Political Analysis (2025), **0**, 1–22
doi:10.1017/pan.2025.10003

ARTICLE

What to Observe When Assuming Selection on Observables

Kevin M. Quinn[1], Guoer Liu[2], Lee Epstein[3] and Andrew D. Martin[4]

[1]Department of Quantitative Theory & Methods and School Law, Emory University, Atlanta, GA, USA; [2]Department of Political Science, University of California, San Diego, CA, USA; [3]Department of Political Science, Washington University, St. Louis, MO, USA; [4]Department of Political Science and Department of Statistics and Data Science and School of Law, Washington University, St. Louis, MO, USA

Corresponding author: Kevin M. Quinn; Email: kevin.michael.quinn@emory.edu

(Received 8 March 2024; revised 12 November 2024; accepted 22 November 2024)

Abstract

Political scientists regularly rely on a selection-on-observables assumption to identify causal effects of interest. Once a causal effect has been identified in this way, a wide variety of estimators can, in principle, be used to consistently estimate the effect of interest. While these estimators are all justified by appeals to the same causal identification assumptions, they often differ greatly in how they make use of the data at hand. For instance, methods based on regression rely on an explicit model of the outcome variable but do not explicitly model the treatment assignment process, whereas methods based on propensity scores explicitly model the treatment assignment process but do not explicitly model the outcome variable. Understanding the tradeoffs between estimation methods is complicated by these seemingly fundamental differences. In this paper we seek to rectify this problem. We do so by clarifying how most estimators of causal effects that are justified by an appeal to a selection-on-observables assumption are all special cases of a general weighting estimator. We then explain how this commonality provides for diagnostics that allow for meaningful comparisons across estimation methods—even when the methods are seemingly very different. We illustrate these ideas with two applied examples.

Keywords: causal inference; selection on observables; regression; propensity scores; matching; inverse probability weighting; diagnostics

Edited by: Jeff Gill

1. Introduction

Political scientists regularly rely (either implicitly or explicitly) on a selection-on-observables assumption to justify causal inferences. Informally, this assumption (which is also referred to as unconfoundedness or conditional ignorability of treatment assignment) states that (a) the potential outcomes are conditionally independent of treatment status given observed covariates and (b) every unit has nonzero probability of being assigned to each treatment condition. Estimation approaches as seemingly different as regression Hainmueller and Hangartner (2013), matching (Ichino and Nathan 2013), inverse propensity score weighting (Burgess and Tyburski 2020), and balancing weights methods (Grier *et al.* 2024; Truex 2014) are all justified by an appeal to selection on observables.[1]

These estimation methods rely on very different *modeling assumptions* which we distinguish from the *causal identification assumptions* that they share. For instance, regression approaches explicitly model

[1] As we explain in more detail below, these approaches also rely on an additional identification assumption, typically the so-called stable unit treatment value assumption (SUTVA) (Imbens and Rubin 2015, 9–13), that guarantees that the potential outcomes are well defined.

the outcome variable (often with fairly strong functional form assumptions) while not explicitly modeling the treatment assignment process. On the other hand, propensity score weighting methods explicitly model the treatment assignment process (often with fairly strong functional form assumptions) but do not explicitly model the outcome variable.

A problem for the applied researcher arises because it is difficult to know which estimation approach, using which set of modeling assumptions, is at least approximately correct within a particular application. Diagnostic checks are of some use here. However, while numerous diagnostics have been proposed over the years, for the most part the proposed diagnostics have not been general in the sense of being available for essentially all estimation methods that are justified by a selection-on-observables assumption.

Recent work in statistics has begun to change this situation.[2] In this paper, we build on these ideas, especially the work of Li, Morgan, and Zaslavsky (2018) and Chattopadhyay and Zubizarreta (2023), to clarify to political scientists (a) how estimators as seemingly different as linear regression, propensity score weighting, 1:M matching with replacement, coarsened exact matching (CEM), and entropy balancing are all special cases of a general weighting estimator and (b) how this commonality allows for a range of diagnostics that can be computed for all of these estimation approaches and meaningfully compared across approaches. We hope this helps applied researchers understand the consequences of their modeling assumptions when using different approaches. By reporting their results more openly and honestly, readers can better see how much the results depend on the models used.

Before proceeding, we want to be clear that it is not our goal in this paper to interrogate the credibility of the selection on observables assumption in particular applications or contexts. Such an enterprise is, in our opinion, best done by subject-matter experts on a case-by-case basis. Instead, we provide useful diagnostic tools along with guidance for using those tools that will be helpful to applied researchers in political science who may not be aware of related work in other fields.

2. Putting Estimation Methods that Assume Selection on Observables on a Common Footing

The basic idea of this article is simple: To compare results across different estimation methods, those different estimation methods must be put on a common footing. Once this common footing is found, we can examine the results to better understand the particular (dis)advantages of each approach. A focus on weighting estimators of Weighted Average Treatment Effects (WATEs; see Hirano, Imbens, and Ridder 2003; Li *et al.* 2018) allows us to make these comparisons.

Here estimation entails reweighting the sample units to make them representative of the target population—just as survey researchers do when they weight their respondents. Importantly, most commonly used estimators of causal effects can be viewed as weighting the observations in this way. This reweighting also serves to balance the covariates between treated and control groups.

In what follows, we provide a brief non-technical introduction to these ideas for researchers in political science who may be unfamiliar with the relevant work in economics and statistics. We begin with a description of the general WATE estimand and then note that many common estimands such as ATE and ATT (as well as many others) are particular examples of WATEs. We then summarize existing results on how weighting estimators—similar to those from survey sampling—can be used to consistently estimate WATE estimands under standard assumptions. This section concludes by pulling together results from disciplines outside political science that show how estimators as seemingly different as regression estimators, matching estimators, balancing weights estimators, and inverse propensity score weighting estimators can all be seen as special cases of a particular type of weighting estimator. This last observation is what allows us, in Section 3, to compare seemingly very different estimation approaches using common diagnostic tools. Readers who have a strong grasp of Li *et al.* (2018), Chattopadhyay and Zubizarreta (2023), and related work may wish to skip ahead to Section 3.

[2] See earlier work on diagnostics for causal inference such as Morgan and Todd (2008).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

2.1. The General WATE Estimand

Our discussion of the WATE estimand closely follows Li *et al.* (2018) who in turn build on work by Hirano *et al.* (2003). We refer interested readers to these articles for more detailed discussion.

Consider a random sample of size n composed of independent and identically distributed draws from an infinite super-population:

$$\{Y_i(0), Y_i(1), Z_i, X_i\}_{i=1}^{n}.$$

Here $Y_i(0)$ and $Y_i(1)$ are the potential outcomes for unit i under control and active treatment respectively, Z_i is a binary treatment indicator with the convention that $Z_i = 1$ indicates that unit i was exposed to active treatment and $Z_i = 0$ indicates that unit i was exposed to the control condition, and $X_i = (X_{i1}, \ldots, X_{ik})$ are measured covariates specific to unit i. The observed outcome Y_i is given by $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. See Imbens and Rubin (2015, Chapter 3) (2015). Finally, let $n^{(1)}$ and $n^{(0)}$ denote the number of treated and control units respectively.

The average treatment effect conditional on x is defined as:

$$\tau(x) \equiv \mathbb{E}[Y(1) - Y(0) | X = x].$$

This is the average unit-specific causal effect among units with $X = x$.

Although such conditional effects may be of direct interest in particular applications, the focus here is on weighted averages of these effects in the super-population. Assume that the density of X, $f_X(x)$, exists.[3] Following Li *et al.* (2018, 392) and Hirano *et al.* (2003,1163), define

$$\tau_h \equiv \frac{\int \tau(x) f_X(x) h(x) dx}{\int f_X(x) h(x) dx} \tag{1}$$

to be the *WATE*. Here $h(x)$ is a weight function that defines the target population (i.e., the target population distribution is the super-population distribution weighted by $h(x)$) and the estimand. τ_h is simply a weighted average of covariate-specific conditional effects.

A wide range of commonly used causal estimands can be written as a WATE; the only difference is in the choice of the weight function $h(x)$ (Li *et al.* 2018, 392). In this paper, we focus on ATE and ATT as they are the estimands most commonly used in political science. The ATE estimand corresponds to a situation where $h(x) = 1$ for all x and the ATT estimand corresponds to a situation where $h(x)$ is equal to the propensity score function, i.e., $h(x) = e(x) = \Pr(Z = 1 | X = x)$. That said, the ideas and diagnostics in this paper apply much more broadly.

2.2. A Hájek-Type Estimator of WATE

In this section, we describe a Hájek-type estimator of the WATE estimand and then note that many commonly used estimators of causal effects can be viewed as special cases of this weighting estimator. This provides the basis for comparing these estimators using the diagnostics discussed in Section 3.

The particular estimator that we focus on is the following:

$$\hat{\tau}_h = \frac{\sum_{i=1}^{n} w_h^{(1)}(x_i) Z_i Y_i}{\sum_{i=1}^{n} w_h^{(1)}(x_i) Z_i} - \frac{\sum_{i=1}^{n} w_h^{(0)}(x_i)(1 - Z_i) Y_i}{\sum_{i=1}^{n} w_h^{(0)}(x_i)(1 - Z_i)}, \tag{2}$$

where $w_h^{(1)}(x)$ and $w_h^{(0)}(x)$ are weight functions for treated and control units respectively. This is based on the estimator proposed by Hájek (1971) within the context of survey sampling. Accordingly, we refer to this as a *Hájek-type estimator*. This estimator goes by other names in the causal inference literature—most commonly *stabilized IPW estimator* (Aronow and Miller 2019, 228, 266).

[3] In what follows, we deal with the case of continuous X for simplicity. But we could easily define $f_X(x)$ in terms of a base measure that would allow for both discrete and continuous components of X or only discrete X, as did Li *et al.* (2018).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

Li *et al.* (2018) show that $\hat{\tau}_h$ is a consistent estimator of τ_h under the following conditions. Treatment assignment is conditionally ignorable given X, SUTVA holds,[4] and the weights are given by:

$$w_h^{(1)}(x) \propto \frac{f_X(x)h(x)}{f_X(x)e(x)} = \frac{h(x)}{e(x)}, \tag{3}$$

and

$$w_h^{(0)}(x) \propto \frac{f_X(x)h(x)}{f_X(x)(1-e(x))} = \frac{h(x)}{1-e(x)}, \tag{4}$$

where $e(x) = \Pr(Z = 1|X = x)$ is the propensity score function. As Li *et al.* (2018) also show, these weights have the property of balancing the covariates between treated and control groups such that these weighted conditional distributions are equal to each other and also equal to $f_X(x)h(x)$ (the covariate distribution of the target population):

$$f_{X|Z}(x|1)w_h^{(1)}(x) = f_{X|Z}(x|0)w_h^{(0)}(x) = f_X(x)h(x). \tag{5}$$

While we have written the weights $w_h^{(1)}(x)$ and $w_h^{(0)}(x)$ as explicit functions of x, it will often be more notationally convenient to simply write them as values indexed by the units ($i = 1,\dots,n$). For example, as $w_{hi}^{(1)}$ and $w_{hi}^{(0)}$ or, when the estimand is clear by context, simply $w_i^{(1)}$ and $w_i^{(0)}$.

2.3. Common Estimators are Equivalent to Hájek-Type Estimators

In this section, we briefly discuss the equivalences between a variety of estimators of ATE and ATT and the Hájek-type estimator discussed in the previous section. In some cases—for instance propensity score weighting methods and methods using balancing weights—these equivalences are obvious and well known within political science. In other cases—such as regression estimators of ATE and ATT—these equivalences have been proven in other fields but are not well known within political science. It is important to note that these are exact equivalences, not approximations.

2.3.1. Propensity Score Weighting Estimators

The group of estimators that can most obviously be written in the form of Equation (2) are those that directly estimate[5] the propensity score function $e(x)$ and then substitute that estimate, $\hat{e}(x)$, into Equations (3) and (4). The general form of the resulting estimator is

$$\hat{\tau}_h^{\text{ps}} = \frac{\sum_{i=1}^n \hat{w}_{hi}^{(1)} Z_i Y_i}{\sum_{i=1}^n \hat{w}_{hi}^{(1)} Z_i} - \frac{\sum_{i=1}^n \hat{w}_{hi}^{(0)} (1-Z_i) Y_i}{\sum_{i=1}^n \hat{w}_{hi}^{(0)} (1-Z_i)}, \tag{6}$$

where

$$\hat{w}_{hi}^{(1)} = \frac{h(x_i)}{\hat{e}(x_i)},$$
$$\hat{w}_{hi}^{(0)} = \frac{h(x_i)}{1-\hat{e}(x_i)}.$$

When $h(x)$ for the estimand in question depends on $e(x)$, the estimate $\hat{e}(x)$ is substituted into the expression for $h(x)$ as well. Both ATE and ATT can be consistently estimated with this approach.

[4] Formally, treatment assignment being conditionally ignorable given x means $[Y(0), Y(1)] \perp\!\!\!\perp Z|X$ and $0 < e(x) < 1$ where $e(x) = \Pr(Z = 1|X = x)$ is the propensity score (Imbens and Rubin 2015). See Imbens and Rubin (2015, 9–13) (2015, pp. 9-13) for a discussion of SUTVA.

[5] A wide variety of estimation methods are possible—including machine learning methods. For example, see Lee, Lessler, and Stuart (2010) and Setoguchi *et al.* (2008).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

As noted above, the ATE estimand has $h(x) = 1$ for all x. For the ATT estimand, $h(x) = e(x)$ so we set $\hat{h}(x) = \hat{e}(x)$. This produces the following weights:

$$\hat{w}_{hi}^{(1)} = \frac{\hat{h}(x_i)}{\hat{e}(x_i)} = \frac{\hat{e}(x_i)}{\hat{e}(x_i)} = 1,$$

and

$$\hat{w}_{hi}^{(0)} = \frac{\hat{h}(x_i)}{1-\hat{e}(x_i)} = \frac{\hat{e}(x_i)}{1-\hat{e}(x_i)}.$$

2.3.2. Methods Using Balancing Weights

Another group of estimators approach the selection of weights for Equation (2) as an explicit constrained optimization problem. We refer to these estimators as *balancing weights estimators* (for examples of such estimators, see Hainmueller 2012; Hazlett 2020; Wang and Zubizarreta 2020; Zubizarreta 2015). The approaches differ in how they define the optimization problem, but in general they attempt to find weights that create (at least) approximate balance between treated and control groups subject to constraints. That is, the weights they find equate the weighted covariate distribution among the treated units with the weighted covariate distribution among the control units as in Equation (5).

These approaches are most often used to estimate the ATT, which is typically written as:

$$\hat{\tau}_{ATT}^{\text{bw}} = \frac{1}{n^{(1)}}\sum_{i=1}^{n} Z_i Y_i - \frac{\sum_{i=1}^{n} w_i^{(0)}(1-Z_i)Y_i}{\sum_{i=1}^{n} w_i^{(0)}(1-Z_i)},$$

where $n^{(1)} = \sum_{i=1}^{n} Z_i$ and $w_i^{(0)}$ $i = 1,\ldots,n$ are weights chosen to maximize balance between the treated and control groups subject to constraints. It is obvious that this expression for $\hat{\tau}_{ATT}^{\text{bw}}$ is a special case of the WATE estimator in Equation (2).[6]

2.3.3. Regression Estimators

In an important recent work, Chattopadhyay and Zubizarreta (2023) have shown that regression estimators of ATE and ATT are also equivalent to the Hájek-type estimator in Equation (2).[7] This important fact is not widely known within political science.

The Multi-Regression Imputation (MRI) Estimator. To begin, consider the following regression estimator for ATE:

$$\hat{\tau}_{ATE}^{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_1(x_i) - \hat{m}_0(x_i)\right) \tag{7}$$

where $\hat{m}_1(x)$ is the ordinary least squares (OLS) regression estimator of $\mathbb{E}[Y|X = x, Z = 1]$ constructed by subsetting the data to the $Z = 1$ units and fitting an OLS regression of Y on X to those data and $\hat{m}_0(x)$ is the OLS regression estimator of $\mathbb{E}[Y|X = x, Z = 0]$ constructed by subsetting the data to the $Z = 0$ units and fitting an OLS regression of Y on X to those data. Chattopadhyay and Zubizarreta (2023) refer to this as the MRI estimator of ATE.[8]

[6] In principle, it is possible to use balancing weights estimators to estimate ATE. These estimators can also be written in the form of Equation (2).

[7] Chattopadhyay and Zubizarreta (2023) derives a number of other important results related to regression estimators of causal effects. Some, but not all, of these results are further discussed in this paper. We encourage readers to read Chattopadhyay and Zubizarreta (2023) for more details.

[8] This estimator was well known prior to Chattopadhyay and Zubizarreta (2023) and goes by a variety of names such as: regression imputation (Hazlett and Shinkre 2024), the t-learner (Künzel *et al.* 2019), and parametric g-estimation (Hernán and Robins 2024, Chapter 13), and simply regression estimation (Imbens 2004, 12–13).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

Let $\mathbf{X}$ denote the $n \times (k+1)$ matrix of covariates including a constant. Similarly, let $\mathbf{X}_1$ and $\mathbf{X}_0$ denote the submatrices that correspond to the portions of $\mathbf{X}$ from the treated and control units respectively. The treatment indicator, Z, is not included in $\mathbf{X}$, $\mathbf{X}_0$, or $\mathbf{X}_1$. Relatedly, let $\mathbf{y}$ denote the $n \times 1$ vector of outcomes for the full sample and $\mathbf{y}_1$ and $\mathbf{y}_0$ be the observed outcomes for the treated and control units respectively.

The regression estimator in Equation (7) is equivalent to the Hájek-type estimator in Equation (2) where the weights applied to the treated units are $\mathbf{w}^{(1)'} = \mathbf{1}_n'\mathbf{X}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'$ and the weights applied to the control units are $\mathbf{w}^{(0)'} = \mathbf{1}_n'\mathbf{X}\left(\mathbf{X}_0'\mathbf{X}_0\right)^{-1}\mathbf{X}_0'$ (Chattopadhyay and Zubizarreta 2023).[9] Here $\mathbf{1}_n$ is an n-vector of ones. Note that these weights can be negative.[10] To get some intuition about these weights, note that the imputed treated outcomes are given by

$$\hat{\mathbf{m}}_1(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}}_1 = \mathbf{X}(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1$$

which is a weighted average of the outcomes for the treated units. The imputed control outcomes are constructed similarly.

Closely related arguments to those above for ATE can be used to express the following MRI estimator of ATT:

$$\hat{\tau}_{ATT}^{\text{reg}} = \frac{1}{n^{(1)}} \sum_{i:Z_i=1} \left(y_i - \hat{m}_0(x_i)\right)$$

in the form of Equation (2) (Chattopadhyay and Zubizarreta 2023).[11]

The Uni-Regression Imputation (URI) Estimator One may also attempt to estimate an average treatment effect with a single regression. Let $\mathbf{z}$ be the $n \times 1$ vector of treatment indicators. Then, as shown by Chattopadhyay and Zubizarreta (2023), the commonly used OLS estimator of τ in the regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\tau + \boldsymbol{\epsilon}$ (what Chattopadhyay and Zubizarreta (2023) refer to as the URI estimator) is equivalent to the Hájek-type estimator of Equation (2) with a particular choice of weights (Chattopadhyay and Zubizarreta 2023). See also Section B of the Supplementary Material for an alternative derivation of these weights and discussion of their properties.[12] See Hazlett and Shinkre (2024) for a discussion of the URI estimator and its drawbacks compared to other estimators (including the MRI estimator).

Bivariate Weighted Least Squares (WLS) on the Treatment Indicator. A third approach using regression is primarily of interest to us because it underlies our DFBETA diagnostic of Section 3.2.

[9] The exact form of the Chattopadhyay and Zubizarreta (2023) weights is slightly different in that those weights sum to 1 within both the treated and control groups whereas the regression weights above sum to n within both treated and control groups. The Chattopadhyay and Zubizarreta (2023) weights and the weights above are equivalent up to this difference in normalization. The fact that Hájek-type estimators normalize the weights to sum to 1 makes this difference inconsequential. Section B of the Supplementary Materialprovides a derivation of our weights which is slightly different from the derivation in Chattopadhyay and Zubizarreta (2023).

[10] Note that it is not necessary for the weights to be given by Equations (3) and (4) in order for the estimator in Equation (2) to be a consistent estimator of a causal estimand. For instance, the MRI estimator discussed in Section 2.3.3 is a consistent estimator of ATE if: treatment assignment is conditionally ignorable given X, $\mathbb{E}[Y|X=x, Z=1] = \mathbf{X}_1\boldsymbol{\beta}_1$, $\mathbb{E}[Y|X=x, Z=0] = \mathbf{X}_0\boldsymbol{\beta}_0$, and SUTVA holds. As shown by Chattopadhyay and Zubizarreta (2023), the weights implied by this estimator can be both positive and negative. This implies that they cannot be written in the form of Equations (3) and (4) since $h(x) = 1$ for ATE and $0 < e(x) < 1$ for all x.

[11] See also Section B of the Supplementary Material.

[12] Again, the exact form of the Chattopadhyay and Zubizarreta (2023) weights is slightly different in that those weights sum to 1 within both the treated and control groups whereas our URI weights sum to n within both treated and control groups. The Chattopadhyay and Zubizarreta (2023) weights and the weights above are equivalent up to this difference in normalization. The fact that Hájek-type estimators normalize the weights to sum to 1 makes this difference inconsequential. Section B of the Supplementary Material provides a derivation of our weights which is slightly different from the derivation in Chattopadhyay and Zubizarreta (2023).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

As Imbens (2004) notes, the Hájek-type estimator of $\hat{\tau}_h$ in Equation (2) is equivalent to the WLS estimator of τ_h in the following bivariate regression of the observed outcomes on the treatment indicator:

$$Y_i = \beta_0 + Z_i \tau_h + \epsilon_i, \quad i = 1, \ldots, n \tag{8}$$

with the ith regression weight equal to:

$$\omega_i = Z_i w_h^{(1)}(x_i) + (1 - Z_i) w_h^{(0)}(x_i).$$

As Section 3.2 shows, this equivalence enables us to use standard regression diagnostics to understand the influence of particular observations on any Hájek-type estimate of a WATE—even those that do not explicitly model the outcome via regression (e.g., entropy balancing estimates).[13]

2.3.4. Matching Estimators: 1:M and CEM

Matching methods attempt to obtain balance between treated and control groups by matching treated units to control units so that the matched sample is balanced on relevant measured pre-treatment covariates (X). Numerous variants of matching exist, and there are also many ways to estimate causal effects once the matched sample has been constructed. In this subsection, we look at two broad classes of matching methods that are commonly used in political science. For each, we consider estimators of causal effects that can be written as weighted differences of means or equivalently a WLS regression of the outcomes on the treatment indicator.

1:M Matching with Replacement. As demonstrated in Section C of the Supplementary Material, the 1:M matching with replacement estimator of ATT that is written as a difference of means (see Abadie and Imbens, 2006, 241, 2006, p. 241) is equivalent to the Hájek-type estimator in Equation (2) where the weights given to the treated and control units are: $w_i^{(1)} = 1$ and $w_i^{(0)} = \frac{K_i}{M}$, respectively. Here $K_i = \sum_{l=1}^{n} \mathbb{I}(i \in \mathcal{J}_l)$ and $\mathcal{J}_l$ is the set of M unit indices of the units matched to unit i.[14]

CEM. CEM (Iacus, King, and Porro 2011, 2012) is a matching method that coarsens the covariate space into a finite number of strata and then exact matches treated and control units within each stratum.[15] CEM produces unit-specific weights that can be used for post-matching estimation of causal effects. The estimand of interest is initially ATT although the actual estimand will diverge from ATT if unmatched treated units are discarded.

The most straightforward way to estimate the ATT using CEM is to fit the WLS regression in Equation (8) above with weights equal to the weights produced by the CEM procedure (Iacus *et al.* 2012, 5). It follows that this estimator can be rewritten as the Hájek-type estimator in Equation (2) with the weights equal to the CEM weights.

3. Diagnostic Tools for Hájek-Type Estimators of Causal Effects

We now turn to diagnostics applicable to any estimator that can be written in the form of Equation (2). At the outset, we want to be clear that these diagnostic tools do not take the place of careful thinking about, and evaluation of, the modeling / specification decisions that are required for a particular approach to be a consistent estimator of an effect of interest.[16] The diagnostics we propose are most

[13] It is important to note that the WLS point estimate of τ in Equation (8) and some regression diagnostics such as DFBETA (see Section 3.2) are not affected by multiplying $w^{(0)}$ by a constant and/or multiplying $w^{(1)}$ by a different constant. However, the estimated standard error of $\hat{\tau}$ that appears in the WLS regression output will be affected by such a rescaling.

[14] Section C of the Supplementary Material also shows that the 1:M matching with replacement estimator of ATE that is written as a difference of means (again see Abadie and Imbens, 2006, 241, 2006, p. 241) is equivalent to the Hájek-type estimator in Equation (2) where the weights given to the treated and control units are: $w_i^{(1)} = w_i^{(0)} = \left(1 + \frac{K_i}{M}\right)$.

[15] CEM can also be conceptualized as a stratification or subclassification estimator (Cochran 1968; Iacus, King, and Porro 2019).

[16] For discussions of these conditions, see Imbens (2004), Zhao and Percival (2016), and Hazlett and Shinkre (2024).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

useful when applied to credible estimators, where this credibility is based on the plausibility of the needed identification and modeling assumptions.

These diagnostic tools are designed to help researchers and their readers: (1) understand the effective sample size of various estimators, (2) discover the influence a given sample point exerts on a causal estimate and how that influence varies across estimators, (3) examine the extent to which an estimator extrapolates outside the range of observed data, and (4) assess mean balance as well as distributional balance between treated and control groups. Upon publication, we will make these diagnostic tools available in an `R` package.

3.1. Effective Sample Size & Effective Sample Size Ratio

In the survey sampling literature, it is common place to calculate and report the "effective sample size" of a weighted estimator (Kish 1965). The basic idea is to note the variance of the unweighted sample mean, $\bar{Y}$, of Y is $\mathbb{V}[\bar{Y}] = \mathbb{V}[Y]/n$ where n is the sample size. If the variance of a weighted estimator can be written as $\mathbb{V}[Y]/c$ for some constant c, then the effective sample size is c. While $\mathbb{V}[Y]$ appears in this motivation, the effective sample size can be calculated before outcome data is collected. We return to this point at the end of this subsection.

The Hájek-type estimator in Equation (2) depends on the weights $w_{hi}^{(1)}$ and $w_{hi}^{(0)}$. Thus the effective sample size of this estimator will also depend on these weights. To acknowledge that, we use the notation n_{eff_w} to denote the effective sample size of this estimator.

Under the assumption that $\mathbb{V}[Y|Z=1] = \mathbb{V}[Y|Z=0] = \mathbb{V}[Y]$, some simple variance calculations and algebraic manipulations give us the following expression for n_{eff_w} (see Section D of the Supplementary Material)[17]:

$$n_{\text{eff}_w} = \frac{n_{\text{eff}}^{(1)} n_{\text{eff}}^{(0)}}{n_{\text{eff}}^{(0)} + n_{\text{eff}}^{(1)}},$$

where

$$n_{\text{eff}}^{(0)} = \frac{\left(\sum_{i:Z_i=0} w_{hi}^{(0)}\right)^2}{\sum_{i:Z_i=0} (w_{hi}^{(0)})^2}$$

and

$$n_{\text{eff}}^{(1)} = \frac{\left(\sum_{i:Z_i=1} w_{hi}^{(1)}\right)^2}{\sum_{i:Z_i=1} (w_{hi}^{(1)})^2}.$$

If the weights are estimated, these estimated weights can be inserted into the above formulas in place of $w_{hi}^{(1)}$ and $w_{hi}^{(0)}$. If at least one $w_{hi}^{(1)}$ value is non-zero and one $w_{hi}^{(0)}$ value is non-zero, then

$$0 < n_{\text{eff}}^{(0)} \leq n^{(0)}$$

$$0 < n_{\text{eff}}^{(1)} \leq n^{(1)}$$

[17] Chattopadhyay and Zubizarreta (2023) present related versions of the effective sample size for the treated and control units which are also based on the work of Kish (1965). The major differences are that they use an absolute value of the weights in the numerators of their formulas and they only compute effective sample sizes within the treated and control groups (the equivalent of what we term $n_{\text{eff}}^{(1)}$ and $n_{\text{eff}}^{(0)}$ below).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

and

$$0 < n_{\text{eff}_w} \leq \frac{n^{(1)}n^{(0)}}{n^{(0)} + n^{(1)}} \leq n/4.$$

Full details are presented in Section D of the Supplementary Material.

The effective sample size is most useful when comparing two or more estimators. It is also useful to express the effective sample size as a fraction of the maximum possible effective sample size, $\frac{n^{(1)}n^{(0)}}{n^{(0)}+n^{(1)}}$, or what we term the *effective sample size ratio*. The applications in Section 4 provide examples.

Since the weights for all of the estimators considered in this paper do not depend on the outcome data, the effective sample size can be calculated prior to observing or recording any outcome data. This is especially relevant in situations where researchers may want to adjust model specifications based on diagnostic information. Making such adjustments prior to observing outcome information minimizes the chance of bias from researcher degrees of freedom (Rubin 2008).

3.2. *Influential Data Points*

Understanding the extent to which some observations may exert a great deal of influence over one's estimates is an important part of any data analysis. Conceptually, one can gauge the influence of observation i by examining the difference between an estimate calculated with the full dataset and an estimate calculated with observation i removed. Ideally, one would like a closed form expression for this measure of influence, as repeatedly deleting observations and re-estimating the quantity of interest can be extremely time-consuming with large datasets and sophisticated estimation methods.

Chattopadhyay and Zubizarreta (2023) have made progress on this goal for the MRI and URI estimators discussed above. What they term the sample influence curve for observation i (denoted SIC_i) is proportional to this difference in estimates with and without observation i.[18] While this provides an exact measure of influence for URI and MRI estimates of ATE, it is not directly useful for non-regression estimates of ATE.

We provide a measure of influence that works for any estimator that can be written as the Hájek-type estimator of Equation (2).

Recall from Section 2.3.3 that any Hájek-type estimator of the form given in Equation (2) is equivalent to a bivariate WLS regression of the outcome on the treatment indicator. Since the coefficient on the treatment indicator in this bivariate WLS regression is the causal effect of interest, a measure of how much this coefficient changes after deleting observation i is a measure of influence that will work for any Hájek-type estimator. In the literature on regression diagnostics, this change in a coefficient vector after deleting observation i is referred to as DFBETA_i (Belsley, Kuh, and Welsch 1980) and is closely related to the approach of Chattopadhyay and Zubizarreta (2023) for URI and MRI estimators.

Directly relevant to our goal of finding a closed form expression for DFBETA_i for WLS regression is the work of Li and Valliant (2011) who studied the analogous problem of calculating DFBETA_i in regressions with survey weights. Their work provides a closed form expression for DFBETA_i for WLS regression. We use this to assess the influence of each observation on the Hájek-type estimator in Equation (2).[19] Note that we are interested in the second element of DFBETA_i from the bivariate WLS regression as it reveals how much the causal effect estimate would change if observation i were deleted.

[18] In the case of estimating ATE via URI, SIC_i is exactly equal to this difference after dividing by $(1-n)$. In the case of estimating ATE via MRI, SIC_i is exactly equal to this difference after dividing by $(1-n^{(1)})$ if i is a treated observation and dividing by $(1-n^{(0)})$ if i is a control observation.

[19] When the weights are data-dependent, which will be the case in most applications, this expression for DFBETA_i is only approximately equal to the change in the estimate after deleting observation i. Still, the fact that this is a closed form expression that can be evaluated nearly instantaneously for any estimator that can be written as a Hájek-type estimator is extremely useful in applications in which estimating the effect of interest within a single dataset may take hours (as is the case in many variants of matching estimators) and there are hundreds or thousands of observations to consider discarding.

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

Full details of how DFBETA_i can be calculated for any Hájek-type estimator are presented is Section E of the Supplementary Material.

3.3. Extrapolation

A key question to ask when using either a URI or MRI estimator to estimate a causal effect is "Are any of the elements of $\mathbf{w}^{(1)}$ or $\mathbf{w}^{(0)}$ negative?"[20] If all of the weights are non-negative, then each of the two weighted averages on the right-hand side of Equation (2) will be a convex combination of the observed treated and control outcomes respectively. In other words, the estimator is interpolating the observed data. If some of the weights are negative, then the estimator is extrapolating outside the range of the observed data.[21,22]

Although researchers should be concerned about any amount of extrapolation, the dangers of extrapolation vary with the extremity of the extrapolation. A simple measure of the amount of extrapolation in the estimation of the mean of $Y(0)$ is

$$\text{EXTRAP}^{(0)} = \frac{\sum_{i=1}^{n} |w_i^{(0)}| \mathbb{I}(w_i^{(0)} < 0)(1-z_i)}{\sum_{i=1}^{n} w_i^{(0)} \mathbb{I}(w_i^{(0)} \geq 0)(1-z_i)},$$

where $\mathbb{I}(\cdot)$ is the indicator function and $|\cdot|$ is the absolute value function. $\text{EXTRAP}^{(0)}$ is the ratio of the sum negative control group weights to the sum of the positive control group weights. The amount of extrapolation in the estimation of the mean of $Y(1)$ is defined analogously as:

$$\text{EXTRAP}^{(1)} = \frac{\sum_{i=1}^{n} |w_i^{(1)}| \mathbb{I}(w_i^{(1)} < 0) z_i}{\sum_{i=1}^{n} w_i^{(1)} \mathbb{I}(w_i^{(1)} \geq 0) z_i}.$$

For the regression estimators discussed in Section 2.3.3 $\text{EXTRAP}^{(0)}$ and $\text{EXTRAP}^{(1)}$ take values between 0 and 1.[23] See Section B.2 of the Supplementary Material for details.

3.4. Balance

We can assess balance between the treated and control groups by weighting the covariate distributions within the treated and control groups by $w_h^{(1)}(x)$ and $w_h^{(0)}(x)$ respectively and checking to see whether Equation (5) holds in the sample.[24] There are two broad approaches to this. The first checks whether the weighted means of the distributions in Equation (5) are equal. The second attempts to assess the overall equality of these distributions.[25]

3.4.1. Weighted Mean Balance

A standard measure of balance is the *absolute standardized mean difference* (ASMD). Following Chattopadhyay *et al.* (2020) this is defined for a particular covariate x as:

[20] Because propensity score weighting, matching, and other methods using balancing weights all constrain the weights to be non-negative, this section is most relevant for regression estimators.

[21] Chattopadhyay and Zubizarreta (2023) express similar concerns about negative weights in URI and MRI estimators.

[22] An additional point to make about negative WATE weights is that the weighted means calculated in Equation (2) cannot be expressed as sample means of an unequal probability sample with probabilities of inclusion proportional to the WATE weights. This means that some of the balance-checking methods discussed in Section 3.4 that reweight the covariate distributions as if the WATE weights are sampling weights are not applicable.

[23] Other definitions of extrapolation are possible. For example, see King and Zeng (2006).

[24] See Austin and Stuart (2015), Imbens and Rubin (2015, Chapter 14), and Chattopadhyay, Hase, and Zubizarreta (2020) among others.

[25] Note that all of the methods discussed here only examine one variable at a time. As such, they may fail to diagnose more complicated forms of imbalance involving nonlinear combinations of variables. See e.g., Hazlett (2020).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

$$\mathrm{ASMD}(x) = \frac{\left|\bar{x}_w^{(1)} - \bar{x}_w^{(0)}\right|}{\sqrt{(s^{2(1)} + s^{2(0)})/2}},$$

where

$$\bar{x}_w^{(z)} = \frac{\sum_i^n w_i^{(z)} \mathbb{I}(z_i = z) x_i}{\sum_i^n w_i^{(z)} \mathbb{I}(z_i = z)}$$

is the weighted mean of covariate x within treatment group z with weights given by $\mathbf{w}^{(z)}$ and $s^{2(z)}$ is the sample variance of x within treatment group z. See also Imbens and Rubin (2015, 310, 311) (2015, pp. 310–311) who propose using the standardized mean difference to check balance.

Chattopadhyay *et al.* (2020) also propose using the *target ASMD* (TASMD) as a measure of balance. For a particular covariate x within treatment group z (denoted $x^{(z)}$), this is defined as:

$$\mathrm{TASMD}(x^{(z)}) = \frac{\left|\bar{x}_w^{(z)} - \bar{x}^*\right|}{s^{(z)}},$$

where $\bar{x}_w^{(z)}$ is as above (the weighted mean of $x^{(z)}$ with weights given by $\mathbf{w}^{(z)}$), $\bar{x}^*$ is the sample mean of x within the target population, and $s^{(z)}$ is the unweighted sample standard deviation of x within treatment group z. The target population is simply the population that defines the estimand. For instance, if the estimand is ATE, then the target population weights each observation in the sample equally and $\bar{x}^*$ is the simple sample mean of x. If the estimand is ATT, then $\bar{x}^*$ is the simple sample mean of x within the treated units. With binary treatment, one would calculate $\mathrm{TASMD}(x^{(0)})$ and $\mathrm{TASMD}(x^{(1)})$ unless one of these quantities is trivially equal to 0 as with ATT.

Recall the balance condition from Equation (5):

$$f_{X|Z}(x|1) w_h^{(1)}(x) = f_{X|Z}(x|0) w_h^{(0)}(x) = f_X(x) h(x).$$

ASMD assesses the first equality, $f_{X|Z}(x|1) w_h^{(1)}(x) = f_{X|Z}(x|0) w_h^{(0)}(x)$, whereas TASMD assesses $f_{X|Z}(x|1) w_h^{(1)}(x) = f_X(x) h(x)$ and $f_{X|Z}(x|0) w_h^{(0)}(x) = f_X(x) h(x)$.

It is common to use an arbitrary threshold to assess weighted mean covariate balance. For instance, values of ASMD and/or TASMD below 0.1 or 0.2 are often viewed as indicating adequate covariate balance (Chattopadhyay *et al.* 2020, 15). Rather than using these arbitrary thresholds, Chattopadhyay *et al.* 2020 recommend using an asymptotically informed threshold of $\min(0.1, k^{-\frac{1}{2}})$, where k is the number of covariates (including functions of covariates such as squared terms) that one is attempting to balance (p. 15).

A major advantage of ASMD and TASMD is that they are not affected by sample size as a z-statistic or t-statistic would be (Imai, King, and Stuart 2008; Imbens and Rubin 2015). A disadvantage is that they only assess mean balance and situations can exist where mean balance is achieved but confounding still exists because of imbalance on higher moments of the covariate distribution.

As proven by Chattopadhyay and Zubizarreta (2023), the MRI estimators of ATE and ATT and the URI estimator of ATE will, by construction, always produce perfect weighted mean balance (both in terms of of ASMD and TASMD) on covariates included in the regression model(s)(Chattopadhyay and Zubizarreta 2023, Proposition 3). As they note, this implies that it is important to check balance on covariates—and transformations of covariates—that were not included in the model(s).[26] Another way to view the fact that the MRI and URI estimators produce perfect weighted mean balance is that either

[26] Indeed, mean balance is not everything. All of the estimation methods discussed in this paper rely on modeling assumptions in addition to the shared causal identification assumptions. For a discussion of how regression misspecification affects causal inference see Hazlett and Shinkre (2024).

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

weighted mean balance is less important than commonly thought or that regression estimators of ATE and ATT are better than commonly thought.

We agree, and we encourage practitioners to go even further to assess more general distributional balance rather than just weighted mean balance. In the next subsection we discuss two approaches can be used to assess more general distributional equality of covariates, beyond simple mean balance, across treated and control groups.

3.4.2. Distributional Balance

The balance conditions in Equation (5) are statements about the equality of distributions—not just the equality of the means of those distributions. Accordingly, one may well want to look beyond ASMD and TASMD when checking balance.

For a particular covariate, a measure of the discrepancy between the weighted distribution of the covariate for treated units and the weighted distribution for control units is a Kolmogorov–Smirnov (KS) test statistic for weighted data.[27] Equation (5) suggests two variants of this KS diagnostic. One compares the weighted distribution of $x^{(1)}$ to the weighted distribution of $x^{(0)}$. The other involves two KS statistics, one from a comparison of the weighted distribution of $x^{(1)}$ to the target distribution of x and the other from a comparison of the weighted distribution of $x^{(0)}$ to the target distribution of x. Larger values of the KS statistic indicate greater covariate imbalance.

A graphical approach to inspecting distributional balance of the weighted univariate covariate distributions that is arguably easier to interpret than the KS diagnostic is a weighted QQ-plot. This is constructed much like a standard QQ-plot except that the various sample quantiles are constructed from the weighted covariate distributions formed by weighting $x^{(z)}$ with weights proportional to $\mathbf{w}^{(z)}$ for $z = 0, 1$. Based on Equation (5), two variations on this ideas are possible.

One version plots the quantiles of the weighted $x^{(1)}$ on the quantiles of the weighted $x^{(0)}$. The other variant consistes of two plots, one of which plots the quantiles of the weighted $x^{(1)}$ on the quantiles of the target distribution of x and the other which plots quantiles of the weighted $x^{(0)}$ on the quantiles of the target distribution of x. A departures away from the 45-degree line indicate a discrepancy between the two distributions.

Both the KS and QQ-plot diagnostics require the weights for treated and control units to be non-negative. They are thus always available for matching and explicit weighting approaches (such as inverse propensity score weighting and balancing-weight methods). When there is no extrapolation (no negative weights) they can also be useful for assessing the balance produced by MRI and URI approaches.

4. Applying the Diagnostic Tools

The diagnostic tools discussed above are just that—diagnostic tools. They are diagnostic in that they provide clues about potential problems with a particular estimation method applied to a particular dataset. They are tools in that they require skill and judgment to be used appropriately.

With that in mind, we encourage practitioners to use these diagnostic tools in the following way when estimating a causal effect under a selection on observables assumption.

1. Choose an estimand that is appropriate for the research question of interest.
2. Collect data and decide on model specifications for which the selection on observables and SUTVA assumptions are plausible or at least defensible.
3. Using multiple estimators, calculate and evaluate the diagnostics discussed above.

[27] We do not report the p-value of this test as this depends on sample size.

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

4. Optionally adjust model specifications to address problems revealed by the diagnostics.[28] For example, if a lack of balance is observed for the square of a covariate, include the squared covariate in the model(s).
5. If the diagnostics reveal serious problems for all estimators and model specifications, conclude that these results do not allow you to reliably estimate the causal effect of interest.
6. If the diagnostics reveal no serious problems for at least some estimators and model specifications, report the results from all estimators and model specifications along with the associated diagnostics. Providing estimates and diagnostics for all estimators and specifications helps readers judge the credibility of the reported results.

To illustrate the use and value of the diagnostics above, we turn to two real-world examples—the effect of judicial vacancies on judicial votes (Black and Owens 2016) and the effect of winning office on wealth (Eggers and Hainmueller 2009).

4.1. Promotion-Seeking Judges

Black and Owens (2016) ask whether U.S. lower-court judges are more likely to cast votes in favor of the president when a vacancy exists on the U.S. Supreme Court.[29] The researchers hypothesize that judges with a chance of promotion to the U.S. Supreme Court ("contenders") are more likely to favor the president when a vacancy exists than when a vacancy does not exist. "Non-contenders" (judges with no chance of promotion) are not expected to adjust their votes in response to a vacancy. In other words, Black and Owens (2016) frame their main research question as a causal question: What is the causal effect of a Supreme Court vacancy on pro-president votes cast by judges who are contenders and by judges who are not contenders?

Black and Owens's approach to answering this question is typical of much applied work. They calculate and report a *single estimate* of a *single causal estimand* (in their case, an estimate of ATT produced by CEM).[30] And that estimate supports their hypothesis: the treatment (a vacancy) causes contender judges to vote differently—such that when a vacancy exists, contenders are more likely to vote for the president to improve their promotion prospects, while non-contenders do not change their behavior.

We re-analyze the Black and Owens data using multiple estimators of ATT and report the results in Table 1.

Looking at the ATT estimates and associated standard errors in the *CEM* rows of Table 1, we see support for Black and Owens hypotheses. The ATT estimate for contenders is positive (0.075) and significant at conventional levels ($z = 0.075/0.024 = 3.125$), while the ATT estimate for non-contenders is not significantly different from 0 at conventional levels ($z = 0.021/0.014 = 1.5$). Taken at face value, these results suggest that contenders change their voting behavior in the expected pro-president direction when a Supreme Court vacancy arises, but non-contenders do not change their behavior in these circumstances.

Then again, questions about the credibility of this result emerge if one follows our guidance and examines the diagnostics for several estimators of ATT.

[28] This is most appropriate when outcome information has (ideally) not yet been collected or at least not been examined by the researcher (see Rubin 2008). With the exception of DFBETA, it is possible to calculate all of the diagnostics without outcome data. See Chattopadhyay and Zubizarreta (2023) for a similar point with respect to the MRI and URI weights and associated diagnostics.

[29] Black and Owens (2016) also examine several other outcomes. We focus here and throughout on pro-presidential votes to keep the example manageable.

[30] In the Supplementary Material to their article, Black and Owens report some additional estimates. These results, however, also rely on CEM. As we demonstrate below, the un-interrogated reliance on coarsened exact matching is consequential.

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

Table 1. *Re-analysis of Black and Owens (2016) with multiple approaches to estimating the ATT.*

	ATT	SE	Effective	Effective	DFBETA minimum		DFBETA aximum		Extrapolation	
	estimate		sample	sample size	Value	Judge	Value	Judge	Control	Treatment
			size	ratio					units	units
The contender judges										
Regression (MRI)	0.066	0.013	1543.8	0.616	−0.000	Merrick Garland	0.000	Merrick Garland	0.099	0.000
Regression (URI)	0.068	0.011	1823.6	0.727	−0.000	Cornelia G. Kennedy	0.000	Merrick Garland	0.027	0.000
Propensity score weighting	0.018	0.012	1249.6	0.498	−0.001	John Roberts	0.001	Emilio M. Garza		
Entropy balancing	0.028	0.010	1295.7	0.517	−0.001	Merrick Garland	0.001	Merrick Garland		
Nearest-neighbor matching	0.033	0.010	363.8	0.145	−0.015	Edith Clement	0.004	Cornelia G. Kennedy		
Coarsened xact matching	0.075	0.024	53.1	0.021	−0.038	J. Harvie Wilkinson III	0.025	Emilio M. Garza		
The non-contender judges										
Regression (MRI)	0.043	0.009	3399.9	0.979	−0.000	NA	0.000	NA	0.000	0.000
Regression (URI)	0.040	0.009	3413.8	0.983	−0.000	NA	0.000	NA	0.000	0.000
Propensity score weighting	0.041	0.009	3408.0	0.981	−0.000	NA	0.000	NA		
Entropy balancing	0.042	0.008	3396.1	0.978	−0.000	NA	0.000	NA		
Nearest-neighbor matching	0.053	0.011	1401.2	0.403	−0.002	NA	0.004	NA		
Coarsened exact matching	0.021	0.014	995.4	0.287	−0.001	NA	0.001	NA		

Note: The *Regression (MRI)* rows correspond to the MRI estimator of ATT. The *Regression (URI)* rows correspond to the URI estimator which is also an estimate of ATE under the constant effects assumption. The *Propensity Score Weighting* rows correspond to the propensity score weighting estimator of ATT in which the propensity scores are estimated via logistic regression. The *Entropy Balancing* rows correspond to the entropy balancing estimator of ATT as implemented in the `WeightItR` package. The *Nearest-Neighbor Matching* rows correspond to 1:1 nearest neighbor propensity score matching using the `Match` function in the `MatchingR` package and the same estimated propensity scores as above. The *Coarsened Exact Matching* rows correspond to the coarsened exact matching estimator of ATT as implemented in the `cemR` package. The minimum and maximum DFBETA values actually correspond to votes by particular judges on particular cases (the units in the study). The listed judge names are the names of the judges from those influential judge-case combinations. Judge names were not reported in the non-contender dataset.

We begin by examing the diagnostics for Black and Owens' preferred CEM estimator. Although Black and Owens (2016, 36) report a sample size of 11,787 contender-judge observations,[31] note that the effective sample size for the contender analysis is 53.1 and the effective sample size ratio is 0.021—meaning that their CEM analysis discards approximately 98% of the possible data for the contenders. The CEM analysis of the non-contender judges also discards a fair amount of data but the effective sample size ratio is not as low as it is for contender judges.

In addition, the DFBETA values for the CEM analysis of the contender judges reveal that some judge-votes exert a very large influence on the estimated effect. The negative DFBETA with the largest magnitude belongs to a vote by Judge J. Harvie Wilkinson III and is equal to −0.038. Dropping this vote by Judge Wilkinson would have increased the estimate of ATT by approximately 50%.[32] The largest positive DFBETA is equal to 0.025 and is from a vote by Judge Emilio Garza. The inclusion of this vote by Judge Garza produces approximately one third of the total estimated effect.[33]

Taken on their own, these diagnostics applied to the CEM analysis provide cause for questioning whether the data support Black and Owens's hypotheses about contender judges. Even more red flags go up when we examine the diagnostics for the other methods of estimating ATT (all of which are based on the same causal identification assumptions of SUTVA and conditional ignorability of treatment assignment given the same covariates).

Looking at Table 1 we see that the other estimation methods all have much larger effective sample sizes, and smaller DFBETA values for contender judges than does CEM. The regression methods do result in some extrapolation for the contender judges, which would be a reason to prefer another estimation method.

We also investigate the covariate balance produced by all of these estimation methods. Figure 1 presents diagnostic plots for contender judges.[34] Specifically, Figure 1a plots the TASMD values and Figure 1b the weighted KS statistics, both for control units. Recall that for a particular estimation method and covariate, the TASMD value and for the controls is the standardized absolute difference between the weighted mean of the covariate within the control observations (with weights given by $w_h^{(0)}(x)$) and the mean of that covariate in the target distribution. Because the estimand in this example is the ATT, the covariate means under the target distribution are just the sample means of the covariates within the set of all treated units.

Looking at Figure 1a, we see that the non-CEM methods, with the exception of the URI estimator, all produce good to very good balance as evidenced by TASMD values below 0.2 and in most cases below 0.1. The MRI estimator of ATT, the propensity score weighting estimator of ATT, and the entropy balancing estimator of ATT produce especially good balance—although it is important to note that the MRI estimator achieves this balance with negative weights, i.e., via extrapolation.[35]

Similarly, the weighted KS test statistics compare the weighted distribution of control units to the target distribution, which is the treated group in this case. KS statistics indicate the maximum absolute difference between the empirical cumulative distribution functions (ECDFs) of the two distributions. As a normalized measure, the distance ranges from 0 (identical distributions) to 1 (completely different distributions). Figure 1b plots the KS test statistics for the covariates, comparing raw and weighted control units and treated groups. As we see, almost all methods reduce the KS statistics compared to the raw data, thereby improving distributional balance.

The comparison in Figure 1 clarifies how different methods handle data at hand and provides guidance for examining specific covariate of interest. For example, `Ideo. Distance` measures ideological distance between the judge and the president, is an important confounder given the

[31] Given the covariates specified in the original paper, there are 10,171 complete judge-case combinations for contenders, which is the data we analyzed.

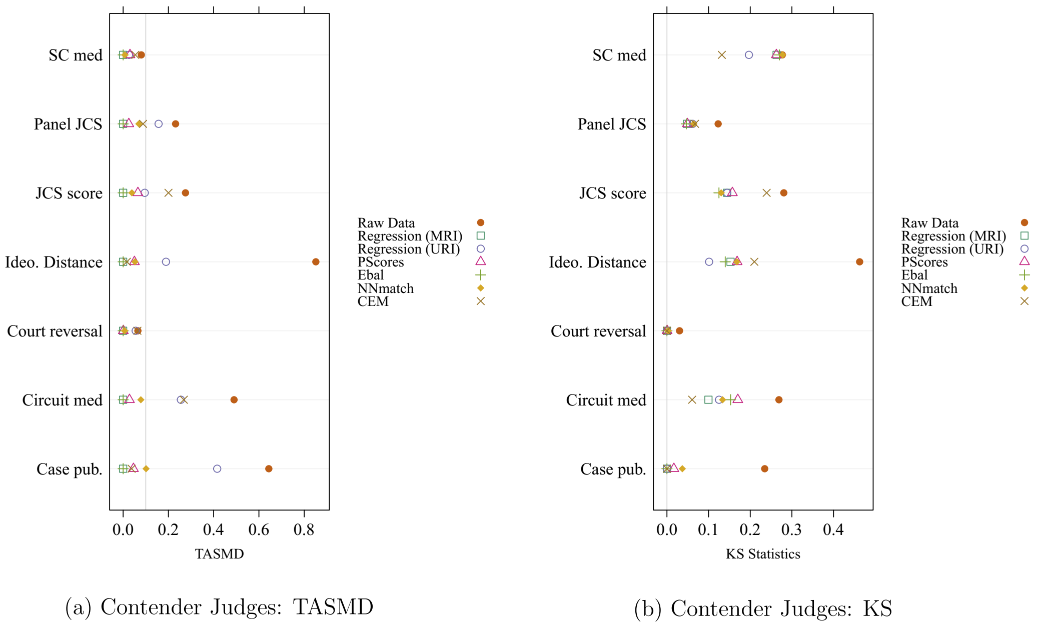
[32] $100 \times (0.075 + 0.038)/0.075 = 150.7$.

[33] $0.025/0.075 = 1/3$.

[34] Section F of the Supplementary Material presents the results for non-contender judges.

[35] Note also that when balance is assessed by TASMD where the target distribution is *all* treated units—not just the treated units that remain after the CEM coarsening procedure—the CEM weights produce fairly poor balance as judged by TASMD.

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press



(a) Contender Judges: TASMD

(b) Contender Judges: KS

Figure 1. Weighted mean covariate balance (assessed by TASMD) and KS test statistic for control units using multiple ATT estimation methods in the re-analysis of Black and Owens (2016).
Note: In the TASMD plot, each symbol shows the standardized difference between the weighted mean of the control and treated data for each covariate and method. The gray vertical line marks where the TASMD value is 0.1. In the KS statistic plot, each symbol shows the maximum absolute difference in the ECDFs of the control and treated groups, using both raw and weighted control data. The gray vertical line marks KS statistics at 0. `SC med` is the JCS score of the median Supreme Court justice; `Panel JCS` is the ideological distance between a judge and the remaining panelists; `JCS score` is each judge's JCS score; `Ideo. Distance` is the ideological distance between the judge and the president; `Court reversal` is whether the circuit court reversed the lower court; `Circuit med` is the JCS score of the median judge on the circuit; and `Case pub.` is whether the case was published.

research question. Figure 1 shows the unweighted control and the treated group exhibit poor mean and distributional balance. Among all methods, entropy balancing performs the best in achieving both mean and distributional balance.[36]

Looking at the various weighted QQ-plots for these covariates helps to clarify how balance is being improved. Given space limitations, we present a single weighted QQ-plot for the `Ideo. Distance` covariate among contender judges in Figure 2. Here we see that in the raw data, `Ideo. Distance` is quite different among treated and control units. Entropy-balancing weights help to shrink this difference. That said, differences on this covariate between treated and control units remain after applying the entropy-balancing weights. While one might favor the entropy balancing results over the other results, one might also wonder whether any of the estimation approaches provide sufficient control of confounding.

Given these diagnostics, it seems that the estimation approaches with the fewest obvious problems (as revealed by the diagnostics) are propensity score weighting and entropy balancing. If these approaches produced results that were consistent with Black and Owens' hypothesis of promotion-seeking behavior among judges, then we might conclude that the overall takeaway from Black and Owens (2016) is correct, even if the estimation method they preferred in that paper doesn't work well in that application. Going back to Table 1, we see that this is not the case. Neither the propensity score weighting nor the entropy balancing estimates show that contenders respond differently to a vacancy than do non-contenders. If anything, the point estimates from propensity score weighting and entropy balancing (as

[36] Entropy balancing and MRI achieve perfect mean balance. URI and Entropy balancing perform better than others in achieving distributional balance. However, both MRI and URI extrapolate and produce negative weights.

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

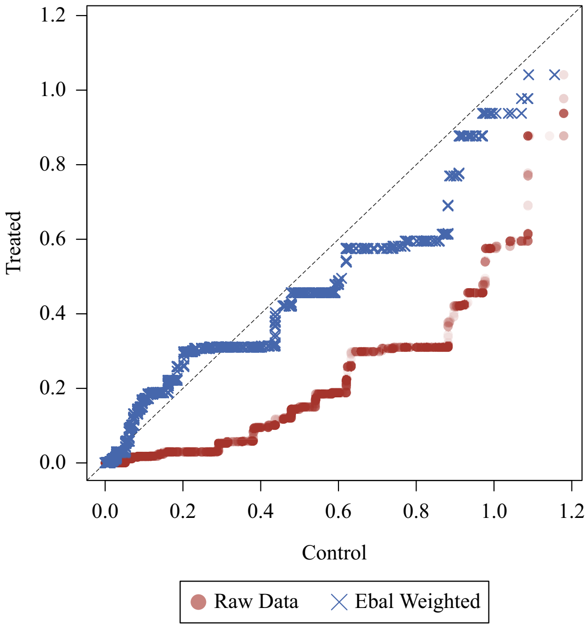


Figure 2. Quantile-quantile plot for `Ideo. Distance` with raw and entropy-balancing-weighted data for contender judges.
Note: The *X*-axis depicts weighted quantiles of ideological distance between the judge and the president for the control units, while the *Y*-axis shows depicts weighted quantiles of ideological distance for the treated group, which is the target distribution in this case. A point on the 45-degree line indicates equality of the quantiles represented by that point.

well as nearest-neighbor matching) are actually larger for the non-contenders than for the contenders (though none of the differences are statistically significant).

To conclude, based on Table 1 and Figure 1, when it comes to estimating ATT for contender judges, propensity score weighting and entropy balancing are better choices than the default application of CEM[37] for this application. At the least, these approaches use the data more efficiently, do not generate extreme influential data points, do not extrapolate, and produce good covariate balance. Unfortunately, though, neither propensity score weighting nor entropy balancing produces results that support the authors' hypotheses. For the contenders, the causal effect of a vacancy on voting is not statistically different from 0 based on the propensity score weighting estimator and substantively fairly trivial based on the entropy balancing estimate. As for the non-contenders, there is evidence that a vacancy exerts a small, significant effect on voting behavior—against the authors' expectation.

4.2. *Wealth-Maximizing Politicians*

In our second applied example, we turn to a study by Eggers and Hainmueller (2009). They ask whether members of the British parliament, relative to losing parliamentary candidates, accumulate greater wealth. Or, more succinctly: What is the causal effect of serving in parliament on wealth accumulation? The authors provide theoretical reasons to think that the effect of office on wealth accumulation differs for Conservative MPs compared to Labour MPs and so condition their analysis on party.

In contrast to Black and Owens (2016), Eggers and Hainmueller (2009) report estimates of three different causal estimands (ATE, ATT, and the local average effect from a regression discontinuity design); and they also make use of three different estimation methods (OLS regression, difference of

[37] CEM requires users to set a number of parameters, such as how covariates will be coarsened, and these choices affect performance. It is possible that other choices for these tuning parameters will perform better in this application.

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

means after 1:1 genetic matching with replacement, and local regression for the regression discontinuity estimates). Although small differences emerge across these various methods, the main substantive point is remarkably consistent: Holding office has a substantial, positive effect on the wealth of Conservative MPs, whereas "no discernible financial benefits [accrue] for Labour MPs" (Eggers and Hainmueller 2009, 513).

Considering the stability of Eggers and Hainmueller's results, it would seem that our proposed diagnostics would add little value to their work. And, in fact, our tools supply no cause to question Eggers and Hainmueller's general substantive conclusions. But they do clarify how various methods use the data at hand and which observations (MPs) exert substantial influence on the reported results. So even in this application, which is close to current "best practices," our diagnostics may still aid researchers and their readers by shedding new light on the analyses and aiding proper interpretation of the findings.

For purposes of illustrating the diagnostics discussed above, we focus, in Table 2, on the ATT and ATE for Conservative Party politicians (see also Section G of the Supplementary Material). Looking at this table we see that the two regression approaches (MRI and URI) tend to have relatively large effective sample sizes and the most influential observations under these methods exert no more influence on the results than the most influential observations under other methods. These regression methods do extrapolate to some extent which could be a reason to prefer other methods. That said, the amount of extrapolation is relatively small.

As judged by our diagnostics, propensity score weighting and entropy balancing exhibit no serious problems. They have relatively large effective sample sizes, only moderately influential observations, and (by construction) they do not extrapolate.

Genetic matching results in a slightly lower effective sample size than the methods above and, when estimating ATT among Conservative Party members, it gives a good deal of influence to one observation (William How). Dropping just this one politician reduces the genetic matching ATT estimate from 1.107 to 0.723 and the ATE estimate drops from 0.621 to 0.361. Put another way, William How is responsible for roughly 38% of the estimated effect size.

In contrast to the other approaches, use of CEM with default settings results in an effective sample size of just two politicians which corresponds to an effective sample size ratio of just 0.036. In other words, CEM is discarding approximately 96% of the available data. This alone suggests that this application of CEM to this dataset should be questioned.

Figure 3 presents diagnostic plots for covariate[38] balance in ATT estimation for Conservative MPs.[39] The TASMD results in Figure 3a suggests that MRI, propensity score, entropy balancing, and genetic matching achieve mean balance for most covariates below the threshold. In contrast, CEM and, to a lesser extent, URI weighted data fail to balance many covariates and sometimes perform worse than the unweighted data. The KS statistics results in Figure 3b show that almost all methods contribute to reducing distributional imbalance. The test statistics also have less dispersion, partly because many covariates are binary variables.

Figure 3 suggests that genetic matching, the preferred method in the original paper, produces the best covariate balance without extrapolation. Examining weighted QQ-plots provides additional information on covariate balance. Again, for reasons of space, we limit attention to how genetic matching adjusts the distribution for the two continuous variables, `Birth Year` and `Death Year` among Conservative politicians. Figure 4 compares the distributional balance for these two covariates between control units and treated units. For both the Birth year and Death Year, the raw data already display good distributional balance. Genetic matching weights brings control units even closer to the target distribution, further improving balance on these covariates.

While there are reasons to think that that the propensity score weighting and entropy balancing results might be the most credible, the particular choice of estimation method matters relatively little

[38] A detailed description of the covariates can be found in Eggers and Hainmueller (2009, 517–521).

[39] The ATE results are provided in Section G of the Supplementary Material.

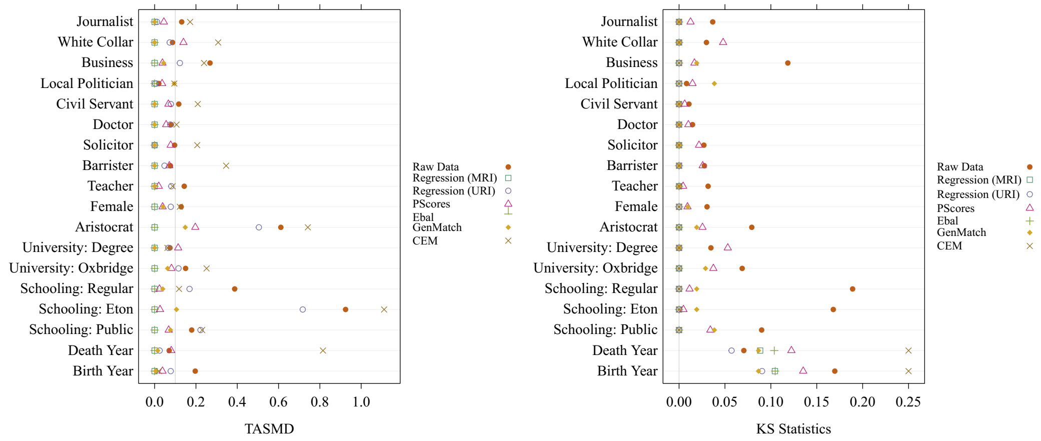
https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

Table 2. *Re-analysis of Eggers and Hainmueller (2009) with multiple estimation methods.*

	Estimate	SE	Effective	Effective	DFBETA minimum		DFBETA maximum		Extrapolation	
			sample	sample Size	Value	Politician	Value	Politician	Control	Treatment
			size	ratio					units	units
ATT (Conservative party)										
Regression (MRI)	0.888	0.355	30.6	0.553	−0.120	Malcolm St Clair	0.210	William How	0.028	0.000
Regression (URI)	0.541	0.201	45.2	0.817	−0.053	Richard Lamb	0.076	William How	0.001	0.006
Propensity score weighting	1.055	0.433	30.4	0.549	−0.121	Malcolm St Clair	0.240	William How		
Entropy balancing	1.094	0.209	28.6	0.518	−0.131	Malcolm St Clair	0.275	William How		
Genetic Matching	1.107	0.210	24.1	0.437	−0.103	Peter Boydell	0.384	William How		
Coarsened exact matching	−0.198	0.510	2.0	0.036	−0.266	Graham Partridge	0.255	William Shearer		
ATE (Conservative party)										
Regression (MRI)	0.629	0.234	42.3	0.765	−0.062	Malcolm St Clair	0.117	William How	0.000	0.002
Regression (URI)	0.541	0.201	45.2	0.817	−0.053	Richard Lamb	0.076	William How	0.001	0.006
Propensity score weighting	0.750	0.276	41.6	0.752	−0.061	Malcolm St Clair	0.133	William How		
Entropy balancing	0.709	0.178	41.8	0.756	−0.064	Malcolm St Clair	0.128	William How		
Genetic matching	0.621	0.135	33.6	0.607	−0.082	Anthony Courtney	0.260	William How		

Note: The *Regression (MRI)* rows correspond to the MRI estimator of either ATT or ATE. The *Regression (URI)* rows correspond to the URI estimator which is an estimate of both ATT and ATE under the constant effects assumption. The *Propensity Score Weighting* rows correspond to the propensity score weighting estimator of either ATT or ATE in which the propensity scores are estimated via logistic regression. The *Entropy Balancing* rows correspond to the entropy balancing estimator of either ATT or ATE as implemented in the `WeightItR` package. The *Genetic Matching* rows correspond to 1:1 matching on all X variables using the `GenMatch` function in the `MatchingR` package. The *Coarsened Exact Matching* row corresponds to the coarsened exact matching estimator of ATT as implemented in the `CEMR` package. Note that the genetic matching results are slightly different from what is reported in Eggers and Hainmueller (2009). This appears to be due to the stochastic nature of the matching procedure. These slight differences do not affect the conclusions reached in Eggers and Hainmueller (2009).

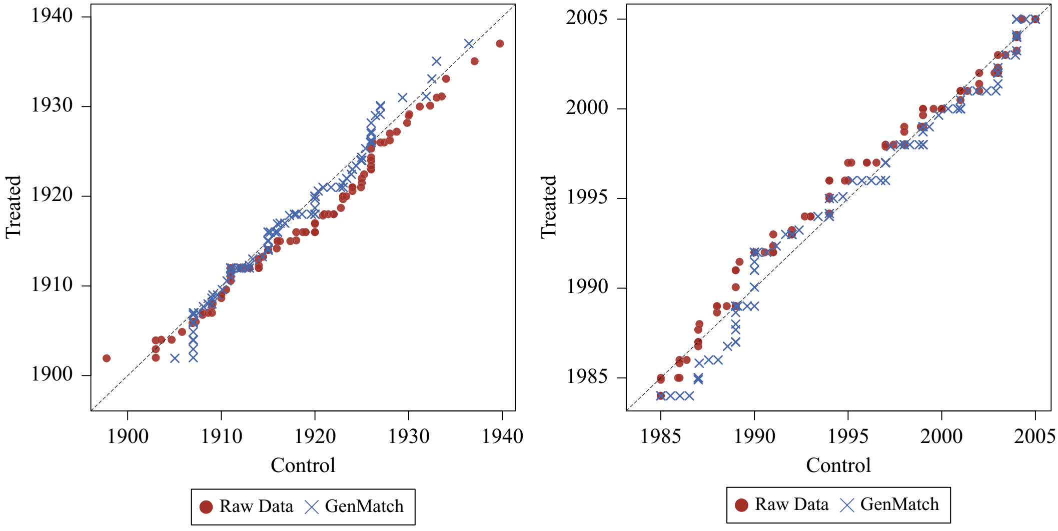


(a) Conservative MPs: TASMD

(b) Conservative MPs: KS Statistics

Figure 3. Weighted mean covariate balance (assessed by TASMD) and KS test statistics for control units using multiple ATT estimation methods in the re-analysis of Eggers and Hainmueller (2009).

Note: In the TASMD plot, each symbol represents the standardized difference between the weighted mean of the control data and the mean of the target data (i.e., the sample mean of that covariate among all treated units in the sample). The gray vertical line marks where the TASMD value is 0.1. In KS statistic plot, each symbol shows the maximum absolute difference in the ECDFs of the control and treated groups, using both raw and weighted control data. The gray vertical line marks KS statistics at 0.



(a) Covariate Distribution: Birth Year

(b) Covariate Distribution: Death Year

Figure 4. Quantile-quantile plot for `Birth Year` and `Death Year` with raw data and genetic matching weighted data for ATT estimation among conservative politicians.

Note: The *X*-axes show the weighted birth and death year quantiles for the control units, while the *Y*-axes shows the corresponding quantiles for the treated group, which is the target distribution in this case. Points on the 45-degree line correspond to equality of the quantiles represented by that point.

as all approaches (with the exception of CEM) produce results that are qualitatively similar and in line with the authors' hypothesis.

5. Discussion

If one believes that selection on observables and SUTVA hold for a particular application, then a very large number of estimation methods can, in principle, be used to consistently estimate the causal effect of interest. How should an applied researcher decide which method(s) to use and which result(s) to report?

In this paper, we have attempted to provide some guidance. First, we recommend that the applied researcher consider many estimation methods and all model specifications for which the selection on observables and SUTVA assumptions are plausible. Second, we encourage researchers to calculate, examine, and report the diagnostics discussed in this paper. Such open and honest reporting would do much to help readers understand how model-dependent the reported results are.

We reiterate that the diagnostic tools discussed in this paper require skill and judgment to be used appropriately. While they can provide clues about potential problems with a particular estimation method applied to a particular dataset, they do not provide guarantees about the accuracy of any given estimate. But our hope is that this paper will help researchers produce more credible results when making selection-on-observables assumptions.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2025.10003.

Data Availability Statement. Replication materials for this article are available at Quinn *et al.* (2025). A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/69RUAD.

References

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

Aronow, P., and B. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge: Cambridge University Press.

Austin, P. C., and E. A. Stuart. 2015. "Moving Towards Best Practice When Using Inverse Probability of Treatment Weightin (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies." *Statistics in Medicine* 34 (28): 3661–3679.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Black, R. C., and R. J. Owens. 2016. "Courting the President: How Circuit Court Judges Alter Their Behavior for Promotion to the Supreme Court." *American Journal of Political Science* 60 (1): 30–43.

Burgess, K., and M. D. Tyburski. 2020. "When Parties Go Abroad: Explaining Patterns of Extraterritorial Voting." *Electoral Studies* 66: 102169.

Chattopadhyay, A., C. H. Hase, and J. R. Zubizarreta. 2020. "Balancing vs. Modeling Approaches to Weighting in Practice." *Statistics in Medicine* 39 (24): 3227–3254.

Chattopadhyay, A., and J. R. Zubizarreta. 2023. "On the Implied Weights of Linear Regression for Causal Inference." *Biometrika* 110 (3): 615–629.

Cochran, W. G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24 (2): 295–313.

Eggers, A. C., and J. Hainmueller. 2009. "MPs for Sale? Returns to Office in Postwar British Politics." *American Political Science Review* 103 (4): 513–533.

Grier, K., R. Grier, and H. J. Moncrieff. 2024. "Uncertain Times: The Causal Effects of Coups on National Income." *American Journal of Political Science* 69: 560–577.

Hainmueller, J. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46.

Hainmueller, J., and D. Hangartner. 2013. "Who Gets a Swiss Passport? A Natural Experiment in Immigrant Discrimination." *American Political Science Review* 107 (1): 159–187.

Hájek, J. 1971. "Comment on 'An Essay on the Logical Foundations of Survey Sampling, Part One'." In *The Foundations of Survey Sampling*, edited by V. Godambe, and D. Sprott. Toronto: Holt, Rinehart, and Winston.

https://doi.org/10.1017/pan.2025.10003 Published online by Cambridge University Press

Hazlett, C. 2020. "Kernel Balancing: A Flexible Non-Parametric Weighting Procedure for Estimating Causal Effects." *Statistica Sinica* 30: 1155–1189.

Hazlett, C., and T. Shinkre. 2024. "Understanding and Avoiding the"Weights of Regression": Heterogeneous Effects, Misspecification, and Longstanding Solutions." https://doi.org/10.48550/arXiv.2403.03299.

Hernán, M. A., and J. M. Robins. 2024. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

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

Iacus, S. M., G. King, and G. Porro. 2011. "Multivariate Matching Methods That are Monotone Imbalance Bounding." *Journal of the American Statistical Association* 106 (493): 345–361.

Iacus, S. M., G. King, and G. Porro. 2012. "Causal Inference Without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20 (1): 1–24.

Iacus, S. M., G. King, and G. Porro. 2019. "A Theory of Statistical Inference for Matching Methods in Causal Research." *Political Analysis* 27 (1): 46–68.

Ichino, N., and N. L. Nathan. 2013. "Do Primaries Improve Electoral Performance? Clientelism and Intra-Party Conflict in Ghana." *American Journal of Political Science* 57 (2): 428–441.

Imai, K., G. King, and E. A. Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502.

Imbens, G. W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics* 86: 4–29.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

King, G., and L. Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14 (2): 131–159.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–4165.

Lee, B., J. Lessler, and E. Stuart. 2010. "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine* 29 (3): 337–346.

Li, F., K. L. Morgan, and A. M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400.

Li, J., and R. Valliant. 2011. "Linear Regression Influence Diagnostics for Unclustered Survey Data." *Journal of Official Statistics* 27 (1): 99–119.

Morgan, S. L., and J. J. Todd. 2008. "A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects." *Sociological Methodology* 38 (1): 231–281.

Quinn, K. M., G. Liu, L. Epstein, and A. D. Martin. 2025. "Replication Data for "What to Observe When Assuming Selection on Observables" V1." https://doi.org/10.7910/DVN/69RUAD.

Rubin, D. B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2 (3): 808–840.

Setoguchi, S., S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook. 2008. "Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study." *Pharmacoepidemiology and Drug Safety* 17 (6): 546–555.

Truex, R. 2014. "The returns to Office in a "Rubber Stamp" Parliament." *American Political Science Review* 108 (2): 235–251.

Wang, Y., and J. R. Zubizarreta. 2020. "Minimal Dispersion Approximately Balancing Weights: Asymptotic Properties and Practical Considerations." *Biometrika* 107 (1): 93–105.

Zhao, Q., and D. Percival. 2016. "Entropy Balancing is Doubly Robust." *Journal of Causal Inference* 5 (1): 20160010. arXiv:1501.03571.

Zubizarreta, J. R. 2015. "Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data." *Journal of the American Statistical Association* 110 (511): 910–922.

Cite this article: Quinn, K. M., Liu, G., Epstein, L. and Martin, A. D., (2025). What to Observe When Assuming Selection on Observables. *Political Analysis*, 1–22. https://doi.org/10.1017/pan.2025.10003