

Residual Balancing: A Method of Constructing Weights for Marginal Structural Models

Xiang Zhou

Geoffrey T. Wodtke

Harvard University

University of Chicago

December 13, 2019

Abstract

When making causal inferences, post-treatment confounders complicate analyses of time-varying treatment effects. Conditioning on these variables naively to estimate marginal effects may inappropriately block causal pathways and may induce spurious associations between treatment and the outcome, leading to bias. To avoid such bias, researchers often use marginal structural models (MSMs) with inverse probability weighting (IPW). However, IPW requires models for the conditional distributions of treatment and is highly sensitive to their misspecification. Moreover, IPW is relatively inefficient, susceptible to finite-sample bias, and difficult to use with continuous treatments. We introduce an alternative method of constructing weights for MSMs, which we call “residual balancing.” In contrast to IPW, it requires modeling the conditional means of the post-treatment confounders rather than the conditional distributions of treatment, and it is therefore easier to use with continuous treatments. Numeric simulations suggest that residual balancing is both more efficient and more robust to model misspecification than IPW and its variants in a variety of scenarios. We illustrate the method by estimating (a) the cumulative effect of negative advertising on election outcomes and (b) the controlled direct effect of shared democracy on public support for war. Open source software is available for implementing the proposed method.

Direct all correspondence to Xiang Zhou, Department of Sociology, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA; email: xiang_zhou@fas.harvard.edu. The authors benefited from communications with Justin Esarey, Kosuke Imai, Gary King, José Zubizarreta, and participants of the Applied Statistics Workshop at Harvard University, the Political Methodology Speaker Series at MIT, and the 35th Annual Meeting of the Society for Political Methodology at Brigham Young University.

1 Introduction

Social scientists are often interested in estimating the marginal, or population average, effects of treatment in the presence of post-treatment confounding. Post-treatment confounding is common in studies of time-varying treatments, where confounders of future treatments may be affected by prior treatments. For example, political scientists study how the timing and frequency of negative advertising during political campaigns affect election outcomes (e.g., Lau, Sigelman and Rovner 2007; Blackwell 2013). In this context, the decision to run negative advertisements at any given point during a campaign is affected by a candidate’s position in recent polling data, which itself is affected by negative advertising conducted previously. Post-treatment confounding is also common in analyses of causal mediation, where confounders for the effect of the mediator on the outcome may be affected by treatment. For example, when assessing the role of morality in mediating the effects of shared democracy on public support for war, post-treatment variables, such as beliefs about the threat posed by the adversary, may affect both the perceived morality of war and support for military action (Tomz and Weeks 2013).

Adjusting for post-treatment confounders using conventional methods, for example, by naively conditioning, stratifying, or otherwise balancing on them, may engender two different types of bias (Robins 1986, 1999). First, adjusting naively for post-treatment confounders leads to bias from over-control of intermediate pathways because it blocks, or “controls away,” the effect of treatment on the outcome that operates through these variables. Second, adjusting naively for post-treatment confounders can lead to collider-stratification bias if these variables are also affected by unobserved determinants of the outcome, as conditioning on a variable generates a spurious association between its common causes even when these common causes are unconditionally independent (Pearl 2009).

Marginal structural models (MSMs) and the associated method of inverse probability weighting (IPW) avoid these biases and are capable of consistently estimating treatment effects under fairly general conditions (Robins 1999; Robins, Hernan and Brumback 2000; VanderWeele 2015). Compared with more traditional models for time-series cross-sectional data (e.g., fixed effects regression models), MSMs with IPW can better accommodate dynamic causal relationships (Imai and Kim 2019). Specifically, unlike conventional methods, this approach allows past treatments to affect current outcomes (i.e., “carryover effects”) and past outcomes to affect current treatment (i.e., “feedback effects”).

Because of this flexibility, political scientists have increasingly used MSMs with IPW to draw causal inferences from longitudinal data (e.g., Ladam, Harden and Windett 2018; Simmons and Creamer 2019; Zhukov 2017).

Nevertheless, IPW has several important limitations. First, IPW requires models for the conditional distributions of exposure to treatment and/or the mediator, and prior research indicates that it is highly sensitive to their misspecification (e.g., Kang and Schafer 2007; Lefebvre, Delaney and Platt 2008). Second, even if these models are correctly specified, IPW is relatively inefficient, and it is susceptible to large finite-sample biases when confounders strongly predict the exposures of interest (Wang et al. 2006; Cole and Hernán 2008).¹ Finally, when the exposures of interest are continuous, IPW tends to perform poorly because estimates of conditional densities are often unreliable (e.g., Vansteelandt 2009; Naimi et al. 2014).

Several remedies have been proposed to improve the efficiency and robustness of IPW. For example, Cole and Hernán (2008) suggest truncating or censoring extreme weights to obtain more precise estimates. With this approach, however, the improved precision comes at the cost of greater bias. Recently, Imai and Ratkovic (2014, 2015) propose constructing weights for an MSM with covariate balancing propensity scores (CBPS). By integrating a large set of balancing conditions when estimating propensity scores, this method is less sensitive to model misspecification. But estimating CBPS can be computationally demanding, and because of the practical difficulties associated with modeling conditional densities, this method remains challenging to use with continuous exposures, even in the cross-sectional setting (Fong et al. 2018).

In this paper, we propose an alternative method of constructing weights for MSMs, which we call “residual balancing.” Briefly, the method is implemented in two stages. First, a model for the conditional mean of each post-treatment confounder, given past treatments and confounders, is estimated and then used to construct residual terms. Second, a set of weights is constructed using Hainmueller’s (2012) entropy balancing method such that, in the weighted sample, (a) the residualized confounders are orthogonal to future exposures, past treatments, and past confounders, and (b) their discrepancy with a set of base weights (e.g., survey sampling weights) is minimized. Thus, our proposed method is an extension of Hainmueller’s (2012) entropy balancing procedure to the longitudinal setting. It exactly balances sample moments for each of the post-treatment confounders across future expo-

¹For expositional simplicity, we occasionally use the term “exposures” to generally refer to treatments or mediators.

tures, conditional on the observed past, without explicit models for the conditional distributions of exposure to treatment and/or a mediator.²

Residual balancing has a number of advantages over both conventional methods of covariate adjustment and over IPW and its variants. First, by appropriately residualizing the post-treatment confounders, the proposed method avoids bias due to over-control and collider stratification, unlike conventional methods that condition, stratify, or otherwise balance on these variables naively. Second, residual balancing is relatively robust to the model misspecification bias that commonly afflicts IPW and its variants. Third, residual balancing is also more efficient than IPW because it tends to avoid highly variable and extreme weights by minimizing their relative entropy with respect to a set of base weights. Fourth, in contrast to CBPS, residual balancing is computationally attractive in that the weighting solution is quickly obtained even with a large number of confounders, time periods, and observations. Finally, because it does not require models for the conditional distributions of the exposures, residual balancing is easy to use when treatments and/or mediators are continuous. This advantage may be especially important in political science applications, where continuous exposures commonly arise in analyses of time-series cross-sectional data (e.g., Blackwell 2013). An open source R package, `rbw`, is available for implementing the proposed method, as is a Stata package with similar functionality.

In the sections that follow, we first briefly review MSMs and the method of IPW. Next, we introduce the method of residual balancing and conduct a set of simulation studies to evaluate its performance relative to IPW and its variants. We then illustrate the method empirically by estimating the cumulative effect of negative advertising on election outcomes as well as the controlled direct effect (CDE) of shared democracy on public support for war. We conclude by discussing the method's limitations along with possible remedies.

²Our method of residual balancing should not be confused with the method of "approximate residual balancing" proposed in Athey, Imbens and Wager (2018). Despite their similar names, the two methods are very different in both their goals and mechanics. The goal of our method is to adjust for post-treatment confounding when estimating the effects of time-varying treatments or assessing causal mediation, whereas the goal of "approximate residual balancing" is to remove bias introduced by penalized regression adjustments when estimating the effects of point-in-time treatments from high-dimensional linear models. Consequently, with our method, the residuals come from regression models of the time-varying confounders, and a set of weights are constructed to balance the residualized confounders across future exposures, past treatments, and past confounders. With "approximate residual balancing," by contrast, a set of weights are constructed first to balance the (unresidualized) confounders between static treatment and control groups, and then they are used to re-weight the residuals from a penalized regression model for the outcome to remove bias introduced by penalization.

2 MSMs and IPW: A Review

In this section, we briefly review MSMs and the method of IPW (Robins 1999; Robins, Hernan and Brumback 2000). Consider first a study with $T \geq 2$ time points where interest is in the effect of a time-varying treatment, A_t ($1 \leq t \leq T$), on an end-of-study outcome, Y . At each time point, there is also a vector of observed time-varying confounders, L_t , that may be affected by prior treatments. Following convention, we use overbars to denote the treatment history, $\bar{A}_t = (A_1, \dots, A_t)$, and confounder history, $\bar{L}_t = (L_1, \dots, L_t)$, up to time t . Similarly, we denote an individual's complete treatment and confounder histories through the end of follow-up by $\bar{A} = \bar{A}_T$ and $\bar{L} = \bar{L}_T$, respectively. Finally, we use $Y(\bar{a})$ to denote the potential outcome under the particular treatment history \bar{a} .

An MSM is a model for the marginal mean of the potential outcomes, which can be expressed in general form as follows:

$$\mathbb{E}[Y(\bar{a})] = \mu(\bar{a}; \beta), \quad (1)$$

where $\mu(\cdot)$ is some function of treatment history, \bar{a} , and a parameter vector, β , that captures the marginal effects of interest. For example, with a large number of time points and a binary treatment, a common parameterization is

$$\mathbb{E}[Y(\bar{a})] = \beta_0 + \beta_1 \text{cum}(\bar{a}), \quad (2)$$

where $\text{cum}(\bar{a}) = \sum_{t=1}^T a_t$ denotes the total number of time periods on treatment and β_1 captures the marginal effect of one additional wave on treatment. Of course, many other parameterizations are possible.

An MSM can be identified from observed data under three key assumptions:

1. consistency, which requires that, for any unit, if $\bar{A} = \bar{a}$, then $Y = Y(\bar{a})$;
2. sequential ignorability, which requires that treatment at each time point must not be confounded by unobserved factors conditional on past treatments and observed confounders, or formally, that $Y(\bar{a}) \perp\!\!\!\perp A_t | \bar{A}_{t-1}, \bar{L}_t$ for any treatment sequence \bar{a} ; and
3. positivity, which requires that treatment assignment must not be deterministic, or formally, that $f(A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_t = \bar{l}_t) > 0$ for any treatment condition a_t if $f(\bar{A}_{t-1} =$

$\bar{a}_{t-1}, \bar{L}_t = \bar{l}_t) > 0$, where $f(\cdot)$ denotes a probability mass or density function.

When these assumptions are satisfied, an MSM can be consistently estimated using the method of IPW.

IPW estimation involves fitting a model for the conditional mean of the observed outcome given an individual's treatment history using weights that balance, in expectation, past confounders across treatment at each time point. The inverse probability weight for individual i is defined as

$$w_i = \prod_{t=1}^T \frac{1}{f(A_t = a_{i,t} | \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t})}, \quad (3)$$

where the $\bar{A}_{t-1} = \bar{a}_{i,t-1}$ term can be ignored when $t = 1$. Since the denominator of equation (3) can be very small, some units may end up with extremely large weights, leading to highly variable estimates. To mitigate this problem, Robins, Hernan and Brumback (2000) suggest using a so-called “stabilized” weight, which is defined as

$$sw_i = \prod_{t=1}^T \frac{f(A_t = a_{i,t} | \bar{A}_{t-1} = \bar{a}_{i,t-1})}{f(A_t = a_{i,t} | \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t})}. \quad (4)$$

Sometimes, the probabilities in both the numerator and denominator are also made conditional on a set of baseline or time-invariant confounders X :

$$sw_i = \prod_{t=1}^T \frac{f(A_t = a_{i,t} | \bar{A}_{t-1} = \bar{a}_{i,t-1}, X = x)}{f(A_t = a_{i,t} | \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t}, X = x)}. \quad (5)$$

In such cases, these variables need to be included in the MSM to properly adjust for confounding, which is unproblematic because they cannot be affected by treatment and thus conditioning on them will not engender bias due to over-control or collider stratification.

In practice, both the numerator and the denominator of the stabilized weight need to be estimated. When treatment is binary, the denominator is typically estimated using a generalized linear model (GLM), with the logit or probit link function, for treatment at each time point, while the numerator is estimated using a constrained version of this model that omits the time-varying confounders. When treatment is continuous, models are needed to estimate the conditional densities in both the numerator and the denominator of the weight. After weights have been computed, the marginal effects of

interest are estimated by fitting a model for the conditional mean of Y given \bar{A}_t (and also possibly X) with weights equal to sw_i . When both this model and the models for treatment assignment are correctly specified, this procedure yields consistent estimates for all marginal means of the potential outcomes, $\mathbb{E}[Y(\bar{a})]$, and thus for any marginal effect of interest, provided that the identification assumptions outlined previously are satisfied.

As shown in prior studies (e.g., Kang and Schafer 2007), IPW estimates of marginal effects can be highly sensitive to misspecification of the models used to construct the weights. To address this limitation, Imai and Ratkovic (2014, 2015) developed the method of CBPS to estimate the denominator in equation (4) for binary treatments. With a logit model for treatment at each time point, this method augments the score conditions of the likelihood function with a set of covariate balance conditions. Because the total number of score and balance conditions exceeds the number of model parameters to be estimated, the generalized method of moments (GMM) is used to minimize imbalance in the weighted sample. This method of incorporating balance conditions into model-based estimation of the weights tends to reduce the bias that results when the treatment models are misspecified.

MSMs and IPW estimation can also be used to examine causal mediation (VanderWeele 2015). Consider now a study with a point-in-time treatment, A , a putative mediator measured at some point following treatment, M , and an end-of-study outcome, Y . Suppose that both treatment and the mediator are confounded by a vector of observed baseline covariates, denoted by X , and that the mediator is additionally confounded by a vector of observed post-treatment covariates, denoted by Z , which may be affected by the treatment received earlier. In this setting, the potential outcomes of interest are denoted by $Y(a, m)$.

As before, an MSM models the marginal mean of the potential outcomes. If, for example, treatment and the mediator are both binary, a saturated MSM can be expressed as follows:

$$\mathbb{E}[Y(a, m)] = \alpha_0 + \alpha_1 a + \alpha_2 m + \alpha_3 am. \quad (6)$$

From this model, the controlled direct effect of treatment is given by $\text{CDE}(m) = \mathbb{E}[Y(1, m) - Y(0, m)] = \alpha_1 + \alpha_3 m$, which measures the strength of the causal relationship between treatment and the outcome when the mediator is fixed at a given value, m , for all individuals (Pearl 2001; Robins 2003). This estimand is useful for assessing causal mediation because it helps to adjudicate between

alternative explanations for a treatment effect. For example, the difference between a total effect and the $CDE(m)$ may be interpreted as the degree to which the mediator contributes to a causal mechanism that transmits the effect of treatment on the outcome (Acharya, Blackwell and Sen 2016; Zhou and Wodtke 2019).

MSMs for the joint effects of a treatment and mediator, like equation (6), can be identified under essentially the same assumptions as outlined previously. In this context, the consistency assumption requires that $Y = Y(a, m)$ if $A = a$ and $M = m$; sequential ignorability requires that both treatment and the mediator must be unconfounded conditional on the observed past, or formally, that $Y(a, m) \perp\!\!\!\perp A|X$ and $Y(a, m) \perp\!\!\!\perp M|X, A, Z$; and positivity requires that both treatment and the mediator are not deterministic functions of past variables. Similarly, the stabilized inverse probability weights are here defined as

$$sw_i^* = \frac{f(A = a_i)}{f(A = a_i|X = x_i)} \times \frac{f(M = m_i|A = a_i)}{f(M = m_i|X = x_i, A = a_i, Z = z_i)}, \quad (7)$$

and they must be estimated using appropriate models for the conditional probabilities and/or densities that compose this expression. After weights have been computed, the marginal effects of interest – here, the $CDE(m)$ – are estimated by fitting a model for the conditional mean of Y given A and M with weights equal to sw_i^* . Alternatively, it is also possible to define the weights as $sw_i^\dagger = \frac{f(M=m_i|X=x_i, A=a_i)}{f(M=m_i|X=x_i, A=a_i, Z=z_i)}$, in which case X must be included in the MSM to properly adjust for confounding. Adjusting for X in the MSM is unproblematic because these variables are not post-treatment confounders, unlike Z .

3 Residual Balancing

In this section, we motivate and explain the method of residual balancing. We first focus on analyses of time-varying treatment effects, and then we outline how the method is easily adapted for studies of causal mediation. Finally, we discuss the advantages and limitations of residual balancing compared with IPW as well as the similarities and differences between residual balancing and the CBPS method.

3.1 Rationale

To explain the method of residual balancing, it is useful to begin with Robins' (1986) g-computation formula. The g-computation formula factorizes the marginal mean of the potential outcome, $Y(\bar{a})$, as follows:

$$\mathbb{E}[Y(\bar{a})] = \int \cdots \int \mathbb{E}[Y|\bar{A} = \bar{a}, \bar{L} = \bar{l}] \prod_{t=1}^T f(l_t|\bar{l}_{t-1}, \bar{a}_{t-1}) d\mu(l_t). \quad (8)$$

In contrast, the conditional mean of the observed outcome Y given $\bar{A} = \bar{a}$ can be factorized into

$$\mathbb{E}[Y|\bar{A} = \bar{a}] = \int \cdots \int \mathbb{E}[Y|\bar{A} = \bar{a}, \bar{L} = \bar{l}] \prod_{t=1}^T f(l_t|\bar{l}_{t-1}, \bar{a}) d\mu(l_t). \quad (9)$$

A comparison of equation (8) with equation (9) indicates that weighting the observed population by

$$W_l = \prod_{t=1}^T \frac{f(L_t|\bar{L}_{t-1}, \bar{A}_{t-1})}{f(L_t|\bar{L}_{t-1}, \bar{A})} \quad (10)$$

would yield a pseudo-population in which $f^*(l_t|\bar{l}_{t-1}, \bar{a}) = f^*(l_t|\bar{l}_{t-1}, \bar{a}_{t-1}) = f(l_t|\bar{l}_{t-1}, \bar{a}_{t-1})$ and thus $\mathbb{E}^*[Y|\bar{A} = \bar{a}] = \mathbb{E}^*[Y(\bar{a})] = \mathbb{E}[Y(\bar{a})]$, where the asterisk denotes quantities in the weighted pseudo-population.³ Because L_t is often high-dimensional, estimation of the conditional densities in equation (10) is practically difficult.

Nevertheless, the condition that $f^*(l_t|\bar{l}_{t-1}, \bar{a}) = f^*(l_t|\bar{l}_{t-1}, \bar{a}_{t-1}) = f(l_t|\bar{l}_{t-1}, \bar{a}_{t-1})$ implies that, in the pseudo-population, the following moment condition would hold for any scalar function $g(\cdot)$ of L_t :

$$\mathbb{E}^*[g(L_t)|\bar{L}_{t-1}, \bar{A}] = \mathbb{E}^*[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}] = \mathbb{E}[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}]. \quad (11)$$

³In fact, the "stabilized" weight in equation (4) is just a different way of writing equation (10):

$$\begin{aligned} W_l &= \prod_{t=1}^T \frac{f(L_t|\bar{L}_{t-1}, \bar{A}_{t-1})}{f(L_t|\bar{L}_{t-1}, \bar{A})} = \frac{\prod_{t=1}^T f(L_t|\bar{L}_{t-1}, \bar{A}_{t-1})}{f(\bar{L}|\bar{A})} = \frac{f(\bar{A}) \prod_{t=1}^T f(L_t|\bar{L}_{t-1}, \bar{A}_{t-1})}{f(\bar{L}, \bar{A})} \\ &= \frac{f(\bar{A}) \prod_{t=1}^T f(L_t|\bar{L}_{t-1}, \bar{A}_{t-1})}{\prod_{t=1}^T f(L_t|\bar{L}_{t-1}, \bar{A}_{t-1}) f(A_t|\bar{L}_t, \bar{A}_{t-1})} = \frac{\prod_{t=1}^T f(A_t|\bar{A}_{t-1})}{\prod_{t=1}^T f(A_t|\bar{L}_t, \bar{A}_{t-1})} \end{aligned}$$

This moment condition can be equivalently expressed as

$$\mathbb{E}^*[\delta(g(L_t))|\bar{L}_{t-1}, \bar{A}] = 0, \quad (12)$$

where $\delta(g(L_t)) = g(L_t) - \mathbb{E}[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}]$ is a residual transformation of $g(L_t)$ with respect to its conditional mean given the observed past. The moment condition in equation (12) in turn implies that for any scalar function $h(\cdot)$ of \bar{L}_{t-1} and \bar{A} , $\delta(g(L_t))$ and $h(\bar{L}_{t-1}, \bar{A})$ are uncorrelated, that is,

$$\mathbb{E}^*[\delta(g(L_t))h(\bar{L}_{t-1}, \bar{A})] = \mathbb{E}^*[\delta(g(L_t))]\mathbb{E}^*[h(\bar{L}_{t-1}, \bar{A})] = 0, \quad (13)$$

where the second equality follows from the fact that $\mathbb{E}^*[\delta(g(L_t))] = \mathbb{E}^*\mathbb{E}^*[\delta(g(L_t))|\bar{L}_{t-1}, \bar{A}] = 0$.

The method of residual balancing emulates the moment conditions (13) that would hold in the pseudo-population were it possible to weight by W_l . In other words, it emulates the moment conditions (13) that would be expected in a *sequentially* randomized experiment. Specifically, this is accomplished by (a) specifying a set of $g(\cdot)$ functions, $G(L_t) = \{g_1(L_t), \dots, g_{J_t}(L_t)\}$, and a set of $h(\cdot)$ functions, $H(\bar{L}_{t-1}, \bar{A}) = \{h_1(\bar{L}_{t-1}, \bar{A}), \dots, h_{K_t}(\bar{L}_{t-1}, \bar{A})\}$; (b) computing a set of residual terms, $\delta(g(L_t)) = g(L_t) - \mathbb{E}[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}]$, from the observed data; and then (c) finding a set of weights such that, for any j, k , and t , the cross-moment of $\delta(g_j(l_{it}))$ and $h_k(\bar{l}_{i,t-1}, \bar{a}_i)$ is zero in the weighted data. Hence, it involves finding a set of non-negative weights, denoted by rbw_i , subject to the following balancing conditions:

$$\sum_{i=1}^n rbw_i \delta(g_j(l_{it})) h_k(\bar{l}_{i,t-1}, \bar{a}_i) = 0, \quad 1 \leq j \leq J_t; 1 \leq k \leq K_t, \quad (14)$$

or, expressed more succinctly,

$$\sum_{i=1}^n rbw_i c_{ir} = 0, \quad 1 \leq r \leq n_c, \quad (15)$$

where c_{ir} is the r th element of $\mathbf{c}_i = \{\delta(g_j(l_{it}))h_k(\bar{l}_{i,t-1}, \bar{a}_i); 1 \leq j \leq J_t, 1 \leq k \leq K_t, 1 \leq t \leq T\}$ and $n_c = \sum_{t=1}^T J_t K_t$ is the total number of balancing conditions. The conditions in equation (14) stipulate that the residualized confounders at each time point are balanced across future treatments, past treatments, and past confounders, or some function thereof. In this way, the proposed method

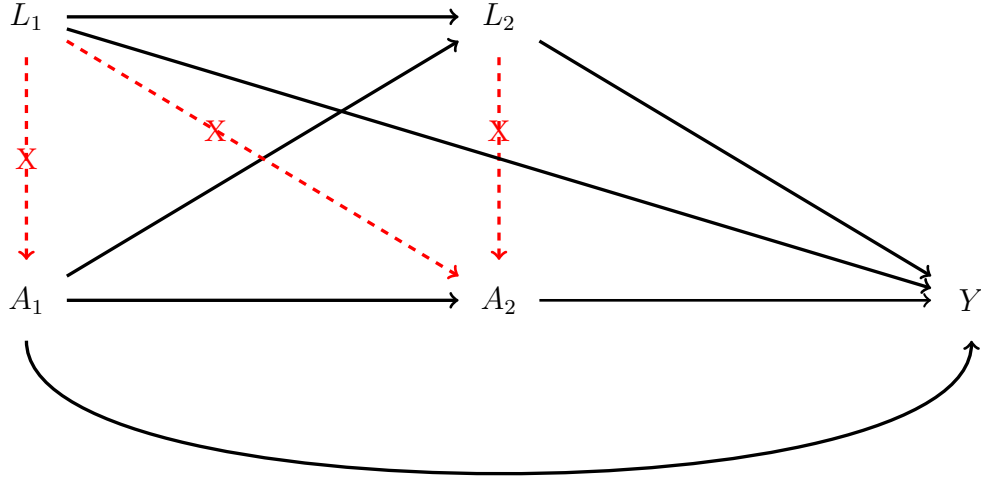


Figure 1: The Logic of Residual Balancing

Note: A_t denotes treatment at time t , L_t denotes time-varying confounders at time t , Y denotes the end-of-study outcome.

adjusts for post-treatment confounding without engendering bias due to over-control or collider-stratification, as the residualized confounders are balanced across future treatments while (appropriately) remaining orthogonal to the observed past.

As long as the convex hull of $\{c_i; 1 \leq i \leq n\}$ contains $\mathbf{0}$, finding the weighting solution is an under-identified (or just-identified) problem. Following Hainmueller (2012), we minimize the relative entropy between rbw_i and a set of base weights q_i (e.g., a vector of ones or survey sampling weights),⁴

$$\min_{rbw_i} \sum_i rbw_i \log(rb w_i / q_i), \quad (16)$$

subject to the n_c balancing conditions. This is a constrained optimization problem that can be solved using Lagrange multipliers. Technical details can be found in Supplementary Material A (see also Hainmueller 2012).

In Figure 1, we illustrate the logic of residual balancing with a directed acyclic graph (DAG), which describes the causal relationships between a time-varying treatment A_t , a vector of time-varying con-

⁴Alternative loss functions, such as the empirical likelihood (Fong et al. 2018) or the variance (Zubizarreta 2015), could also be used to construct the weights. We use the relative entropy metric because it can easily accommodate a set of base weights. Moreover, in contrast to the empirical likelihood, the relative entropy metric is convex and thus computationally convenient.

founders L_t , and an end-of-study outcome Y with two time periods $t = 1, 2$. Weighting is intended to create a pseudo-population in which the confounding arrows $L_1 \rightarrow A_1$, $L_1 \rightarrow A_2$, and $L_2 \rightarrow A_2$ are “broken,” that is, a pseudo-population in which (a) L_1 no longer predicts A_1 or A_2 and (b) L_2 no longer predicts A_2 given L_1 and A_1 . The first condition requires L_1 to be marginally independent of both A_1 and A_2 . Thus, any function of L_1 should be uncorrelated with any function of A_1 and A_2 in the weighted population. The second condition, by contrast, requires L_2 to be *conditionally independent* of A_2 given L_1 and A_1 . To this end, we could divide the original population into a number of strata defined by L_1 and A_1 and then balance L_2 across levels of A_2 within each stratum. This approach, however, becomes impractical when L_1 and A_1 are continuous and/or multidimensional. To circumvent this problem, our method invokes a model for the conditional mean of L_2 (or some function of L_2) given L_1 and A_1 , and it then balances the residuals from this model across levels of A_2 and levels of (L_1, A_1) . This procedure breaks the confounding arrow $L_2 \rightarrow A_2$ but preserves the causal arrow $A_1 \rightarrow L_2$, thereby adjusting properly for the observed post-treatment confounders while avoiding bias due to over-control and collider stratification. Taken together, the balancing conditions for both L_1 and L_2 yield a weighted population in which all the confounding arrows ($L_1 \rightarrow A_1$, $L_1 \rightarrow A_2$, and $L_2 \rightarrow A_2$) are “broken” and all the other arrows are left intact. A marginal structural model can then be fit to this population in order to estimate the average causal effects of A_1 and A_2 on Y .

3.2 Implementation

In practice, residual balancing requires specifying a set of $g(\cdot)$ functions that constitute $G(L_t)$. A natural choice is to set $g_j(L_t) = L_{jt}$, where L_{jt} is the j th element of the covariate vector L_t . If there is concern about confounding by higher-order or interaction terms, they can also be included in $G(L_t)$. Then, the residual terms, $\delta(g(L_t))$, need to be estimated from the data. Because $\delta(g(L_t)) = g(L_t) - \mathbb{E}[g(L_t) | \bar{L}_{t-1}, \bar{A}_{t-1}]$, they can be estimated by fitting GLMs for $g(L_t)$ and then extracting the response residuals, $\hat{\delta}(g(L_t)) = g(L_t) - m(\hat{\beta}_t^T r(\bar{L}_{t-1}, \bar{A}_{t-1}))$, where $r(\bar{L}_{t-1}, \bar{A}_{t-1}) = [r_1(\bar{L}_{t-1}, \bar{A}_{t-1}), \dots, r_{L_t}(\bar{L}_{t-1}, \bar{A}_{t-1})]$ is a vector of regressors and $m(\cdot)$ denotes the inverse link function of the GLM.

In addition, residual balancing requires specifying a set of $h(\cdot)$ functions that constitute $H(\bar{L}_{t-1}, \bar{A})$. Because weighting is intended to neutralize the relationship between L_t and future

treatments, we suggest including all future treatments, A_t, A_{t+1}, \dots, A_T , in $H(\bar{L}_{t-1}, \bar{A})$. However, if it is reasonable to assume that the effects of L_t on future treatments stop at $A_{t'}$, where $t \leq t' < T$, treatments beyond time t' may be excluded from $H(\bar{L}_{t-1}, \bar{A})$. Equation (13) additionally indicates that $\delta(g(L_t))$ should be uncorrelated with past treatments, \bar{A}_{t-1} , and past confounders, \bar{L}_{t-1} , in the weighted pseudo-population. Because $\mathbb{E}[\delta(g(L_t)) | \bar{L}_{t-1}, \bar{A}_{t-1}] = 0$ by construction, zero correlation is guaranteed in the original unweighted population, and when the GLMs for $g(L_t)$ are Gaussian, binomial, or Poisson regressions with canonical links, the score equations ensure that the response residuals, $\hat{\delta}(g(L_t))$, are orthogonal to the regressors $r(\bar{L}_{t-1}, \bar{A}_{t-1})$ in the original sample. But to ensure that the response residuals, $\hat{\delta}(g(L_t))$, are also orthogonal to the regressors in the weighted sample, we suggest including all elements of $r(\bar{L}_{t-1}, \bar{A}_{t-1})$ in $H(\bar{L}_{t-1}, \bar{A})$.

In general, then, $H(\bar{L}_{t-1}, \bar{A})$ should include all future treatments as well as all regressors in the GLMs for $g(L_t)$, including an intercept. A reassuring property of this specification for $H(\bar{L}_{t-1}, \bar{A})$ is that if the GLMs for $g(L_t)$ are Gaussian, binomial, or Poisson regressions with canonical links and they are fit to the weighted sample with all future treatments, A_t, A_{t+1}, \dots, A_T , as additional regressors, the coefficients on future treatments will all be exactly zero and the coefficients on $r(\bar{L}_{t-1}, \bar{A}_{t-1})$ will be the same as those in the original sample. Therefore, when the GLMs for $g(L_t)$ are correctly specified, the first moments of $g(L_t)$ are guaranteed to be balanced across future treatments, conditional on past treatments and confounders, as would be expected in a scenario where treatment is unconfounded by \bar{L}_t .

In sum, a typical implementation of residual balancing for estimating the marginal effects of a time-varying treatment proceeds in two steps:

1. At each time point t and for each confounder j , fit a linear, logistic, or Poisson regression of l_{ijt} , as appropriate given its level of measurement, on $\bar{l}_{i,t-1}$ and $\bar{a}_{i,t-1}$, and then compute the response residuals, $\hat{\delta}(l_{ijt})$.
2. Find a set of weights, rbw_i , such that:
 - (a) in the weighted sample, the residuals, $\hat{\delta}(l_{ijt})$, are orthogonal to all future treatments and the regressors of l_{ijt} ; and
 - (b) the relative entropy between rbw_i and the base weights, q_i , is minimized.

The weighting solution can then be used to fit any MSM of interest.

3.3 Application to Causal Mediation

Residual balancing can also be used to estimate an MSM for the joint effects of a point-in-time treatment, A , and mediator, M , in the presence of both baseline confounders, X , and a set of post-treatment confounders, Z , for the mediator-outcome relationship. In this setting, residual balancing is implemented using essentially the same procedure as outlined previously but with several minor adaptations. First, for each baseline confounder X_j , compute the response residuals, $\hat{\delta}(x_{ij})$, by centering it around its sample mean. Then, for each post-treatment confounder Z_j , fit a linear, logistic, or Poisson regression of z_{ij} , depending on its level of measurement, on x_i and a_i , and then compute the response residuals, $\hat{\delta}(z_{ij})$. Finally, find a set of weights, rbw_i , such that, in the weighted sample, the baseline residuals $\hat{\delta}(x_{ij})$ are orthogonal to both treatment a and the mediator m ; the post-treatment residuals $\hat{\delta}(z_{ij})$ are orthogonal to treatment, the mediator, and the pre-treatment confounders x_{ij} ; and the relative entropy between rbw_i and the base weights q_i is minimized. The weighting solution can then be used to fit any MSM for the joint effects of the treatment and mediator on the outcome, from which the controlled direct effects of interest are constructed. Alternatively, it is also possible to skip the first step and construct weights that only balance the residualized post-treatment confounders, in which case the baseline confounders X must be included as regressors in the MSM.

3.4 Comparison with Existing Methods

Compared with IPW, residual balancing has both advantages and limitations. On the one hand, because it does not require explicit models for the conditional distribution of exposure to treatment and/or a mediator, residual balancing is robust to the bias that results when these models are misspecified, and it is easy to use with both binary and continuous exposures. Also, by minimizing the relative entropy between the balancing weights and the base weights, the method tends to avoid highly variable and extreme weights, thus yielding more stable estimates of causal effects.

On the other hand, residual balancing requires models for the conditional means of the post-treatment confounders (or transformations thereof). When these models are misspecified, the mo-

ment condition in equation (11) is only partially achieved. In this case, equation (12) implies

$$\mathbb{E}^*[g(L_t)|\bar{L}_{t-1}, \bar{A}] = \mathbb{E}^*[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}] \neq \mathbb{E}[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}],$$

where future treatments (i.e., A_t, A_{t+1}, \dots, A_T) may still be unconfounded in the weighted pseudo-population but the pseudo-population no longer mimics the original unweighted population. As a result, estimates of marginal effects based on residual balancing weights may be biased. In addition, even when models for $\mathbb{E}[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}]$ are correctly specified, residual balancing estimates of marginal effects may still be biased if the balancing conditions are insufficient. For example, if both the treatment and outcome are affected by the product of two confounders, say $L_{1t}L_{2t}$, but L_{1t} and L_{2t} are only included separately in the $G(L_t)$ functions, uncontrolled confounding may still be present in the weighted sample, leading to bias.

Residual balancing is similar to the CBPS method (Imai and Ratkovic 2015) in that it seeks a set of weights that balance time-varying confounders across future treatments by explicitly specifying a set of balancing conditions. Residual balancing differs from CBPS, however, in two important respects. First, unlike CBPS, residual balancing can easily accommodate continuous treatments and/or mediators. As mentioned previously, this is because residual balancing does not require parametric models for exposure to treatment and/or a mediator, and thus it can balance confounders across both binary and continuous treatments using a common set of balancing conditions (equation 14). CBPS, by contrast, is based on a parametric logistic model for the propensity score, and it is therefore limited to settings with binary treatments and/or mediators.

Second, residual balancing allows for the specification of more flexible and parsimonious balancing conditions than those specified with the CBPS method. In fact, the balancing conditions specified by CBPS can also be generated within the residual balancing framework. To see the connection, note that CBPS attempts to balance the time-varying confounders across *all* possible sequences of future treatments within *each* possible history of past treatments. Thus, for each confounder j , there are $2^{t-1} \times (2^{T-t+1} - 1) = 2^T - 2^{t-1}$ balancing conditions at time t . Summing over t and j , the total number of balancing conditions associated with CBPS is $n_c^{\text{CBPS}} = J[(T-1)2^T + 1]$. Because $n_c^{\text{CBPS}} \sim O(J \cdot T \cdot 2^T)$, the number of balancing conditions can easily exceed the sample size, in which case they are at best approximated (even without the method's parametric constraints).

With residual balancing, the number of balancing conditions $n_c = \sum_{t=1}^T J_t K_t$ depends on the specification of $G(L_t)$ and $H(\bar{L}_{t-1}, \bar{A})$. As mentioned previously, a natural specification of $G(L_t)$ is $\{L_{1t}, L_{2t}, \dots, L_{jt}\}$. If $\mathbb{E}[g_j(L_t) | \bar{L}_{t-1}, \bar{A}_{t-1}]$ is then modeled with a saturated GLM of L_{jt} on \bar{A}_{t-1} only, and $H(\bar{L}_{t-1}, \bar{A})$ is defined as a set of dummy variables for each possible sequence of future treatments interacted with each possible history of past treatments, the balancing conditions in equation (14) would be equivalent to those for the CBPS method.

With residual balancing, however, $G(L_t)$, $\mathbb{E}[g_j(L_t) | \bar{L}_{t-1}, \bar{A}_{t-1}]$, and $H(\bar{L}_{t-1}, \bar{A})$ can be specified more flexibly. For example, when a parsimonious GLM is used to fit $\mathbb{E}[g_j(L_t) | \bar{L}_{t-1}, \bar{A}_{t-1}]$, and only the L_t regressors of $g_j(L_t)$ and $T - t + 1$ future treatments are included in $H(\bar{L}_{t-1}, \bar{A})$, the number of balancing conditions will be $n_c = J \sum_{t=1}^T (T - t + 1 + L_t)$, which is substantially smaller than n_c^{CBPS} . In large and even moderately sized samples, these balancing conditions can often be satisfied exactly.

4 Simulation Experiments

In this section, we conduct a set of simulation studies to assess the performance of residual balancing for estimating marginal effects with (a) a binary time-varying treatment under correct model specification, (b) a binary time-varying treatment under incorrect model specification, (c) a continuous time-varying treatment under correct model specification, and (d) a continuous time-varying treatment under incorrect model specification. In each of these four settings, we compare residual balancing with four variants of IPW: conventional IPW with weights estimated from GLMs (IPW-GLM), IPW with weights estimated from GLMs and then censored (IPW-GLM-Censored), IPW with weights estimated from CBPS (IPW-CBPS), and as a benchmark, IPW with weights based on the true exposure probabilities (IPW-Truth). Because the CBPS method has not been extended for continuous treatments in the time-varying setting, we assess the performance of IPW-CBPS only for binary treatments.

The data generating process (DGP) in our simulations is similar to that of Imai and Ratkovic (2015). It involves four time-varying covariates measured at $T = 3$ time periods with a sample of $n = 1,000$. At each time t , the covariates L_t are determined by treatment at time $t - 1$ and a multiplicative error: $L_t = (U_t \epsilon_{1t}, U_t \epsilon_{2t}, |U_t \epsilon_{3t}|, |U_t \epsilon_{4t}|)$, where $U_1 = 1$, $U_t = (5/3) + (2/3)A_{t-1}$

for $t > 1$ and $\epsilon_{jt} \sim N(0, 1)$ for $1 \leq j \leq 4$. Treatment at each time t depends on prior treatment at time $t - 1$ and the covariates L_t . Specifically, when treatment is binary, it is generated as a Bernoulli draw with probability $p = \text{logit}^{-1}[-A_{t-1} + \gamma^T L_t + (-0.5)^t]$, and when treatment is continuous, it is generated as $A_t \sim N(\mu_t = -A_{t-1} + \gamma^T L_t + (-0.5)^t, \sigma_t^2 = 2^2)$, where $A_0 = 0$ and $\gamma = \alpha(1, -0.5, 0.25, 0.1)^T$. Here, we use the α parameter to control the level of treatment-outcome confounding. We consider two values of α , 0.4 and 0.8, corresponding to scenarios where treatment-outcome confounding is weak and strong, respectively. Finally, the outcome is generated as $Y \sim N(\mu = 250 - 10 \sum_{t=1}^3 A_t + \sum_{t=1}^3 \delta^T L_t, \sigma^2 = 5^2)$, where $\delta = (27.4, 13.7, 13.7, 13.7)^T$. To assess the impact of model misspecification, we use the same DGP, but we recode the “observed” covariates as nonlinear transformations of the “true” covariates: specifically, $L_t^* = (L_{1t}^3, 6 \cdot L_{2t}, \log(L_{3t} + 1), 1/(L_{4t} + 1))^T$. We then use only the transformed covariates, L_t^* , to implement IPW, its variants, and residual balancing. For IPW and its variants, using the transformed covariates leads to misspecification of the treatment assignment model. For residual balancing, the conditional mean model for L_{jt}^* is still correct when treatment is binary but incorrect when treatment is continuous. However, in both cases, using the transformed covariates (instead of the original covariates) leads to misspecification of the balancing conditions.

For each scenario described previously, we generate 2,500 random samples. Then, for each sample, we construct weights using IPW-GLM, IPW-GLM-Censored, IPW-CBPS, and residual balancing. With IPW-GLM, we estimate the weights using logistic regression for binary treatments and normal linear models for continuous treatments, assuming homoskedastic errors. With IPW-GLM-Censored, we follow Cole and Hernán’s (2008) example and censor weights at the 1st and 99th percentiles. With IPW-CBPS, we estimate weights using the methods proposed by Imai and Ratkovic (2015) with the function `CBMSM()` in the R package `CBPS`. With residual balancing, $G(L_t) = L_t$, and the residual terms are estimated from linear models for L_t with prior treatment A_{t-1} as a regressor, and $H(\bar{L}_{t-1}, \bar{A})$ includes A_t as well as the regressors in the model for L_t (i.e., 1 and A_{t-1}). Finally, with each set of weights, we fit an MSM by regressing the outcome Y on the three treatment variables $\{A_1, A_2, A_3\}$ and denote their coefficient estimates as $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. We obtain the true values of these coefficients by simulating potential outcomes with the g-computation formula, regressing them on the treatment variables, and averaging their coefficients over a large number of simulations. The performance of each method is evaluated using the simulated sampling distributions of $\hat{\beta}_1$, $\hat{\beta}_2$, and

$\hat{\beta}_3$.

Figure 2 presents results from simulations with a binary treatment. Specifically, this figure displays a set of violin plots, which show the sampling distributions of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ centered at the true values of these coefficients. In these plots, black dots represent means of the sampling distributions, and the shaded distributions highlight the estimator with the smallest root mean squared error (RMSE) in each scenario. In this figure, the first two panels show the sampling distributions of the parameter estimates under correct model specification. Comparing the first and second panels, we see that IPW and its variants suffer from finite-sample bias and may have skewed sampling distributions, especially when the covariates are strongly predictive of treatment. By contrast, residual balancing is roughly unbiased, and its estimates appear approximately normally distributed, regardless of the level of confounding. Second, the results indicate that residual balancing is much more efficient than IPW-GLM, especially when the level of confounding is high. In addition, with a high level of confounding, both IPW-GLM-Censored and IPW-CBPS yield much less variable estimates than IPW-GLM, but this gain in precision comes at the expense of greater bias. Residual balancing, by contrast, improves efficiency without inducing bias.

The last two panels of Figure 2 show the sampling distributions of parameter estimates under misspecified models where L_t is measured incorrectly. In these simulations, the treatment assignment models for IPW and the balancing conditions for residual balancing are misspecified. As indicated by its extreme level of sampling variation, IPW-GLM is highly unstable when models for the conditional probability of treatment are misspecified. Consistent with Imai and Ratkovic (2015), IPW-CBPS appears more robust to model misspecification, as reflected in its substantially smaller sampling variation compared with IPW-GLM. However, this improvement in precision comes at the cost of greater bias. In addition, censoring the inverse probability weights also appears to substantially improve the method's performance in the presence of misspecification. In fact, IPW-GLM-Censored outperforms IPW-CBPS in these simulations. Nevertheless, despite the improvements achieved by censoring the weights or using CBPS, residual balancing consistently produces the most accurate and efficient estimates across nearly all scenarios, even though its balancing conditions are incorrectly specified.

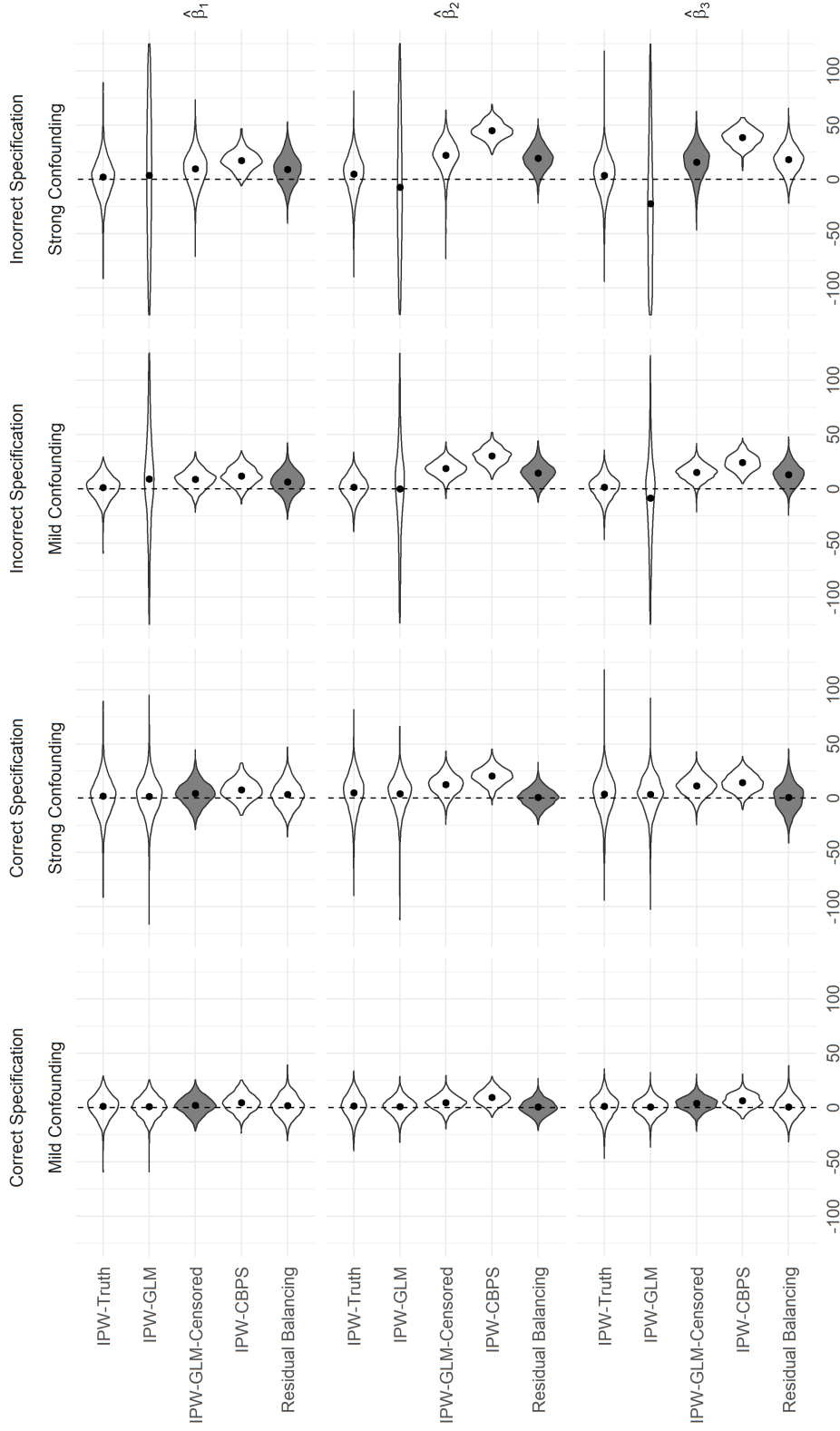


Figure 2: Simulation results for a binary treatment. “Mild confounding” and “strong confounding” corresponds to $\alpha = 0.4$ and $\alpha = 0.8$, respectively. Four different methods are compared: IPW based on the standard logistic regression (IPW-GLM), IPW based on the standard logistic regression with weights censored at the 1st and 99th percentiles (IPW-GLM-Censored), IPW based on the CBPS (IPW-CBPS), and residual balancing. As a benchmark, results from IPW based on true treatment probabilities (IPW-Truth) are also reported. The violin plots show the sampling distributions (from 2500 random samples) of different estimators centered at the true values of corresponding parameters, and the shaded violin plots highlight the estimator with the smallest root mean squared error (RMSE) in each scenario.

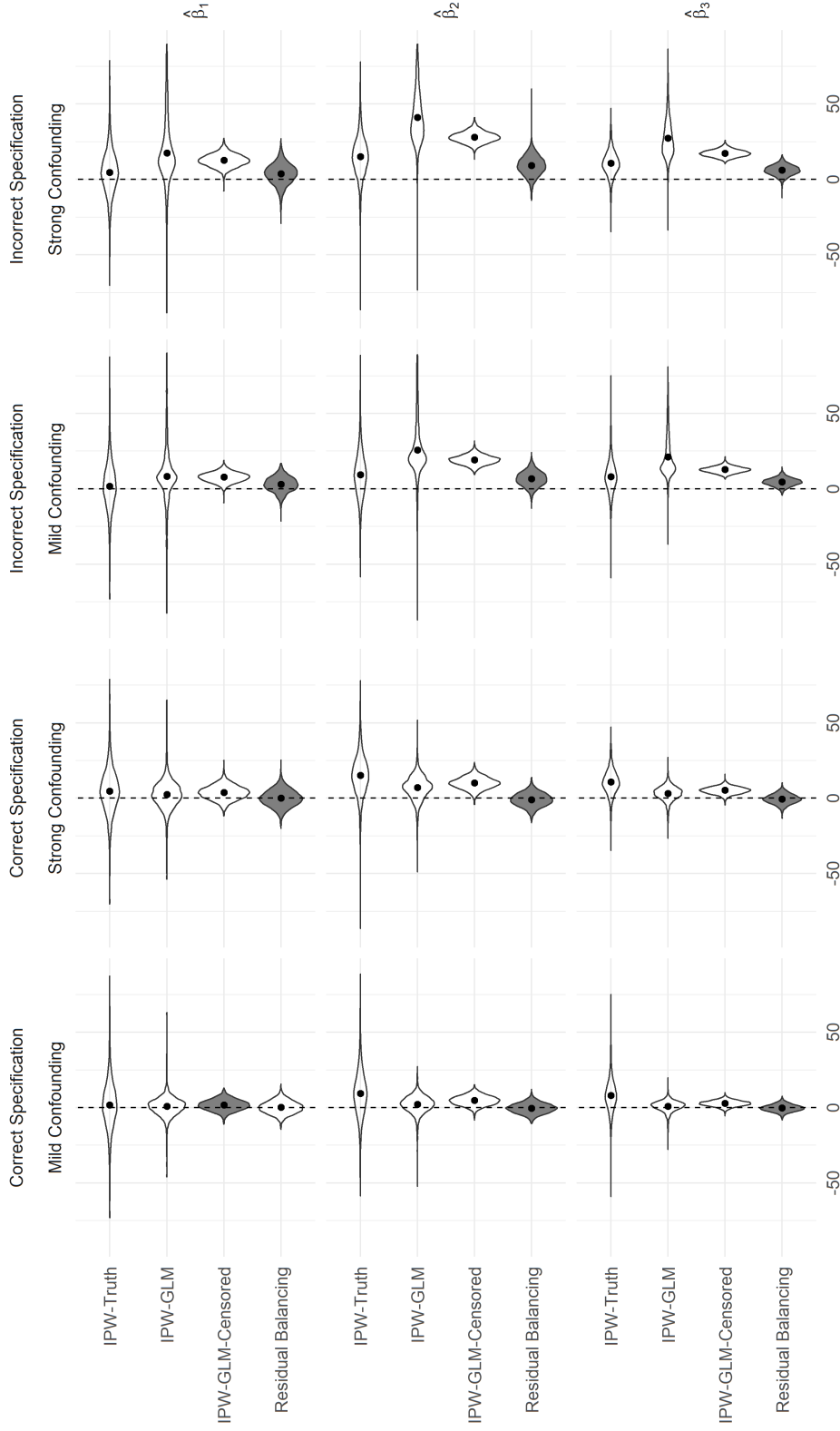


Figure 3: Simulation results for a continuous treatment. “Mild confounding” and “strong confounding” corresponds to $\alpha = 0.4$ and $\alpha = 0.8$, respectively. Three different methods are compared: IPW based on the standard logistic regression (IPW-GLM), IPW based on the standard logistic regression with weights censored at the 1st and 99th percentiles (IPW-GLM-Censored), and residual balancing. As a benchmark, results from IPW based on true treatment probabilities (IPW-Truth) are also reported. The violin plots show the sampling distributions (from 2500 random samples) of different estimators centered at the true values of corresponding parameters, and the shaded violin plots highlight the estimator with the smallest root mean squared error (RMSE) in each scenario.

Figure 3 presents another set of violin plots based on simulations with a continuous treatment . As shown in the first two panels, when both the treatment assignment models and the confounder models are correctly specified, the bias for IPW and its variants increases substantially with the level of confounding. Residual balancing, by contrast, is approximately unbiased across both levels of confounding. Moreover, residual balancing consistently outperforms IPW and its variants in terms of efficiency. For example, residual balancing is the most accurate and precise estimator for β_2 and β_3 under both high and low levels of confounding, and for β_1 , the performance of residual balancing is comparable to that of IPW-GLM-Censored.

The last two panels of Figure 3 present sampling distributions under misspecified models where L_t is measured incorrectly. In these simulations, the treatment assignment models for IPW are misspecified, as are both the confounder models and the balancing conditions used with residual balancing. Consistent with the results discussed previously, this figure also indicates that IPW-GLM is extremely biased and inefficient under incorrect models for treatment, that censoring the weights reduces bias and improves efficiency, and that residual balancing yields by far the most accurate and efficient estimator among all methods. Residual balancing even outperforms IPW based on the true treatment densities, even though its confounder models and balancing conditions are both misspecified.

5 The Cumulative Effect of Negative Advertising on Vote Shares

In this section, we illustrate residual balancing by estimating the cumulative effect of negative campaign advertising on election outcomes (Lau, Sigelman and Rovner 2007; Blackwell 2013; Imai and Ratkovic 2015). Drawing on U.S. senate and gubernatorial elections from 2000 to 2006, Blackwell (2013) used MSMs with IPW to evaluate the cumulative effects of negative campaign advertising on election outcomes for 114 Democratic candidates. MSMs are appropriate for this problem because campaign advertising is a dynamic process plagued by post-treatment confounding. For example, candidates adjust their campaign strategies on the basis of current polling results, where trailing candidates are more likely to “go negative” than leading candidates. At the same time, polling results

change over time and are likely affected by a candidate’s previous use of negative advertising.

Treatment, A_t , in this analysis is the proportion of campaign advertisements that are “negative” (i.e., that mention the opposing candidate) in each campaign-week. Because IPW tends to perform poorly with continuous treatments, we also consider a binary version of treatment, B_t , for which the proportion of negative advertisements is dichotomized using a cutoff of 10%, as in Blackwell (2013). The time-varying confounders, L_t , included in this analysis are the Democratic share in the polls and the share of undecided voters in the previous campaign-week. This analysis also uses a set of baseline confounders, X , including total campaign length, election year, incumbency status, and whether the election is for the senate or governor’s office. The outcome, Y , is the Democratic share of the two-party vote.

Following Imai and Ratkovic (2015), we focus on the final five weeks preceding the election and estimate an MSM for the binary version of treatment with form

$$\mathbb{E}[Y(\bar{b})|X] = \theta_0 + \theta_1 \text{cum}(\bar{b}) + \theta_2 V \cdot \text{cum}(\bar{b}) + \theta_3^T X, \quad (17)$$

and an MSM for the continuous treatment with form

$$\mathbb{E}[Y(\bar{a})|X] = \beta_0 + \beta_1 \text{avg}(\bar{a}) + \beta_2 V \cdot \text{avg}(\bar{a}) + \theta_3^T X. \quad (18)$$

In these models, $\text{cum}(\bar{b})$ denotes the total number of campaign-weeks for which more than 10% of the candidate’s advertising was negative, $\text{avg}(\bar{a})$ denotes the average proportion of advertisements that were negative over the final five weeks of the campaign, V is an indicator of incumbency status used to construct interaction terms that allow the effect of negative advertising to differ between incumbents and nonincumbents.⁵ Thus, the effect of an additional week with more than 10% negative advertising for nonincumbents is θ_1 , and for incumbents, it is $\theta_1 + \theta_2$. Similarly, β_1 and $\beta_1 + \beta_2$ correspond to the effects of a 1 percentage point increase in negative advertising for nonincumbents and incumbents, respectively. To facilitate comparison of results across the different versions of treatment, we report estimates for the effects of a 10 percentage point increase in negative advertising—that is, $10\beta_1$ and $10(\beta_1 + \beta_2)$.

We estimate these models with both IPW methods and residual balancing. Specifically, we first

⁵In equations (17) and (18), the “main” effect of V is captured in the term $\theta_3^T X$.

Table 1: Estimated Marginal Effects of Negative Advertising on the Candidate’s Vote Share

Estimator	Dichotomized Treatment		Continuous Treatment	
	Nonincumbent	Incumbent	Nonincumbent	Incumbent
IPW-GLM	1.42 (0.43; 0.49)	-1.73 (0.47; 0.55)	0.80 (0.28; 0.32)	-1.15 (0.31; 0.34)
IPW-CBPS	0.78 (0.89; 0.87)	-2.03 (0.41; 0.53)		
Residual Balancing	0.98 (0.54; 0.68)	-1.67 (0.46; 0.68)	0.49 (0.32; 0.43)	-0.99 (0.36; 0.43)

Note: For the dichotomized treatment, results represent the estimated marginal effects of an additional week with more than 10% negative advertising. For the continuous treatment, results represent the estimated marginal effects of a 10 percentage point increase in the average proportion of negative advertisements across all campaign-weeks. The two numbers in each parenthesis are the robust (i.e., “sandwich”) and jackknife standard errors, respectively.

implement IPW-GLM by fitting, at each time point, a logistic regression of the dichotomized treatment on both time-varying confounders and baseline confounders, and then constructing the inverse probability weights using equation (5). Second, we implement IPW-CBPS with the same treatment assignment model using the function `CBMSM()` in the R package `CBPS`. Finally, we implement residual balancing by, first, fitting linear models for each covariate in L_t ($t \geq 2$) with lagged values of treatment and the time-varying confounders as regressors, and then extracting residual terms $\hat{\delta}(L_t)$. For each covariate in L_1 , the residual term is computed as the deviation from its sample mean. Next, we find a set of minimum entropy weights such that, in the weighted sample, $\hat{\delta}(L_t)$ is orthogonal to treatment at time t and the regressors of L_{jt} . We compute estimates of standard errors using both the robust (i.e., “sandwich”) variance estimator⁶ and the jackknife method.⁷ R code for implementing residual balancing in this analysis is available in Part C of the Supplementary Material.

Results from these analyses are presented in Table 1, where the first two columns contain IPW-GLM, IPW-CBPS, and residual balancing estimates based on the dichotomized version of treatment.

⁶In Part B of the Supplementary Material, we report a set of simulation results on the performance of the robust variance estimator for IPW-GLM, IPW-GLM-Censored, IPW-CBPS, and residual balancing. We find that the robust variance estimator is consistently conservative for residual balancing. For IPW and its variants, the robust variance estimator appears to sometimes over-estimate and other times under-estimate the true sampling variance, depending on the particular scenario.

⁷When possible, the nonparametric bootstrap can also be used with residual balancing and IPW. However, because of the small sample size of the campaign advertising dataset, the residual balancing algorithm does not converge in about 25% of the bootstrapped samples, likely because the convex hull of $\{c_i; 1 \leq i \leq n\}$ does not contain $\mathbf{0}$ in those cases. Because it is dubious to use a variance estimate based on a nonrandom fraction of bootstrapped samples, we report standard errors from only the robust variance estimator and the jackknife method.

For nonincumbent candidates, these results suggest that the effect of negative advertising is positive. However, both IPW-CBPS and residual balancing yield point estimates that are considerably smaller than IPW-GLM. While IPW-GLM suggests that an additional week with more than 10% negative advertising increases a candidate's vote share by 1.42 percentage points, on average, the estimated effect is reduced to 0.78 percentage points for IPW-CBPS and 0.98 percentage points for residual balancing. For incumbent candidates, all three methods indicate that negative advertising has a substantively large negative effect on vote shares. Residual balancing, for example, suggests that an additional week with more than 10% negative advertising decreases a candidate's vote share by 1.67 percentage points, on average.

The last two columns of Table 1 present results based on the continuous version of treatment. Because IPW-CBPS has not been extended for continuous treatments in the time-varying setting, we focus on estimates from IPW-GLM and residual balancing. Overall, these results are quite consistent with those based on the dichotomized treatment. For nonincumbents, the effect of negative advertising appears to be positive, although the estimate from residual balancing is relatively small. For incumbents, both methods suggest a sizable negative effect. According to the residual balancing estimate, a 10 percentage point increase in the proportion of negative advertising throughout the final five weeks of the campaign reduces a candidate's vote share by about one percentage point, on average.

6 The Controlled Direct Effect of Shared Democracy on Public Support for War

In this section, we reanalyze data from Tomz and Weeks (2013) to estimate the controlled direct effect (CDE) of shared democracy on public support for war, controlling for a respondent's perceived morality of war. With a nationally representative sample of 1,273 US adults, Tomz and Weeks (2013) conducted a survey experiment to analyze the role of public opinion in the democratic peace, that is, the empirical regularity that democracies almost never fight each other. In this experiment, they presented respondents with a situation in which a country was developing nuclear weapons and, when describing the situation, they randomly and independently varied three characteristics of the

country: its political regime (whether it was a democracy), alliance status (whether it had signed a military alliance with the United States), and economic ties (whether it had high levels of trade with the United States). They then asked respondents about their levels of support for a preventive military strike against the country’s nuclear facilities. The authors found that individuals are substantially less supportive of military action against democracies than against otherwise identical autocracies.

To investigate the causal mechanisms through which shared democracy reduces public support for war, Tomz and Weeks (2013) also measured each respondent’s beliefs about the threat posed by the potential adversary (*threat*), the cost of military intervention (*cost*), and the likelihood of victory (*success*). In addition, the authors assessed each respondent’s moral concerns about using military force (*morality*). With these data, they conducted a causal mediation analysis and found that shared democracy reduces public support for war primarily by changing perceptions of the threat and morality of using military force. In this analysis, the authors examined the role of each mediator separately by assuming that they operate independently and do not influence one another. However, it is likely that one’s perception of morality is partly influenced by beliefs about the threat, cost, and likelihood of success, which also affect support for war directly. Thus, in the following analysis, we treat these variables as post-treatment confounders and reassess the mediating role of morality accordingly.

In these data, the outcome, Y , is a measure of support for war on a five-point scale; treatment, A , denotes whether the country developing nuclear weapons was presented as a democracy; the mediator, M , is a dummy variable indicating whether the respondent thought it would be morally wrong to strike; the baseline covariates X include dummy variables for each of the two other randomized treatments (alliance status and economic ties) as well as a number of demographic and attitudinal controls; and the post-treatment confounders Z include measures of the respondent’s beliefs about threat, cost, and likelihood of success.⁸ We estimate the CDE of shared democracy, controlling for perceptions of morality, using an MSM with form

$$\mathbb{E}[Y(a, m)|X] = \alpha_0 + \alpha_1 a + \alpha_2 m + \alpha_3 am + \alpha_4^T X. \quad (19)$$

In this model, we control for baseline covariates because, although treatment is randomly assigned, they may still confound the mediator-outcome relationship.⁹ The controlled direct effect is given by

⁸For detailed descriptions of the variables included in L and Z , see Tomz and Weeks (2013, Table 5).

⁹Alternatively, these pretreatment confounders can be adjusted for using IPW or residual balancing weights. We adjust

Table 2: Estimated CDE of Shared Democracy on Support for War using IPW and Residual Balancing

	Total Effect	IPW	Residual Balancing
intercept	2.39 (0.05; 0.05)	3.12 (0.05; 0.06)	2.76 (0.05; 0.05)
shared democracy	-0.35 (0.07; 0.07)	-0.20 (0.07; 0.08)	-0.36 (0.08; 0.08)
moral concerns		-1.63 (0.14; 0.15)	-1.20 (0.13; 0.13)
shared democracy * moral concerns		-0.05 (0.16; 0.16)	0.14 (0.16; 0.16)

Note: Coefficients of pretreatment covariates are omitted. For ease of interpretation, all pretreatment covariates are centered at their means. The two numbers in each parenthesis are robust (i.e., “sandwich”) standard errors and jackknife standard errors, respectively.

$CDE(m) = \alpha_1 + \alpha_3 m$, where α_1 measures the effect of shared democracy on support for war if none of the respondents had moral reservations about military intervention and $\alpha_1 + \alpha_3$ measures the effect of shared democracy on support for war if all respondents thought it would be morally wrong to strike.

We estimate this model with both IPW-GLM and residual balancing weights. Specifically, we first implement IPW-GLM by fitting a logit model for M with X , A , and Z as regressors, by fitting a second logit model for M with only X and A as regressors, and then by using the fitted values from these models to estimate a set of weights with the following form: $sw_i^\dagger = \frac{\mathbb{P}(M=m_i|X=x_i, A=a_i)}{\mathbb{P}(M=m_i|X=x_i, A=a_i, Z=z_i)}$. Second, we implement residual balancing by fitting a linear model for each post-treatment confounder in Z with X and A as regressors, computing residual terms $\hat{\delta}(Z)$, and then finding a set of minimum entropy weights such that, in the weighted sample, $\hat{\delta}(Z)$ is orthogonal to M and the regressors of Z . Standard errors are computed using the robust (i.e., “sandwich”) variance estimator and the jackknife method. R code for implementing residual balancing in this analysis is available in Part C of the Supplementary Material.

As a benchmark, the first column of Table 2 presents an estimate of the total treatment effect from a regression of Y on X and A . Consistent with the original study, we find that shared democracy significantly reduces public support for war—specifically, by 0.35 points on the five-point scale, or about 0.25 standard deviations. The next two columns present IPW and residual balancing estimates,

for them directly in the MSM for the sake of statistical efficiency.

respectively, for model (19). In this model, the “main effect” of shared democracy represents the estimated CDE if respondents had no moral reservations about military intervention, and the sum of this coefficient and the interaction term represents the estimated CDE if respondents did have moral reservations.

IPW and residual balancing yield somewhat different estimates of these effects. According to IPW, the estimated CDE of shared democracy is -0.20 if respondents had no moral concerns about war, and it is -0.25 if respondents thought it was morally wrong to strike. According to residual balancing, by contrast, the estimated CDE of shared democracy is -0.36 if respondents had no moral concerns about war, and it is -0.22 if respondents thought military intervention was morally wrong. Notwithstanding these differences, however, both IPW and residual balancing suggest that most of the total effect is “direct,” that is, transmitted through pathways other than morality.

7 Discussion and Conclusion

Post-treatment confounding arises in analyses of both time-varying treatments and causal mediation, where it complicates the use of conventional regression, matching, and balancing methods for causal inference. To adjust for this type of confounding, researchers most often use MSMs along with the associated method of IPW estimation (Robins 1999; Robins, Hernan and Brumback 2000; VanderWeele 2015). IPW, however, is highly sensitive to model misspecification, relatively inefficient, susceptible to finite-sample bias, and difficult to use with continuous treatments. Several remedies for these problems have been proposed, such as censoring the weights (Cole and Hernán 2008) or constructing them with CBPS (Imai and Ratkovic 2014; 2015), but these corrections are not without their own limitations.

In this article, we proposed the method of residual balancing for constructing weights that can be used to estimate MSMs. Like IPW, residual balancing avoids the bias that afflicts conventional methods of covariate adjustment when some or all of the covariates are post-treatment confounders. In contrast to IPW, residual balancing does not require models for the conditional distribution of exposure to treatment and/or a mediator. Rather, it entails modeling only the conditional means of the post-treatment confounders, and because it simultaneously imposes covariate balancing and minimum entropy conditions on the weights, the method is both more efficient and more robust

to model misspecification than IPW. It is also much easier to use with continuous treatments, which obviates the need for arbitrary quantile binning as is often employed in practice (e.g., Wodtke, Harding and Elwert 2011; Blackwell 2013).

Residual balancing also appears to outperform IPW even when the weights are constructed with CBPS, which also incorporate explicit balancing conditions when estimating the conditional probabilities of exposure. The reason, we believe, is that IPW with CBPS is torn between two conflicting goals. On the one hand, it imposes a parametric logistic model on the propensity score, which limits the number of balancing conditions that can be satisfied with inverse probability weights. On the other hand, it attempts to balance the time-varying confounders across all possible sequences of future treatments within all possible histories of prior treatments, generating an extremely large number of balancing conditions. Therefore, the search for covariate balancing weights is almost always an over-identified problem with CBPS, leading to weights that can at best satisfy the balancing conditions approximately. In this situation, IPW with CBPS may remain biased if certain important balancing conditions are not well satisfied in the weighted sample. By contrast, residual balancing does not impose a parametric structure on the conditional probability/density of the exposure. Moreover, it models the conditional means of the time-varying confounders and balances only their residuals across a parsimonious representation of future treatments and the observed past. Therefore, the search for residual balancing weights is often an under-identified problem, leading to exact, rather than approximate, balance in the weighted sample.

Despite its many advantages, residual balancing is still limited in several ways. First, it requires modeling the conditional means of the post-treatment confounders (or transformations thereof). As noted earlier, when these models are misspecified, the pseudo-population created by the residual balancing weights will no longer mimic the original unweighted population, making estimates of marginal effects biased for the target quantities of interest. This problem might be mitigated in practice by combining residual balancing with a sensitivity analysis to assess the robustness of estimates to different parametric models for the post-treatment confounders. Another remedy might involve fitting non- or semi-parametric models for $\mathbb{E}[g(L_t)|\bar{L}_{t-1}, \bar{A}_{t-1}]$, although this may potentially engender inferential problems (e.g., a lack of \sqrt{n} -consistency; see Newey 1994) and thus additional research is needed to better understand the method's performance with these types of models for the post-treatment confounders.

Second, even when models for the conditional means of the post-treatment confounders are correctly specified, residual balancing estimates of marginal effects may still be biased if the balancing conditions are insufficient. In practice, this bias can be mitigated by including more functions (e.g., cross-product and higher-order terms) in $G(L_t)$. Nevertheless, if there are a large number of time-varying confounders, inclusion of their cross-product and higher-order terms would multiply the number of balancing conditions, making exact balance more difficult to achieve. In those cases, the balancing conditions in equation (15) may need to be relaxed to allow for approximate, rather than exact, balance (e.g., Wang and Zubizarreta Forthcoming). We leave this extension for future work.

Another important direction for future research will be to further investigate the theoretical properties of residual balancing. For example, consistency may be established if the method can be recast as a form of IPW with treatment probabilities/densities estimated from a proper scoring rule (an objective function that is not necessarily the log-likelihood). As Zhao and Percival (2017) show, when treatment is binary and the estimand is the average treatment effect on the treated (ATT), entropy balancing weights can be recast as inverse probability weights estimated from a tailored objective function that differs from the Bernoulli likelihood. However, this relationship does not hold when the estimand is the average treatment effect (ATE). Specifically, Zhao (2019) shows that inverse probability weights for the ATE can be viewed as a set of covariate balancing weights only when a different loss function ($\sum_i (w_i - 1) \log(w_i - 1) - w_i$), rather than the entropy loss ($\sum_i w_i \log w_i$), is used in the optimization problem. This result suggests that alternative loss functions may be required to establish a formal link between residual balancing and IPW. Future work should therefore explore the properties and performance of residual balancing with a variety of loss functions, including but not limited to the entropy loss on which we focus in the present study.

These limitations notwithstanding, residual balancing appears to provide an efficient and robust method of constructing weights for MSMs. It should therefore find wide application in analyses of time-varying treatments and causal mediation, wherever post-treatment confounding presents itself. To facilitate its implementation in practice, we have developed an open-source R package, `rbw`, for constructing residual balancing weights, which is available from GitHub: <https://github.com/xiangzhou09/rbw>. A Stata package with similar functionality is also available from GitHub: <https://github.com/gtwodtke/rbw>. In addition, Part C of the Supplementary Material provides R code illustrating the use of `rbw` in our two empirical examples.

References

- Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3):512–529.
- Athey, Susan, Guido W Imbens and Stefan Wager. 2018. "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(4):597–623.
- Blackwell, Matthew. 2013. "A Framework for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57(2):504–520.
- Cole, Stephen R and Miguel A Hernán. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168(6):656–664.
- Fong, Christian, Chad Hazlett, Kosuke Imai et al. 2018. "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements." *The Annals of Applied Statistics* 12(1):156–177.
- Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.
- Imai, Kosuke and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63(2):467–490.
- Imai, Kosuke and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.
- Imai, Kosuke and Marc Ratkovic. 2015. "Robust Estimation of Inverse Probability Weights for Marginal Structural Models." *Journal of the American Statistical Association* 110(511):1013–1023.
- Kang, Joseph DY and Joseph L Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22(4):523–539.

- Ladam, Christina, Jeffrey J Harden and Jason H Windett. 2018. "Prominent Role Models: High-Profile Female Politicians and the Emergence of Women as Candidates for Public Office." *American Journal of Political Science* 62(2):369–381.
- Lau, Richard R, Lee Sigelman and Ivy Brown Rovner. 2007. "The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment." *Journal of Politics* 69(4):1176–1209.
- Lefebvre, Genevieve, Joseph AC Delaney and Robert W Platt. 2008. "Impact of Mis-specification of the Treatment Model on Estimates from a Marginal Structural Model." *Statistics in medicine* 27(18):3629–3642.
- Naimi, Ashley I, Erica EM Moodie, Nathalie Auger and Jay S Kaufman. 2014. "Constructing Inverse Probability Weights for Continuous Exposures: a Comparison of Methods." *Epidemiology* 25(2):292–299.
- Newey, Whitney K. 1994. "The Asymptotic Variance of Semiparametric Estimators." *Econometrica: Journal of the Econometric Society* pp. 1349–1382.
- Pearl, Judea. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. pp. 411–420.
- Pearl, Judea. 2009. *Causality (2nd Edition)*. Cambridge University Press.
- Robins, James. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period-Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7(9-12):1393–1512.
- Robins, James M. 1999. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." *Statistical Models in Epidemiology: The Environment and Clinical Trials* .
- Robins, James M. 2003. "Semantics of Causal DAG models and the Identification of Direct and Indirect effects." *Highly Structured Stochastic Systems* pp. 70–81.
- Robins, James M, Miguel Angel Hernan and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.

- Simmons, Beth A and Cosette D Creamer. 2019. "Do Self-Reporting Regimes Matter? Evidence From the Convention Against Torture." *International Studies Quarterly* pp. 19–16.
- Tomz, Michael R and Jessica L Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107:849–865.
- VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.
- Vansteelandt, Stijn. 2009. "Estimating Direct Effects in Cohort and Case–control Studies." *Epidemiology* 20(6):851–860.
- Wang, Yixin and José R Zubizarreta. Forthcoming. "Minimal Approximately Balancing Weights: Asymptotic Properties and Practical Considerations." *Biometrika* .
- Wang, Yue, Maya L Petersen, David Bangsberg and Mark J van der Laan. 2006. "Diagnosing Bias in the Inverse Probability of Treatment Weighted Estimator Resulting from Violation of Experimental Treatment Assignment."
- Wodtke, Geoffrey T, David J Harding and Felix Elwert. 2011. "Neighborhood Effects in Temporal Perspective: The impact of Long-term Exposure to Concentrated Disadvantage on High School Graduation." *American Sociological Review* 76(5):713–736.
- Zhao, Qingyuan. 2019. "Covariate Balancing Propensity Score by Tailored Loss Functions." *The Annals of Statistics* 47(2):965–993.
- Zhao, Qingyuan and Daniel Percival. 2017. "Entropy Balancing is Doubly Robust." *Journal of Causal Inference* 5(1).
- Zhou, Xiang and Geoffrey T. Wodtke. 2019. "A Regression-with-Residuals Method for Estimating Controlled Direct Effects." *Political Analysis* .
- Zhukov, Yuri M. 2017. "External Resources and Indiscriminate Violence: Evidence from German-Occupied Belarus." *World Politics* 69(1):54–97.
- Zubizarreta, José R. 2015. "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association* 110(511):910–922.

Supplementary Materials for “Residual Balancing: A Method of Constructing Weights for Marginal Structural Models”

Xiang Zhou Geoffrey T. Wodtke
Harvard University University of Chicago

December 13, 2019

A. Minimization of Relative Entropy

Following Hainmueller (2012), we use the method of Lagrange multipliers to find a set of weights rbw_i that minimize their relative entropy with the base weights q_i subject to the balancing constraints. Substituting $\hat{\delta}(g_j(l_{it}))$ for $\delta(g_j(l_{it}))$ in equation (14) in the main text, the balancing constraints can be written as

$$\sum_{i=1}^n rbw_i \hat{c}_{ir} = 0, \quad 1 \leq r \leq n_c,$$

where \hat{c}_{ir} is the r th element of $\hat{c}_i = \{\hat{\delta}(g_j(l_{it}))h_k(\bar{l}_{i,t-1}, \bar{a}_i); 1 \leq j \leq J_t, 1 \leq k \leq K_t, 1 \leq t \leq T\}$. In addition, we impose a normalization constraint $\sum_i rbw_i = n$ such that the residual balancing weights sum to the sample size. Thus, the primal optimization problem is

$$\min_{rbw_i} L^p = \sum_{i=1}^n rbw_i \log \frac{rbw_i}{q_i} + \sum_{r=1}^{n_c} \lambda_r \sum_{i=1}^n rbw_i c_{ir} + \lambda_0 \left(\sum_{i=1}^n rbw_i - n \right), \quad (1)$$

where $\{\lambda_1, \dots, \lambda_{n_c}\}$ are the Lagrange multipliers for the balancing constraints and λ_0 is the Lagrange multiplier for the normalization constraint. Since the loss function L^p is strictly convex, the first order condition of equation (1) implies that the solution for each weight is

$$rbw_i^* = \frac{nq_i \exp(-\sum_{r=1}^{n_c} \lambda_r c_{ir})}{\sum_{i=1}^N q_i \exp(-\sum_{r=1}^{n_c} \lambda_r c_{ir})}. \quad (2)$$

Inserting equation (2) into L^p leads to the dual problem given by

$$\max_{\lambda_r} L^d = -\log \left(\sum_{i=1}^n q_i \exp \left(- \sum_{r=1}^{n_c} \lambda_r c_{ir} \right) \right),$$

or equivalently,

$$\min_Z L^d = \log \left(Q' \exp (CZ) \right),$$

where $Q = [q_1, q_2, \dots, q_n]'$, $C = [c_1, c_2, \dots, c_n]'$, and $Z = -[\lambda_1, \lambda_2, \dots, \lambda_{n_c}]'$. Since both the gradient and the Hessian have closed-form expressions, this problem can be solved using Newton's method. Inserting the solutions for λ_r into equation (2) yields the residual balancing weights.

B. Performance of the Robust (“Sandwich”) Variance Estimator

In most applications of marginal structural models (MSMs), standard errors are computed with the robust (“sandwich”) variance estimator. In this section, we present a simulation study that evaluates the performance of the robust variance estimator for MSM coefficients estimated via IPW-GLM, IPW-GLM-Censored, IPW-CBPS, and residual balancing (under the same setup described in Section 4 of the main text). The results are shown in Figures S1-S4, where the box plots display the sampling distributions of the robust standard errors divided by the true standard errors estimated from the 2,500 random samples. Across nearly all scenarios, and especially when the confounder models are correctly specified, the robust variance estimator is conservative for residual balancing, that is, it tends to overestimate the true sampling variance. Consequently, as Tables S1-S2 show, when the confounder models are correctly specified, confidence intervals constructed with these standard errors typically ensure true coverage rates that are at least equal to, and often exceed, the nominal coverage rate. By contrast, results from this simulation study suggest that the robust variance estimator may underestimate the true sampling variance under IPW-GLM in many different situations, even though it is expected to be conservative in large samples (Robins 1999; Robins, Hernan and Brumback 2000). As a result, confidence intervals constructed with these standard errors often fall short of the nominal coverage rate, even when the propensity score models are correctly specified.

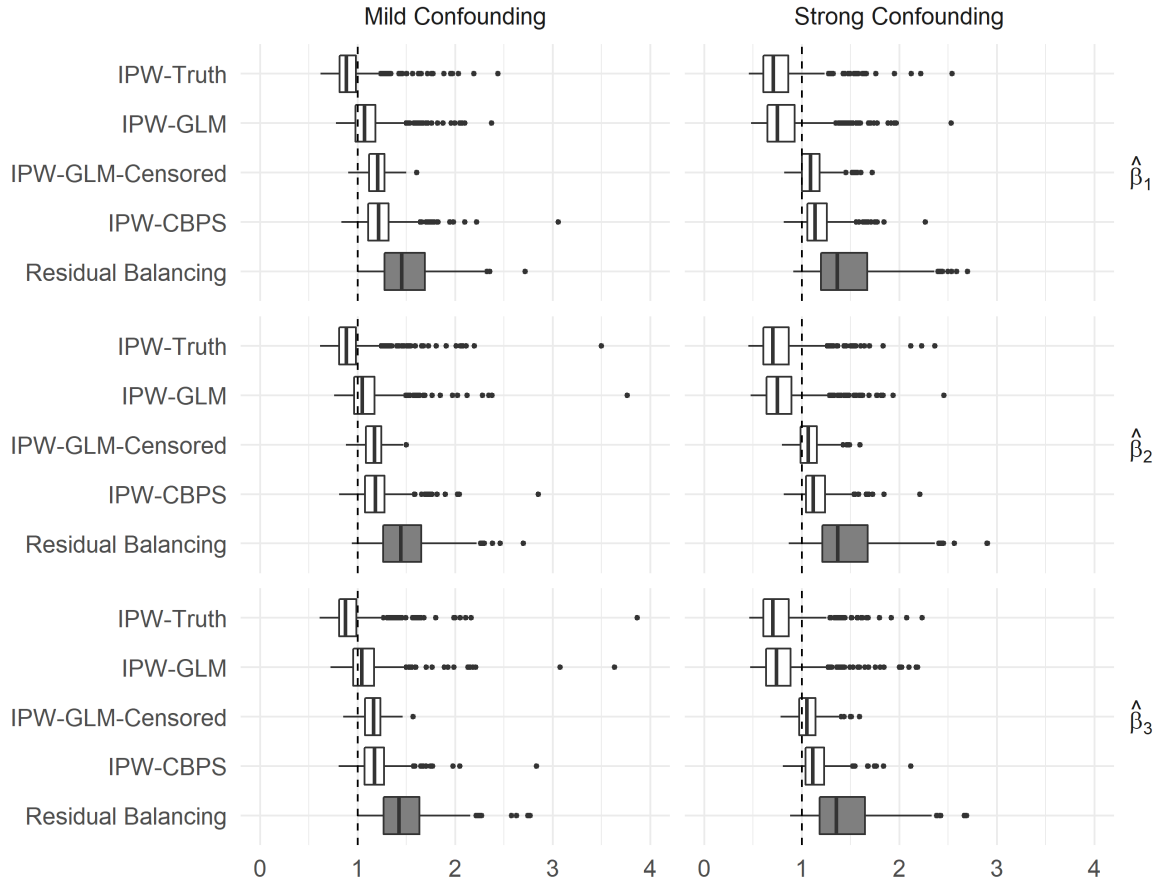


Figure S1: Performance of the robust (“sandwich”) variance estimator for a binary treatment with correct model specification. The left and right panels correspond to the settings of “mild confounding” ($\alpha = 0.4$) and “strong confounding” ($\alpha = 0.8$) respectively. Four different methods are compared: IPW based on the standard logistic regression (IPW-GLM), IPW based on the standard logistic regression with weights censored at the 1st and 99th percentiles (IPW-GLM-Censored), IPW based on the CBPS (IPW-CBPS), and residual balancing. As a benchmark, results from IPW based on true treatment probabilities (IPW-Truth) are also reported. The box plots show the sampling distributions (from 2500 random samples) of the robust standard errors divided by the true standard errors (estimated via the 2500 random samples).

Table S1: Coverage of 95% confidence intervals constructed with robust (“sandwich”) standard errors for a binary treatment with correct model specification.

	Mild Confounding			Strong Confounding		
	β_1	β_2	β_3	β_1	β_2	β_3
IPW-Truth	0.94	0.92	0.93	0.90	0.85	0.88
IPW-GLM	0.95	0.97	0.97	0.92	0.90	0.90
IPW-GLM-Censored	0.94	0.95	0.97	0.93	0.79	0.82
IPW-CBPS	0.90	0.83	0.95	0.87	0.40	0.67
Residual Balancing	0.98	1.00	0.99	0.98	1.00	0.98

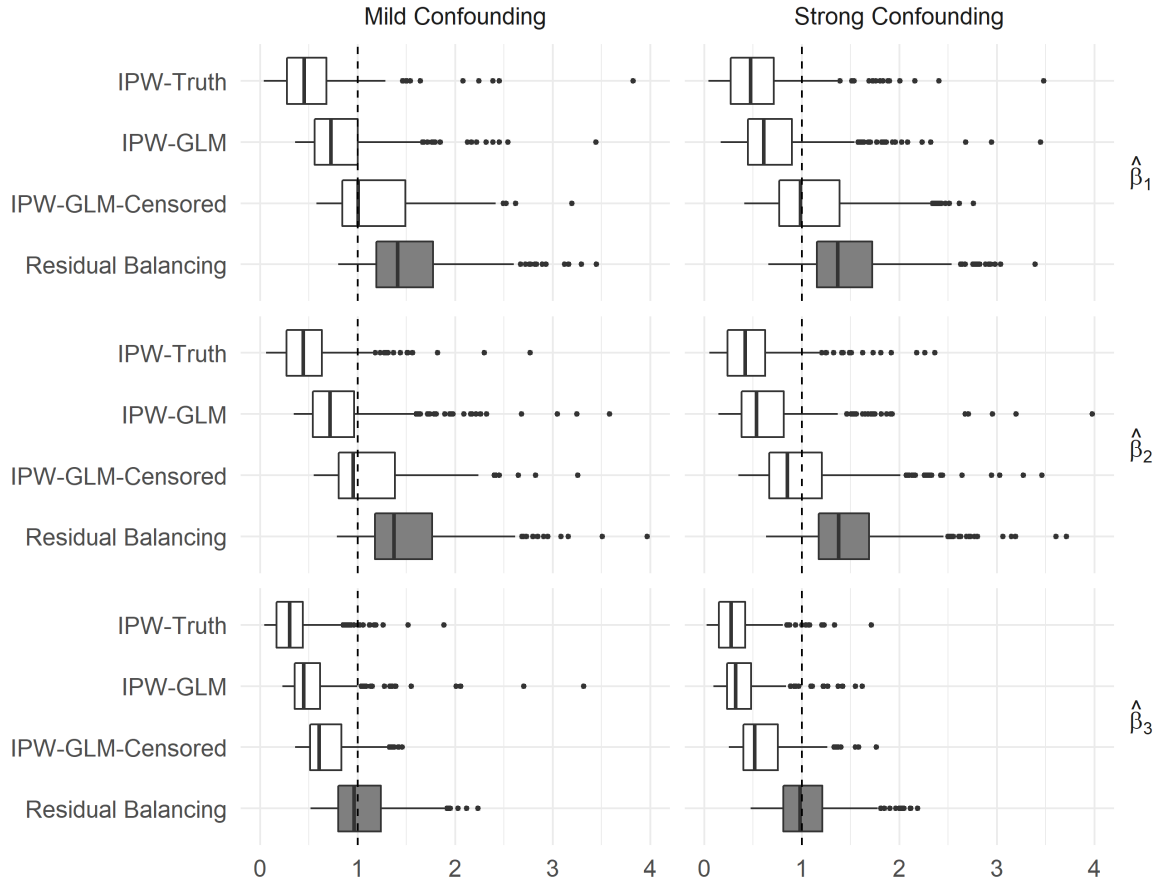


Figure S2: Performance of the robust (“sandwich”) variance estimator for a continuous treatment with correct model specification. The left and right panels correspond to the settings of “mild confounding” ($\alpha = 0.4$) and “strong confounding” ($\alpha = 0.8$) respectively. Three different methods are compared: IPW based on the standard logistic regression (IPW-GLM), IPW based on the standard logistic regression with weights censored at the 1st and 99th percentiles (IPW-GLM-Censored), and residual balancing. As a benchmark, results from IPW based on true treatment probabilities (IPW-Truth) are also reported. The box plots show the sampling distributions (from 2500 random samples) of the robust standard errors divided by the true standard errors (estimated via the 2500 random samples).

Table S2: Coverage of 95% confidence intervals constructed with robust (“sandwich”) standard errors for a continuous treatment with correct model specification.

	Mild Confounding			Strong Confounding		
	β_1	β_2	β_3	β_1	β_2	β_3
IPW-Truth	0.72	0.63	0.56	0.68	0.39	0.34
IPW-GLM	0.91	0.85	0.88	0.8	0.58	0.66
IPW-GLM-Censored	0.91	0.72	0.77	0.83	0.27	0.40
Residual Balancing	0.98	0.99	1.00	0.97	0.99	0.99

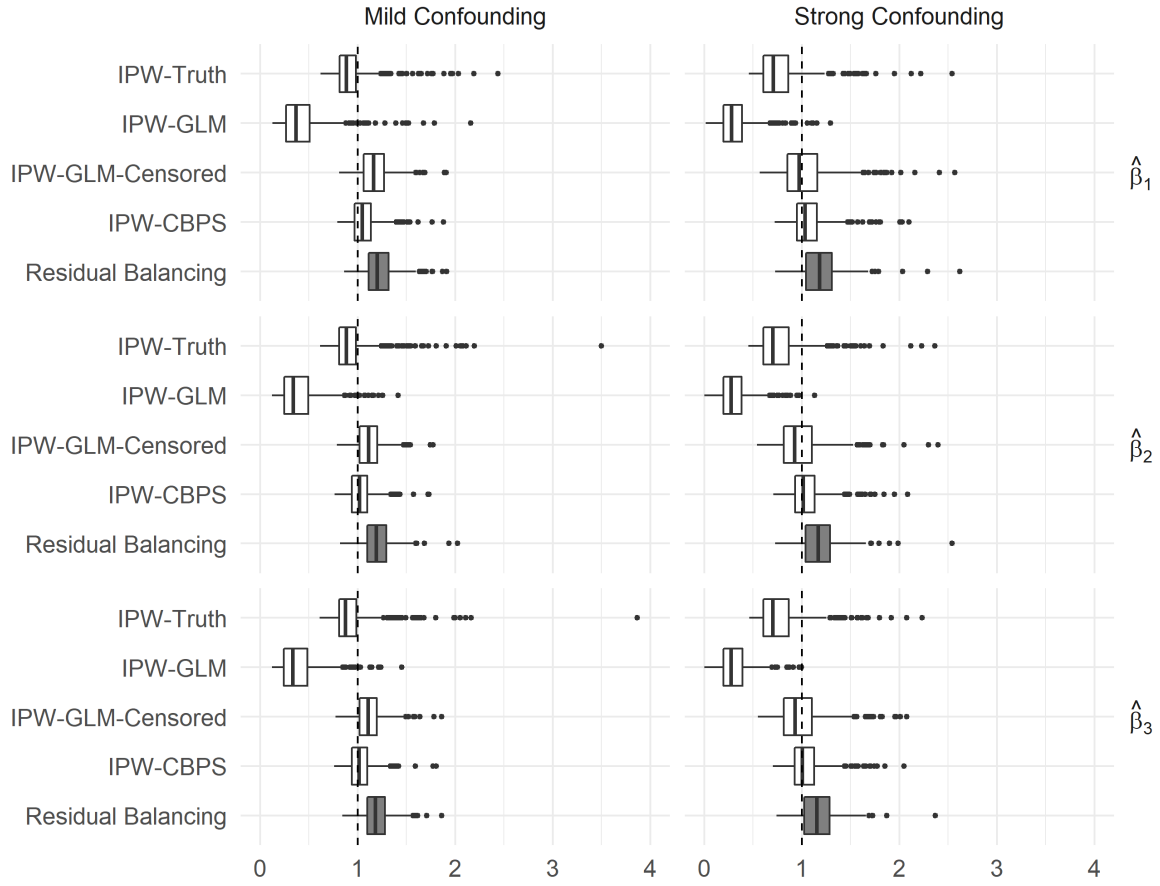


Figure S3: Performance of the robust (“sandwich”) variance estimator for a binary treatment with incorrect model specification. The left and right panels correspond to the settings of “mild confounding” ($\alpha = 0.4$) and “strong confounding” ($\alpha = 0.8$) respectively. Four different methods are compared: IPW based on the standard logistic regression (IPW-GLM), IPW based on the standard logistic regression with weights censored at the 1st and 99th percentiles (IPW-GLM-Censored), IPW based on the CBPS (IPW-CBPS), and residual balancing. As a benchmark, results from IPW based on true treatment probabilities (IPW-Truth) are also reported. The box plots show the sampling distributions (from 2500 random samples) of the robust standard errors divided by the true standard errors (estimated via the 2500 random samples).

Table S3: Coverage of 95% confidence intervals constructed with robust (“sandwich”) standard errors for a binary treatment with incorrect model specification.

	Mild Confounding			Strong Confounding		
	β_1	β_2	β_3	β_1	β_2	β_3
IPW-Truth	0.94	0.92	0.93	0.90	0.85	0.88
IPW-GLM	0.64	0.69	0.69	0.44	0.49	0.44
IPW-GLM-Censored	0.84	0.39	0.57	0.88	0.60	0.74
IPW-CBPS	0.69	0.03	0.13	0.47	0.00	0.01
Residual Balancing	0.94	0.80	0.82	0.90	0.75	0.76

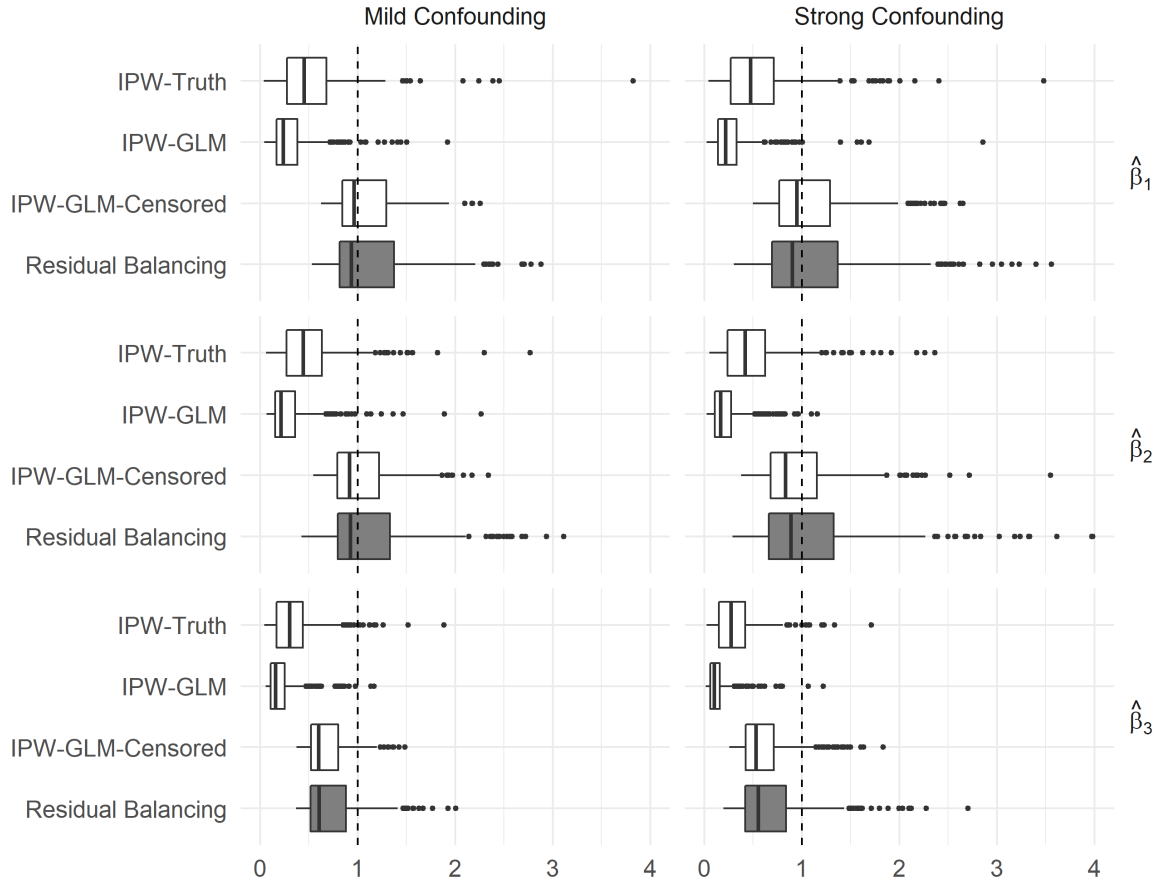


Figure S4: Performance of the robust (“sandwich”) variance estimator for a continuous treatment with incorrect model specification. The left and right panels correspond to the settings of “mild confounding” ($\alpha = 0.4$) and “strong confounding” ($\alpha = 0.8$) respectively. Three different methods are compared: IPW based on the standard logistic regression (IPW-GLM), IPW based on the standard logistic regression with weights censored at the 1st and 99th percentiles (IPW-GLM-Censored), and residual balancing. As a benchmark, results from IPW based on true treatment probabilities (IPW-Truth) are also reported. The box plots show the sampling distributions (from 2500 random samples) of the robust standard errors divided by the true standard errors (estimated via the 2500 random samples).

Table S4: Coverage of 95% confidence intervals constructed with robust (“sandwich”) standard errors for a continuous treatment with incorrect model specification.

	Mild Confounding			Strong Confounding		
	β_1	β_2	β_3	β_1	β_2	β_3
IPW-Truth	0.72	0.63	0.56	0.68	0.39	0.34
IPW-GLM	0.48	0.11	0.06	0.29	0.02	0.02
IPW-GLM-Censored	0.33	0.00	0.00	0.10	0.00	0.00
Residual Balancing	0.89	0.69	0.72	0.87	0.64	0.66

C. Illustrative R Code

In this appendix, we illustrate the implementation of residual balancing using the R package `rbw` for the two empirical examples.

```
devtools::install_github("xiangzhou09/rbw")
library(rbw); library(survey)

## Example 1: The Cumulative Effect of Negative Advertising on Candidate's Voteshare ##
# models for time-varying confounders
m1 <- lm(dem.polls ~ (d.gone.neg.l1 + dem.polls.l1 + undother.l1) * factor(week),
        data = campaign_long)
m2 <- lm(undother ~ (d.gone.neg.l1 + dem.polls.l1 + undother.l1) * factor(week),
        data = campaign_long)
xmodels <- list(m1, m2)
# residual balancing weights
fit <- rbwPanel(exposure = d.gone.neg, xmodels = xmodels, id = id, time = week,
               data = campaign_long)
campaign_wide <- merge(campaign_wide, fit$weights, by = "id")
# fitting a marginal structural model
rbw_design <- svydesign(ids = ~ 1, weights = ~ rbw, data = campaign_wide)
msm_rbw <- svyglm(demprcnt ~ cum_neg * deminc + camp.length + factor(year) + office,
                 design = rbw_design)

## Example 2: The Controlled Direct Effect of Shared Democracy on Public Support for War ##
haven::read_dta("peace.dta")
# models for post-treatment confounders
m1 <- lm(threatc ~ ally + trade + h1 + i1 + p1 + e1 + r1 + male + white + age + ed4 + democ,
        data = peace)
m2 <- lm(cost ~ ally + trade + h1 + i1 + p1 + e1 + r1 + male + white + age + ed4 + democ,
        data = peace)
m3 <- lm(successc ~ ally + trade + h1 + i1 + p1 + e1 + r1 + male + white + age + ed4 + democ,
        data = peace)
# residual balancing weights
fit <- rbwMed(treatment = democ, mediator = immoral, zmodels = list(m1, m2, m3),
             data = peace)
peace$rbw <- fit$weights
# fitting a marginal structural model
rbw_design <- svydesign(ids = ~ 1, weights = ~ rbw, data = peace)
msm_rbw <- svyglm(strike ~ ally + trade + h1 + i1 + p1 + e1 + r1 + male + white +
                 age + ed4 + democ + democ * immoral, design = rbw_design)
```

References

- Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.
- Robins, James M. 1999. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." *Statistical Models in Epidemiology: The Environment and Clinical Trials* .
- Robins, James M, Miguel Angel Hernan and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.