, H., M. FRÖLICH, M. HUBER, AND M. LECHNER (2020): “Endogeneity and non-response bias in treatment evaluation–nonparametric identification of causal effects by instruments,” *Journal of Applied Econometrics*, 35, 481–504.

FRÖLICH, M. AND M. HUBER (2014): “Treatment evaluation with multiple outcome periods under endogeneity and attrition,” *Journal of the American Statistical Association*, 109, 1697–1711.

HAHN, J. (1998): “On the role of the propensity score in efficient semiparametric estimation of average treatment effects,” *Econometrica*, 315–331.

HECKMAN, J., H. CHILINGURU, J. SMITH, AND P. TODD (1998): “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, 66, 1017–1098.

HECKMAN, J. J. AND V. J. HOTZ (1989): “Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training,” *Journal of the American statistical Association*, 84, 862–874.
Hirano, K. and G. W. Imbens (2001): “Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization,” Health Services and Outcomes Research Methodology, 2, 259–278.

Horvitz, D. G. and D. J. Thompson (1952): “A generalization of sampling without replacement from a finite universe,” Journal of the American Statistical Association, 47, 663–685.

Hotz, V. J., G. W. Imbens, and J. A. Klerman (2006): “Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program,” Journal of Labor Economics, 24, 521–566.

Huber, M. (2014): “Treatment evaluation in the presence of sample selection,” Econometric Reviews, 33, 869–905.

Huber, M. and B. Melly (2015): “A test of the conditional independence assumption in sample selection models,” Journal of Applied Econometrics, 30, 1144–1168.

Kane, T. J. and D. O. Staiger (2008): “Estimating teacher impacts on student achievement: An experimental evaluation,” Tech. rep., National Bureau of Economic Research.

Koenker, R. and G. Bassett (1978): “Regression Quantiles,” Econometrica, 46, 33–50.

Komunjer, I. (2005): “Quasi-maximum likelihood estimation for conditional quantiles,” Journal of Econometrics, 128, 137 – 164.

LaLonde, R. J. (1986): “Evaluating the econometric evaluations of training programs with experimental data,” The American Economic Review, 604–620.

Little, R. J. and D. B. Rubin (2019): Statistical analysis with missing data, vol. 793, John Wiley & Sons.

Negi, A. and J. M. Wooldridge (2020): “Revisiting regression adjustment in experiments with heterogeneous treatment effects,” Econometric Reviews, 0, 1–31.

Newey, W. K. and D. McFadden (1994): “Large sample estimation and hypothesis testing,” Handbook of Econometrics, 4, 2111–2245.

Prokhorov, A. and P. Schmidt (2009): “GMM redundancy results for general missing data problems,” Journal of Econometrics, 151, 47–55.

Robins, J. M. and A. Rotnitzky (1995): “Semiparametric efficiency in multivariate regression models with missing data,” Journal of the American Statistical Association, 90, 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): “Estimation of regression coefficients when some regressors are not always observed,” Journal of the American Statistical Association, 89, 846–866.

Rosenbaum, P. R. and D. B. Rubin (1983): “The central role of the propensity score in observational studies for causal effects,” Biometrika, 70, 41–55.
SHADISH, W. R., M. H. CLARK, AND P. M. STEINER (2008): “Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments,” *Journal of the American statistical association*, 103, 1334–1344.

SŁOCZYŃSKI, T. AND J. M. WOOLDRIDGE (2018): “A general double robustness result for estimating average treatment effects,” *Econometric Theory*, 34, 112–133.

WHITE, H. (1982): “Maximum likelihood estimation of misspecified models,” *Econometrica: Journal of the Econometric Society*, 1–25.

WOOLDRIDGE, J. M. (2007): “Inverse probability weighted estimation for general missing data problems,” *Journal of Econometrics*, 141, 1281–1301.

——— (2010): *Econometric analysis of cross section and panel data*, MIT press.
A Tables and figures for main text

Figure A.1: Empirical distribution of estimated ATE for N=5,000

Case 1: Misspecified CEF, correct weights
Case 2: Misspecified CEF, misspecified weights

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 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.

Case 3: Correct CEF, misspecified weights

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 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.
Figure A.2: Bias in the estimated LP relative to the true LP to CQTE as a function of $X_1$ for $N=5,000$

a) $\tau = 0.25$

Case 1: Misspecified CQF, correct weights  
Case 2: Misspecified CQF, misspecified weights

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 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 both the treatment and missing outcomes probabilities.
Figure A.3: Estimated CQTE with true CQTE as a function of $X_1$ for $N = 5,000$

a) $\tau = 0.25$

Case 3: Correct CQF, misspecified weights

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 $N_0 = 5,000 \times (1 - 0.41) \times 0.38 = 1,121$. 

27
Figure A.4: Empirical distribution of estimated UQTE for N=5,000

a) $\tau = 0.25$

Case 1: Correct weights

Case 2: Misspecified weights

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 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.
Figure A.5: Relative estimated bias in UQTE estimates at different quantiles of the 1979 earnings distribution

a) PSID-1 control group

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.

Table A.1: Covariate means and p-values from the test of equality of two means, by treatment status

| Covariates                  | Treatment | Control | P(|T| > |t|) | PSID-1 | P(|T| > |t|) | PSID-2 | P(|T| > |t|) |
|-----------------------------|-----------|---------|--------|--------|--------|--------|--------|
| Age in years                | 33.37     | 33.64   | 0.46   | 36.73  | 0.00   | 34.41  | 0.11   |
|                            (7.42) | (7.19)   |         |        | (10.60) | 0.00   | (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) | 0.00   | (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) | 0.00   | (0.49) |        |
| Proportion Married          | 0.02     | 0.04    | 0.03   | 0.02   | 0.05   | 0.01   | 0.08   |
|                            (0.15) | (0.20)   |         |        | (0.13) | 0.00   | (0.10) |        |
| Proportion Black            | 0.84     | 0.82    | 0.29   | 0.66   | 0.00   | 0.87   | 0.13   |
|                            (0.37) | (0.39)   |         |        | (0.47) | 0.00   | (0.34) |        |
| Proportion Hispanic         | 0.12     | 0.13    | 0.59   | 0.02   | 0.00   | 0.02   | 0.00   |
|                            (0.32) | (0.33)   |         |        | (0.12) | 0.00   | (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) | 0.00   | (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) | 0.00 | (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, column 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 A.2: Covariate means and p-values from the test of equality of two means for the observed and missing samples

| Covariates                      | Control Missing | Control Observed | Treatment Missing | Treatment Observed | PSID-1 Missing | PSID-1 Observed | PSID-2 Missing | PSID-2 Observed |
|---------------------------------|-----------------|------------------|-------------------|--------------------|----------------|----------------|----------------|----------------|
| Age                             | 33.36           | 33.74            | 0.51              |                    | 34.00          | 37.07          | 0.01           | 33.32          | 34.54          | 0.62           |
|                                 | (7.30)          | (7.15)           |                   |                    | (10.50)        | (10.57)        |                | (10.81)        | (9.34)         |                |
| Years of education              | 10.29           | 10.26            | 0.85              |                    | 11.44          | 11.30          | 0.60           | 11.05          | 10.49          | 0.18           |
|                                 | (1.93)          | (2.03)           |                   |                    | (2.17)         | (2.77)         |                | (1.73)         | (2.13)         |                |
| Proportion of high school dropouts | 0.70            | 0.68             | 0.57              |                    | 0.43           | 0.45           | 0.73           | 0.55           | 0.59           | 0.68           |
|                                 | (0.46)          | (0.47)           |                   |                    | (0.50)         | (0.50)         |                | (0.51)         | (0.49)         |                |
| Proportion married              | 0.05            | 0.04             | 0.61              |                    | 0.00           | 0.02           | 0.00           | 0.00           | 0.01           | 0.16           |
|                                 | (0.21)          | (0.19)           |                   |                    | (0.00)         | (0.14)         |                | (0.00)         | (0.10)         |                |
| Proportion black                | 0.81            | 0.82             | 0.81              |                    | 0.74           | 0.65           | 0.10           | 0.91           | 0.86           | 0.50           |
|                                 | (0.39)          | (0.39)           |                   |                    | (0.44)         | (0.48)         |                | (0.29)         | (0.35)         |                |
| Proportion hispanic             | 0.12            | 0.13             | 0.87              |                    | 0.01           | 0.02           | 0.82           | 0.05           | 0.02           | 0.62           |
|                                 | (0.33)          | (0.33)           |                   |                    | (0.11)         | (0.12)         |                | (0.21)         | (0.15)         |                |
| Number of children in 1975      | 2.33            | 2.23             | 0.34              |                    | 1.54           | 1.71           | 0.33           | 2.41           | 2.97           | 0.05           |
|                                 | (1.29)          | (1.34)           |                   |                    | (1.45)         | (1.78)         |                | (1.14)         | (1.79)         |                |
| Real earnings in 1975           | 621.54          | 879.28           | 0.12              |                    | 6927.95        | 7510.92        | 0.50           | 896.56         | 2211.45        | 0.02           |
|                                 | (1,523.00)      | (2,194.93)       |                   |                    | (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.
| Comparison group | Unadjusted | Post-training earnings estimates | Adjusted | Adjusted |
|-----------------|------------|---------------------------------|----------|----------|
|                 |            | Unweighted | PS-weighted | D-weighted | Unweighted | PS-weighted | D-weighted | Unweighted | PS-weighted | D-weighted |
| NSW             | 821        | 848        | 824        | 845        | 852        | 828        | 864        | 850        | 826        |
| N=1,185         | (307.22)   | (304.04)   | (304.61)   | (303.60)   | (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)   | (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)  | (664.46)   | (961.74)   | (953.80)   | (784.28)   | (1264.18)  | (1217.46)  |

### Bias estimates using NSW control

|                 | Pre-training earnings (1975) | Age | Age² | Education | High school dropout | Black | Hispanic | Marital status | Number of Children (1975) |
|-----------------|-------------------------------|-----|------|-----------|---------------------|-------|----------|-----------------|--------------------------|
| NSW             | ✓                             | ✓   | ✓    | ✓         | ✓                   | ✓     | ✓        | ✓               | ✓                        |
| N=1,185         |                               |     |      |           |                     |       |          |                 |                          |
| 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) |

### 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 which combines 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 appendix E.
Online Appendix

Akanksha Negi†

November 23, 2020

Abstract

In this online appendix, section A provides details of the simulation study. Section B discusses an extension of the doubly weighted framework to the case of estimating unconditional quantile treatment effects using recentered influence functions. Section C provides a simple extension to the case when treatment assumes multiple values. Section D provides the asymptotic variance expressions for the average treatment effect under the first and second half of asymptotic theory. Section E provides some background information on the National supported work demonstration along with augmenting Calhino and Smith (2017)’s sample for missing information and trimming rules for the probability weights. Section F contains proofs for results in the main text. Finally sections G and H provide supplementary tables and figures, respectively.

A Simulation details

This section outlines details of the simulation study for evaluating the finite sample behavior of unweighted, ps-weighted, and d-weighted (doubly weighted) estimators of ATE and QTE parameters. For each data generating process, the population is generated using a million observations. The empirical distributions of ATE and QTE estimands are simulated from drawing random vectors \( \{(Y_i, X_i, W_i, S_i); i = 1, 2, \ldots, N\} \) of size \( N \) a thousand times without replacement from the population. This is done to mimic the setting of “random sampling” from an infinite population.

A.1 Average treatment effect

To allow for possible misspecification of the regression functions \( \mathbb{E}[Y(g)|X] \), I simulate two binary potential outcomes generated using a probit as follows

\[
Y(g) = \begin{cases} 
1, & Y^*(g) > 0 \\
0, & Y^*(g) \leq 0 
\end{cases}
\]

\[
Y^*(g) = X\theta^0_g + U(g)
\]
Note that \( X \) here includes an intercept. The linear index, \( X\theta^0 \), is parameterized to have covariates be only mildly predictive of the potential outcomes with \( R^2_0 = 0.19 \) and \( R^2_1 = 0.14 \) in the population. The two covariates and the two latent errors are drawn from two independent bivariate normal distributions as follows,

\[
\begin{pmatrix}
X_1 \\
X_2
\end{pmatrix}
\sim N
\left(
\begin{pmatrix}
1 \\
2
\end{pmatrix},
\begin{pmatrix}
3 & 0.2 \\
0.2 & 2
\end{pmatrix}
\right)
\quad \text{and} \quad
\begin{pmatrix}
U(0) \\
U(1)
\end{pmatrix}
\sim N
\left(
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
1 & 0.2 \\
0.2 & 1
\end{pmatrix}
\right)
\] (A.1)

The assignment and missing outcome mechanisms have been simulated to ensure that unconfoundedness and MAR are satisfied

\[
W = \begin{cases}
1, & W^* > 0 \\
0, & W^* \leq 0
\end{cases}
\quad \text{and} \quad
S = \begin{cases}
1, & S^* > 0 \\
0, & S^* \leq 0
\end{cases}
\] (A.2)

where

\[
W^* = X\gamma_0 + \nu \quad \quad \quad S^* = Z\delta_0 + \upsilon
\]

with the errors \( \nu \) and \( \upsilon \) drawn from two independent standard logistic distributions. Misspecification in the true assignment and missing outcome distributions is allowed in both the functional form and linear index dimension where for the misspecified cases, I estimate a probit with \( X_1 \) omitted from the linear index. For scenarios where the conditional mean is misspecified, I estimate a linear model with a correct index. The parameters, \( \gamma_0 \) and \( \delta_0 \), indexing the assignment and missingness mechanisms have been chosen to ensure average propensity of assignment to be 41% and average propensity of being observed to be 38%. The missing data have been simulated to imitate empirical settings where a significant portion of the outcomes are missing. The following table gives an estimation summary for the eight different cases of misspecification,

| Scenario | CEF | \( G(\cdot) \) | \( R(\cdot) \) |
|----------|-----|----------------|----------------|
|          | Model | Estimation | Model | Estimation | Model | Estimation |
| 1        | M    | X\theta_0^g | C    | \Lambda(X\gamma_0) | C    | \Lambda(Z\delta_0) |
| 2        | M    | X\theta_0^g | M    | \Phi(X^{(1)}\gamma^{(1)}_0) | M    | \Phi(Z^{(1)}\gamma^{(1)}_0) |
| 3        | C    | \Phi(X\theta_0^g) | M    | \Phi(X^{(1)}\gamma^{(1)}_0) | M    | \Phi(Z^{(1)}\gamma^{(1)}_0) |

Notes: C and M correspond to whether the estimated model is correctly specified or misspecified. \( X \) and \( Z \) both include an intercept. \( X^{(1)} \) and \( Z^{(1)} \) are the subsets of \( X \) and \( Z \) left after omitting \( X_1 \). \( G(\cdot) \) refers to the propensity score model and \( R(\cdot) \) refers to the missing outcomes probability model.

1Here \( \theta_0^0 = (0, 1, 1)' \) and \( \theta_1^0 = (-1, 1, 1)' \). With cross-sectional data, covariates are typically seen to be mildly predictive of the outcome. For example, in the National Supported Work dataset from Calónico and Smith (2017), baseline factors explain about 26-50 percent of the variation in the non-experimental sample and about .04-.2 percent in the experimental sample depending upon the included subset of covariates.

2This implies that \( P(W = 1|X) \equiv p(X) = \Lambda(X\gamma_0) \) and \( P(S = 1|W, X) \equiv r(X, W) = \Lambda(Z\delta_0) \) where \( \Lambda(\cdot) \) is the standard logistic CDF.

3Here \( \gamma_0 = (0.05, -0.2, -0.11)' \), \( \delta_0 = (0.01, 0.03, 0.05, -0.28)' \) and \( Z = (1, W, X_1, X_2) \)
A.2 Quantile treatment effects

To ensure that the marginal quantiles of the potential outcome distributions are unique with no flat spots, I simulate two continuous non-negative outcomes as follows,

\[ Y(g) = \exp[ X\theta^0_g + U(g)] \]

where \( \theta^0_1 = (0.1, -0.36, -0.1)' \) and \( \theta^0_0 = (0.2, 0.24, -0.45)' \) are parameterized to ensure \( R^2_0 = 0.15 \) and \( R^2_1 = 0.13 \) in the population. The two covariates and the two latent errors are drawn from two independent normal distributions following (A.1). The missing outcomes and the treatment assignment mechanisms are also generated according to eq (A.2). Since \( \exp(\cdot) \) is an increasing continuous function, the equivariance property of quantiles implies that

\[ Q_\tau[ Y(g) | X] = Q_\tau[ \exp( X\theta^0_g + U(g)) | X] = \exp[ Q_\tau( X\theta^0_g + U(g)) | X] = \exp[ \exp( X\theta^0_g + Q_\tau(U(g)) | X] = \exp[ X\theta^0_g + \Phi^{-1}(\tau)] \]

where \( \Phi^{-1}(\tau) \) is the inverse standard normal CDF evaluated at \( \tau \). This equivariance property helps to characterize and estimate CQTE for cases when the CQF is correct. The three different cases of misspecification are enumerated in Table A.2 below. Case 1 corresponds to the situation for which results are derived in section 4, Case 2 allows for misspecification in both conditional quantile function and the probability weights. Even though the theory in this paper does not address that specific case, the simulation results show that the proposed estimator has the lowest bias among all three alternatives. Finally, case 3 relates to situations considered in section 5; correct CQF but misspecified weights.

| Scenario | CQF | G(·) | R(·) |
|----------|-----|------|------|
| 1        | \( M \) \( X\theta^0_g(\tau) \) | \( C \) \( \Lambda(X\gamma) \) | \( C \) \( \Lambda(Z\gamma) \) |
| 2        | \( M \) \( X\theta^0_g(\tau) \) | \( M \) \( \Phi(X^{(1)}\gamma^{(1)}) \) | \( M \) \( \Phi(X^{(1)}\gamma^{(1)}) \) |
| 3        | \( C \) \( \exp(X\theta^0_g(\tau)) \) | \( M \) \( \Phi(X^{(1)}\gamma^{(1)}) \) | \( M \) \( \Phi(X^{(1)}\gamma^{(1)}) \) |

Notes: C and M denote whether the estimated model is correctly specified or misspecified. \( X \) and \( Z \) both include an intercept. \( X^{(1)} \) and \( Z^{(1)} \) are the subsets of \( X \) and \( Z \) left after omitting \( X_1 \). Therefore, the probability models have been misspecified in both the functional form and the linear index dimension. \( G(·) \) refers to the propensity score model and \( R(·) \) refers to the missing outcomes probability model.

For plotting the estimated and true CQTE functions, I first collect the estimates that solve the unweighted, ps-weighted, and doubly weighted CQR problem (defined in (19)) corresponding to a particular quantile level, \( \tau = 0.25, 0.50, 0.75 \) across 1,000 Monte Carlo simulation draws. I then
draw a linearly spaced vector of values for $X_1$ and simulate the CQTE using the 1,000 estimated conditional quantile coefficients. Averaging these 1,000 functions at each point on the $X_1$ vector gives me the estimated average CQTE function. I plot this along with the 1,000 individual functions and the true CQTE, which is calculated using the population conditional quantile parameters, $\theta^0_g$.

### B Unconditional quantile treatment effect using recentered influence functions

This section discusses an alternative method of estimating UQTE using Firpo et al. (2009)’s (FFL, thereafter) recentered influence function (RIF) methodology.

Following FFL, let $v(F)$ be a real valued functional such that $v : \mathcal{F} \to \mathbb{R}$ whose domain $\mathcal{F}$ is a class of distribution functions such that $F \in \mathcal{F}$ if $|v(F)| < +\infty$. One may define $v(\cdot)$ to be any distributional statistic of interest like mean, variance, quantiles, inequality indices etc. We can define various treatment effects as the difference in the functionals of the marginal outcome distributions

$$\Delta_v = v_1 - v_0$$

(B.1)

where $v_g \equiv v(F_g)$ is the functional of the distribution function for $Y(g)$.\(^4\) As defined in FFL, the RIF is nothing but the influence function which has been centered at the statistic $v_g$. Formally,

$$\text{RIF}(Y(g); v, F_g) = v(F_g) + \text{IF}(Y(g); v, F_g)$$

(B.2)

where $\text{IF}(Y(g); v, F_g)$ captures the change in $v_g$ as a result of an infinitesimal change in the distribution of $X$. FFL introduce the idea of running a standard regression of RIF on $X$ with the objective of estimating the function

$$\mathbb{E} \left[ \text{RIF}(Y(g); v, F_g) \big| X \right] = X\theta^0_g$$

One can then use the law of iterated expectations to express $v_g$ in terms of the regression function as follows,

$$\mathbb{E} [\mathbb{E}(\text{RIF}(Y(g); v, F_g) | X)] = v_g$$

(B.3)

For $v_g = Q_{\tau,g}$, equation B.2 defines the UQTE for the $\tau^{th}$ quantile. We know that the RIF for $Q_{\tau,g}$ is given as:

$$\text{RIF}(Y(g); Q_{\tau}, F_g) = Q_{\tau,g} + \frac{\tau - 1\{Y(g) \leq Q_{\tau,g}\}}{f_g(Q_{\tau,g})}$$

(B.4)

where $f_g(\cdot)$ is the density of $Y(g)$.\(^5\) Then estimation of doubly weighted UQTE using RIFs involves the following steps:

---

\(^4\)Note that Firpo and Pinto (2016) use the above formulation to consider inequality treatment effects by exclusively considering $v$ to be different inequality measures.

\(^5\)Note that Firpo et al. (2009) express the conditional RIF expectation as $\mathbb{E} [\text{RIF}(Y(g); Q_{\tau}, F_g) | X] = c_{1,\tau,g} \cdot \mathbb{P}[Y(g) > Q_{\tau,g} | X] + c_{2,\tau,g}$ where $c_{1,\tau,g} = 1/f_g(Q_{\tau,g})$ and $c_{2,\tau,g} = Q_{\tau,g} - c_{1,\tau,g} \cdot (1 - \tau)$ for the $\tau^{th}$ quantile of $Y(g)$.
a.  \[ \hat{\theta}_g = \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{ig} X'_i X_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{ig} X'_i \cdot \widehat{\text{RIF}}(Y_i; \hat{Q}_r, \hat{F}_g) \right) \]

b.  \[ \widehat{\text{RIF}}(Y(g); \hat{Q}_r, \hat{F}_g) = \hat{Q}_{r,g} + \tau - 1 \left\{ Y(g) \leq \hat{Q}_{r,g} \right\} \frac{\hat{f}_g(\hat{Q}_{r,g})}{\hat{f}_g(\hat{Q}_{r,g})} \]

where \( \hat{f}_g(y) \) is the non-parametric kernel density estimator with bandwidth \( h_g \).

c.  \[ \hat{f}_g(\hat{Q}_{r,g}) = \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot \frac{1}{h_g} \cdot K_g \left( \frac{\hat{Q}_{r,g}}{h_g} \right) \]

d.  \[ \hat{Q}_{r,g} = \arg\min_{Q_g} \sum_{i=1}^{N} \hat{\omega}_{ig} \cdot c_r(Y_i - Q_g) \]

e.  \[ \hat{\omega}_{i1} = \frac{S_i \cdot W_i}{R(X_i, W_i, \hat{\delta}) \cdot G(X_i, \gamma_g)} \quad \text{and} \quad \hat{\omega}_0 = \frac{S_i \cdot (1 - W_i)}{R(X_i, W_i, \hat{\delta}) \cdot (1 - G(X_i, \gamma_g))} \]

where double weighting has to be performed at each stage that uses the observed sample. This implies that for ensuring consistency of UQTE, the weights would necessarily have to be correctly specified. One may estimate the weights nonparametrically using sieves to sidestep this issue of misspecification. Estimating UQTE in this manner also has the advantage initially put forth in FFL which is that one can directly estimate the effect of covariates on UQTE.

C  Multivalued Treatments

One can easily extend the binary treatment case considered here to the case when there are multiple treatment values. Let \( Y(g) \) denote the potential outcome for treatment level \( g \) where \( g = 0, 1, \ldots, T \) and \( W_g \) be a binary indicator for receiving treatment level \( g \) such that

\[ W_0 + W_1 + \ldots + W_T = 1 \]

\[ \mathbb{P}(W_g = 1) \equiv \rho_g > 0 \]

Also, let \( W = \left( W_0, W_1, \ldots, W_T \right) \). Then the observed outcome is

\[ Y = W_0 \cdot Y(0) + W_1 \cdot Y(1) + \ldots + W_T \cdot Y(T) \]

Let \( \rho_g(x) \equiv \mathbb{P}(W_g = 1|X = x) \) be the propensity score and \( r(x, w) \equiv \mathbb{P}(S = 1|X = x, W_g = w) \) be the missing outcomes probability for treatment level \( g \). One may then consider solving the same population problem, \( Q_0(\theta_0) \) but with true weights given as

\[ \omega_g = \frac{S \cdot W_g}{r(X, W_g) \cdot \rho_g(X)} \]

To construct the doubly weighted estimator, we would assume unconfoundedness and MAR along with assuming parametric models for the two probability weights; \( R(X, W_g, \delta) \) and \( G(X, \gamma_g) \).
D Asymptotic variance for ATE

Given \( \sqrt{N} \) consistent and asymptotically normal estimators, \( \hat{\theta}_1 \) and \( \hat{\theta}_0 \), the estimated average treatment effect

\[
\hat{\Delta}_{ate} = \frac{1}{N} \sum_{i=1}^{N} m(X, \hat{\theta}_1) - \frac{1}{N} \sum_{i=1}^{N} m(X, \hat{\theta}_0)
\]

is easily shown to also be \( \sqrt{N} \)-consistent and asymptotically normal [Wooldridge (2010) chapter 21]. Regularity conditions for such an asymptotic result would require that the parametric model, \( m(X, \theta_g) \), is continuously differentiable on the parameter space \( \Theta_g \subset \mathbb{R}^p \) and \( \theta_g^0 \) is in the interior of \( \Theta_g \). Then, by the continuous mapping theorem and Slutsky’s theorem,

\[
\sqrt{N} \left( \hat{\Delta}_{ate} - \Delta_{ate} \right) \xrightarrow{d} N(0, V)
\]

where \( V = \mathbb{E} \left[ \psi(X_i)\psi(X_i)’ \right] \). Let’s denote \( \mathbb{E} \left[ \nabla_{\theta_g} m(X_i, \theta_g^0) \right] = J_g^0 \), then

\[
\psi(X_i) = \left\{ m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \right\} - J_1^0 \cdot H_1^{-1} u_{1i} + J_0^0 \cdot H_0^{-1} u_{0i}
\]

where \( H_g \) is the Hessian for the treatment group \( g \), and \( u_{ig} \) is the residual from the regression of the weighted score on the scores of two probability models. For the case when the conditional mean model is correctly specified, the variance expression simplifies to

\[
V = \mathbb{E} \left[ (m(X_i, \theta_1^0) - m(X_i, \theta_0^0)) - \Delta_{ate} \right]^2 + J_1^0 \cdot V_1 \cdot J_1^0’ + J_0^0 \cdot V_0 \cdot J_0^0’ \tag{D.1}
\]

Here \( V_1 \) and \( V_0 \) are the asymptotic variances of the doubly weighted estimator that solve the treatment and control group problems, respectively. The above formula makes it clear that it better to use more efficient estimators of \( \hat{\theta}_g \). But we know from the results in section 5 that when the conditional mean model is correctly specified, using estimated weights is as efficient as using known weights. Another alternative in this case is to use unweighted estimators of \( \theta_g^0 \) since under GCIME, unweighted estimators is more efficient than the doubly weighted estimators of \( \theta_g^0 \).

For the case when the mean model is misspecified, the asymptotic variance of the ATE is given as follows

\[
V = \mathbb{E} \left[ (m(X_i, \theta_1^0) - m(X_i, \theta_0^0)) - \Delta_{ate} \right]^2 + J_1^0 \cdot V_1 \cdot J_1^0’ + G_1^0 \cdot V_0 \cdot J_0^0’ - 2 \mathbb{E} \left\{ m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \right\} u_{1i}’ \cdot H_1^{-1} J_1^0’ + 2 \mathbb{E} \left\{ m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \right\} u_{0i}’ \cdot H_0^{-1} J_0^0’ \tag{D.2}
\]

In this case, the variance expression is a bit more complicated than the previous case. Even though it is better to have more efficient estimators of \( \theta_g^0 \) in this case as well, it is not obvious whether that would help obtain a smaller variance for the ATE since we now have cross correlation terms in the variance expression.
D.1 Proofs

Asymptotic variance expression for ATE: Correctly specified mean model. Assuming continuous differentiability of \(m(X_i, \theta_g)\) on \(\Theta_g\), mean value expansion around \(\theta_g^0\) gives

\[
\frac{1}{N} \sum_{i=1}^{N} m(X_i, \hat{\theta}_g) \approx \frac{1}{N} \sum_{i=1}^{N} m(X_i, \theta_g^0) + \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta_g} m(X_i, \tilde{\theta}_g) \cdot (\hat{\theta}_g - \theta_g^0)
\]

where \(\tilde{\theta}_g\) lies between \(\hat{\theta}_g\) and \(\theta_g^0\). Since \(\hat{\theta}_g \overset{p}{\to} \theta_g^0\), so does \(\tilde{\theta}_g\). Hence, using the weak law of large numbers, we obtain

\[
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} m(X_i, \hat{\theta}_g) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} m(X_i, \theta_g^0) + J_g^0 \cdot \sqrt{N}(\hat{\theta}_g - \theta_g^0) + o_p(1)
\]

Adding and subtracting \(\sqrt{N} \cdot \mathbb{E}[m(X_i, \theta_g^0)]\) on both sides gives us

\[
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left( m(X_i, \hat{\theta}_g) - \mathbb{E}[m(X_i, \theta_g^0)] \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left( m(X_i, \theta_g^0) - \mathbb{E}[m(X_i, \theta_g^0)] \right) + J_g^0 \cdot \sqrt{N}(\hat{\theta}_g - \theta_g^0) + o_p(1)
\]

Then, using the asymptotic results from section 5, where we posit that the conditional feature of interest is correctly specified, we have

\[
\sqrt{N} \left( \hat{\theta}_1 - \theta_1^0 \right) = -H_1^{-1} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} l_{i1} \right\} + o_p(1)
\]

\[
\sqrt{N} \left( \hat{\theta}_0 - \theta_0^0 \right) = -H_0^{-1} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} l_{i0} \right\} + o_p(1)
\]

Therefore,

\[
\sqrt{N} \left( \hat{\Delta}_{ate} - \Delta_{ate} \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left( m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \right) - J_1^0 \cdot H_1^{-1} l_{i1} + J_0^0 \cdot H_0^{-1} l_{i0} + o_p(1)
\]

We may rewrite the above using the influence function representation as

\[
\sqrt{N} \left( \hat{\Delta}_{ate} - \Delta_{ate} \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(X_i) + o_p(1) \text{ where } \mathbb{E} \left[ \psi(X_i) \right] = 0
\]

Then, provided that \(\mathbb{E} \left[ \psi(X_i)\psi(X_i)' \right]\) exists,

\[
\text{Avar} \left[ \sqrt{N}(\hat{\Delta}_{ate} - \Delta_{ate}) \right] = \mathbb{E} \left[ \left( m(X_i, \theta_1^0) - m(X_i, \theta_0^0) \right) - \Delta_{ate} \right]^2 + J_1^0 \cdot V_1 \cdot J_1^0 + J_0^0 \cdot V_0 \cdot J_0^0'
\]

Note that the covariance term involving \(l_{ig}\) is zero since they denote scores for the treatment and control group problems. The covariance terms involving \(\{ m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \}\) and \(l_{ig}\)
will also be zero. This is because $\theta_0^0$ solves the conditional problem. However, using that fact that $\mathbb{E}[h(Y_i(g), X_i, \theta_0^0)|X_i] = 0$ along with LIE, those covariance terms can be shown to be zero.

**Misspecified mean model**  In the case of a misspecified mean model, we still have

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ m(X_i, \hat{\theta}_g) - \mathbb{E} \left( m(X_i, \theta_0^0) \right) \right\} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ m(X_i, \theta_0^0) - \mathbb{E}[m(X_i, \theta_0^0)] \right\} + J_0^0 \cdot \sqrt{N}(\hat{\theta}_g - \theta_0^0) + o_p(1)$$

Now using results from section 4

$$\sqrt{N} \left( \hat{\theta}_1 - \theta_1^0 \right) = -H_1^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ l_{i1} - \mathbb{E} \left( l_{i1} b_i' \right) \mathbb{E} \left( b_i b_i' \right)^{-1} b_i - \mathbb{E}(l_{i1} d_i')\mathbb{E}(d_i d_i')^{-1}d_i \right\} + o_p(1)$$

$$= -H_1^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} u_{i1} + o_p(1)$$

$$\sqrt{N} \left( \hat{\theta}_0 - \theta_0^0 \right) = -H_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ l_{i0} - \mathbb{E} \left( l_{i0} b_i' \right) \mathbb{E} \left( b_i b_i' \right)^{-1} b_i - \mathbb{E}(l_{i0} d_i')\mathbb{E}(d_i d_i')^{-1}d_i \right\} + o_p(1)$$

$$= -H_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} u_{i0} + o_p(1)$$

Then,

$$\sqrt{N} \left( \hat{\Delta}_{ate} - \Delta_{ate} \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left( \left\{ m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \right\} + J_0^0 \cdot H_1^{-1} u_{i1} \right) + o_p(1)$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(X_i) + o_p(1)$$

Then,

$$\text{Avar} \left[ \sqrt{N} \left( \hat{\Delta}_{ate} - \Delta_{ate} \right) \right] = \mathbb{E} \left[ \left( m(X_i, \theta_1^0) - m(X_i, \theta_0^0) \right) - \Delta_{ate} \right]^2 + J_1^0 \cdot V_1 \cdot J_1^{0'} + J_0^0 \cdot V_0 \cdot J_0^{0'}$$

$$- 2\mathbb{E} \left[ \left\{ m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \right\} u_{i1}' \right] H_1^{-1} J_1^{0'}$$

$$+ 2\mathbb{E} \left[ \left\{ m(X_i, \theta_1^0) - m(X_i, \theta_0^0) - \Delta_{ate} \right\} u_{i0}' \right] H_0^{-1} J_0^{0'}$$

\]

**D.2  Practical advice for obtaining doubly weighted ATE estimates**

An easy way to obtain the doubly weighted estimates, $\hat{\theta}_g$, for estimating ATE, is to combine the treatment and control group problems into a one-step GMM procedure. Essentially, this means that one would stack the moment conditions from the first and second steps, which can then be
solved jointly via GMM. Since there are no over-identifying restrictions in the doubly weighted framework, one-step estimation of $\theta_0^g$ is equivalent to two-step estimation. Then, suppressing explicit dependence on data,

$$m(\theta_0, \theta_1, \gamma, \delta) = \frac{1}{N} \sum_{i=1}^{N} m_i(\theta_0, \theta_1, \gamma, \delta) = N^{-1} \left( \frac{N}{N_0} \sum_{i=1}^{N} m_{i_0}(\theta_0, \gamma, \delta) \right)$$

$$= \frac{N}{N_0} \sum_{i=1}^{N} m_{i_1}(\theta_1, \gamma, \delta)$$

$$= \frac{N}{N_0} \sum_{i=1}^{N} m_{i_2}(\gamma)$$

$$= \frac{N}{N_0} \sum_{i=1}^{N} m_{i_3}(\delta)$$

where,

$$m_{i_0}(\theta_0, \gamma, \delta) = \frac{S_i \cdot (1 - W_i)}{R(X_i, W_i, \hat{\delta}) \cdot (1 - G(X_i, \hat{\gamma}))} \cdot \nabla_{\theta_0} q(Y_i(0), X_i, \theta_0)'$$

$$m_{i_1}(\theta_1, \gamma, \delta) = \frac{S_i \cdot W_i}{R(X_i, W_i, \hat{\delta}) \cdot G(X_i, \hat{\gamma})} \cdot \nabla_{\theta_1} q(Y_i(1), X_i, \theta_1)'$$

$$m_{i_2}(\gamma) = \nabla_{\gamma} G(X_i, \gamma)' \cdot \frac{W_i - G(X_i, \gamma)}{G(X_i, \gamma) \cdot (1 - G(X_i, \gamma))}$$

$$m_{i_3}(\delta) = \nabla_{\delta} R(X_i, W_i, \delta)' \cdot \frac{S_i - R(X_i, W_i, \delta)}{R(X_i, W_i, \delta) \cdot (1 - R(X_i, W_i, \delta))}$$

The example code below uses STATA’s `gmm` command to estimate the doubly weighted ATE estimate.

**Example code using STATA’s `gmm`**

```
local Rhat="exp(b31+b32*w+b33*x1+b34*x2)/(1+exp(b31+b32*w+b33*x1+b34*x2))"
local Ghat="exp(b21+b22*x1+b23*x2)/(1+exp(b21+b22*x1+b23*x2))"
gmm ((-2*s*(1-w)/('Rhat'*(1-'Ghat')))*(y-b00-b01*x1-b02*x2)*(n/nc)) ///
((-2*s*w/('Rhat'*'Ghat')))*(y-b10-b11*x1-b12*x2)*(n/nt)) ///
(w-exp(b21+b22*x1+b23*x2)/(1+exp(b21+b22*x1+b23*x2)) )///
(s-exp(b31+b32*w+b33*x1+b34*x2)/(1+exp(b31+b32*w+b33*x1+b34*x2))), ///
imstruments(1 2 3: x1 x2) instruments(4: w x1 x2) winitial(identity) ///
nocommonesample onestep from(b00 0.1 b01 0.1 b02 0.1 b10 0.1 b11 0.1 b12 ///
0.1 b21 0.1 b22 0.1 b23 0.1 b31 0.1 b32 0.1 b33 0.1 b34 0.1)
```

Then using the GMM estimates, one can estimate the average treatment effect as

```
gen y0hat = _b[b00: _cons]+_b[b01: _cons]*x1+_b[b02: _cons]*x2
gen y1hat = _b[b10: _cons]+_b[b11: _cons]*x1+_b[b12: _cons]*x2
egen ate = mean(y1hat-y0hat)
```

Since I am estimating the two probability models as logits, the last two moments simplify to
\[m_{i2}(\gamma) = X_i' \cdot (W_i - \Lambda(X_i\gamma))\]
\[m_{i3}(\delta) = Z_i' \cdot (S_i - \Lambda(Z_i\delta))\]

where \(Z_i \equiv (X_i, W_i)\). Even though this one-step estimation allows us to obtain variance estimates \(\hat{\sigma}_1^2\) and \(\hat{\sigma}_0^2\) for \(\hat{\theta}_1\) and \(\hat{\theta}_0\) respectively, obtaining analytically correct standard errors for estimated ATE requires additional work. A command that implements the correct standard errors is still in the works. Meanwhile, one can use bootstrapped standard errors, which provide asymptotically correct inference.

\[\text{E Appendix to CS (2017) Application}\]

\textbf{E.1 Description of National Supported Work Program}

The NSW was a transitional and subsidized work experience program that was mainly intended to target four sub-populations; ex-offenders, former drug addicts, women on AFDC welfare and high school dropouts.\(^6\) The program became operational in 1975 and continued until 1979 at fifteen locations in the United States. In ten of these sites, the program operated as a randomized experiment where individuals who qualified for the training program were randomly assigned to either the treatment or control group.\(^7\) At the time of enrollment in April 1975, individuals were given a retrospective baseline survey which was then followed by four follow-up interviews conducted at nine month intervals each. The survey data was collected using these baseline and follow-up interviews over a period of four years. The data included measurement on baseline covariates like age, years of education, number of children in 1975, high school dropout status, marital status, two race indicators for black and Hispanic sub-populations and other demographic and socio-economic information. The main outcome of interest was real earnings for the post-training year of 1979.

\textbf{E.2 Augmenting the CS sample to account for missing earnings in 1979}

I obtain the data from CS’s supplementary data files in the Journal of Labor Economics where the authors recreate the experimental sample on AFDC women using the raw public use data files maintained by the Inter-University Consortium for Political and Social Research (ICPSR). Then, I use the PSIDcross file provided by CS along with other supplementary data files to add back the individuals whom CS originally dropped from the analysis for not having valid earnings information between 1975-1979. For this, I apply the same filters applied by CS who use them to match their PSID samples to the ones used by LaLonde (1986). These filters involve keeping all female household heads continuously from 1975-1979 who were between 20 and 55 years of age, \(\text{AFDC program is administered and funded by the federal and state governments and is meant to provide financial assistance to needy families. Source: US Census Bureau. Beyond the main eligibility criteria that was applied to all four target populations, the AFDC group was subjected to two additional criteria which were, a) no child below 6 years of age and b) on AFDC welfare for at least 30 of the last 36 months.}\)

\(^6\)Out of the 10 sites, 7 served AFDC women with random assignment at one or more of these sites in operation from Feb 1976-Aug 1977 (CS (2017)).
This constitutes the first non-experimental sample that CS use in their analysis, which they call the PSID-1 sample. The second PSID sample, which they label PSID-2 further restricts the PSID-1 sample to include only those women who received AFDC welfare in 1975.\textsuperscript{9} In order to compare my sample with the original sample used by CS, I first apply all the above mentioned filters and create a dummy variable which I call “cs”. Next, I remove the filter which requires the women to be continuous household heads and instead only impose that filter for 1975 and 1976. The reason this filter is imposed for both years 1975 and 1976 but not for any other years is because in the PSID datasets, the income information in a particular year corresponds to the previous calendar year. This ensures that merging the cross-file with the separate single-year files for 1975 and 1976 guarantee that only those women are included who do not have any missing earnings information for the pre-training year of 1974 and 1975. This is important since pre-training earnings are treated as any other baseline covariate in this paper, on which I do not allow any missing information.

After merging cross year individual file with the single year family files, I then merge this PSID dataset with the NSW dataset using CS’s .do files and generate the various sample dummies essentially in the same manner as they do. After this, I further restrict the sample to include only those women who have valid earnings information in 1975, which is the pre-training year for AFDC women. I also drop the cases where the measured age or education is less than zero. In order to make sure that any observations not used by CS only correspond to the ones that have missing post-program earnings, I also drop observations that do not satisfy the CS criteria but have observed earnings in 1979.

\textbf{E.3 Treatment and missing outcome probability specifications and sample trimming}

In this application, I estimate three sets of treatment assignment and missing outcomes probability models depending upon which comparison group is used for obtaining the estimates. For the experimental estimates, I use the experimental treatment and control groups to estimate the propensity score model. For the PSID-1 estimates, I consider the NSW experimental observations to be the treatment group and use PSID-1 as the control group. For estimating the PSID-2 propensity score model, I switch to PSID-2 as being the comparison control group. For estimating the missing outcome probability models, I include the treatment indicator depending upon the comparison group as mentioned above. The probability models are estimated as logits and include the following covariates in their specification. For the treatment probability, I include the real earnings in 1974 and 1975 along with an indicator variable for whether the individual had any zero earnings in 1974 and 1975. Beyond these, I also include Age, Age-squared, Education, High school dropout status, the race indicators of black and Hispanic along as well as the number of children in 1975. CS also add some interaction terms in their propensity score specification which I do not. I noticed that allowing for those terms in my specifications drove the final weights for many women in the sample.

\textsuperscript{8} For the additional filters that CS impose, see their supplementary material provided in JLE.

\textsuperscript{9} Even though the two PSID comparison groups are not perfectly representative of women who would have proven eligible for NSW, there is no clear alternative since the PSID data lacks detailed covariate information that would be needed to impose the full eligibility criteria on the PSID sample.
too close to a 0 or 1. For the missing outcomes probability, I include the treatment indicator along with the same covariates. I kept the specifications to be the same for the three sets of probabilities I estimated. However, my regression specifications include the same covariates as CS to allow for some comparison across the analyses. These comparisons should be made with some caution. Except the estimates that use the NSW control group, all other estimates are obtained using samples that are different than the CS samples.

The final sample used to obtain estimates for the PSID-1 comparison group is trimmed in order to ensure common support for the weights in the treatment and comparison groups. For the PSID-1 group, this meant dropping observations with final weight either less than 0.03 or greater than 0.8. For the PSID-2 sample, this meant dropping observations with final weight that was either less than 0.1 or greater than 0.86. These final weights are the weights that are specified in the regression commands in Stata and are constructed as follows:

\[ \text{weight} = \left( \frac{w}{Ghat} + \frac{1-w}{1-Ghat} \right) \times \frac{s}{Rhat} \]

The trimming threshold for ps-weighted estimates is kept the same as for computing the doubly weighted estimates since the overlap problem was relatively more severe when using the composite weights than when using propensity scores only. The graphs below plot the kernel density for the probabilities \( Rhat \times Ghat \) for the treatment group and \( Rhat \times (1-Ghat) \) for the control group. The common support problem due to which the samples were appropriately trimmed can be seen in figure E.1.

Additionally, figures E.2 and E.3 plot the estimated distributions for the propensity score and missing outcomes probability, where panel (a)-(c) display these for the three treatment and comparison group combinations. A couple of points emerge from the estimated graphs. For figure E.2, panel (a), we see that the treatment and control distributions appear very similar, confirming the strong role of randomization in producing groups that are balanced in terms of covariates. For panel (b), we see that the experimental observations have a relatively high probability of being treated whereas the control group have low probabilities. Note, however, that the common support condition holds quite strongly for the PSID-1 group. In panel (c), while the estimated distribution for the treated units still has a higher mean, the PSID-2 comparison group distribution is relatively similar than PSID-1 in panel (b). These findings suggest that nonrandom assignment is predicted well by the covariates in the propensity score distributions. The same cannot be said for the estimated missing outcomes probabilities where panel (b) and (c) reveal a strong overlap problem. Moreover, we see that the treated units are less likely to be missing outcomes compared to the comparison groups.
Figure E.1: Kernel density plots for the composite probability

a) Experimental treatment and control groups
b) Experimental treatment and PSID-1 group
c) Experimental treatment and PSID-2 group

Notes: The weights here correspond to the product of the estimated assignment and missing outcomes probabilities. Following CS (2017), I exploit the efficiency gain from combining the experimental treatment and control groups for estimating the treatment and missing outcome probability models. For the PSID-1 group, this means using the full experimental group to be the treatment group and the PSID-1 as the control group. Similarly, to construct weights for the PSID-2 group, this means using the full experimental group along with the PSID-2 as the control group.
Figure E.2: Kernel density plots for the estimated propensity score

a) Experimental treatment and control groups

b) Experimental treatment and PSID-1 group

c) Experimental treatment and PSID-2 group

Notes: Following CS (2017), I exploit the efficiency gains from combining the experimental treatment and control groups for estimating the propensity scores. For the PSID-1 group, this means using the full experimental group to be the treatment group and the PSID-1 as the control group. Similarly, to construct weights for the PSID-2 group, this means using the full experimental group along with the PSID-2 as the control group.
Figure E.3: Kernel density plots for the estimated missing outcomes probability

a) Experimental treatment and control groups

b) Experimental treatment and PSID-1 group

c) Experimental treatment and PSID-2 group

Notes: Following CS (2017), I exploit the efficiency gains from combining the experimental treatment and control groups for estimating the missing outcome probability. For the PSID-1 group, this means using the full experimental group to be the treatment group and the PSID-1 as the control group. Similarly, to construct weights for the PSID-2 group, this means using the full experimental group along with the PSID-2 as the control group.
F Proofs

Proof of Lemma 1. Let us first consider the argument for \( \theta_0^0 \). By LIE and using the fact that 
\[ q(Y, X, \theta) = W \cdot q(Y(1), X, \theta_1) + (1 - W) \cdot q(Y(0), X, \theta_0) \]  
we can write,
\[
E [\omega_1 \cdot q(Y, X, \theta)] = E \left[ E \left( \frac{S}{r(X, W)} \cdot \frac{W}{p(X)} \cdot q(Y(1), X, \theta_1) \right| Y(1), X, W \right] 
\]
\[
= E \left[ \frac{W}{r(X, W) \cdot p(X)} \cdot q(Y(1), X, \theta_1) \cdot \mathbb{P}(S = 1|Y(1), X, W) \right] 
\]
\[
= E \left[ \frac{W}{r(X, W) \cdot p(X)} \cdot q(Y(1), X, \theta_1) \cdot \mathbb{P}(S = 1|X, W) \right] 
\]
Using another application of LIE along with unconfoundedness, we obtain
\[
E \left[ \frac{W}{p(X)} \cdot q(Y(1), X, \theta_1) \right] = E [q(Y(1), X, \theta_1)]
\]
where the third equality follows from MAR and fourth follows from part ii) of Assumption 3. The proof for \( \theta_0^0 \) follows analogously. \( \square \)

Proof of Theorem 1. It has already been established that
\[
E [\omega_g \cdot q(Y, X, \theta)] = E [\omega_g \cdot q(Y(g), X, \theta_g)] = E [q(Y(g), X, \theta_g)]
\]
for both \( g = 0, 1 \). By iii) \( \omega_g(\gamma, \delta) \) is continuous in \( \gamma \) and \( \delta \) and is bounded in absolute value by Assumptions 4 and 5. Moreover, \( \omega_g(\gamma, \delta)q(\gamma, \theta) \) is continuous with probability one. Then, along with v), DCT, and boundedness of \( \omega_g(\gamma, \delta) \) we obtain,
\[
\sup_{(\theta_g, \gamma, \delta) \in (\Theta_g, \Gamma, \Delta)} \left| \frac{1}{N} \sum_{i=1}^{N} \omega_{ig}(\gamma, \delta) \cdot q(Y_i(g), X_i, \theta_g) - E [\omega_g(\gamma, \delta) \cdot q(Y(g), X, \theta_g)] \right| \xrightarrow{p} 0 \quad (F.1)
\]
by Lemma 2.4 in Newey and McFadden (1994).\(^{10} \) Then, by triangle inequality,
\[
\sup_{\theta_g \in \Theta_g} \left| \frac{1}{N} \sum_{i=1}^{N} \tilde{\omega}_{ig} \cdot q(Y_i(g), X_i, \theta_g) - E [\omega_g \cdot q(Y(g), X, \theta_g)] \right| 
\]
\[
\leq \sup_{\theta_g \in \Theta_g} \left| \frac{1}{N} \sum_{i=1}^{N} \tilde{\omega}_{ig} \cdot q(Y_i(g), X_i, \theta_g) - E [\tilde{\omega}_g \cdot q(Y(g), X, \theta_g)] \right| 
\]
\[
+ \sup_{\theta_g \in \Theta_g} \left| E [\tilde{\omega}_g \cdot q(Y(g), X, \theta_g)] - E [\omega_g \cdot q(Y(g), X, \theta_g)] \right| \quad (F.2)
\]
\[ (A.2) \text{ is } o_p(1) \text{ because of (A.1). (A.3) is } o_p(1) \text{ due to } \hat{\gamma} \overset{p}{\rightarrow} \gamma_0, \hat{\delta} \overset{p}{\rightarrow} \delta_0 \text{ and uniform continuity of } E [\omega_g \cdot q(Y(g), X, \delta_g)] \text{ on } \Theta_g \times \hat{\Gamma} \times \hat{\Delta}. \text{ Then consistency of } \hat{\theta}_g \text{ for } \theta_g^0 \text{ follows from Theorem 2.1} \]
\[ \text{\cite{NeweyMcFadden94}} \]

\(^{10}\)\( \hat{\Gamma} \) and \( \hat{\Delta} \) are compact neighborhoods around \( \gamma_0 \) and \( \delta_0 \).
Proof of Theorem 2. Explicit dependence on data is suppressed for notational simplicity. Then expanding \( \tilde{\omega}_{ig} \) around \( \omega_{ig} \),

\[
\tilde{\omega}_{ig} \approx \omega_{ig} - \bar{\omega}_{ig} b_i' (\delta) \cdot (\delta - \delta_0) - \bar{\omega}_{ig} d_i' (\gamma) \cdot (\gamma - \gamma_0)
\]

where \( \tilde{\delta} \) lies between \( \delta \) and \( \delta_0 \) and \( \tilde{\gamma} \) lies between \( \gamma \) and \( \gamma_0 \). Then, consider

\[
N^{-1/2} \sum_{i=1}^{N} \tilde{\omega}_{ig} \cdot h_{ig}
\]

\[
= N^{-1/2} \sum_{i=1}^{N} \left\{ \omega_{ig} h_{ig} - \bar{\omega}_{ig} h_{ig} \cdot b_i' (\delta) \cdot (\delta - \delta_0) - \bar{\omega}_{ig} h_{ig} \cdot d_i' (\gamma) \cdot (\gamma - \gamma_0) \right\}
\]

\[
= N^{-1/2} \sum_{i=1}^{N} \omega_{ig} h_{ig} - N^{-1} \sum_{i=1}^{N} \bar{\omega}_{ig} h_{ig} b_i' (\delta) \cdot \sqrt{N} (\tilde{\delta} - \delta_0) - N^{-1} \sum_{i=1}^{N} \bar{\omega}_{ig} h_{ig} d_i' (\gamma) \cdot \sqrt{N} (\tilde{\gamma} - \gamma_0)
\]

Now let, \( (\theta^*_g, \delta^*) = \arg \sup_{\theta_g \in \Theta_g, \delta \in \Delta} \| h(\theta_g) \cdot b'(\delta) \| \). Then,

\[
(E\|h(\theta^*_g) b'(\delta^*)\|^2) \leq E\|h(\theta^*_g)\|^2 E\|b'(\delta^*)\|^2 \leq E \left( \sup_{\theta_g \in \Theta_g} \| h(\theta_g) \|^2 \right) E \left( \sup_{\theta_g \in \Theta_g} \| b'(\delta) \|^2 \right) < \infty \tag{F.4}
\]

where first inequality holds by cauchy-schwartz, second holds due to the definition of supremums, and third by conditions iv) and vi). Then,

\[
E \left( \sup_{\theta_g \in \Theta_g, \delta \in \Delta} \| h(\theta_g) b'(\delta) \| \right) \leq \left( E \left( \sup_{\theta_g \in \Theta_g} \| h(\theta_g) b'(\delta) \| \right) \right)^2 < \infty
\]

where the first inequality holds trivially and second inequality holds because of (F.4). An analogous argument may be made for showing \( E \left( \sup_{\theta_g \in \Theta_g, \gamma \in \Gamma} \| h(\theta_g) d'(\gamma) \| \right) < \infty \). Using the fact that \( \omega_g (\gamma, \delta) \) is continuous and bounded along with continuity of \( 1(\theta_g) \) (condition ii)), \( b(\delta) \), \( d(\gamma) \) (condition iii) of theorem 1), we obtain

\[
\frac{1}{N} \sum_{i=1}^{N} \tilde{\omega}_{ig} h_{ig} b_i' (\delta) = E \left[ \omega_{ig} h_{ig} b_i' \right] + o_p(1) \tag{F.5}
\]

\[
\frac{1}{N} \sum_{i=1}^{N} \tilde{\omega}_{ig} h_{ig} d_i' (\gamma) = E \left[ \omega_{ig} h_{ig} d_i' \right] + o_p(1)
\]

using Lemma 4.3 in Newey and McFadden (1994) as \( \gamma \to p \gamma_0 \) and \( \tilde{\delta} \to p \delta_0 \). Rewriting (7) using
influence function representations for \( \hat{\gamma} \) and \( \hat{\delta} \) along with (F.5)

\[
N^{-1/2} \sum_{i=1}^{N} \hat{\omega}_{ig} h_{ig} = N^{-1/2} \sum_{i=1}^{N} \left\{ l_{ig} - \mathbb{E} \left[ l_{ig} b'_i \right] \cdot \mathbb{E} \left[ b_i b'_i \right]^{-1} b_i - \mathbb{E} \left[ l_{ig} d'_i \right] \cdot \mathbb{E} \left[ d_i d'_i \right]^{-1} d_i \right\} + o_p(1)
\]

\[
\equiv N^{-1/2} \sum_{i=1}^{N} u_{ig} + o_p(1)
\]

\[
\overset{d}{\to} N(0, \Omega_g)
\] (F.6)

where \( u_{ig} \equiv l_{ig} - \mathbb{E} \left[ l_{ig} b'_i \right] \cdot \mathbb{E} \left[ b_i b'_i \right]^{-1} b_i - \mathbb{E} \left[ l_{ig} d'_i \right] \cdot \mathbb{E} \left[ d_i d'_i \right]^{-1} d_i \). Since \( \mathbb{E}(u_{ig}) = 0 \),

\[
\Omega_g = \mathbb{E} \left( l_{ig} l'_{ig} \right) - \mathbb{E} \left( l_{ig} b'_i \right) \mathbb{E} \left( b_i b'_i \right)^{-1} \mathbb{E} \left( b_i l'_{ig} \right) - \mathbb{E} \left( l_{ig} d'_i \right) \mathbb{E} \left( d_i d'_i \right)^{-1} \mathbb{E} \left( d_i l'_{ig} \right)
\]

Next part of the proof uses the theory of empirical processes for obtaining asymptotic normality of the doubly weighted estimator. Using the definition in (11) along with the fact that \( \mathbb{E}[\hat{\omega}_{ig} h_i(\theta_g)] \overset{d}{\to} \mathbb{E}[\omega_{ig} h_i(\theta_g)] \) (by continuity of \( \omega(\gamma, \delta) h(\theta_g) \), condition iv) and DCT as \( (\hat{\gamma}, \hat{\delta}) \overset{p}{\to} (\gamma_0, \delta_0) \), rewrite

\[
v_N(\theta_g) = v^*_N(\theta_g) + o_p(1)
\] (F.7)

where \( v^*_N(\theta_g) \equiv \frac{1}{N} \sum_{i=1}^{N} \left\{ \hat{\omega}_{ig} h_i(\theta_g) - \mathbb{E} \left[ \omega_{ig} h_i(\theta_g) \right] \right\} \). Let

\[
\hat{m}_N(\theta_g) = \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{ig} h_i(\theta_g)
\]

\[
m^*_N(\theta_g) = \mathbb{E} \left[ \omega_{ig} h_i(\theta_g) \right]
\]

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

\[
0 = \sqrt{N} m^*_N(\theta^0_g) = \sqrt{N} m^*_N(\hat{\theta}_g) - \nabla_{\theta_g} m^*_N(\hat{\theta}_g) \cdot \sqrt{N}(\hat{\theta}_g - \theta^0_g)
\]

where \( \hat{\theta}_g \) lies between \( \hat{\theta}_g \) and \( \theta^0_g \). Since the population first order condition is zero at the truth

\[
0 = \nabla_{\theta_g} \mathbb{E} \left[ \omega_g \cdot q(Y(g), X, \theta^0_g) \right]
\]

\[
= \mathbb{E} \left[ \omega_g \cdot h(Y(g), X, \theta^0_g) \right] \equiv m^*_N(\theta^0_g)
\]

The second equality follows from dominance condition iv) and application of Lemma 3.6 in Newey and McFadden (1994). Then, by the continuity of \( \nabla_{\theta_g} \mathbb{E}[\omega_{ig} h_i(\theta_g)] \) (condition vii)

\[
\nabla_{\theta_g} m^*_N(\hat{\theta}_g) \overset{p}{\to} H_g
\]

By continuous mapping theorem and condition viii),

\[
\sqrt{N}(\hat{\theta}_g - \theta^0_g) = (H^{-1}_g + o_p(1)) \cdot \sqrt{N} m^*_N(\hat{\theta}_g)
\] (F.8)
Consider,

\[-\sqrt{N}m_N^*(\hat{\theta}_g) = v_N^*(\hat{\theta}_g) - \sqrt{N}m_N^*(\hat{\theta}_g)\]

\[= v_N^*(\hat{\theta}_g) - v_N^*(\theta_g^0) + v_N^*(\theta_g^0) - \sqrt{N}m_N^*(\hat{\theta}_g)\]

\[= v_N^*(\theta_g^0) + o_p(1)\]

since \(v_N^*(\hat{\theta}_g) - v_N^*(\theta_g^0) = o_p(1)\) by asymptotic equivalence in (F.7) and stochastic equicontinuity by condition ix). Moreover, \(\sqrt{N}m_N^*(\hat{\theta}_g) = o_p(1)\) by condition iii). Therefore,

\[v_N^*(\theta_g^0) = \frac{1}{N} \sum_{i=1}^{N} \hat{w}_{ig} h_{ig} \overset{d}{\rightarrow} N(0, \Omega_g)\]

by (F.6). Then using (F.8) along with Slutsky’s theorem, \(\sqrt{N}(\hat{\theta}_g - \theta_g^0) \overset{d}{\rightarrow} N(0, H_g^{-1}\Omega_g H_g^{-1})\).

\[\Box\]

**Proof of Corollary 1.** Consider,

\[\Sigma_g - \Omega_g = \mathbb{E}(1_{ig}1_{ig}') - \mathbb{E}(1_{ig}b_i') \mathbb{E}(b_ib_i')^{-1} \mathbb{E}(b_i1_{ig}') - \mathbb{E}(1_{ig}d_i') \mathbb{E}(d_i1_{ig}')\]

\[= \mathbb{E}(1_{ig}b_i') \mathbb{E}(b_ib_i')^{-1} \mathbb{E}(b_i1_{ig}') + \mathbb{E}(1_{ig}d_i') \mathbb{E}(d_i1_{ig}')\]

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

\[\Box\]

**Proof of Theorem 3.** It has already been established that \(\theta_g^0\) solves

\[\mathbb{E} \left[ B_g^* \cdot q(Y(g), X, \theta_g) \right]\]

The proof of uniform convergence follows similar to the proof of theorem 1 where we replace \(\omega_g\) by \(\omega_g^*\). Then, consistency of \(\hat{\theta}_g\) for \(\theta_g^0\) follows from Theorem 2.1 in Newey and McFadden (1994).

\[\Box\]

**Proof of Theorem 4.** The proof follows in the manner of Theorem 2 where we replace \(\omega_g\) by \(\omega_g^*\). Also, \(\Omega_g\) now denotes the variance of the score of the objective function, \(1_{ig}\), without the first stage adjustment for the estimated weights. This is because, \(\mathbb{E}(1_{ig}b_i') = \mathbb{E}(1_{ig}d_i') = 0\) because the conditional score of \(1_{ig}\), \(\mathbb{E}[h(Y(g), X, \theta_g^0)|X] = 0\) due to strong identification of \(\theta_g^0\).

\[\Box\]

**Proof of Corollary 2.** This proof follows from the proof of theorem 4, and the asymptotic variance of the estimator that uses known weights which is

\[\text{Avar} \left[ \sqrt{N}(\hat{\theta}_g - \theta_g^0) \right] = H_g^{-1}\Omega_g H_g^{-1}\]

where \(\Omega_g = \mathbb{E}(1_{ig}1_{ig}')\). The result follows immediately.

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

\[ \mathbb{E} \left[ \frac{S_i \cdot W_i}{R(X_i, W_i, \delta^*) \cdot G(X_i, \gamma^*)} \cdot q(Y_i(1), X_i, \theta_1^0) \right] = \mathbb{E} \left[ \frac{r(X_i, 1)}{R(X_i, 1, \delta^*) \cdot G(X_i, \gamma^*)} \cdot \frac{p(X_i)}{G(X_i, \gamma^*)} \cdot q(Y_i(1), X_i, \theta_1^0) \right] \]

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

\[ \mathbb{E} \left[ \frac{r(X_i, 1)}{R(X_i, 1, \delta^*)} \cdot \frac{p(X_i)}{G(X_i, \gamma^*)} \right] \cdot \mathbb{E} \left\{ q(Y_i(1), X_i, \theta_1^0) \bigg| X_i \right\} \]

Then,

\[ H_1 = \mathbb{E} \left[ \frac{r(X_i, 1)}{R(X_i, 1, \delta^*)} \cdot \frac{p(X_i)}{G(X_i, \gamma^*)} \right] \cdot \nabla_{\theta_1} \mathbb{E} \left\{ h(Y_i(1), X_i, \theta_1^0) \bigg| X_i \right\} \]

\[ = \mathbb{E} \left[ \frac{r(X_i, 1)}{R(X_i, 1, \delta^*)} \cdot \frac{p(X_i)}{G(X_i, \gamma^*)} \right] \cdot A(X_i, \theta_1^0) \]

Similarly, I use LIE to express \( \Omega_1 \) as

\[ \Omega_1 = \mathbb{E} \left[ \frac{r(X_i, 1)}{R^2(X_i, 1, \delta^*)} \cdot \frac{p(X_i)}{G^2(X_i, \gamma^*)} \cdot \mathbb{E} \left\{ h(Y_i(1), X_i, \theta_1^0)h(Y_i(1), X_i, \theta_1^0) \bigg| X_i \right\} \right] \]

\[ = \sigma_{\theta_1}^2 \cdot \mathbb{E} \left[ \frac{r(X_i, 1)}{R^2(X_i, 1, \delta^*)} \cdot \frac{p(X_i)}{G^2(X_i, \gamma^*)} \right] \cdot A(X_i, \theta_1^0) \]

For the unweighted estimator, the variance simplifies, and this happens precisely due to the GCIME. To see this, consider \( H_1^u \). Then using LIE, I can rewrite

\[ H_1^u = \mathbb{E} \left[ r(X_i, 1) \cdot p(X_i) \cdot \nabla_{\theta_1} \mathbb{E} \left\{ h(Y_i(1), X_i, \theta_1^0) \bigg| X_i \right\} \right] \]

\[ = \mathbb{E} \left[ r(X_i, 1) \cdot p(X_i) \cdot A(X_i, \theta_1^0) \right] \]

and similarly we can rewrite \( \Omega_1^u \) using LIE as

\[ \Omega_1^u = \mathbb{E} \left[ r(X_i, 1) \cdot p(X_i) \cdot \mathbb{E} \left\{ h(Y_i(1), X_i, \theta_1^0)h(Y_i(1), X_i, \theta_1^0) \bigg| X_i \right\} \right] \]

\[ = \sigma_{\theta_1}^2 \cdot \mathbb{E} \left[ r(X_i, 1) \cdot p(X_i) \cdot A(X_i, \theta_1^0) \right] \]

Therefore, the asymptotic variance simplifies to simply

\[ \text{Avar} \left[ \sqrt{N} \left( \hat{\theta}_1^u - \theta_1^0 \right) \right] = \sigma_{\theta_1}^2 \cdot \left( \mathbb{E} \left[ r(X_i, 1) \cdot p(X_i) \cdot A(X_i, \theta_1^0) \right] \right)^{-1} \]
For showing that the two variances are positive semi-definite consider the following

\[
\left[ \text{Avar} \left\{ \sqrt{N} (\hat{\theta}_u - \theta_0) \right\} \right]^{-1} - \left[ \text{Avar} \left\{ \sqrt{N} (\hat{\theta}_1 - \theta_0) \right\} \right]^{-1}
\]

\[
= \frac{1}{\sigma^2_{01}} \cdot \left\{ \E \left( r_{i1} \cdot p_i \cdot A_i \right) - \E \left( \frac{r_{i1} \cdot p_i}{R_{i1} \cdot G_i} \cdot A_i \right) \cdot \E \left( \frac{r_{i1} \cdot p_i}{R_{i1} \cdot G_i^2} \cdot A_i \right)^{-1} \cdot \E \left( \frac{r_{i1} \cdot p_i}{R_{i1} \cdot G_i} \cdot A_i \right) \right\}
\]

Let \( B_i = \frac{r_{i1}^{1/2} \cdot p_i^{1/2} \cdot A_i^{1/2}}{1} \) and \( D_i = \left( \frac{r_{i1}^{1/2}}{R_{i1}} \right) \cdot \left( \frac{p_i^{1/2}}{G_i} \right) \cdot A_i^{1/2} \)

\[
= \frac{1}{\sigma^2_{01}} \cdot \left\{ \E \left( B_i^T B_i \right) - \E \left( B_i^T D_i \right) \cdot \E \left( D_i^T D_i \right)^{-1} \cdot \E \left( D_i^T B_i \right) \right\}
\]

where the quantity inside the brackets is nothing but the variance of the residuals from the population regression of \( B_i \) on \( D_i \). Hence, the difference is positive semi-definite. The results for \( g = 0 \) can be proven analogously.

\[ \square \]

**F.1 Identification of ATE using pooled and separate slopes mean functions under second half of DR**

**Pooled slopes.** Let us assume that \( m(X, \theta) = h(X\theta + \eta W) \) is the chosen mean function for \( \E[Y(g)|X] \). Then, in the presence of nonrandom sampling, we have the following first order conditions

\[
\sum_{i=1}^{N} S_i \cdot \left( \frac{W_i}{\hat{R}_i \cdot \hat{G}_i} + \frac{(1 - W_i)}{(1 - \hat{G}_i)} \right) \cdot [Y_i - h(X_i \hat{\theta} + \hat{\eta} W_i)] = 0
\]

\[
\sum_{i=1}^{N} S_i \cdot \hat{R}_i \cdot \hat{G}_i \cdot [Y_i - h(X_i \hat{\theta} + \hat{\eta} W_i)] = 0
\]

\[
\sum_{i=1}^{N} S_i \cdot \left( \frac{W_i}{\hat{R}_i \cdot \hat{G}_i} + \frac{(1 - W_i)}{(1 - \hat{G}_i)} \right) \cdot X_i^T \left[ Y_i - h(X_i \hat{\theta} + \hat{\eta} W_i) \right] = 0
\]

where \( \hat{R} = R(X, W, \hat{\theta}) \) and \( \hat{G} = G(X, \hat{\gamma}) \). Ignoring the last set of moment conditions, the population counterpart to the FOCs above are:

\[
\E \left[ S \cdot \left( \frac{W}{R \cdot G} + \frac{(1 - W)}{R \cdot (1 - G)} \right) \cdot [Y - h(X\theta^* + \eta^* W)] \right] = 0 \quad (F.9)
\]

\[
\E \left[ S \cdot \frac{W}{R \cdot G} \cdot [Y - h(X\theta^* + \eta^* W)] \right] = 0 \quad (F.10)
\]
where $\theta^*$ and $\eta^*$ are the probability limits of QMLE estimators $\hat{\theta}$ and $\hat{\eta}$. Rearranging (F.9) and (F.10) gives us

$$
E \left[ \frac{S}{R} \cdot \left( \frac{W}{G} + \frac{(1-W)}{(1-G)} \right) \cdot Y \right] = E \left( \frac{S}{R} \cdot \left( \frac{W}{G} + \frac{(1-W)}{(1-G)} \right) \cdot h(X\theta^* + \eta^*W) \right) \quad (F.11)
$$

$$
E \left[ \frac{S \cdot W}{R \cdot G} \cdot Y \right] = E \left[ \frac{S \cdot W}{R \cdot G} \cdot h(X\theta^* + \eta^*W) \right] \quad (F.12)
$$

Now, $Y = Y(1) \cdot W + Y(0) \cdot (1-W)$ which implies that we can replace $Y$ in the above two equations to obtain the LHS of (F.11) equal to

$$
E \left[ \frac{S}{R} \cdot \left\{ \frac{W}{G} \cdot Y(1) + \frac{(1-W)}{(1-G)} \cdot Y(0) \right\} \right]
$$

By using iterated expectations we can rewrite the above equation as

$$
E \left[ \frac{W}{G \cdot R} \cdot E(S \cdot Y(1)|X, W) + \frac{(1-W)}{(1-G) \cdot R} \cdot E(S \cdot Y(0)|X, W) \right]
$$

Due to MAR, we can split the conditional expectation into parts.

$$
E \left[ \frac{W}{G \cdot R} \cdot E(S|X, W) \cdot E(Y(1)|X, W) + \frac{(1-W)}{(1-G) \cdot R} \cdot E(S|X, W) \cdot E(Y(0)|X, W) \right]
$$

Note that, $W \cdot E(S|X, W) = W \cdot R$. Similarly, $(1-W) \cdot E(S|X, W) = (1-W) \cdot R$ and due to unconfoundedness we have, $E[Y(1)|X, W] = E[Y(1)|X]$ and $E[Y(0)|X, W] = E[Y(0)|X]$. Therefore, we can simplify the above expression into

$$
E \left[ \frac{W \cdot R}{G \cdot R} \cdot E(Y(1)|X) + \frac{(1-W) \cdot R}{(1-G) \cdot R} \cdot E(Y(0)|X) \right]
$$

Another application of iterated expectation gives us

$$
E \left[ \frac{E(Y(1)|X)}{G} \cdot E[W|X] + \frac{E(Y(0)|X)}{(1-G)} \cdot E[(1-W)|X] \right] = E \left[ E(Y(1)|X) + E(Y(0)|X) \right] = E[Y(1)] + E[Y(0)]
$$
Manipulating the RHS of (F.11) using iterated expectations gives us

\[
\mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \left\{ \frac{W_1}{G} \cdot \frac{1}{R} \cdot \mathbb{E}(S|\mathbf{X}, W_1) + \frac{(1 - W)}{(1 - G)} \cdot \frac{1}{R} \cdot \mathbb{E}(S|\mathbf{X}, W) \right\} \right]
\]

\[
= \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \left\{ \frac{W}{G} + \frac{(1 - W)}{(1 - G)} \right\} \right]
\]

\[
= \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{W}{G} \right] + \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{(1 - W)}{(1 - G)} \right]
\]

Therefore, combining the LHS and RHS give the result

\[
\mathbb{E}[Y(1)] + \mathbb{E}[Y(0)] = \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{W}{G} \right] + \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{(1 - W)}{(1 - G)} \right] \quad (F.13)
\]

Now, consider the LHS of F.12.

\[
\mathbb{E} \left[ \frac{S \cdot \mathbf{W}}{R \cdot G} \cdot \mathbf{Y} \right] = \mathbb{E} \left[ \frac{S \cdot \mathbf{W}}{R \cdot G} \cdot Y(1) \right]
\]

\[
= \mathbb{E}[Y(1)] \quad \text{(by LIE)}
\]

Similarly using LIE, the RHS of F.12 can be re-written as

\[
\mathbb{E} \left[ \frac{S \cdot \mathbf{W}}{R \cdot G} \cdot h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \right] = \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{W}{G} \cdot \frac{1}{R} \cdot \mathbb{E}(S|\mathbf{X}, W) \right]
\]

\[
= \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{W}{G} \right]
\]

Therefore combining the LHS and RHS give us

\[
\mathbb{E}[Y(1)] = \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{W}{G} \right] \quad (F.14)
\]

Then using F.14 along with F.13 implies that

\[
\mathbb{E}[Y(0)] = \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot \frac{(1 - W)}{(1 - G)} \right] \quad (F.15)
\]

Consider

\[
\mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^* \mathbf{W}) \cdot W|\mathbf{X} \right]
\]

\[
= \mathbb{E} \left[ h(\mathbf{X}\theta^* + \eta^*) \right] \cdot P(W = 1|\mathbf{X})
\]
Therefore, \( \mathbb{E} \left[ h(X\theta^* + \eta^*W) \cdot \frac{W}{G} \right] = \mathbb{E} \left[ h(X\theta^* + \eta^*) \right] \). Similarly, we can also show that

\[
\mathbb{E} \left[ h(X\theta^* + \eta^*W) \cdot \frac{(1 - W)}{(1 - G)} \right] = \mathbb{E} \left[ h(X\theta^*) \right]
\]

Hence, the pooled regression adjustment estimator can be written as

\[
\Delta_{ate}^p = \mathbb{E} \left[ h(X\theta^*) \right] - \mathbb{E} \left[ h(X\theta^*) \right]
\]

so a consistent estimator of the QMLE pooled regression adjustment estimator can be obtained by replacing the population expectation by the sample average in the above expression and weighting by the appropriate probabilities to recover the balance of the random sample which gives us

\[
\hat{\Delta}_{ate} = \frac{1}{N} \sum_{i=1}^{N} h(X_i\hat{\theta} + \hat{\eta}) - \frac{1}{N} \sum_{i=1}^{N} h(X_i\hat{\theta})
\]

\(\square\)

**Separate slopes.** Let us assume that \( m(X, \theta_g) = h(X\theta_g) \) is the chosen mean function for \( \mathbb{E} \left[ Y(g) | X \right] \). Then the population FOCs are

\[
\mathbb{E} \left[ S \cdot \frac{W}{R \cdot G} \cdot \left[ Y - h(X\theta^*_1) \right] \right] = 0
\]

\( F.16 \)

\[
\mathbb{E} \left[ S \cdot \frac{(1 - W)}{R \cdot (1 - G)} \cdot \left[ Y - h(X\theta^*_0) \right] \right] = 0
\]

\( F.17 \)

where \( \theta^*_g \) are the probability limits of QMLE estimators \( \hat{\theta}_g \). Rearranging F.16 and F.17 just like in the pooled case gives us the following equalities.

\[
\mathbb{E} \left[ S \cdot \frac{W}{R \cdot G} \cdot Y \right] = \mathbb{E} \left[ S \cdot \frac{W}{R \cdot G} \cdot h(X\theta^*_1) \right]
\]

\[
\mathbb{E} \left[ S \cdot \frac{(1 - W)}{R \cdot (1 - G)} \cdot Y \right] = \mathbb{E} \left[ S \cdot \frac{(1 - W)}{R \cdot (1 - G)} \cdot h(X\theta^*_0) \right]
\]

Proceeding with the above two equations in the same way as in the pooled case gives us the results

\[
\mathbb{E}[Y(1)] = \mathbb{E} \left[ h(X\theta^*_1) \right]
\]

\[
\mathbb{E}[Y(0)] = \mathbb{E} \left[ h(X\theta^*_0) \right]
\]

Therefore, \( \Delta_{ate}^f = \mathbb{E} \left[ h(X\theta^*_1) \right] - \mathbb{E} \left[ h(X\theta^*_0) \right] \) and a consistent estimator of the QMLE separate regression adjustment estimator can be obtained as

\[
\hat{\Delta}_{ate} = \frac{1}{N} \sum_{i=1}^{N} h(X_i\hat{\theta}_1) - \frac{1}{N} \sum_{i=1}^{N} h(X_i\hat{\theta}_0)
\]
G Supplementary Tables

Table G.1: Proportion of missing earnings in the experimental sample

| Earnings in 1979 | Treated | Control | Total |
|-----------------|---------|---------|-------|
| Missing         | 196     | 210     | 406   |
| Observed        | 600     | 585     | 1185  |
| **Total**       | 796     | 795     | 1591  |

Table G.2: Proportion of missing data in the PSID samples

| Earnings in 1979 | PSID-1 | PSID-2 |
|-----------------|--------|--------|
| Missing         | 81     | 22     |
| Observed        | 648    | 182    |
| **Total**       | 729    | 204    |
Table G.3: Unweighted and weighted earnings comparisons and estimated training effects using NSW and PSID comparison groups

| Comparison group | Unadjusted | Pre-training estimates | Adjusted |
|------------------|------------|------------------------|----------|
|                  |            | Unweighted | PS-weighted | D-weighted | Unweighted | PS-weighted | D-weighted |
| NSW              | -18        | (123.45)   | (51.07)     | 1          | (48.76)    | (124.70)   | (51.34)    |
|                  |            |            | (48.97)     |            |            |            |            |
| N=1,185          |            | (123.45)   | (51.07)     | 1          | (48.76)    | (124.70)   | (51.34)    |
| PSID-1           | -2,534     | (283.95)   | (213.57)    | -255       | (205.59)   | (281.49)   | (212.55)   |
|                  |            |            | (205.45)    |            |            |            |            |
| N=1,016          |            | (283.95)   | (213.57)    | -255       | (205.59)   | (281.49)   | (212.55)   |
| PSID-2           | -2,080     | (411.23)   | (331.41)    | -1,357     | (317.41)   | (427.24)   | (359.98)   |
|                  |            |            | (342.16)    |            |            |            |            |
| N=720            |            | (411.23)   | (331.41)    | -1,357     | (317.41)   | (427.24)   | (359.98)   |
| Bias using NSW control |        |            |            |            |            |            |            |
|                  |            | -2,517     | 289         | 236        | -2,760     | 334        | 287        |
|                  |            | (279.38)   | (256.93)    | (247.18)   | (283.09)   | (257.50)   | (248.20)   |
|                  |            | (279.38)   | (256.93)    | (247.18)   | (283.09)   | (257.50)   | (248.20)   |
| N=1,001          |            | -2,063     | -1,249      | -1,255     | -2,144     | -1,306     | -1,297     |
|                  |            | (416.53)   | (323.36)    | (310.59)   | (435.74)   | (354.12)   | (337.68)   |
|                  |            | (416.53)   | (323.36)    | (310.59)   | (435.74)   | (354.12)   | (337.68)   |
| Adjusted covariates |        |            |            |            |            |            |            |
|                  |            |            |            |            |            |            |            |
| Pre-training earnings (1975) |        | ✓          | ✓          | ✓          |            |            |            |
| Age              |            | ✓          | ✓          | ✓          |            |            |            |
| Age2             |            | ✓          | ✓          | ✓          |            |            |            |
| Education        |            | ✓          | ✓          | ✓          |            |            |            |
| High school dropout|        | ✓          | ✓          | ✓          |            |            |            |
| Black            |            | ✓          | ✓          | ✓          |            |            |            |
| Hispanic         |            | ✓          | ✓          | ✓          |            |            |            |
| Marital status   |            | ✓          | ✓          | ✓          |            |            |            |
| Number of Children (1975) |      | ✓          | ✓          | ✓          |            |            |            |

Notes: This table reports unadjusted and adjusted pre-training earnings differences where the first row reports the experimental estimates which combines the NSW treatment and control groups. The second and third row reports non-experimental earnings estimates computed from using the PSID-1 and PSID-2 comparison groups respectively. The second panel of the table reports bias estimates computed from combining the NSW control and PSID-1 and PSID-2 comparison groups respectively. Both the pre-training estimates and the bias estimates should be compared to zero. Bootstrapped standard errors are given in parentheses and have been constructed from using 10,000 replications. All values are in 1982 dollars. The samples used for estimating the training and bias estimates using PSID-1 and PSID-2 comparison groups have been trimmed to ensure common support in the distribution of weights for the NSW-treatment and comparison groups. For more detail, see appendix E.
Table G.4: Unconditional quantile treatment effect (UQTE) using PSID-1 comparison group

| Quantile | Experimental | Unweighted | PS-weighted | D-weighted |
|----------|--------------|------------|-------------|------------|
| 0.1      | 0            | 0          | 0           | 0          |
|          | (0)          | (0)        | (0)         | (0)        |
| 0.2      | 0            | 0          | 0           | 0          |
|          | (0)          | (0)        | (0)         | (0)        |
| 0.3      | 0            | 0          | 0           | 0          |
|          | (0)          | (12.91)    | (0)         | (0)        |
| 0.4      | 0            | -1124.61   | 0           | 0          |
|          | (11.17)      | (552.97)   | (207.14)    | (174.89)   |
| 0.5      | 993.52       | -2227.26   | 2076.58     | 1847.04    |
|          | (695.93)     | (983.43)   | (851.09)    | (829.42)   |
| 0.6      | 2004.40      | -860.55    | 3602.76     | 3535.85    |
|          | (1112.82)    | (964.97)   | (1299.08)   | (1284.64)  |
| 0.7      | 2129.93      | 428.01     | 3415.47     | 3340.84    |
|          | (716.04)     | (728.22)   | (988.24)    | (992.95)   |
| 0.8      | 1753.27      | -190.60    | 2019.44     | 2019.44    |
|          | (372.37)     | (519.63)   | (984.59)    | (999.47)   |
| 0.9      | 1134.21      | -1563.27   | -385.45     | -385.45    |
|          | (449.86)     | (952.85)   | (1059.43)   | (1056.09)  |

Notes: This table reports unweighted, ps-weighted and d-weighted UQTE estimates for three different comparison groups, namely, NSW control, PSID-1 and PSID-2. The estimates are reported at every 10th quantile of the 1979 earnings distribution. The experimental and PSID-1 estimates have been constructed using N=1,185 and N=1,016 observations respectively. Bootstrapped standard errors are given in parentheses and have been constructed using 1,000 replications. All values are in 1982 dollars. The samples used for constructing these estimates have been trimmed to ensure common support across the treatment and comparison groups.
Table G.5: Unconditional quantile treatment effect (UQTE) using PSID-2 comparison group

| Quantile | Experimental | Unweighted | PS-weighted | D-weighted |
|----------|--------------|------------|------------|------------|
| 0.1      | 0            | 0          | 0          | 0          |
|          | (0)          | (0)        | (0)        | (0)        |
| 0.2      | 0            | 0          | 0          | 0          |
|          | (0)          | (0)        | (10.07)    | (10.07)    |
| 0.3      | 0            | 0          | 0          | 0          |
|          | (0)          | (111.74)   | (136.31)   | (129.77)   |
| 0.4      | 0            | -795.71    | 0          | 0          |
|          | (13.25)      | (672.87)   | (573.22)   | (546.78)   |
| 0.5      | 993.52       | -237.98    | 378.98     | 372.07     |
|          | (693.73)     | (1232.63)  | (1312.93)  | (1267.28)  |
| 0.6      | 2004.40      | 193.77     | 1480.47    | 1294.77    |
|          | (1114.65)    | (1426.40)  | (1647.31)  | (1659.69)  |
| 0.7      | 2129.93      | 1857.64    | 2616.22    | 2599.73    |
|          | (710.26)     | (943.38)   | (1217.80)  | (1209.60)  |
| 0.8      | 1753.27      | 1148.85    | 2010.87    | 1990.37    |
|          | (371.73)     | (1152.92)  | (1541.14)  | (1553.67)  |
| 0.9      | 1134.21      | -237.08    | 1089.10    | 1089.10    |
|          | (452.08)     | (1888.06)  | (3321.56)  | (3246.78)  |

Notes: This table reports unweighted, ps-weighted and d-weighted UQTE estimates for three different comparison groups, namely, NSW control, PSID-1 and PSID-2. The estimates are reported at every 10th quantile of the 1979 earnings distribution. The experimental and PSID-2 estimates have been computed using N=1,185 and N=720 observations respectively. Bootstrapped standard errors are given in parentheses and have been constructed using 1,000 replications. All values are in 1982 dollars. The samples used for constructing these estimates have been trimmed to ensure common support across the treatment and comparison groups.
H Supplementary Figures

Figure H.1: Estimated CQTE with true CQTE as a function of $X_1$ for N=5,000

Case 3: Correct CQF, misspecified weights

b) $\tau = 0.50$

c) $\tau = 0.75$

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 H.2: Bias in the estimated LP relative to the true LP of CQTE as a function of $X_1$ for $N=5,000$

Case 1: Misspecified CQF, correct weights

b) $\tau = 0.50$

c) $\tau = 0.75$

Notes: This figure plots the bias in the unweighted, ps-weighted, and d-weighted LP of the true CQTE relative to the true population LP of CQTE. The average treated sample is $N_1 = 5,000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5,000 \times (1 - 0.41) \times 0.38 = 1,121$. 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.
Figure H.3: Bias in the estimated linear projection relative to the true linear projection for N=5,000

Case 3: Misspecified CQF, misspecified weights

b) $\tau = 0.50$

c) $\tau = 0.75$

Notes: This figure plots the bias in the unweighted, ps-weighted, and d-weighted LP of the true CQTE relative to the true population LP of CQTE. The average treated sample is $N_1 = 5,000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5,000 \times (1 - 0.41) \times 0.38 = 1,121$. 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.
Figure H.4: Empirical distribution of estimated UQTE for N=5,000 when weights are wrong

b) $\tau = 0.50$

c) $\tau = 0.75$

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 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.
Figure H.5: Empirical distribution of estimated UQTE for N=5,000 when weights are correct

b) $\tau = 0.50$

c) $\tau = 0.75$

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 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.