Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data

José R. Zubizarreta

Abstract

Weighting methods that adjust for observed covariates, such as inverse probability weighting, are widely used for causal inference and estimation with incomplete outcome data. Part of the appeal of such methods is that one set of weights can be used to estimate a range of treatment effects based on different outcomes, or a variety of population means for several variables. However, this appeal can be diminished in practice by the instability of the estimated weights and by the difficulty of adequately adjusting for observed covariates in some settings. To address these limitations, this paper presents a new weighting method that finds the weights of minimum variance that adjust or balance the empirical distribution of the observed covariates up to levels prespecified by the researcher. This method allows the researcher to balance very precisely the means of the observed covariates and other features of their marginal and joint distributions, such as variances and correlations and also, for example, the quantiles of interactions of pairs and triples of observed covariates, thus balancing entire two- and three-way marginals. Since the weighting method is based on a well-defined convex optimization problem, duality theory provides insight into the behavior of the variance of the optimal weights in relation to the level of covariate balance adjustment, answering the question, how much does tightening a balance constraint increases the variance of the weights? Also, the weighting method runs in polynomial time so relatively large data sets can be handled quickly. An implementation of the method is provided in the new package `sbw` for `R`. This paper shows some theoretical properties of the resulting weights and illustrates their use by analyzing both a data set from the 2010 Chilean earthquake and a simulated example.

Keywords: causal inference; observational study; propensity score; quadratic programming; unit nonresponse; weight adjustment

José R. Zubizarreta, Division of Decision, Risk and Operations, and Statistics Department, Columbia University, 3022 Broadway, 417 Uris Hall, New York, NY 10027; `zubizarreta@columbia.edu`). The author acknowledges very helpful comments from three referees, an associate editor, and Joseph Ibrahim. The author also thanks Kosuke Imai, Marshall Joffe, Peter Lynn, Arian Maleki, Luke Miratrix, Marc Ratkovic, María de los Ángeles Resa, James Robins, Paul Rosenbaum and Dylan Small for valuable conversations and help at different stages of this paper. This paper was supported by a grant from the Alfred P. Sloan Foundation.

1 Introduction

Weighting methods that adjust for observed covariates are widely used for causal inference and estimation with incomplete outcome data. For example, under the assumption of selection on observed covariates or the assumption that outcomes are missing at random (MAR), weighting methods are used in observational studies for estimating the effect of interventions and events, and in sample surveys and panel data for estimating the mean of an outcome variable in the presence of unit nonresponse. Part of the practical appeal of weighting methods is that weights do not require explicit modeling of the response surfaces of the outcomes (Rosenbaum 1987), and that one set of weights can be used to estimate a range of treatment effects or population means based on different outcomes (Little and Rubin 2002).

In these contexts, the goal of weights is often twofold: to adjust or balance the empirical distributions of the observed covariates (to remove biases due to observed confounders or recover the observed structure of the target population) and to yield stable estimates of the parameters of interest (very large weights may overly influence the results and highly variable weights produce results with high variance; see Little and Rubin 2002). A widespread method uses logistic regression to estimate the probabilities of sample selection and then inverts these probabilities to calculate the weights. However, there is nothing in this procedure that directly targets covariate balance or that explicitly restrains the variability of the weights. In fact, the resulting weights can vary substantially and lead to instability in the estimates (Robins and Wang 2000; Kang and Schafer 2007). Of course, if the probability model is correctly specified, then it is correct to have highly variable weights; however this is hard to determine in practice. In view of this limitation, common practice is to trim the extreme weights, but this is often done in an ad hoc manner that may introduce bias in the estimates (see Elliott 2008 and Crump et al. 2009 for discussions and interesting alternative methods). Also, prediction methods from the machine learning literature have been used to more flexibly estimate the probability model and obtain weights that are less sensitive to model misspecification (e.g. Lee et al. 2010).

Unlike model-based approaches that aim at the goal of weights indirectly, in this paper we propose a weighting method that directly constrains covariate imbalances and explicitly optimizes the stability of the weights. In other words, by design this method optimizes and constrains the criteria used to evaluate the weights. Specifically, by solving a convex optimization problem, this method finds the weights of minimum variance that balance the empirical distribution of the observed covariates up to levels prespecified by the researcher. This weighting method allows the researcher to adjust very precisely for means of the observed covariates and beyond this for other features of their marginal and joint distributions, such as variances and correlations and also, for example, quantiles of interactions of pairs and triples of observed covariates, thus balancing entire two- and three-way marginals. With this weighting method, the researcher does not have to specify a probability model, but rather the standards for covariate balance only (which are necessary anyway to assess the performance of a weighting method). Since this weighting method is based on a convex optimization problem, duality theory provides useful insight into the behavior of the optimal weights, answering the question, how much does tightening a covariate balance constraint increases the variance of the weights? In this manner the proposed method allows the researcher to explore and explicitly handle the trade-off between the stability and the level of covariate adjustment of the weights. Finally, this method is potentially valuable for practitioners as the optimal weights are found in polynomial time, so relatively large data sets can be handled quickly. In practice, this method is easy to use as currently implemented in the new package `sbw` for `R`.

This paper is organized as follows. Section 2 explains the basic estimation problem, introduces the notation, and emphasizes that, under simple conditions, the variance of the weighted estimator is determined by the coefficient of variation of the weights. In view of this fact, section 3 poses a quadratic program that minimizes the coefficient of variation of the weights that directly adjust for the empirical distributions of the observed covariates up to levels prespecified by the researcher. This section also discusses computational properties, the `R` implementation of the method, and the duality and post optimality analysis that can be conducted after finding the optimal weights, giving useful insight into the weighting problem at hand. Afterwards, based on model approximation theory, section 4 provides guidelines on which aspects of the empirical distributions of the observed covariates are desirable to balance to obtain accurate estimates. Section 5 analyzes data from the 2010 Chilean earthquake, and section 6 shows the comparative performance of the method using Kang and Schafer's (2007) simulated data set. Section 7 provides some guidance for implementing the method in practice. Finally, section 8 concludes with a summary and remarks.

2 The estimation problem

For simplicity, consider the problem of estimating a population mean from a sample with incomplete outcome data. As discussed by Rubin (1978) this problem is very similar to the one of causal inference. Essentially, we can view the treatment indicator as an indicator for outcome response and under the assumption that the outcome is conditionally independent of the response (or treatment) indicator given the observed covariates (specifically, under the strong ignorability assumption; Rosenbaum and Rubin 1983), we can proceed with the estimation of population means (or average treatment effects) from sample averages after adjusting for covariates (see Gelman and Meng 2004 and Rubin 2005 for reviews of causal inference from a missing data perspective, and Kang and Schafer 2007 for a detailed example of the connection between the causal inference estimation with incomplete data problems).

Let $\mathcal{P}$ be a target population with N elements, $\mathcal{S}$ be a random sample of $\mathcal{P}$ of size n, and $\mathcal{R}$ be the sample of r respondents of $\mathcal{S}$. For each unit i in $\mathcal{S}$ define the response indicator Z_i where $Z_i = 1$ if unit i responds and $Z_i = 0$ otherwise. Denote Y_i as the outcome variable of unit i and X_{ip} as the p^{th} covariate of the same unit measured at baseline, with $p = 1, ..., P$.

The parameter that we wish to estimate is the population mean

$$\overline{Y}_N = \frac{\sum_{i=1}^{N} Y_i}{N}.$$

If the missing responses in $\mathcal{S}$ are missing completely at random (this is, if they are independent both of observed and unobserved covariates; Little and Rubin 2002), then the average of the Y_i's of the respondents,

$$\hat{Y}_r = \frac{\sum_{i=1}^{r} Y_i}{r},$$

is an unbiased and consistent estimator of $\overline{Y}_N$. Of course, the assumption that the nonrespondents are missing completely at random is very strong and unrealistic in typical applications. If instead the nonrespondents are missing at random (meaning that both Z_i and Y_i are systematically related to the observed covariates $\boldsymbol{X}_i$ but not to an unobserved covariate U_i; Little and Rubin 2002), then the weighted estimator

$$\hat{Y}_w = \frac{\sum_{i=1}^{r} w_i Y_i}{r}$$

can be used to estimate without bias and consistently $\overline{Y}_N$, provided that the weights $\boldsymbol{w}$ appropriately adjust for the $\boldsymbol{X}_i$'s.

In practice, perhaps the most common way of calculating the weights is fitting a model to estimate the probabilities of response and then inverting these probabilities to obtain the weights. This probability is called the propensity score (Rosenbaum and Rubin 1983) and it has the appealing property that weighting (or, more generally, balancing) on the propensity score tends to adjust or balance the distributions of the observed covariates. However, this is a stochastic property that relies on the law of large numbers (in other words, a property that holds over repeated realizations of the data generation mechanism), and for a given data set, even if the true treatment assignment or missing data mechanism is known, there is no guarantee that the propensity score will balance the observed covariates. Actually, in a randomized experiment covariate balance is attained by means of a random device (such as flipping a coin), but even in a randomized experiment covariates may be imbalanced (especially if the covariates are sparse or have many categories) and a modification to the random device is needed if ones wants to ensure balance (see Shadish et al. 2002 for a review of these methods including Efron's 1971 biased coin design). Similarly, in observational studies propensity scores balance covariates with the help of the laws of probability, but one needs large numbers for the laws to act. In view of this, it is often desirable to base the design of an observational study on an optimization method so that the laws of probability can act faster (see Yang et al. 2012 and Zubizarreta et al. 2011 for related discussions).

An added difficulty is that in practice the true assignment mechanism is unknown and this makes the task of balancing the observed covariates difficult, even if one only wants to balance their means. Moreover, in some settings it is desirable to balance other features of the empirical distributions of the observed covariates beyond means, such as entire marginal distributions, especially if the true response surface of the outcome is nonlinear function of these covariates (see section 4).

Another drawback of this common approach is that the resulting weights can be highly variable or unstable and considerably increase the variance of the weighted estimator. In fact, under simple conditions —that is, assuming simple random sampling, ignoring sampling variation in the weights and scaling the weights to average one— if the outcome Y has constant variance σ^2, then $\mathrm{Var}\left(\hat{Y}_w\right) = (\sigma^2/r)\left(1 + (\mathrm{cv}(w_i))^2\right)$ where $\mathrm{cv}(w_i)$ is the coefficient of variation of the weights. In other words, the square of the coefficient of variation of the weights yields a simplified measure of the proportional added variance induced by weighting (see section 3.3.2 of Little and Rubin 2002). For example, the added variance induced by weights that have a coefficient of variation of 1 instead of 0.5 (as seen in the simulation study in Section 6 below) is equal to $\sigma^2/r \times 1^2 - \sigma^2/r \times 0.5^2 = 0.75 \times \sigma^2/r$, and if the weights have a coefficient of variation of 2 instead of 0.5, then the added variance is equal to $3.75 \times \sigma^2/r$. In this way, if a few units have very high weights (resulting from estimated probabilities very close to zero), then the added variance induced by weighting will be considerable.

3 Stable weights that balance covariates

3.1 A convex optimization problem

To handle the potential instability of the estimated weights and adjust in a precise manner for observed covariates, instead of modeling the probability of treatment assignment or nonresponse, we take a design-based approach and directly find the weights of minimum variance that balance the covariates as specified by the researcher. For this, using standard notation in convex optimization (Boyd and

Vandenberghe 2004), we solve the optimization problem

$$
\begin{aligned}
\underset{\boldsymbol{w}}{\text{minimize}} \quad & \|\boldsymbol{w} - \overline{\boldsymbol{w}}\|_2^2 && (1.1) \\
\text{subject to} \quad & |\boldsymbol{w}^\top \boldsymbol{X}_{\mathcal{R}p} - \overline{X}_{\mathcal{S}p}| \leq \delta_p,\ p = 1, ..., P && (1.2) \\
& \mathbf{1}^\top \boldsymbol{w} = 1, && (1.3) \\
& \boldsymbol{w} \succeq 0, && (1.4)
\end{aligned}
\qquad (1)
$$

where $\boldsymbol{w}$ is the $r \times 1$ vector of weights, $\overline{\boldsymbol{w}}$ is the mean value vector of the weights, and $\| \cdot \|_2$ is the ℓ_2 norm. In this way, (1.1) minimizes the variance of the weights. In the constraints (1.2), $\boldsymbol{X}_{\mathcal{R}p}$ is the vector of covariate p for the sample of respondents $\mathcal{R}$, $\overline{X}_{\mathcal{S}p}$ is the mean value of covariate p for the whole sample $\mathcal{S}$, and δ_p is a scalar specified by the researcher. As a result, (1.2) constrains the absolute difference in means of the weighted covariates to be less or equal than specified δ_p's. Finally, (1.3) normalizes the weights to sum up to one and (1.4) constraints each of the weights to be greater or equal than zero. By imposing these last two sets of constraints (1) actually minimizes the coefficient of variation of the weights. In fact, (1.3) fixes the mean weight, so (1.1) is optimizing the variance (for fixed mean) and hence also optimizing the coefficient of variation (for fixed mean).

We will call the weights obtained from solving this problem stable balancing weights (henceforth SBW). These weights are stable in the sense that they have minimum variance (or, more precisely, minimum coefficient of variation), and balance because they adjust for covariates up to levels specified by the researcher.

It is important to note that by adequately augmenting the covariate matrix $\boldsymbol{X}_{\mathcal{R}}$.

the constraints (1.2) can be used to balance statistics other than means. The basic idea is to augment $\boldsymbol{X}_{\mathcal{R}\cdot}$ with additional covariates $\widetilde{\boldsymbol{X}}_{\mathcal{R}\cdot}$, which are transformations of $\boldsymbol{X}_{\mathcal{R}\cdot}$. For example, if we mean center $\boldsymbol{X}_{\mathcal{R}p}$ and let $\widetilde{\boldsymbol{X}}_{\mathcal{R}\cdot}= \boldsymbol{X}^2_{\mathcal{R}p}$, then mean balancing $(\boldsymbol{X}_{\mathcal{R}p}, \widetilde{\boldsymbol{X}}_{\mathcal{R}\cdot})$ will balance both the mean and the variance of covariate p. Similarly, if we mean center $\boldsymbol{X}_{\mathcal{R}p_1}$ and $\boldsymbol{X}_{\mathcal{R}p_2}$ for two covariates p_1 and p_2, and let $\widetilde{\boldsymbol{X}}_{\mathcal{R}\cdot}= (\boldsymbol{X}^2_{\mathcal{R}p_1}, \boldsymbol{X}^2_{\mathcal{R}p_2}, \boldsymbol{X}_{\mathcal{R}p_1} \cdot \boldsymbol{X}_{\mathcal{R}p_2})$, then mean balancing $(\boldsymbol{X}_{\mathcal{R}p_1}, \boldsymbol{X}_{\mathcal{R}p_2}, \widetilde{\boldsymbol{X}}_{\mathcal{R}\cdot})$ will balance the means and variances of the covariates, and also its correlation. Also, if we define $\widetilde{\boldsymbol{X}}_{\mathcal{R}\cdot}$ as the matrix of column indicators of quantiles of $\boldsymbol{X}_{\mathcal{S}p}$ (say, its deciles), then (1) will find the minimum variance weights that balance the entire empirical distribution of covariate p. These tactics can be used simultaneously for a number of covariates (see Zubizarreta 2012 for a related discussion in matching in observational studies).

In (1) we ask the weights to recover or represent the covariate structure of the sample before nonresponse. This is somewhat similar to estimating the average treatment effect (ATE) in an observational study. Specifically, for estimating the ATE we would weight the treated units to represent the structure of the sample of the treated and controls units together before treatment, and weight the controls again to represent the treated and controls units together before treatment. Then we would take the difference in means between the weighted treated and control groups. Similarly, (1) can be used in observational studies for estimating the average treatment on the treated (ATT), this by weighting the controls to represent the treated units, and for estimating average treatment on the controls (ATC) by weighting the treated units to represent the controls. In principle, with (1) an average treatment effect can be calculated for any target population by defining the target means $\overline{X}_{\mathcal{S}p}$'s adequately.

3.2 Computational properties and implementation

The optimization problem above has several features that make it attractive from a computational standpoint. Specifically, (1) is a convex optimization problem (precisely, a convex quadratic programming problem), where the objective function is defined by a positive definite matrix implying that the problem can be solved efficiently, i.e. in a polynomial number of iterations (Papadimitriou 1994), for instance by means of the ellipsoid method (Wright 1997). The practical meaning of this is that the optimal weights can be computed quickly for relatively large data sets. A solution to (1) is implemented in the new `R` package `sbw` which uses either of the optimization solvers CPLEX, Gurobi or `quadprog`. At the present time, `sbw` is available upon request, and soon it will be publicly available through CRAN or a specialized webpage.

3.3 Duality and post optimality analysis

In a convex optimization problem like (1), each constraint has associated a dual variable that can provide useful insight into the behavior of the optimal solution in relation to the constraints. Formally, the optimal dual variable or shadow price of a constraint is the value of its Lagrange multiplier at the optimum, and it can be interpreted as the infinitesimal change in the optimal value of the objective function due to an infinitesimal change in the constraint. A dual variable provides a quantitative measure per unit of a constraint of how active or constraining is that constraint at the optimum (see chapter 5 of Boyd and Vandenberghe 2004).

In (1) above, the optimal dual variable of a covariate balance constraint in (1.2) is the marginal cost in terms of the variance of the weights of tightening that given balance constraint. Equivalently, it is the rate of improvement (again, in terms of the variance of the weights) of relaxing that constraint at the optimum. In this way, a dual variable equal to zero means that the constraint in question is not binding at the optimum and therefore that it can be tightened at no cost in terms of the variance of the weights. On the other hand, if the value of the dual variable is large it means that if the constraint is tightened the effect on the variance of the weights will be large. If the value of the dual variable is small it means that the constraint can be tightened to an extent without much effect on the optimum (Boyd and Vandenberghe 2004). In this manner the dual variables of our optimization problem tell us how much tightening a balance constraint increases the variance of the weights.

4 Bounding bias under different response surfaces

Throughout we assume that the outcome is conditionally independent of the response (treatment) indicator given the observed covariates. Using Dawid's (1979) notation for conditional independence, this is $Y \perp\!\!\!\perp Z \mid \boldsymbol{X}$. This means that given the observed covariates the distributions of the variable of interest are the same for the respondents and the nonrespondents, and in particular that $\mathrm{E}(Y \mid \boldsymbol{X}) = \mathrm{E}(Y \mid \boldsymbol{X}, Z = 1) = \mathrm{E}(Y \mid \boldsymbol{X}, Z = 0)$. Another common assumption in the causal inference and incomplete outcome data literatures is that the true function that describes $\mathrm{E}(Y \mid \boldsymbol{X})$ is linear in $\boldsymbol{X}$. Proposition 4.1 below shows that if this function is indeed linear and the weights approximately balance the means of the P covariates, then the weighted estimator for the population mean is approximately unbiased.

Proposition 4.1. *Suppose that $Y_i = \alpha + \boldsymbol{X}_i^\top \beta + \varepsilon_i$ with $\mathrm{E}(\varepsilon_i|\boldsymbol{X}_i) = 0$ for all i in the target population, $i = 1, ..., N$, and all i in the sample of respondents, $i = 1, ..., r$. If $\left|\frac{\sum_{i=1}^r w_i X_{i,p}}{r} - \frac{\sum_{i=1}^N X_{i,p}}{N}\right| < \delta$ for each covariate $p = 1, ..., P$, then $\mathrm{E}(\hat{Y}_w - \overline{Y}_N) < \delta \sum_{p=1}^P |\beta_p|$.*

See Appendix A for a proof. Note that in this proposition we can standardize the covariates to get rid of the units. Consequently, Proposition 4.1 is saying is that if we make the weighted differences in standardized covariates small then we can make the bias small, and we can do this before looking at the outcomes.

As noted, Proposition 4.1 assumes that the regression of Y on $\boldsymbol{X}$ is linear. A more general representation is considered in the following proposition, where the regression has a generalized additive form, $\mathrm{E}(Y_i|\boldsymbol{X}_i) = \alpha + f_1(X_{i,1}) + f_2(X_{i,2}) + ... + f_P(X_{i,P})$. Here, the f_p's are unspecified smooth functions, $p = 1, ..., P$. Proposition 4.2 below bounds the bias under a generalized additive regression form by balancing auxiliary covariates $\widetilde{X}_{i,j,p}^{k}$, which are a transformation of the original covariates $X_{i,p}$. Specifically, for each $X_{i,p}$ we break its support $[0, M_p] \in \mathbb{R}_0^+$ into M_p/l_p disjoint intervals of length l_p and midpoint $\xi_{j,p}$, and define the transformed piecewise covariates centered around $\xi_{j,p}$ as $\widetilde{X}_{i,j,p}= (X_{i,p} - \xi_{j,p})\mathbb{1}_{X_{i,p} \in [\xi_{j,p} - l_p/2, \xi_{j,p} + l_p/2]}$. In Proposition 4.2 we balance the first $K - 1$ powers of these transformed covariates $\widetilde{X}_{i,j,p}$.

Proposition 4.2. *Suppose that* $Y_i = \alpha + \sum_{p=1}^{P} f_p(X_{i,p}) + \varepsilon_i$ *where each* f_p *is a* K*-times differentiable function at all* $X_{i,p}$ *with* $f^{(K)} \leq c$ *for some constant* c*, and* $\mathrm{E}(\varepsilon_i|\boldsymbol{X}_i) = 0$*, for all* $i = 1, ..., N$ *and all* $i = 1, ..., r$*. If* $\left|\frac{\sum_{i=1}^{r} w_i \tilde{X}^{k}_{i,p}}{r} - \frac{\sum_{i=1}^{N} \tilde{X}^{k}_{i,p}}{N}\right| < \delta$ *for each* $p = 1, ..., P$ *and each* $k = 1, ..., K-1$*, then* $\mathrm{E}(\hat{Y}_w - \overline{Y}_N) < \delta \sum_{p=1}^{P} \sum_{j=1}^{M_p/l_p} \sum_{k=1}^{K-1} |\gamma_{k,p}| + 2P\frac{M_p}{l_p}L$ *where the* $\gamma_{k,p}$*'s are the coefficients of the Taylor expansion of order* k *around each* $\xi_{j,p}$*,* $j = 1, ..., M_p/l_p$*,* $p = 1, ..., P$*, and* L *is the Lagrange error bound of the expansion,* $L = \left|\frac{f^{(K)}(\xi_{j,p})}{K!}\right| (l_p/2)^K$*.*

A proof is provided in Appendix A. The key insight in Proposition 4.2 is to augment the original vector of covariates $\boldsymbol{X}$ with transformations of these covariates $\widetilde{\boldsymbol{X}}$ and then balance the augmented vector $(\boldsymbol{X}, \widetilde{\boldsymbol{X}})$. Proposition 4.2 shows that if the regression function has a generalized additive form and the weights approximately balance the means of the transformed covariates as defined above, then the weighted estimator for the population mean is approximately unbiased. Specifically, Proposition 4.2 says that if we balance the first $K-1$ powers of each covariate $X_{i,p}$, with each power centered and restricted to an interval of length l_p of the support of $X_{i,p}$, then we approximately get rid of the bias without using the outcome.

Here, we are balancing $\left(\sum_{p=1}^{P} M_p/l_p\right) \times (K-1)$ transformed covariates defined to be piecewise polynomials. Note that a simpler yet less general alternative is to balance piecewise linear functions. In this case the total number of transformed covariates to be balanced is $\sum_{p=1}^{P} M_p/l_p$.

In Proposition 4.2 a natural question that arises is how to choose the length of the approximation intervals l_p and the order $K-1$ of the approximating polynomials.

In principle, both decisions depend on the smoothness of the function and the data available, however from a practical standpoint in the statistical learning literature it is argued that there is little reason to go beyond a cubic approximation unless one wants to obtain smooth derivatives at the knots (see section 5.2 of Hastie et al. 2009). On the other hand, as we decrease l_p we make the bias smaller but the variance grows larger. We think that l_p should be determined with data at hand in view of the number of observations available within each approximation interval (see Tsybakov 2008 for a more theoretical discussion).

As noted, Proposition 4.2 assumes that the regression of Y on $\boldsymbol{X}$ is nonlinear but additive in $\boldsymbol{X}$. A more general representation of the regression is $\mathrm{E}(Y_i|\boldsymbol{X}_i) = \alpha + g_1(\boldsymbol{X}_i) + g_2(\boldsymbol{X}_i) + ... + g_H(\boldsymbol{X}_i)$ where $g_h(\boldsymbol{X}_i) : \mathbb{R}^p \mapsto \mathbb{R}$ is the h^{th} transformation of $\boldsymbol{X}_i$, $h = 1, ..., H$. This representation allows for interactions of covariates by letting $g_h(\boldsymbol{X}_i)$ be the product of two or more covariates. Under this representation, it is conceptually easy but notationally burdensome to extend Proposition 4.2 to bound the bias of nonlinear and nonadditive regression functions by balancing not only powers of the original covariates but also the interactions within certain intervals. However, from a practical standpoint note that the number of auxiliary covariates that need to be balanced grows exponentially with the number of covariates. Specifically, with P original covariates the number of additional auxiliary covariates that need to be balanced is $\widetilde{P}= \sum_{p=1}^{P} \binom{p+P-1}{p} - P$. Therefore, for example, with 3, 4 and 5 original covariates the number of additional auxiliary covariates that need to be balanced is 16, 65 and 246, respectively. Note also that this number is increased further if one uses a piecewise polynomials approximation as in Proposition 4.2. A natural extension of Proposition 4.2 is to use wavelets and other basis expansions, but this goes beyond the scope of this work.

5 Case study: the 2010 Chilean earthquake

5.1 The EPT data

On February 27, 2010, the fourth strongest earthquake in the world in the last 60 years struck the coast of central Chile (USGS 2014). To evaluate its impact on health, housing and socioeconomic outcomes, the Chilean government reinterviewed a subsample of respondents of the National Survey of Social Characterization (CASEN), which is the main household survey in the country and which was completed approximately two months before the earthquake. Approximately two months after the earthquake, this subsample of CASEN was reinterviewed, resulting in the longitudinal survey called Post Earthquake Survey (EPT; see MIDEPLAN 2011 for a description of the survey). To our knowledge, the EPT is the first survey that provides measurements of a representative sample of the same individuals before and after a natural disaster. This longitudinal structure, in addition to a wide range of socioeconomic and health-related measures, and a relatively large sample size, make it a very valuable data set to study the impact of an earthquake. Here, we use the proposed weighting methodology to adjust for unit nonresponse and recover the covariate structure of the 2009 CASEN sample on the 2010 EPT sample of respondents. Specifically, the goal is to make the empirical distributions of the 2009 observed covariates be as similar as possible in the target CASEN sample and the weighted sample of EPT respondents with weights of minimum variance. For an analysis of the effect of the earthquake on posttraumatic stress using the EPT see Zubizarreta et al. (2013a).

5.2 Adjustments with weights that have minimum variance

For illustration, we center our analysis on the southern regions of Chile; specifically, on the regions of Los Lagos, Aysén, and Magallanes y la Antártica Chilena, which comprise 7300 out of the 71460 households in CASEN 2009, and 1542 out of the 22134 households in the EPT. The analysis presented here readily extends to the other regions of the country sampled by the EPT. We compare two weighting schemes: inverse probability weights derived from logistic regression and SBW with different levels of covariate adjustment (for comparisons with other weighting schemes, see the simulation study in section 6). Table 1 shows the covariate balance and coefficient of variation of the weights resulting from these two weighting schemes.

In the table, for each weighting scheme the column labeled "Target" describes the 2009 CASEN sample before the 2010 EPT follow-up and the column "Weighted," the structure of the 2010 follow-up sample after weighting. We observe that the logistic regression weights adjust well for most of the covariates but not for regions. A question with logistic regression is: having found imbalances after weighting, how can we reduce these imbalances? One possible answer would be to fit a different probability model, perhaps adding higher order terms and interactions for the region indicators, and hope that this would result in better balance. In contrast, with (1) one can directly target covariate balance with the constraints (1.2). The second set of columns of Table 1 (denoted by SBW no constraints) shows balance and the coefficient of variation of the weights when the constraints (1.2) are eliminated from the problem, whereas the third and fourth sets of columns (SBW 1e-01sd and SBW 1e-03sd respectively) show the same results when all the differences in means are constrained to be at most 0.1 and 0.001 absolute standard deviations after weighting. The second set of columns is useful for confirmatory purposes, showing that the coefficient of variation of the weights is essentially zero. In the table, we observe how covariate imbalances decrease as we make the constraints more stringent, and also the cost that this has in terms coefficient of variation of the weights. This coefficient increases from 0.68 to 0.79 when the differences in means are constrained from 0.1 to 0.001 standard deviations. On the other hand, the coefficient of variation of the weights derived from logistic regression is 0.34, but it allows for covariate imbalances substantially greater, equal to 0.33 standard deviations (corresponding to the covariate Region of Los Lagos). Here, the choice of δ that produces approximately the same coefficient of variation as the logistic regression weights is 0.21.

It is worth commenting on the precision of the adjustments achieved by the SBW. In the fourth set of columns (where all the differences in means are constrained to be at most 0.001 standard deviations) we observe that the resulting differences in means are essentially zero for all the covariates except for the total monthly income where the differences are approximately equal to 1000 pesos (roughly 2 dollars per month). With the proposed weighting method one can directly target balance (meaning that one can specify the largest imbalance that one is willing to tolerate) and adjust for covariates in a very precise manner with weights that have minimum variance.

Importantly, one can explore and regulate the trade-off there is between covariate balance and the variability of the weights for a given data set. For instance, one can plot the relationship between δ and the coefficient of variation of the weights as in Figure 2. Here, we observe that the coefficient of variation is nearly constant and equal to 0.06 for values of δ between 1 and 0.37 standard deviations, and that after this value it increases approximately linearly with δ. In this example, for values of δ greater than 0.37 standard deviations there is a clear trade-off between balance and variability, but in other applications (such as the Kang and Schafer study below) there may be room in the data to find weights that adjust much better for covariates while not increasing much their variance.

In this example, δ is constant for every covariate, but note that the covariate balance constraints (1.2) are indexed by the covariates, $p = 1, ..., P$, so different covariates may have a different value of δ. The interpretation of δ depends on the units of the covariate to be balanced. Here the covariates are standardized so δ is expressed in standard deviations.

Table 1 shows that the SBW can adjust very precisely for means of covariates and also for entire marginal distributions such as per capita income. This is done by defining indicators for its quantiles and balancing the means (see Zubizarreta 2012 for a related discussion in the context of bipartite matching in observational studies). Similarly, the SBW can adjust for joint distributions by following a similar procedure. Table 2 shows balance for the interactions of region and zone. Here, we constrained the differences in means to be at most 0.001 standard deviations away. Likewise, one can do this for three or more combinations of covariates.

In the last columns of Table 1 all the differences in means were constrained to be at most 0.001 standard deviations apart. A natural question is whether all these differences can be constrained to be exactly equal to zero. With this data set the answer is yes and then the coefficient of variation of the weights equals 0.79. It is important to note that when all the constraints in (1) are equality constraints the resulting problem can be solved analytically by using Lagrange multipliers and solving a linear system of equalities (see, for instance, section 2.8 of Fuller 2009 and Rao and Singh 2009). While this is a useful fact, the more general problem (1) is better suited for practice as one may not need the weighted differences to be exactly zero and one may want to explore the trade-off between covariate balance and stability of the weights.

5.3 How much does tightening a balance constraint increase the variance of the weights?

Figure 1 shows the absolute standardized dual variables associated to each of the covariate balance constraints. In the figure the absolute value of each dual variable is standardized by the value of the objective function at the optimum. In this way, each transformed dual variable in the plot tells us the proportional reduction or increase in the variance of the weights that we would obtain if the corresponding covariate balance constraint was relaxed or tightened in one unit. For example, in the first weighting scheme (which limits all the differences in means to be at most 0.1 standard deviations apart) if the constraint of Region of Aysén was relaxed by 1%, then the variance of the weights would decrease by 6%. Similarly, in the fourth weighting scheme (which forces all the differences in means to be at most 0.001 standard deviations away) if the constraint for decile 10 was relaxed to allow differences in means 1% greater then the variance of the weights would decrease by 1%. With the three weighting schemes we can see that we can tighten the corresponding covariate balance constraints of all the optimal dual variables equal to zero in one unit without increasing the variance of the weights. In principle, one could use these dual variables to build to an automated procedure for optimally balancing covariates with weights that have minimum variance.

6 Simulation study: Kang and Schafer (2007)

In an interesting study, Kang and Schafer (2007) evaluate the performance of various methods that use inverse probability weighting for estimating a population mean from incomplete outcome data. The authors focus on the performance of doubly robust estimators (Robins et al. 1994; Scharfstein et al. 1999) when neither the probability model nor the outcome model are severely misspecified. They find that methods that use inverse probability weights are sensitive to misspecification of the probability model when the estimated weights are highly variable, regardless of whether the methods are doubly robust or not. Robins et al. (2007) reply to Kang and Schafer (2007) and there are alternative doubly robust methods (e.g. Rotnitzky et al. 2012). In their evaluation, Kang and Schafer (2007) use the following simulation study, which has become a standard setting for assessing the performance of weighting methods for causal inference and estimation with incomplete outcome data (e.g. Tan 2010, Rotnitzky et al. 2012, Imai and Ratkovic 2014).

6.1 Study layout

In this simulation study, $\boldsymbol{U}_i^\top = (U_{i1}, U_{i2}, U_{i3}, U_{i4})$ is a vector of unobserved covariates independently sampled from a multivariate normal distribution. Outcomes are generated using the linear model $Y_i = 210 + 27.4U_{i1} + 13.7U_{i2} + 13.7U_{i3} + 13.7U_{i4} + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0,1)$, and the missing data indicators Z_i distribute $\mathcal{Ber}(\pi_i)$ where the probability model is given by $\pi_i = \text{expit}(-U_{i1} + 0.5U_{i2} - 0.25U_{i3} - 0.1U_{i4})$ for $i = 1, ..., n$. The researcher does not observe the actual $\boldsymbol{U}_i$'s but instead a transformation of them, $\boldsymbol{X}_i^\top = (X_{i1}, X_{i2}, X_{i3}, X_{i4}) = (\exp(U_{i1}/2), U_{i1}/(1 + \exp(U_{i1})) + 10, (U_{i1}U_{i3}/25 + 0.6)^3, (U_{i2} + U_{i4} + 20)^2)$. In this way the correct outcome and probability models are linear functions of $\log(X_{i1})$, X_{i2}, $X_{i1}^2 X_{i2}$, $1/\log(X_{i1})$, $X_{i3}/\log(X_{i1})$ and $X_{i4}^{1/2}$. This data generation mechanism produces an average response rate equal to 0.5 and a population mean of 210.

Kang and Schafer (2007) study the performance of estimators when both the outcome and nonresponse models are correctly specified (meaning that the researcher used the correct transformation of the $\boldsymbol{X}_i$'s in each of the models), when both models are misspecified (meaning that the researcher used the observed $\boldsymbol{X}_i$'s in each of the models), and the combinations where only one of the two models is misspecified. The estimators they consider include: the Horvitz-Thompson estimator, $\hat{\mu}_{\text{HT}} = \frac{1}{n}\sum_{i=1}^n Z_i \hat{\pi}_i^{-1} Y_i$ (Horvitz and Thompson 1952); the inverse probability weighting estimator, $\hat{\mu}_{\text{IPW}} = \frac{\sum_{i=1}^n Z_i \hat{\pi}_i^{-1} Y_i}{\sum_{i=1}^n Z_i \hat{\pi}_i^{-1}}$ (see, for instance, Hirano and Imbens 2001); the weighted least squares estimator, $\hat{\mu}_{\text{WLS}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{\text{WLS}}$ where $\hat{\boldsymbol{\beta}}_{\text{WLS}} = \left(\sum_{i=1}^n Z_i \hat{\pi}_i^{-1} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1} \sum_{i=1}^n Z_i \hat{\pi}_i^{-1} \boldsymbol{X}_i Y_i$ (for instance see Kang and Schafer 2007);

and the doubly robust estimator, $\hat{\mu}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{\text{OLS}} + Z_i\pi_i^{-1}(Y_i - \boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{\text{OLS}})\right)$ where $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\sum_{i=1}^{n} Z_i\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)^{-1}\sum_{i=1}^{n} Z_i\boldsymbol{X}_iY_i$ (Robins et al. 1994).

For each of these estimators we evaluate the following weighting schemes: true weights, obtained by inverting the probabilities actually used to generate the missing data; logit weights, obtained by inverting the estimated probabilities of missing data from logit models correctly specified and misspecified; trimmed logit weights, obtained by trimming the previous weights at their 95$^{\text{th}}$ percentile; Covariate Balancing Propensity Score (CBPS) weights, obtained by using the novel method of Imai and Ratkovic (2014) with the default settings of the `CBPS` package in `R` with the models correctly specified and misspecified; SBW, obtained by constraining the absolute differences in weighted means of the covariates to be at most 1e-01, 1e-02, 1e-03 and 1e-04 standardized differences, both for the correct and incorrect covariates.

6.2 Analysis of results

6.2.1 Weight diagnostics

Before looking at the outcomes, and therefore as part of the design of the study, we evaluate the weights based on how well they balance covariates and how stable they are. Table 3 shows weight diagnostics for 1000 simulated data sets of size 1000 for these two criteria. In the table, the symbols ✓ and x denote whether the weights are correctly specified by using the unobserved covariates $\boldsymbol{U}$ or the observed covariates $\boldsymbol{X}$, respectively. The stability of the weights is evaluated in terms of their mean coefficient of variation (CV) and mean 95th and 99th normalized percentiles calculated across simulations. The last four columns of the table show covariate balance expressed in mean absolute standardized differences in means after weighting across simulations.

In the table, the SBW exhibit a better performance both in terms balance and stability. In terms of stability, the coefficient of variation of the SBW is always lower than with other weighting schemes, regardless of the tolerances imposed on balance. Also, the presence of aberrant weights as measured by the 95th and 99th percentiles is lower in the SBW. For instance, with other weighting schemes the 99th percentile of the weights is between 4 and 22.8 whereas with the SBW it is between 1.8 and 2.9. Figure 3 allows us to visualize the presence of outliers using both the correct and incorrect observed covariates. Note that since the SBW are constrained to sum up to one we can interpret them as propensities. In Figure 3 we can see that the propensities implied by the SBW are less dispersed than the true propensities.

While achieving greater stability, the SBW also tend to balance covariates better. In Table 3, we note that with the SBW we can finely-tune covariate balance and as a result observe a gradient in the imbalances as we constrain the weighted differences in means to be smaller and smaller. In fact, either with the correct or incorrect covariates, the absolute standardized differences in means decrease from 0.1 to 0.01 to 0.001, and so on, as we constrain the differences to be no greater than these values (for x_3/u_3 and x_4/u_4 there are a few differences that are not equal to these values, but the differences are smaller, as required). In general, tighter balance on the covariates may come at a cost in terms of stability, although in this simulated study it is not large in comparison to the performance of the other weighting methods. Indeed, with the incorrect covariates, when we constrain the absolute standardized differences in means to be smaller or equal than 0.01, the CV increases the most, jumping from 0.38 to 0.53 approximately, but this CV is still smaller than with other weights.

6.2.2 Outcome assessments

We now evaluate the performance of the different weighting schemes using outcomes and the four estimators presented in subsection 6.1. We focus our discussion on the RMSE of the estimates to emphasize the importance of the stability of the weights in addition to bias adjustment. Table 4 shows the results.

In the table, it is first worth noting the performance of the SBW when both the probability and outcome models are incorrectly specified by using the observed covariates $\boldsymbol{X}$. Remember that this is the situation faced in practice: not knowing the true transformation of the observed covariates that generate the outcomes. In this case, the RMSEs of the SBWs are smaller than those of the other weighting schemes regardless of the estimator, but provided an adjustment smaller than 1e-02 standard deviations. In particular, it is worth highlighting the substantial improvement of the RMSE of the HT and the DR estimators when using the SBW. In fact, the RMSEs of the HT and the DR estimators decrease from 235 and 118 when using the weights derived from logistic regression to a value around 2 when using the SBW.

When the probability model is correctly specified by using the correct transformation of the observed covariates, the results vary to a limited extent depending on the estimator and the level of adjustment of the SBW. With the HT estimator, the RMSE of the SBW is smaller regardless of the level of adjustment for the means of the transformed covariates. With the IPW estimator, the RMSE of the SBW is smaller provided the adjustment is smaller than 1e-02 standard deviations.

When the probability model is correctly specified and the outcome model is incorrectly specified, there is not a very big difference between the different weighting schemes with the DR and WLS. Interestingly, with the DR and WLS, the performance of the SBW is slightly worse with smaller levels of adjustment (with the WLS estimator the trimmed logit weights have the smallest RMSE of 1.34, whereas the SBW that adjust up to 1e-01 standard deviations have a RMSE of 1.36 and the SBW that adjust up to 1e-04 standard deviations have a RMSE of 2.34). Finally, when both models are correctly specified and one uses the DR or WLS estimators, the performance of the different weighting schemes is essentially the same.

In brief, in the Kang and Schafer (2007) study the SBW perform especially well with the observed covariates. With the transformed covariates the performance of the SBW is very good but not markedly different from that of the other weighting schemes, and in two cases a little worse than that of other weighting schemes. In this simulated example, the right level of adjustment for the SBW seems to be somewhat smaller than 0.1 standardized differences, perhaps of 0.01.

7 Guidance for practice

How to choose δ in practice? What forms of covariate balance enforce in the constraints (1.2)? And for which covariates specifically? While more work needs to be done to definitively answer these questions, here we provide some general guidelines.

First, we emphasize that these decisions should be made as part of the design of the study, without using the outcomes. Also, we recall the objective of the proposed methodology: among the weights that balance covariates, find the weights that are least variable. Here, in a spirit similar to matching in observational studies, improving efficiency is subject to reducing biases, so we focus primarily on decisions for covariate balance (see section 8.7 of Rosenbaum 2010 for a related discussion).

As discussed in section 4, bias reduction due to weighting depends on certain covariates $\boldsymbol{X}_p$, transformations of these covariates f_p or g_h, and their relative importance in the regression function. Since weighting is part of the design of the study, we recommend making the corresponding decisions about covariates, transformations and values of δ based on substantive knowledge of the problem at hand, and then fine-tuning δ using the data, but without outcomes.

In relation to the covariates, for reducing biases through weight adjustments, the covariates need to be associated both with the treatment (nonresponse) and the outcome. Broadly, we recommend deciding which covariates to include in the balancing constraints (1.2) using domain-expert knowledge and rather err on the side of including more covariates than fewer. In terms of the form of balance, as propositions 4.1 and 4.2 suggest, some covariates are stronger predictors of the outcome than others, and some covariates may have nonlinear effects. If based on subject-matter knowledge one conjectures a nonlinear relationship between a covariate and the outcome, then one should balance not only the mean of that covariate but also polynomial or spline transformations as discussed in sections 3.1 and 4. Similarly, if one envisions that the interaction of certain covariates may be relevant, then one should mean balance quantiles of interactions of these covariates, thereby balancing their joint distribution. Finally, in relation to δ, in the context of matching in observational studies, general practice is to balance means to 0.1 absolute standardized differences or less (Normand et al. 2001), but again this choice is problem-specific. In some studies, such as the Kang and Schafer (2007) simulation study above, tighter levels of balance perform better. In principle, covariates that are stronger predictors of the outcome should be balanced more tightly. On the other hand, if one is going to use weights as a complement to model-based adjustments (as in Robins et al. 1994), then one might be willing to tolerate greater imbalances.

In brief, we recommend choosing the covariates, form of balance and maximum imbalances δ based on substantive knowledge of the problem at hand, and then, based on the data but without looking at the outcomes, fine-tuning δ. For this, a plot such as Figure 2 as well as the dual variables of the constraints (1.2) may be used. In these analyses, if the coefficient of variation does not increase too much, then it may be worth tightening balance. How to optimally balance covariates is an open-ended question both in the matching and weighting literatures that deserves close attention. Here we propose a method that gives the researcher a fine degree of control over covariate balance adjustments with weights of minimum variance.

8 Summary and concluding remarks

Weighting methods that adjust for observed covariates are common both in causal inference and estimation with incomplete outcome data. In these settings, the goal of weights is (i) to adjust for the empirical distributions of the observed covariates, and (ii) to yield stable estimates for the parameters of interest. This paper proposes a new weighting method that aims at (i) and (ii) directly. Specifically, by solving a convex optimization problem, this method finds the weights of minimum variance that balance the empirical distribution of the observed covariates up to levels prespecified by the researcher. As discussed, this method allows the researcher to adjust very precisely for the observed covariates, and, for example, directly adjust for means of the observed covariates and for other features of their marginal and joint distributions such quantiles of interactions of pairs or triples of observed covariates, thus balancing entire two- and three-way marginals. In this problem, duality theory provides useful insight into the behavior of the variance of the optimal weights in relation to the level of covariate balance adjustment, informing which covariate balance constraints are "expensive" in terms of the variance of the weights and which covariates can be balanced more tightly at no cost at all. Conceptually, this method is based on a well-defined optimization problem that can be solved in polynomial time, so relatively large data sets can be handled quickly. In practice this method is easy to use as implemented in the new package `sbw` for `R`.

The proposed weighting method can be applied to a variety of settings in causal inference and estimation with incomplete outcome data. For instance, in randomized experiments the SBW can be used for adjusting treatment effect estimates by post-stratification (Miratrix et al. 2013). In causal inference in observational studies, these weights can be used to strengthen an instrumental variable (Baiocchi et al. 2010; Zubizarreta et al. 2013b). Also, in longitudinal studies of treatment effects they can be used as an alternative to inverse probability weighting (where weights are multiplied successively tending to achieve very large values; Hogan and Lancaster 2004), and as a complement to marginal structural models (Joffe et al. 2004; Cole and Hernán 2008). In sample surveys, this method can be used to adjust for nonresponse in cross-sectional surveys, attrition in longitudinal surveys, and noncompliance in (broken) randomized experiments. Also, this method can be extended to obtain representativeness of randomized experiments and matched observational studies (Stuart et al. 2011).

In an interesting paper, Stuart (2010) draws a parallel between weighting and matching in observational studies. Unlike some model-based approaches and in a manner similar to matching, the weighting method proposed in this paper forces the researcher to look closely at the data by checking covariate balance and the degree of dispersion of the weights. Also, this method is used as part of the design of the study because it does not require outcomes and thus it prevents the selection of a model that suits the hypotheses of the investigation. In this regard, Rubin (2008) advocates methods that help to separate the design and analysis stages of the study, and takes the position that observational studies should be designed to approximate the template randomized experiment. In a somewhat parallel way, this paper advocates the use of weights for covariate adjustment to approximate the structure of a target sample or population and yield stable estimates. Small predicted probabilities for sample selection are common in practice both in estimation with incomplete outcome data and causal inference and this paper offers an alternative method to build weights in such instances.

References

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010), “Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants,” *Journal of the American Statistical Association*, 105, 1285–1296.

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press.

Cole, S. R. and Hernán, M. A. (2008), “Constructing Inverse Probability Weights for Marginal Structural Models,” *American Journal of Epidemiology*, 168, 656–664.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), “Dealing with Limited Overlap in Estimation of Average Treatment Effects,” *Biometrika*, 96, 187–199.

Dawid, A. P. (1979), “Conditional Independence in Statistical Theory,” *Journal of the Royal Statistical Society: Series B*, 41, 1–31.

Efron, B. (1971), “Forcing a Sequential Experiment to be Balanced,” *Biometrika*, 58, 403–407.

Elliott, M. R. (2008), “Model Averaging Methods for Weight Trimming,” *Journal of Official Statistics*, 24, 517–540.

Fuller, W. A. (2009), *Sampling Statistics*, vol. 560, John Wiley & Sons.

Gelman, A. and Meng, X.-L. (eds.) (2004), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, John Wiley & Sons.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, Springer.

Hirano, K. and Imbens, G. W. (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.

Hogan, J. W. and Lancaster, T. (2004), “Instrumental Variables and Inverse Probability Weighting for Causal Inference from Longitudinal Observational Studies,” *Statistical Methods in Medical Research*, 13, 17–48.

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

Imai, K. and Ratkovic, M. (2014), "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society: Series B*, 76, 243–263.

Joffe, M. M., Have, T. R. T., Feldman, H. I., and Kimmel, S. E. (2004), "Model Selection, Confounder Control, and Marginal Structural Models: Review and New Applications," *The American Statistician*, 58, 272–279.

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

Lee, B. K., Lessler, J., and Stuart, E. A. (2010), "Improving Propensity Score Weighting Using Machine Learning," *Statistics in Medicine*, 29, 337–346.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley, 2nd ed.

MIDEPLAN (2011), "Ficha Técnica Encuesta Post Terromoto," *http://www.ministeriodesarrollosocial.gob.cl/encuesta-post-terremoto/index.html.*

Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013), "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments," *Journal of the Royal Statistical Society: Series B*, 75, 369–396.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., and McNeil, B. J. (2001), "Validating Recommendations for Coronary Angiography Following Acute Myocardial Infarction in the Elderly: A Matched Analysis Using Propensity Scores," *Journal of Clinical Epidemiology*, 54, 387–398.

Papadimitriou, C. (1994), *Computational Complexity*, Reading (Mass.): Addison-Wesley.

Rao, J. N. K. and Singh, A. C. (2009), "Range Restricted Weight Calibration for Survey Data Using Ridge Regression," *Pakistan Journal of Statistics*, 25, 371–384.

Robins, J. and Wang, N. (2000), "Inference for Imputation Estimators," *Biometrika*, 87, 113–124.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

Robins, J. M., Sued, M., Quanhong, L. G., and Rotnitzky, A. (2007), "Comment:

Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable," *Statistical Science*, 22, 544–559.

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

— (2010), *Design of Observational Studies*, Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012), "Improved double-robust estimation in missing data and causal inference models," *Biometrika*, 99, 439–456.

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

— (2005), "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, 100, 322–331.

— (2008), "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2, 808–840.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models: Rejoinder," *Journal of the American Statistical Association*, 94, 1135–1146.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin.

Stuart, E. A. (2010), "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, 25, 1–21.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011), "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials," *Journal of the Royal Statistical Society: Series A*, 174, 369–386.

Tan, Z. (2010), "Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting," *Biometrika*, 97, 661–682.

Tsybakov, A. (2008), *Introduction to Nonparametric Estimation*, Springer.

USGS (2014), "Largest Earthquakes in the World Since 1900," *http://earthquake.usgs.gov/earthquakes/world/10_largest_world.php.*

Wright, S. (1997), *Primal-Dual Interior-Point Methods*, Society for Industrial and Applied Mathematics.

Yang, D., Small, D., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal Matching with Minimal Deviation from Fine Balance in a Study of Obesity and Surgical Outcomes," *Biometrics*, 68, 628–636.

Zubizarreta, J. R. (2012), "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

Zubizarreta, J. R., Cerdá, M., and Rosenbaum, P. R. (2013a), "Effect of the 2010 Chilean Earthquake on Posttraumatic Stress: Reducing Sensitivity to Unmeasured Bias Through Study Design," *Epidemiology*, 24, 79–87.

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2011), "Matching for Several Sparse Nominal Variables in a Case-Control Study of Readmission Following Surgery," *The American Statistician*, 65, 229–238.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S. A., and Rosenbaum, P. R. (2013b), "Stronger Instruments Via Integer Programming in an Observational Study of Late Preterm Birth Outcomes," *Annals of Applied Statistics*, 7, 25–50.

Appendix A: Proofs

Proof. (Proposition 4.1.) The proof is straightforward:

$$\begin{aligned}
\mathrm{E}\left(\hat{Y}_w - \overline{Y}_N\right) &= \mathrm{E}\left(\mathrm{E}\left(\hat{Y}_w - \overline{Y}_N | \boldsymbol{X}\right)\right) \\
&= \mathrm{E}\left(\mathrm{E}\left(\sum_{i=1}^{r} \frac{w_i Y_i}{r} - \sum_{i=1}^{N} \frac{Y_i}{N} \;\middle|\; \boldsymbol{X}\right)\right) \\
&= \mathrm{E}\left(\mathrm{E}\left(\sum_{i=1}^{r} \frac{w_i}{r}\left(\alpha + \sum_{p=1}^{P} \beta_p X_{i,p} + \varepsilon_i\right) - \sum_{i=1}^{N} \frac{1}{N}\left(\alpha + \sum_{p=1}^{P} \beta_p X_{i,p} + \varepsilon_i\right) \;\middle|\; \boldsymbol{X}\right)\right) \\
&= \sum_{p=1}^{P} \beta_p \left(\sum_{i=1}^{r} \frac{w_i X_{i,p}}{r} - \sum_{i=1}^{N} \frac{X_{i,p}}{N}\right) \\
&\leq \delta \sum_{p=1}^{P} |\beta_p|.
\end{aligned}$$

□

Proof. (Proposition 4.2.) Let $\gamma_{k,p}$, $k = 1, ..., K-1$, $p = 1, ..., P$, be the coefficient of the Taylor expansion of order k around each $\xi_{j,p}$. This is $\gamma_{k,p} := \frac{f^{(k)}(\xi_{j,p})}{k!}$. Let $R_{i,j,K,p}$ be the residual of this Taylor expansion. By the Lagrange error bound, $R_{i,j,K,p} \leq \left|\frac{f^{(K)}(\xi_{j,p})}{K!}\right| (l_p/2)^K$ for all $j = 1, ..., M_p/l_p$, $p = 1, ..., P$. Then,

$$
\begin{aligned}
\mathrm{E}\left(\hat{Y}_w - \overline{Y}_N\right) &= \mathrm{E}\left(\mathrm{E}\left(\sum_{i=1}^{r} \frac{w_i Y_i}{r} - \sum_{i=1}^{N} \frac{Y_i}{N} \,\middle|\, \boldsymbol{X}\right)\right) \\
&= \mathrm{E}\left(\mathrm{E}\left(\sum_{i=1}^{r} \frac{w_i}{r}\left(\alpha + \sum_{p=1}^{P} f_p(X_{i,p}) + \varepsilon_i\right) - \sum_{i=1}^{N} \frac{1}{N}\left(\alpha + \sum_{p=1}^{P} f_p(X_{i,p}) + \varepsilon_i\right) \,\middle|\, \boldsymbol{X}\right)\right) \\
&= \sum_{p=1}^{P}\left(\sum_{i=1}^{r} \frac{w_i}{r} f_p(X_{i,p}) - \sum_{i=1}^{N} \frac{1}{N} f_p(X_{i,p})\right) \\
&= \sum_{p=1}^{P}\left(\sum_{i=1}^{r} \sum_{j=1}^{M_p/l_p}\left(\sum_{k=1}^{K-1} \frac{w_i}{r}\left(\gamma_{k,p}\, \widetilde{X}^{k}_{i,j,p} + R_{i,j,K,p}\right)\right)\right. \\
&\quad \left. - \sum_{i=1}^{N} \sum_{j=1}^{M_p/l_p}\left(\sum_{k=1}^{K-1} \frac{1}{N}\left(\gamma_{k,p}\, \widetilde{X}^{k}_{i,j,p} + R_{i,j,K,p}\right)\right)\right) \\
&= \sum_{p=1}^{P} \sum_{j=1}^{M_p/l_p} \sum_{k=1}^{K-1}\left(\sum_{i=1}^{r} \frac{w_i}{r}\left(\gamma_{k,p}\, \widetilde{X}^{k}_{i,j,p} + R_{i,j,K,p}\right) - \sum_{i=1}^{N} \frac{1}{N}\left(\gamma_{k,p}\, \widetilde{X}^{k}_{i,j,p} + R_{i,j,K,p}\right)\right) \\
&= \sum_{p=1}^{P} \sum_{j=1}^{M_p/l_p} \sum_{k=1}^{K-1} \gamma_{k,p}\left(\sum_{i=1}^{r} \frac{w_i}{r}\, \widetilde{X}^{k}_{i,j,p} - \widetilde{X}^{k}_{i,j,p}\right) + \sum_{p=1}^{P} \sum_{j=1}^{M_p/l_p}\left(\sum_{i=1}^{r} \frac{w_i}{r} R_{i,j,K,p} - \frac{1}{N} R_{i,j,K,p}\right) \\
&\leq \delta \sum_{p=1}^{P} \sum_{j=1}^{M_p/l_p} \sum_{k=1}^{K-1} |\gamma_{k,p}| + P \frac{M_p}{l_p} 2 \left|\frac{f^{(K)}(\xi_{j,p})}{K!}\right| (l_p/2)^K. \\
&= \delta \sum_{p=1}^{P} \sum_{j=1}^{M_p/l_p} \sum_{k=1}^{K-1} |\gamma_{k,p}| + 2P \frac{M_p}{l_p} L.
\end{aligned}
$$

□

Table 1: Covariate balance in the EPT survey using logistic regression inverse probability weights and stable balancing weights with different levels of adjustment.

Covariate	Logit weights		SBW no constr.		SBW 1e-01sd		SBW 1e-03sd	
	Target	Weighted	Target	Weighted	Target	Weighted	Target	Weighted
Region of Los Lagos	0.75	0.60	0.75	0.37	0.75	0.70	0.75	0.75
Region of Aysén	0.15	0.23	0.15	0.36	0.15	0.17	0.15	0.15
Region of Magallanes	0.10	0.17	0.10	0.27	0.10	0.13	0.10	0.10
Rural zone	0.52	0.45	0.52	0.34	0.52	0.48	0.52	0.52
Female	0.28	0.28	0.28	0.29	0.28	0.27	0.28	0.28
Indigenous	0.20	0.20	0.20	0.20	0.20	0.24	0.20	0.20
Married or cohabitating	0.69	0.69	0.69	0.68	0.69	0.70	0.69	0.69
Divorced or widow	0.19	0.19	0.19	0.20	0.19	0.19	0.19	0.19
Single	0.12	0.12	0.12	0.12	0.12	0.11	0.12	0.12
Household size	3.28	3.25	3.28	3.20	3.28	3.28	3.28	3.28
Education (years)	7.78	7.94	7.78	8.21	7.78	7.90	7.78	7.78
Employed	0.67	0.68	0.67	0.69	0.67	0.69	0.67	0.67
Unemployed	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Inactive	0.30	0.29	0.30	0.28	0.30	0.28	0.30	0.30
Total income (1000 pesos)	633.87	652.62	633.87	685.59	633.87	632.86	633.87	632.97
Per capita income decile 1	0.10	0.10	0.10	0.09	0.10	0.09	0.10	0.10
Per capita income decile 2	0.10	0.10	0.10	0.09	0.10	0.10	0.10	0.10
Per capita income decile 3	0.10	0.10	0.10	0.09	0.10	0.11	0.10	0.10
Per capita income decile 4	0.10	0.10	0.10	0.09	0.10	0.09	0.10	0.10
Per capita income decile 5	0.10	0.10	0.10	0.09	0.10	0.11	0.10	0.10
Per capita income decile 6	0.10	0.09	0.10	0.08	0.10	0.09	0.10	0.10
Per capita income decile 7	0.10	0.10	0.10	0.10	0.10	0.09	0.10	0.10
Per capita income decile 8	0.10	0.10	0.10	0.11	0.10	0.09	0.10	0.10
Per capita income decile 9	0.10	0.11	0.10	0.12	0.10	0.11	0.10	0.10
Per capita income decile 10	0.10	0.11	0.10	0.13	0.10	0.11	0.10	0.10
Coefficient of variation	0.34		1.15e-14		0.68		0.79	
Time (seconds)	0.11		0.93		0.97		1.79	

Note: For each weighting scheme, the column Target describes the EPT sample in 2009 before the 2010 follow-up; the column Weighted shows the structure of the 2010 follow-up sample after weighting. The terms 1e-01sd and 1e-03sd denote that the corresponding SBW constrain the absolute differences in means to be at most 0.1 and 0.001 standard deviations after weighting.

Table 2: Balance of the two-way marginal of region and zone in the EPT survey using stable balancing weights. All the other covariates are balanced as in the last column of Table 1. Here the coefficient of variation of the weights is 0.8.

	SBW 1e-03sd	
Region/Zone	Target	Weighted
Los Lagos/Urban	0.32	0.32
Los Lagos/Rural	0.43	0.43
Aysén/Urban	0.09	0.09
Aysén/Rural	0.06	0.06
Magallanes/Urban	0.07	0.07
Magallanes/Rural	0.03	0.03

Note: The term 1e-03sd denotes that the stable balancing weights constrain the absolute differences in means to be at most 0.001 standard deviations after weighting.

Figure 1: Absolute standardized dual variables for the covariate balance constraints in the earthquake study. Each transformed dual variable in the plot quantifies the proportional reduction in the variance of the weights that would be obtained if the corresponding covariate balance constraint was relaxed in one unit.

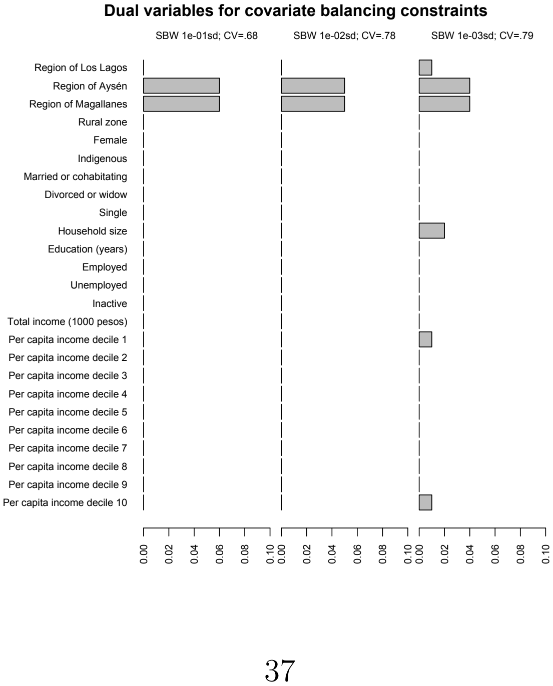

Figure 2: Trade-off between tighter covariate balance and the variability of the weights in the earthquake study.

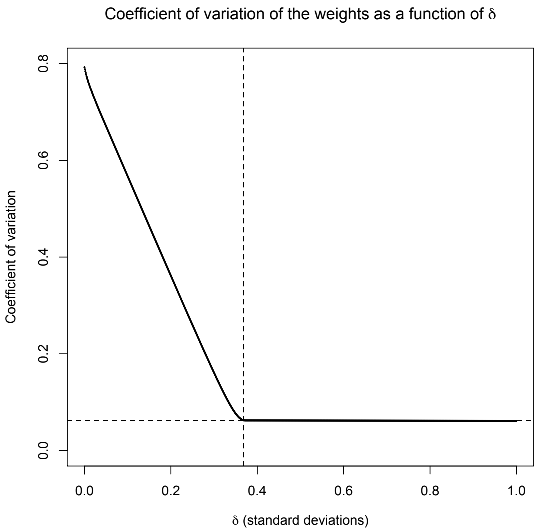

Figure 3: Distribution of normalized weights in the Kang and Schafer (2007) simulation study.

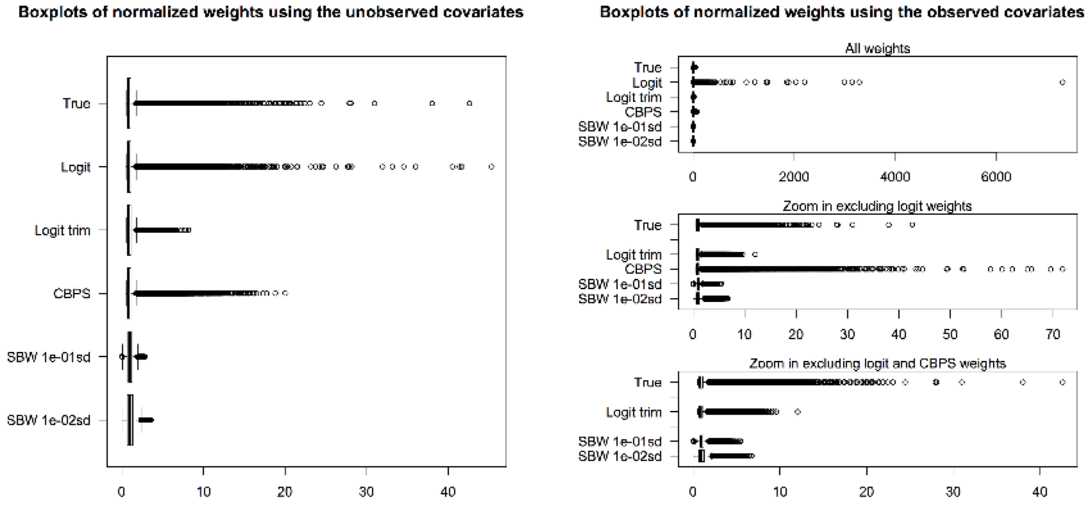

Table 3: Weight diagnostics in the Kang and Schafer (2007) simulation study.

Weighting scheme	Specification	Stability			Balance (mean abs. std. dif.)			
		CV	p95	p99	x_1/u_1	x_2/u_2	x_3/u_3	x_4/u_4
True	✓	1.0645	3.8075	7.6240	0.0486	0.0383	0.0349	0.0359
Logit	✓	1.1207	3.9607	8.1402	0.0330	0.0245	0.0208	0.0210
Logit	x	9.2911	4.6321	22.7846	0.0352	0.0218	0.0170	0.0163
Logit trim	✓	0.6680	3.9588	3.9607	0.4447	0.0242	0.1961	0.0822
Logit trim	x	0.7241	4.6282	4.6321	0.0284	0.0210	0.0526	0.0318
CBPS	✓	0.9952	3.6144	7.0979	0.0367	0.0245	0.0195	0.0190
CBPS	x	6.4058	3.6862	14.6673	0.0710	0.0190	0.0578	0.0306
SBW 1e-01sd	✓	0.3515	1.5691	1.8004	0.1000	0.1000	0.0927	0.0476
SBW 1e-02sd	✓	0.5011	1.8345	2.1766	0.0100	0.0100	0.0100	0.0095
SBW 1e-03sd	✓	0.5178	1.8677	2.2234	0.0010	0.0010	0.0010	0.0010
SBW 1e-04sd	✓	0.5195	1.8711	2.2282	0.0001	0.0001	0.0001	0.0001
SBW 1e-01sd	x	0.3802	1.8229	2.3658	0.1000	0.1000	0.0438	0.0751
SBW 1e-02sd	x	0.5272	2.1049	2.8469	0.0100	0.0100	0.0096	0.0091
SBW 1e-03sd	x	0.5432	2.1332	2.8981	0.0010	0.0010	0.0010	0.0010
SBW 1e-04sd	x	0.5448	2.1360	2.9033	0.0001	0.0001	0.0001	0.0001

Note: For different weighting schemes, the symbols ✓ and x denote whether the weights are correctly specified by using the unobserved covariates $\boldsymbol{U}$ or the observed covariates $\boldsymbol{X}$. CV is the mean coefficient of variation of the weights, and p95 and p99 are the mean normalized 95th and 99th percentiles of the weights across simulations. The terms 1e-01sd, 1e-02sd, and so on, denote that the corresponding SBW constrain the absolute differences in means to be at most 0.1 and 0.01 standard deviations after weighting. In this vein, covariate balance is expressed in terms of mean absolute standardized differences in means between the target and weighted samples for each of the covariates.

Table 4: Performance of estimators using different weighting schemes in the Kang and Schafer (2007) simulation study.

Estimator	Model		Weighting scheme							
	Probability	Outcome	True	Logit	Logit trim	CBPS	SBW 1e-01sd	SBW 1e-02sd	SBW 1e-03sd	SBW 1e-04sd
Bias										
$\hat{\mu}_{HT}$	×	-	-	-50.49	3.32	-1.32	2.21	0.98	0.86	0.85
$\hat{\mu}_{IPW}$	×	-	-	-5.50	1.76	1.21	2.21	0.98	0.86	0.85
$\hat{\mu}_{WLS}$	×	×	-	2.99	2.20	3.00	1.12	0.86	0.85	0.85
$\hat{\mu}_{DR}$	×	×	-	19.21	2.41	3.83	0.81	0.84	0.85	0.85
$\hat{\mu}_{HT}$	✓	-	0.30	0.17	4.25	3.75	3.28	0.42	0.08	0.05
$\hat{\mu}_{IPW}$	✓	-	0.17	0.08	0.89	0.80	3.28	0.42	0.08	0.05
$\hat{\mu}_{WLS}$	×	✓	-	0.04	0.04	0.04	0.04	0.04	0.04	0.04
$\hat{\mu}_{DR}$	×	✓	-	0.06	0.04	0.04	0.04	0.04	0.04	0.04
$\hat{\mu}_{WLS}$	✓	×	-0.13	-0.12	-0.27	-0.04	0.18	-1.68	-1.91	-1.93
$\hat{\mu}_{DR}$	✓	×	-0.04	-0.01	-0.32	-0.07	-0.21	-1.41	-1.54	-1.55
$\hat{\mu}_{WLS}$	✓	✓	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
$\hat{\mu}_{DR}$	✓	✓	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
RMSE										
$\hat{\mu}_{HT}$	×	-	-	235.17	4.56	5.41	2.64	1.79	1.74	1.73
$\hat{\mu}_{IPW}$	×	-	-	13.17	2.33	2.21	2.64	1.79	1.74	1.73
$\hat{\mu}_{WLS}$	×	×	-	3.33	2.59	3.39	1.86	1.74	1.73	1.73
$\hat{\mu}_{DR}$	×	×	-	118.47	2.83	4.55	1.72	1.73	1.73	1.73
$\hat{\mu}_{HT}$	✓	-	10.56	5.08	5.12	5.60	3.52	1.24	1.17	1.17
$\hat{\mu}_{IPW}$	✓	-	2.27	1.85	1.69	1.80	3.52	1.24	1.17	1.17
$\hat{\mu}_{WLS}$	×	✓	-	1.17	1.17	1.17	1.17	1.17	1.17	1.17
$\hat{\mu}_{DR}$	×	✓	-	1.43	1.17	1.17	1.17	1.17	1.17	1.17
$\hat{\mu}_{WLS}$	✓	×	1.47	1.36	1.34	1.38	1.36	2.13	2.32	2.34
$\hat{\mu}_{DR}$	✓	×	1.75	1.69	1.46	1.54	1.42	1.96	2.05	2.06
$\hat{\mu}_{WLS}$	✓	✓	1.17	1.17	1.17	1.17	1.17	1.17	1.17	1.17
$\hat{\mu}_{DR}$	✓	✓	1.17	1.17	1.17	1.17	1.17	1.17	1.17	1.17

Note: For different weighting schemes, the symbols ✓ and × denote whether the weights are correctly specified by using the unobserved covariates $\boldsymbol{U}$ or the observed covariates $\boldsymbol{X}$. The terms 1e-01sd, 1e-02sd, and so on, denote that the corresponding SBW constrain the absolute differences in means to be at most 0.1 and 0.01 standard deviations after weighting.