Harvard Data Science Review • Issue 6.1, Winter 2024

Causation, Comparison, and Regression

Ambarish Chattopadhyay[1] José R. Zubizarreta[2,3,4]

[1]Stanford Data Science, Stanford University, Stanford, California, United States of America,

[2]Department of Health Care Policy, Harvard Medical School, Harvard University, Cambridge, Massachusetts, United States of America,

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, Massachusetts, United States of America,

[4]Department of Statistics, Harvard Medical School, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Jan 31, 2024

DOI: https://doi.org/10.1162/99608f92.87c6125f

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

ABSTRACT

Comparison and contrast are the basic means to unveil causation and learn which treatments work. To build good comparison groups, randomized experimentation is key, yet often infeasible. In such non-experimental settings, we illustrate and discuss diagnostics to assess how well the common linear regression approach to causal inference approximates desirable features of randomized experiments, such as covariate balance, study representativeness, interpolated estimation, and unweighted analyses. We also discuss alternative regression modeling, weighting, and matching approaches and argue they should be given strong consideration in empirical work.

Keywords: causal inference, randomized experiments, observational studies, regression modeling, weighting adjustments, multivariate matching

Media Summary

To go from correlation to causation, and to figure out which treatments are effective, we need to make good comparisons—ideally randomized ones. However, forming such random groups is not always possible due to ethical or practical constraints. In such situations, we assess how well the standard regression approach to causal inference approximates essential features of randomized experiments. We also explore and discuss other regression, weighting, and matching approaches, advocating that they should be given strong consideration in empirical work.

Visit the web version of this article to view interactive content.

Article Video Summary

1. Unweighted Comparisons Work Well in Randomized Experiments

1.1. Learning Which Treatments Work

Comparison and contrast are the basic means to learn the effects of causes—which treatments work—in a complex world of which we only know so much. Especially in fields like medicine and the social sciences, where the distinct heterogeneity of each individual makes them unique, making these comparisons is crucial and requires meticulous attention. But to render these contrasts useful and ‘see’ the effect of treatment, it is crucial that subjects exposed and not exposed to the treatment are similar. This is essential to isolate the effect from observable and unobservable confounding factors.

1.2. Randomized Comparisons

The randomized experiment stands out as the ideal method for building such comparable groups. Randomization tends to produce groups with similar pretreatment observed and unobserved characteristics that differ only on the receipt of treatment, and possibly, if the treatment is effective, on their subsequent outcomes. Randomization also provides a factual basis to quantify whether differences in the outcomes are due to system or chance and to conduct statistical tests for the effect of treatment (Fisher, 1935). To illustrate, consider the following simulated example inspired by the famous RAND Health Insurance Experiment (Newhouse, 1993).[1]

1.3. Running Example

The goal of the investigation is to learn the causal effect of a new health insurance plan (the *treatment* variable) on health expenditures (the *outcome* variable). The purpose of this new plan is to protect patients against high out-of-pocket medical costs while reducing health spending, so we anticipate that this plan will lead to a decrease in average health expenditure. The basic parameter we wish to learn is the average reduction in health expenditure caused by the insurance plan among those who enroll; that is, the average effect of treatment on the treated subjects, henceforth denoted ATT. We focus on the ATT as opposed to the average effect of treatment on all subjects (ATE) for illustrative purposes, but our analyses can be readily extended to encompass the ATE.

To this end, suppose that a randomized experiment is conducted on a random sample of 200 individuals from a population. Half of these individuals are randomly assigned to enroll in the plan—the treatment group—and the other 100 are similarly assigned to a control group. For each individual, there is information on two pretreatment characteristics, or covariates; namely, household per capita income and number of hospital visits in the previous year. Table 1 presents a sample of data from this experiment for 10 typical individuals in each group.

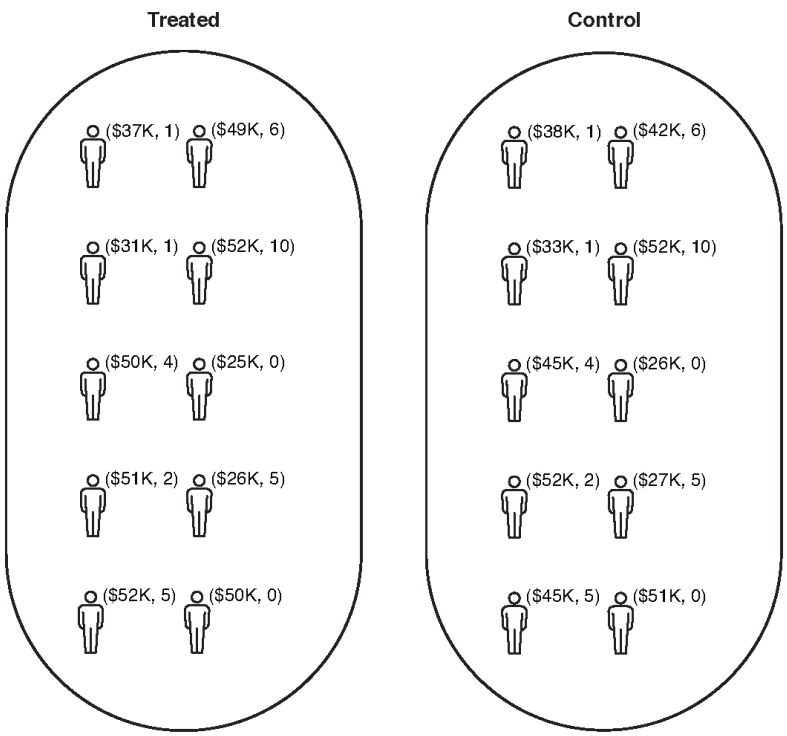
Table 1. Twenty typical individuals in the randomized experiment. The baseline covariates are Income and Visits. The treatment variable is Insurance (1 if insured, 0 otherwise). The outcome variable is Expenditure.

Income	Visits	Insurance	Expenditure
37451	1	1	6,250
48,509	6	1	8,461
31,202	1	1	5,234
52,205	10	1	9,356

49,854	4	1	8,522
24,977	0	1	4,134
51,224	2	1	8,613
26,432	5	1	4,734
52,012	5	1	8,940
50,009	0	1	8,254
37,749	1	0	7,658
41,520	6	0	8,903
32,736	5	0	7,053
52,025	3	0	10,696
45,111	4	0	9,427
25,673	0	0	5,144
52,462	2	0	10,700
27,404	5	0	5,964
45,305	5	0	9,560
50,938	0	0	10,182

Figure 1 portrays the randomized experiment for these 20 individuals. In the figure, the covariate values for each individual are shown over their shoulder. For example, the treated individual in the top-left corner has a household per capita income of $37K and one hospital visit in the previous year. The *diagnostic dashboard* on the rightmost panel tabulates first, under 'covariate balance,' the means of these two covariates in the sample of 100 treated subjects, 100 control subjects, and their absolute standardized mean difference (Rosenbaum & Rubin, 1985).[2] This diagnostic shows that, save for chance imbalances, the groups bear close resemblance to each other in terms of the covariate means. Second, under 'study representativeness,' it shows the corresponding means in the target (here, treated) population as well as the absolute standardized mean differences between each group and the target. Third, under 'sample size,' it presents the effective sample size

(i.e., the number of subjects effectively used for estimation) and the nominal sample size (i.e., the original number of subjects) in each group.[3] Finally, under 'differential weighting,' it displays the minimum, maximum, and quartiles (25th, 50th, and 75th percentiles) of the distribution of the weights (in this case, constant). This diagnostic quantifies the deviation from the ideal scenario of uniform (i.e., constant) weights, which aid interpretability and minimize the variance of general weighting estimators.



Diagnostic Dashboard

Covariate Balance

	Income	Visits
Treated	$38,934	4.2
Control	$39,325	4.3
Difference[1] Treated-Control	0.04	0.03

Study Representativeness

	Income	Visits
Target	$38,934	4.2
Difference[2] Treated-Target	0.00	0.00
Difference[2] Control-Target	0.04	0.03

Sample Size

	Effective	Nominal
Treated	100	100
Control	100	100

Differential Weighting[3]

	Min	Q1	Med	Q3	Max
Treated	1	1	1	1	1
Control	1	1	1	1	1

[1]: Absolute standardized mean difference. [2]: Target absolute standardized mean difference. [3]: The weights are scaled to sum up the size of each treatment group. Min: minimum. Q1, Q3: first and third quartiles. Med: median. Max: maximum.

Figure 1. Sketch of a randomized experiment. Covariate values are displayed over the shoulder of each individual. The diagnostic dashboard shows that randomization has produced balance in the covariates between the groups and in relation to the target population without weighting. As a result, the effective sample size equals the nominal sample size, and the weights are constant and equal to one. Here, the target population is the treated population because the estimand is the average treatment effect among the treated.

1.4. Weighted Contrasts

After randomization, because the groups tend to have similar observed and unobserved characteristics, we can estimate the average effect of treatment by simply taking the mean difference in outcome values y between the

treated and the control groups,

$$\overbrace{\frac{y_{t,1} + y_{t,2} + \ldots + y_{t,10}}{10}}^{\text{Treated group}} - \overbrace{\frac{y_{c,1} + y_{c,2} + \ldots + y_{c,10}}{10}}^{\text{Control group}}.$$

In this expression, each observation contributes equally to the total effect estimate, so we can say that the treated and control groups are unweighted or uniformly weighted (because each observation has a constant weight equal to one) and equivalently write

$$\overbrace{\frac{1 \cdot y_{t,1} + 1 \cdot y_{t,2} + \ldots + 1 \cdot y_{t,10}}{10}}^{\text{Treated group}} - \overbrace{\frac{1 \cdot y_{c,1} + 1 \cdot y_{c,2} + \ldots + 1 \cdot y_{c,10}}{10}}^{\text{Control group}}.$$

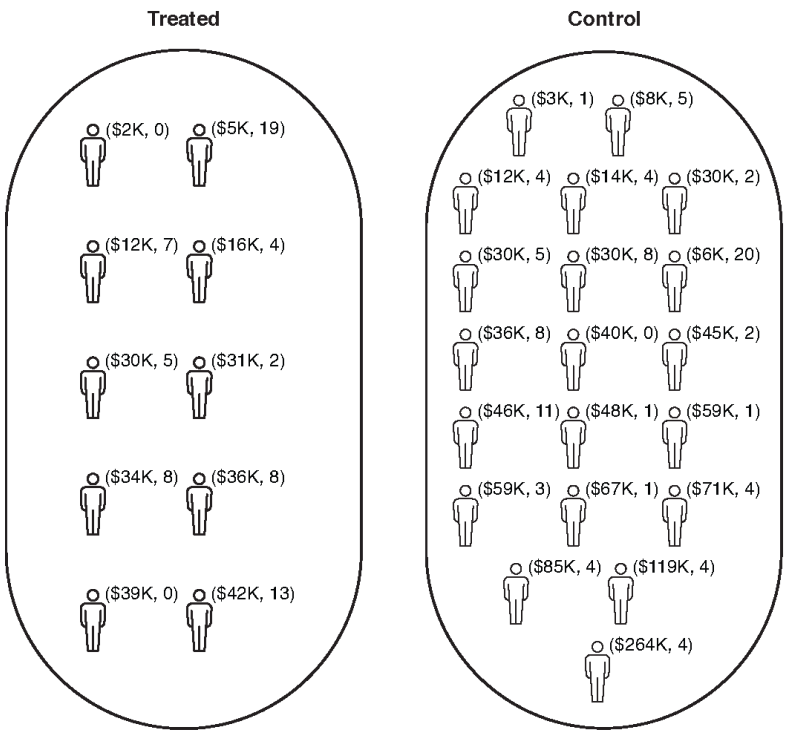
For our simulated example, the average difference in the outcome (health expenditures) between the treatment and control group is roughly −$1.5K, suggesting the insurance plan reduces health expenditure by that amount.

In randomized experiments, without weights the treated and control groups are balanced in expectation, so in principle, we can estimate the average effect of treatment by contrasting the simple averages in the treated and control groups. This observation is important as we move forward to more complex settings and methods.

1.5. Observational Studies

The power of randomized experiments stems from its treatment assignment mechanism—which is controlled by the investigator—and by its resulting data structure, which is transparent and simple: by design, we can clearly see how randomization works, whether the randomly formed groups are comparable or not, and we can talk about it, fostering collaboration and the elaboration of scientific theories. However, due to ethical or practical constraints, we cannot always carry out experiments. In these situations, we must fall back on observational studies, where by means of a process unbeknownst to the investigator, subjects are selected into treatment or control, typically in an unbalanced way.

Following our running example, we now turn our attention to a parallel observational study. This is a simulated study of 300 subjects randomly drawn from a population distinct from the previous randomized experiment. From this population, a process unknown to the investigator has led some subjects to enroll in the new health insurance plan, while others have not. As before, the goal is to learn the average treatment effect on the treated or ATT. In the sample, there are 100 treated subjects and 200 controls. Figure 2 summarizes this for 10 and 20 typical subjects from the treated and control groups, respectively. The average difference in the outcome variable, health expenditures, between the treatment and control group is roughly −$5K, suggesting that the insurance plan reduces the health expenditure by that amount. But how trustworthy is this conclusion?



Diagnostic Dashboard

Covariate Balance	Income	Visits
Treated	$27,216	4.6
Control	$44,388	3.8
Difference.[1] Treated-Control	0.83	0.16

Study Representativeness	Income	Visits
Target	$27,216	4.6
Difference.[2] Treated-Target	0.00	0.00
Difference.[2] Control-Target	1.56	0.17

Sample Size	Effective	Nominal
Treated	100	100
Control	200	200

Differential Weighting[3]	Min	Q1	Med	Q3	Max
Treated	1	1	1	1	1
Control	1	1	1	1	1

[1]: Absolute standardized mean difference. [2]: Target absolute standardized mean difference. [3]: The weights are scaled to sum up the size of each treatment group. Min: minimum. Q1, Q3: first and third quartiles. Med: median. Max: maximum.

Figure 2. Sketch of an observational study before adjustments. Covariate values are displayed alongside each individual's shoulder. The diagnostic dashboard shows covariate imbalances among groups and in comparison to the target population. At this stage, no differential weighting is applied, and the effective sample size equals to the nominal sample size.

As seen in the diagnostic dashboard of Figure 2, the treatment groups differ in their observed characteristics, and likely in the unobserved ones. For instance, the average income in the treated group is $27.2K, which is markedly different from the $44.4K average income among the controls, likely implying that the observed differences in health expenditure are driven by this pretreatment imbalance. Therefore, to estimate the average treatment effect on the treated, some form of adjustments must be made to make the groups comparable, at least in terms of the observed characteristics. To this end, the randomized experiment is the conceptual benchmark that guides most methods of adjustment in an observational study (Cochran, 1965; Dorn, 1953; Hernán & Robins, 2020; Imbens & Rubin, 2015; Rosenbaum, 2010).

2. Regression Models Implicitly Use Unequal Weights for Estimating Treatment Effects

2.1. Implied Adjustments

In observational studies we seek to approximate or emulate as closely as possible randomized experiments in order to estimate the effect of treatments (Rubin, 2007). However, observational studies often rely on regression models to adjust for covariate imbalance and estimate such effects. In these analyses, how well does regression mimic key features of a randomized experiment? Specifically, how does linear regression recover (i) covariate balance, (ii) study representativeness, (iii) interpolated estimation, and (iv) unweighted samples that facilitate transparent analyses?

2.2. Approximating Experiments

In this question, (i) covariate balance refers to making the observed characteristics between the treatment and control groups similar. The influence of likely imbalances in unobserved characteristics must be assessed with a sensitivity analysis (e.g., Cornfield et al., 1959; Rosenbaum, 1987a; Rosenbaum, 2005; VanderWeele & Ding, 2017; Zhao et al., 2019). (ii) Study representativeness is the requirement that the characteristics of each group be similar to (and hence, representative of) a target population. (iii) Interpolated estimation or sample-boundedness means estimating effects that are an interpolation of the observed data rather than an extrapolation beyond the range or support of the actual data (Robins et al., 2007). (iv) Finally, an unweighted or self-weighted sample refers to a simple data structure where the observations have homogeneous weights and do not require further adjustments that can decrease the efficiency and transparency of the study. A recent paper provides answers to this question by deriving and analyzing the implied weights of linear regression (Chattopadhyay & Zubizarreta, 2023).

3. Regression Adjustments in Nonexperimental Comparisons

It is common to fit regression models to nonexperimental data, including linear regression, logistic regression, and proportional hazards regression. Consider the following linear regression model where the outcome of interest y is regressed on observed covariates or characteristics (e.g., age, gender, and risk factors) and the treatment assignment indicator,[4]

$$\text{Outcome variable} = \alpha + \beta \cdot \text{Observed covariates} + \tau \cdot \text{Treatment indicator} + \epsilon.$$

In the model, ε is a random error term with mean zero, distributed mean-independently of the covariates and treatment, and α, β, and τ are the unknown regression coefficients. Under the assumption that the observed covariates encompass all relevant confounders, the objective again is to estimate the parameter τ for the ATT. This parameter is typically estimated by ordinary least squares (OLS) from a random sample of observations from a large (possibly infinite) population, where it is postulated that uncertainty stems from the randomness of the sampling process.

Arguably, this is the traditional regression approach to causal inference. By standard regression theory, if the assumed model is correct, then the traditional OLS estimator is the most efficient, linear unbiased estimator for the ATT. But where is the experiment? Does the regression approach approximate the above features of randomized experiments? And what happens if the model is used when it is incorrect?

4. A Regression Adjustment Is Implicitly a Weighted Comparison

To answer these questions, we must first ask a crucial intermediate question. How does the above regression approach implicitly weight the individual observations in the data at hand? Building on important work by Imbens (2015) (see also Abadie et al., 2015, Gelman and Imbens, 2018), Chattopadhyay & Zubizarreta (2023) show that the traditional regression estimator of τ is equivalent to the following weighted contrast

$$\overbrace{\frac{w_{t,1} \cdot y_{t,1} + w_{t,2} \cdot y_{t,2} + \ldots + w_{t,10} \cdot y_{t,10}}{10}}^{\text{Treated group}} - \overbrace{\frac{w_{c,1} \cdot y_{c,1} + w_{c,2} \cdot y_{c,2} + \ldots + w_{c,10} \cdot y_{c,10}}{10}}^{\text{Control group}} \quad (4.1)$$

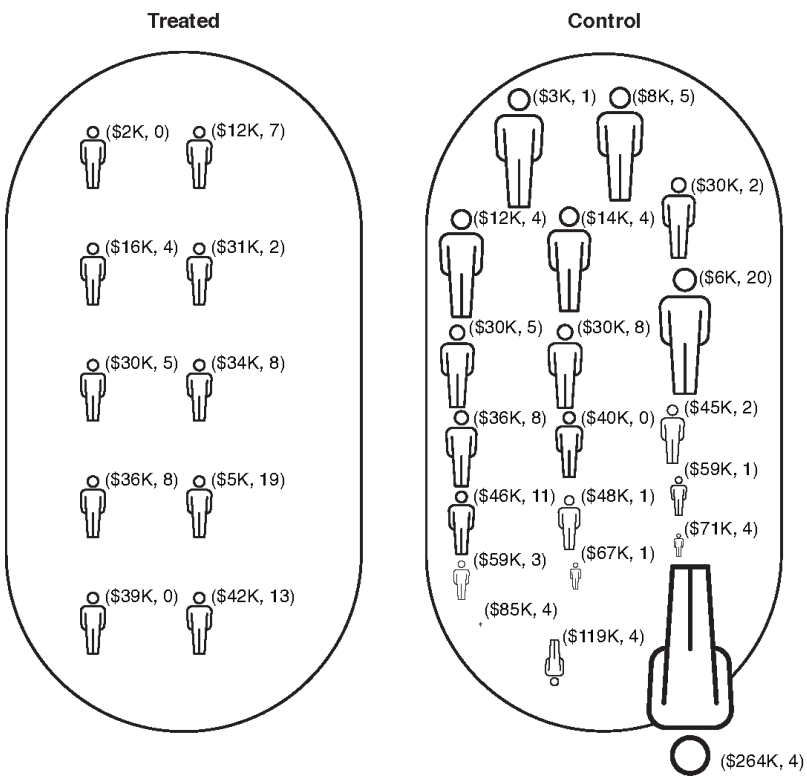
where the weights w in each group sum to the corresponding group size and can be expressed in closed form, that is without approximations.[5] If, in the control group, $w_{c,1} = 2$, then the first control counts as if she were two controls, correcting for the fact that people like this control are underrepresented in the control group and overrepresented in the treated group because of the absence of randomization. In other words, linear regression implicitly creates and contrasts weighted samples of treated and control observations, where the contribution or weight of each observation to the aggregate average treatment effect estimate can be computed and scrutinized. By analyzing these weights, we can transparently assess how linear regression attempts to mimic features of randomized experiments, as we discuss next. In Figure 3, we illustrate the structure of the weighted groups after regression in our running example.[6] The implied regression weights are computed using the `R` package `lmw`; see Chattopadhyay et al. (2023) for details. Here, the size of individuals is proportional to their corresponding weights, with negative weights turning them upside down.

5. The Implied Regression Weights Create a Limited Form of Covariate Balance

5.1. Exact Covariate Balance

Does linear regression balance covariates between the treated and control groups? As it turns out, there is a specific form of covariate balance concealed within the linear regression model. The implied weights of linear regression exactly balance the means of the covariates included in the model, such that the means of those observed characteristics are identical between the weighted treated and weighted control groups. See the diagnostic dashboard in Figure 3: across the two groups, the mean pretreatment income and hospital visits are the same after regression adjustments. Furthermore, if one includes transformations of the covariates in the regression model, such as squares or interactions between them, then the means of these transformations will

also be exactly balanced. However, there is no guarantee of balance for covariate transformations that are not included in the model.[7]



Diagnostic Dashboard

Covariate Balance	Income	Visits
Treated	$28,533	4.5
Control	$28,533	4.5
Difference[1] Treated-Control	0.00	0.00

Study Representativeness	Income	Visits
Target	$27,216	4.6
Difference[2] Treated-Target	0.12	0.03
Difference[2] Control-Target	0.12	0.03

Sample Size	Effective	Nominal
Treated	99	100
Control	164	200

Differential Weighting[3]	Min	Q1	Med	Q3	Max
Treated	0.7	0.9	1.0	1.1	1.2
Control	-3.8	0.8	1.1	1.4	1.9

[1]: Absolute standardized mean difference. [2]: Target absolute standardized mean difference. [3]: The weights are scaled to sum up the size of each treatment group. Min: minimum. Q1, Q3: first and third quartiles. Med: median. Max: maximum.

Figure 3. Sketch of an observational study after regression adjustments. Covariate values are displayed alongside the shoulder of each individual. The size of each person is scaled according to their contribution or weight in the calculation of the average treatment effect. Individuals with a negative weight are depicted upside down.

5.2. A Hidden Population

Does linear regression create treated and control samples that are representative of the target population? While the above linear regression model exactly balances the means of the covariates in the treated and control samples, it may not balance them at the target population. See the diagnostic dashboard in Figure 3. After regression, the average income in both groups is $28.5K, whereas the average income in the target (i.e., treated) population is $27.2K. Similarly, after regression, the average number of hospital visits in both groups is 4.5, which differs from 4.6, the average in the target.

In general, linear regression may not balance the samples relative to the natural populations characterized by the overall subjects, the treated subjects, or the control subjects. Implicitly, it balances them elsewhere, at a different covariate profile (see Chattopadhyay & Zubizarreta, 2023, for an exact formula). In this example, the profile is ($28.5K, 4.5), which does not correspond to any of these natural populations. This profile may in fact characterize a population that does not exist. In other words, here linear regression ends up estimating the average treatment effect on a hypothetical population of individuals whose average income is $28.5K and who has 4.6 hospital visits on average.

Taking a step back, the above linear regression model assumes that the treatment effect is constant everywhere. If this assumption is correct, then it does not really matter which population regression ends up targeting: the average treatment effect is the same as that in any other population. But if this assumption is incorrect, then effect estimates will be biased, especially if the implied weights take large negative values, as we explain next.

5.3. Estimation Beyond the Support of the Data and Negative Weights

Does linear regression produce interpolated or sample-bounded estimates (Robins et al., 2007) of the treatment effects? To answer this question, we first note that the implied weights of the regression model can take negative values. In Figure 3, subjects with negative weights are upside down. The most notable of them is a control subject with an income of $264K and 4 hospital visits, who is assigned a very large negative weight.

From a practical standpoint, negative weights are hard to interpret: while a person with a weight value of 1/2 in the sample means that it has half its weight in the population, and a weight of zero removes the person from the population, what is the meaning of a person with a weight value of -1? From an analytic perspective, these negative weights translate into forming effect estimates that may lie outside the support or range of the observed data, and thus that are not sample-bounded. In other words, one can get estimates that are very distant from any of the measurements in the actual data. As a result, one may end up obtaining rather silly estimates of the treatment effect. In our example, the effect estimate under regression is $147, suggesting that health insurance increases health expenditure. If the assumed regression model is correct, then such extrapolation is not detrimental in large samples, because the regression estimator is consistent for the ATT. However, if the model is incorrect, then this problem may arise even in large samples.

5.4. Differential Weighting of Minimum Variance

Does linear regression produce an evenly weighted or self-weighted sample? In general, unless the covariate means are exactly balanced between the treated and control groups before adjustment, regression ensures such mean balance by enforcing differential weighting of individuals. In our example, differential weighting is evident in the control group and is also present in the treatment group. As shown in the diagnostic dashboard in Figure 3, this differential weighting results in effective sample sizes that are smaller than the nominal sample sizes.

Further, an additional fundamental property is that the implied regression weights have minimum variance. Thus, although the sample individuals are differentially weighted, the weights are 'as stable as possible' subject to their mean balancing property.[8] Hence, the OLS estimator of τ is the best (minimum variance) linear unbiased estimator if the assumed regression model is correct. The issue, again, is that the true model that generates the data is unknown. If it were known, we would not need the data. If the assumed model is incorrect, then this stability property of the regression weights offers limited value, because in that case, regression may end up balancing the wrong functions, targeting the wrong population, and extrapolating beyond the range of the data.

6. Alternative Regression Modeling, Weighting, and Matching Approaches

6.1. Multi Regression

An alternative regression approach extends the previous model by including interactions between the observed covariates and the treatment assignment indicator as follows,[9]

$$\begin{aligned}\text{Outcome variable} = \alpha + \beta \cdot \text{Observed covariates} + \tau \cdot \text{Treatment indicator} \\ + \gamma \cdot \text{Observed covariates} \cdot \text{Treatment indicator} + \epsilon.\end{aligned}$$

This approach can be used to estimate the average treatment effect in the overall population, which corresponds to the parameter τ. See Peters (1941) and Belson (1956) for early discussions of this method. By including the treatment-covariates interactions, this regression model specification now allows for effect modification or heterogeneity. As discussed by Chattopadhyay and Zubizarreta (2023), this approach is equivalent to fitting separate linear regression models of the outcome on the covariates in the treated and control groups, and then taking the mean difference in the predicted outcomes under treatment and control, based on the two fitted models.

For estimating the ATT the multi regression approach proceeds as follows. First fit the regression model

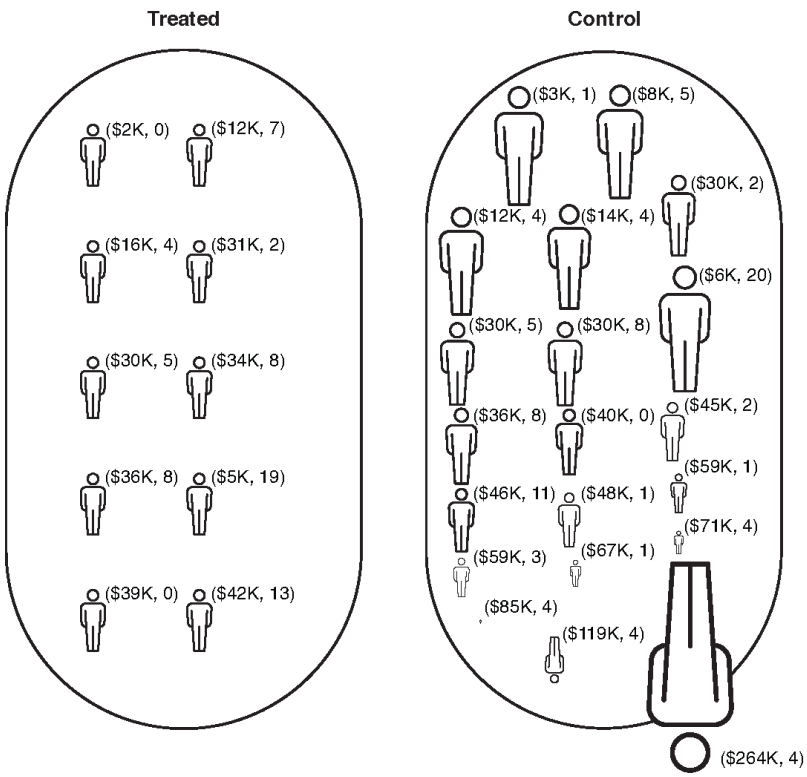
$$\text{Outcome variable} = \alpha + \beta \cdot \text{Observed covariates} + \epsilon$$

in the *control* group. Then, for each unit in the treatment group, predict its outcome in the absence of treatment, by using the previously fitted regression model. Finally, over all the treated units, compute the mean of their observed outcome minus the mean of the predicted outcomes.

To what extent does this approach approximate an experiment? Chattopadhyay and Zubizarreta (2023) show that, under this model specification, the estimator of the ATT can also be written as a weighted contrast,

$$\overbrace{\frac{1 \cdot y_{t,1} + 1 \cdot y_{t,2} + \ldots + 1 \cdot y_{t,10}}{10}}^{\text{Treated group}} - \overbrace{\frac{w_{c,1} \cdot y_{c,1} + w_{c,2} \cdot y_{c,2} + \ldots + w_{c,10} \cdot y_{c,10}}{10}}^{\text{Control group}}$$

with implied weights w that satisfy the following properties. (i) Like the previous approach, the means of the observed covariates are exactly balanced between the treatment and control groups. (ii) The covariate means of each group are centered at the treated population (the target population for the ATT), addressing the representativeness problem of the previous approach. (iii) An issue that persists is that the weights can take negative values, leaving room for biases due to extrapolation from an incorrectly specified model. (iv) As before, regression induces differential weighting on the control subjects such that the weights have the minimum variance among those that exactly balance the means of the covariates at the target population. See Figure 4 for an illustration of these properties.



Diagnostic Dashboard

Covariate Balance	Income	Visits
Treated	$27,216	4.6
Control	$27,216	4.6
Difference[1] Treated-Control	0.00	0.00

Study Representativeness	Income	Visits
Target	$27,216	4.6
Difference[2] Treated-Target	0.00	0.00
Difference[2] Control-Target	0.00	0.00

Sample Size	Effective	Nominal
Treated	100	100
Control	160	200

Differential Weighting[3]	Min	Q1	Med	Q3	Max
Treated	1.0	1.0	1.0	1.0	1.0
Control	-4.1	0.7	1.0	1.4	2.1

[1]: Absolute standardized mean difference. [2]: Target absolute standardized mean difference. [3]: The weights are scaled to sum up the size of each treatment group. Min: minimum. Q1, Q3: first and third quartiles. Med: median. Max: maximum.

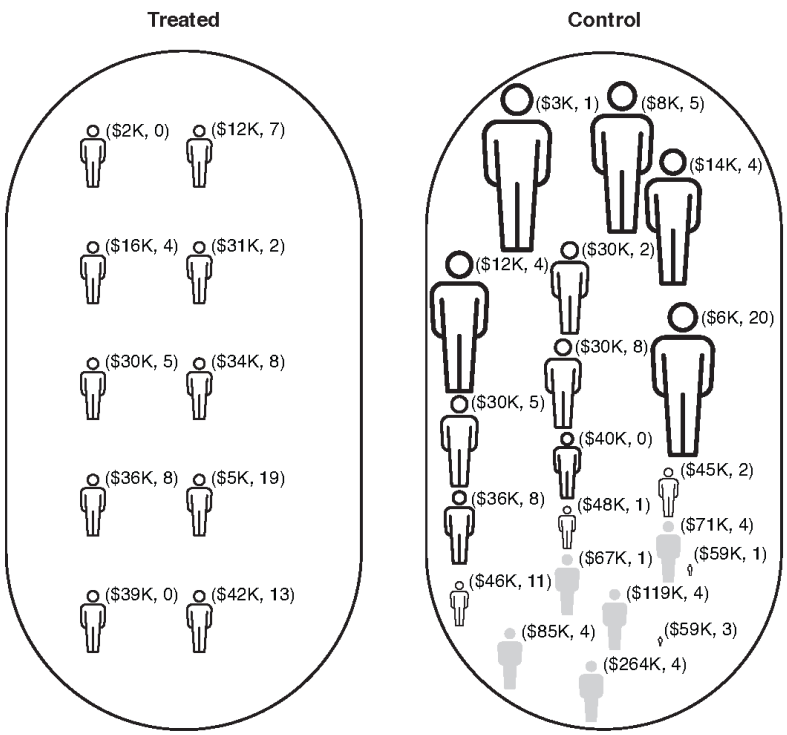
Figure 4. Sketch of an observational study after multi or interacted regression adjustments. The covariate values for each individual are shown over their shoulder. The size of each person is scaled according to their contribution or weight in the calculation of the average treatment effect. Individuals with a negative weight are depicted upside down.

Properties (i) and (ii) are manifest because the covariate profiles of both groups are ($27.2K, 4.6), the same as that of the target. Also, the differential weighting in property (iv) is evident in the control group, while by construction, subjects are weighted uniformly in the treatment group. Finally, similar to the previous regression approach, property (iii) is present in the control group, with the same control subject having an income of $264K and four hospital visits getting a very large negative weight. The effect estimate turns out to be $309 which, like the previous approach, contradicts established substantive knowledge.

Through the lens of the above properties, this multi regression approach comes closer to approximating the ideal randomized experiment than the previous approach. Thus, unless there is a high degree of certainty that treatment effects are homogeneous, the multi regression approach may be preferred. However, a potential

drawback that persists is the bias that may arise from extrapolating a wrong model, coupled with a loss in interpretative simplicity due to departures from an unweighted sample.

Arguably, the performance of both regression methods discussed can be improved by conducting careful exploratory analyses of the data before adjustment and estimation. For instance, in our example, diagnostics may suggest that the control unit with profile ($264K, 4) is an outlier in terms of income and may be discarded before adjustment. However, such checks can be overlooked in routine use of regression in practice, and moreover, with multiple covariates, these checks may not be straightforward. Ideally, one would use methods that have some of these checks built in, and hence, that force the investigator to carefully inspect the data. Along these lines, we now discuss weighting and matching methods for adjustments in observational studies.



Diagnostic Dashboard

Covariate Balance

	Income	Visits
Treated	$27,216	4.6
Control	$27,243	4.6
Difference[1] Treated-Control	0.001	0.00

Study Representativeness

	Income	Visits
Target	$27,216	4.6
Difference[2] Treated-Target	0.00	0.00
Difference[2] Control-Target	0.002	0.00

Sample Size

	Effective	Nominal
Treated	100	100
Control	125	200

Differential Weighting[3]

	Min	Q1	Med	Q3	Max
Treated	1.0	1.0	1.0	1.0	1.0
Control	0.0	0.4	1.0	1.6	2.7

[1]: Absolute standardized mean difference. [2]: Target absolute standardized mean difference. [3]: The weights are scaled to sum up the size of each treatment group. Min: minimum. Q1, Q3: first and third quartiles. Med: median. Max: maximum.

Figure 5. Sketch of an observational study after weighting adjustments (SBW). The covariate values for each individual are shown over their shoulder. The size of each individual is proportional to his/her contribution or weight in the average treatment effect estimate. Individuals with a weight of zero are depicted in grey.

6.2. Weighting

In the two aforementioned regression methods, we started with an outcome model and then considered its implied weights to understand how these methods attempt to mimic a randomized experiment. Alternatively, acknowledging that the true outcome model is unknown, we can instead consider directly weighting the treatment and control groups without explicitly specifying an outcome model (Rosenbaum, 1987b). The most commonly used weighting method is based on modeling the probability of treatment assignment, or propensity score, using the observed covariates (Rosenbaum & Rubin, 1983).[10] If the model for the propensity score is correct, then in large samples we can unbiasedly estimate the average treatment effect by simply weighting the observations by the inverse of their propensity score estimates.

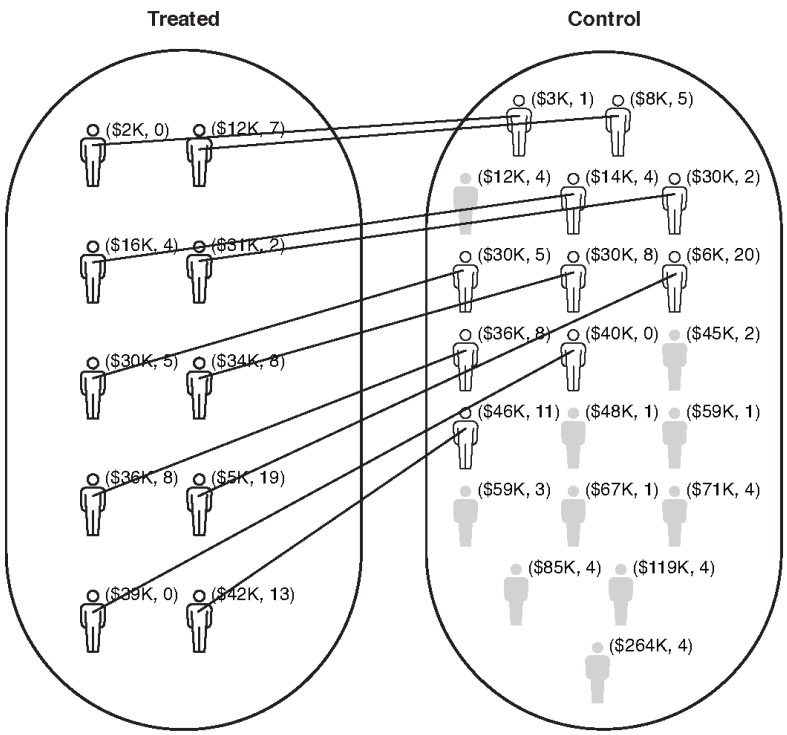
Does the above weighting procedure create a sample that looks as if it came from a randomized experiment? For a correctly specified propensity score model, (i) the weights balance, in large samples, the entire joint distribution of the observed covariates between the two groups. (ii) They also balance each group relative to the target (i.e., treated) population. With an incorrect model, however, large imbalances between the groups and relative to the target may exist for some covariates or their transformations. (iii) The propensity score is, by definition, nonnegative and, hence, for any reasonable propensity score model (correct or incorrect) the resulting effect estimates are sample-bounded. (iv) The resulting weights can be extreme and highly variable across individuals, especially when there is a lack of overlap in covariate distributions between the treatment and control groups.

Overall, weighting is a powerful and flexible method to adjust for covariates. However, in view of the experimental ideal, care must be taken to adequately balance the covariates toward the right target population, with weights that produce stable estimators. Recent weighting methods that emphasize these features are Hainmueller (2012), Imai and Ratkovic (2014), and Zubizarreta (2015), among others. See Austin and Stuart (2015) and Chattopadhyay et al. (2020) for reviews.

In our example, we implement the stable balancing weights (SBW) of Zubizarreta (2015). Figure 5 shows the structure of the resulting weighted sample. Overall, the covariates are reasonably well-balanced relative to the target population, although not exactly (see the balance diagnostic for income). By construction, for the ATT, this method puts differential weights only on the control units, but with weights that are nonnegative and less extreme than the regression methods. Notice that it puts zero weight on the outlying control subject with profile ($264K, 4). With this method, the estimated treatment effect amounts to −$967 in the example. This estimate aligns with existing substantive knowledge, and is closer to the estimate from the randomized experiment. We note that these results need not be identical, because the data for the randomized experiment and the observational study were drawn from different populations. When treatment effects vary across populations (i.e., there is effect modification), these results do not necessarily have to coincide.

6.3. Matching

Matching is a common approach that aims to find the randomized experiment that is 'hidden inside' (Hansen, 2004) the observational study (Imbens, 2015; Rosenbaum, 2020; Stuart, 2010). In its simplest form, matching forms pairs of similar treated and control subjects, such that the matched groups are balanced on aggregate. After matching, the structure of the data is similar to that of a randomized experiment: (i) covariates—the observed ones—are transparently balanced, while the unobserved ones, like with any other method for adjustment, must be dealt with a sensitivity analysis. (ii) Representativeness is overt and easily verifiable, as the target population can simply be compared to the matched sample in balance tables. (iii) By achieving covariate balance, effect estimates mostly are an interpolation grounded in the actual data and not an extrapolation of a potentially incorrect model. (iv) Each observation in the matched sample has the same weight, so the simplest test statistics and graphical displays can be used to analyze the outcomes. Finally, conducting a sensitivity analysis after matching is straightforward using Rosenbaum bounds (Rosenbaum, 1987a; Rosenbaum, 2002, chapter 4), chapter 4). Popular matching methods include Rosenbaum (1989), Iacus et al. (2012), and Diamond and Sekhon (2013). Recent optimal matching methods that capitalize on modern optimization using network flows include Hansen and Klopfer (2006), Pimentel et al. (2015), Yu et al. (2020), and Yu and Rosenbaum (2022). For recent optimization matching methods that leverage modern optimization using generic integer programming, see Zubizarreta (2012), Zubizarreta et al. (2014), and Cohn and Zubizarreta (2022).



Diagnostic Dashboard

Covariate Balance

	Income	Visits
Treated	$27,216	4.6
Control	$27,829	4.5
Difference.[1] Treated-Control	0.03	0.02

Study Representativeness

	Income	Visits
Target	$27,216	4.6
Difference.[2] Treated-Target	0.00	0.00
Difference.[2] Control-Target	0.06	0.03

Sample Size

	Effective	Nominal
Treated	100	100
Control	100	200

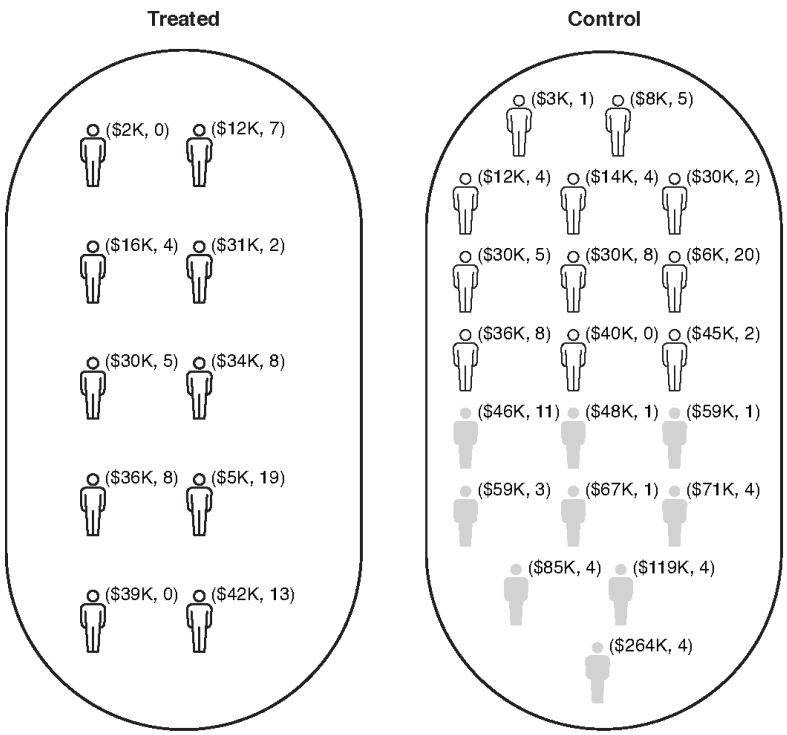
Differential Weighting[3]

	Min	Q1	Med	Q3	Max
Treated	1.0	1.0	1.0	1.0	1.0
Control	0.0	0.0	1.0	1.0	1.0

[1]: Absolute standardized mean difference. [2]: Target absolute standardized mean difference. [3]: The weights are scaled to sum up the size of each treatment group. Min: minimum. Q1, Q3: first and third quartiles. Med: median. Max: maximum.

Figure 6. Sketch of an observational study after pair matching. The covariate values for each individual are shown over their shoulder. The size of each individual is proportional to his/her contribution or weight in the average treatment effect estimate. Individuals with a weight of zero are depicted in grey. Matched individuals are connected.

In our example, we implement two matching methods, optimal pair matching as proposed by Rosenbaum (1989) and profile matching by Cohn and Zubizarreta (2022). The corresponding structures of the matched samples are depicted in Figures 6 and 7. Overall, both methods produce well-balanced and uniformly weighted samples relative to the target population. Finally, the effect estimates under pair matching and profile matching are −$1124 and −$999, respectively, close to the estimate provided by SBW.



Diagnostic Dashboard

Covariate Balance	Income	Visits
Treated	$27,216	4.6
Control	$27,342	4.6
Difference[1] Treated-Control	0.01	0.00

Study Representativeness	Income	Visits
Target	$27,216	4.6
Difference[2] Treated-Target	0.00	0.00
Difference[2] Control-Target	0.01	0.00

Sample Size	Effective	Nominal
Treated	100	100
Control	110	200

Differential Weighting[3]	Min	Q1	Med	Q3	Max
Treated	1.0	1.0	1.0	1.0	1.0
Control	0.0	0.0	1.0	1.0	1.0

[1]: Absolute standardized mean difference. [2]: Target absolute standardized mean difference. [3]: The weights are scaled to sum up the size of each treatment group. Min: minimum. Q1, Q3: first and third quartiles. Med: median. Max: maximum.

Figure 7. Sketch of an observational study after profile matching. The covariate values for each individual are shown over their shoulder. The size of each individual is proportional to his/her contribution or weight in the average treatment effect estimate. Individuals with a weight of zero are depicted in grey.

7. It Is Through Contrast That We See

Comparison and contrast are the basic means for understanding whether a treatment works; and randomization, the ideal mechanism for building similar groups of subjects where the effect of treatment is isolated. In the absence of randomization, covariate imbalance adjustments are needed to ensure that the groups are comparable, and three core methods that attempt to achieve this are regression, weighting, and matching.

These methods can be enhanced using nonparametric machine learning and combined in doubly robust estimators (Robins et al., 1994). For example, as discussed in Chattopadhyay and Zubizarreta (2023, Section 6), some of these doubly robust estimators can be represented as a simpler weighted contrast, as in Equation 4.1. This representation, in turn, allows us to inspect the implied unit-level weights and conduct diagnostic

analyses to evaluate covariate balance, study representativeness, effective sample size, differential weighting, and influence.

As discussed, the extent to which matching, regression, and weighting methods approximate the experimental ideal varies and is not always clear-cut. In observational studies, the essence of regression lies in correctly specifying the true model, in which case the approach is optimal. However, true models are elusive, and although some estimated models are helpful, for causal inference, it is often more practical to concentrate first on building good contrasts. Among others, this means approximating an experiment as closely as possible, with guarantees of robustness, interpretability, and ease to assess generalizability to target populations and sensitivity to hidden biases.

Along these dimensions, the traditional regression method is dominated by the multi regression approach, and matching and weighting approaches should be given strong consideration. We hope that increased use of implied weights diagnostics will improve the practice of causal inference. Upon further development, the diagnostics illustrated in this article may contribute to a better understanding of advanced statistical and machine learning techniques for causal inference. The ultimate goal? More transparent and reliable scientific findings.

Acknowledgments

We thank Isaiah Andrews, Kosuke Imai, Ronald Kessler, Russell Localio, Michael Sobel, Joseph Newhouse, Bijan Niknam, and Paul Rosenbaum for helpful comments and conversations. For graphics, we thank Xavier Alemañy and Esmeralda Aceituno.

Disclosure Statement

This work was supported by the Alfred P. Sloan Foundation (G-2020-13946) and the Patient Centered Outcomes Research Initiative (PCORI, ME-2022C1-25648).

References

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510. https://doi.org/10.1111/ajps.12116

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (pp. 1277–1366, Vol. 3). Elsevier. https://doi.org/10.1016/S1573-4463(99)03004-7

Aronow, P. M., & Samii, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, *60*(1), 250–267. https://doi.org/10.1111/ajps.12185

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *5*(3), 195–202. https://doi.org/10.2307/2985420

Chattopadhyay, A., Greifer, N., & Zubizarreta, J. R. (2023). *lmw: Linear model weights for causal inference*. arXiv. https://doi.org/10.48550/arXiv.2303.08790

Chattopadhyay, A., Hase, C. H., & Zubizarreta, J. R. (2020). Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, *39*(24), 3227–3254. https://doi.org/10.1002/sim.8659

Chattopadhyay, A., & Zubizarreta, J. R. (2023). On the implied weights of linear regression for causal inference. *Biometrika*, *110*(3), 615–629. https://doi.org/10.1093/biomet/asac058

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A (General)*, *128*(2), 234–266. https://doi.org/10.2307/2344179

Cohn, E. R., & Zubizarreta, J. R. (2022). Profile matching for the generalization and personalization of causal inferences. *Epidemiology*, *33*(5), 678–688. https://doi.org/10.1097/EDE.0000000000001517

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, *22*(1), 173–203. https://doi.org/10.1093/jnci/22.1.173

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932–945. https://doi.org/10.1162/REST_a_00318

Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health*, *43*(6), 677–683. https://doi.org/10.2105/ajph.43.6_pt_1.677

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Gelman, A., & Imbens, G. (2018). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447-456. https://doi.org/10.1080/07350015.2017.1366909

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. https://doi.org/10.1093/pan/mpr025

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609–618. https://doi.org/10.1198/016214504000000647

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. https://doi.org/10.1198/106186006X137047

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. CRC Press.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), Article mpr013. https://doi.org/10.1093/pan/mpr013

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243–263. https://doi.org/10.1111/rssb.12027

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373–419.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Newhouse, J. P. (1993). *Free for all?: Lessons from the RAND Health Insurance Experiment*. Harvard University Press.

Peters, C. C. (1941). A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, *34*(8), 606–612. https://doi.org/10.1080/00220671.1941.10881036

Pimentel, S. D., Kelz, R. R., Silber, J. H., & Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, *110*(510), 515–527. https://doi.org/10.1080/01621459.2014.997879

Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science*, *22*(4), 544–559. https://doi.org/10.1214/07-STS227D

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.2307/2290910

Rosenbaum, P. R. (1987a). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*(1), 13–26. https://doi.org/10.2307/2336017

Rosenbaum, P. R. (1987b). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394. https://doi.org/10.1080/01621459.1987.10478441

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032. https://doi.org/10.1080/01621459.1989.10478868

Rosenbaum, P. R. (2002). *Observational studies*. Springer.

Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In Brian S. Everitt & David C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1809–1814). Wiley. https://doi.org/10.1002/9781118445112.stat06358

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, *7*, 143–176. https://doi.org/10.1146/annurev-statistics-031219-041058

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214%2F09-STS313

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine*, *167*(4), 268–274. https://doi.org/10.7326/m16-2607

Yu, R., & Rosenbaum, P. R. (2022). Graded matching for large observational studies. *Journal of Computational and Graphical Statistics*, *31*(4), 1406–1415. https://doi.org/10.1080/10618600.2022.2058001

Yu, R., Silber, J. H., & Rosenbaum, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statistical Science*, *35*(3), 338–355. https://doi.org/10.1214/19-STS699

Zhao, Q., Small, D. S., & Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(4), 735–761. https://doi.org/10.1111/rssb.12327

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, *107*(500), 1360–1371. https://doi.org/10.1080/01621459.2012.703874

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. https://doi.org/10.1080/01621459.2015.1023805

Zubizarreta, J. R., Paredes, R. D., & Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Annals of Applied Statistics*, *8*(1), 204–231. https://doi.org/10.1214/13-AOAS713

©2024 Ambarish Chattopadhyay and José R. Zubizarreta. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. The data and code of this example are available upon request. ↩
2. The absolute standardized mean difference of a covariate x between the treatment and the control group is defined as $\frac{|\bar{x}_t - \bar{x}_c|}{\sqrt{(s_t^2 + s_c^2)/2}}$, where $(\bar{x}_t, s_t)$ and $(\bar{x}_c, s_c)$ are the mean and standard deviations of *x* in the treatment and control group, respectively. Smaller values of this measure indicate better mean balance on x. ↩
3. More formally, for a sample of *n* subjects with weights $w_1, w_2, \ldots, w_n$, the effective sample size is defined as $\frac{(|w_1| + |w_2| + \ldots + |w_n|)^2}{w_1^2 + w_2^2 + \ldots + w_n^2}$. If the weights are uniform, the effective sample size equals the nominal sample size n. ↩
4. The treatment assignment indicator of a subject equals one if the subject receives treatment and equals zero otherwise. ↩
5. See Angrist and Krueger (1999) and Aronow and Samii (2016) for related work. ↩
6. Here, we regress the outcome (expenditure) linearly on the two baseline covariates: income and visits. ↩
7. Unless the true propensity score model is inverse linear on the covariates included in the model (Robins et al., 2007; Chattopadhyay and Zubizarreta, 2023). ↩
8. In other words, the OLS estimator for τ can be equivalently obtained, solving an optimization problem that minimizes the variance of the weights that exactly balance the means of the covariates between the

treatment and control groups, with weights that add up to one and that can take negative values. See Chattopadhyay and Zubizarreta (2023) for a general expression of this optimization problem that also encompasses weighted least squares (WLS), augmented inverse probability weighting (AIPW), and ridge regression estimation, in connection to the stable balancing weights (SBW; Zubizarreta, 2015). ↩

9. In this case, the covariates need to be centered, that is, have mean zero. This can be achieved by subtracting from each covariate its sample mean before fitting the model. Centering allows us to represent the average treatment effect as a single coefficient τ, as opposed to a combination of τ and γ. ↩

10. More formally, for estimating the ATT, a subject with covariate x receives a weight of 1 if treated and a weight $\frac{\hat{e}(x)}{1-\hat{e}(x)}$ if not treated, where $\hat{e}(x)$ is the value of the estimated propensity score for the subject. ↩

References

- Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510. https://doi.org/10.1111/ajps.12116 ↩
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607 ↩
- Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *5*(3), 195–202. https://doi.org/10.2307/2985420 ↩
- Chattopadhyay, A., & Zubizarreta, J. R. (2023). On the implied weights of linear regression for causal inference. *Biometrika*, *110*(3), 615–629. https://doi.org/10.1093/biomet/asac058 ↩
- Chattopadhyay, A., Greifer, N., & Zubizarreta, J. R. (2023). *lmw: Linear model weights for causal inference*. arXiv. https://doi.org/10.48550/arXiv.2303.08790 ↩
- Chattopadhyay, A., Hase, C. H., & Zubizarreta, J. R. (2020). Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, *39*(24), 3227–3254. https://doi.org/10.1002/sim.8659 ↩
- Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A (General)*, *128*(2), 234–266. https://doi.org/10.2307/2344179 ↩
- Cohn, E. R., & Zubizarreta, J. R. (2022). Profile matching for the generalization and personalization of causal inferences. *Epidemiology*, *33*(5), 678–688. https://doi.org/10.1097/EDE.0000000000001517 ↩
- Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, *22*(1), 173–203. https://doi.org/10.1093/jnci/22.1.173

↩

- Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932–945. https://doi.org/10.1162/REST_a_00318 ↩
- Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health*, *43*(6), 677–683. https://doi.org/10.2105/ajph.43.6_pt_1.677 ↩
- Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd. ↩
- Gelman, A., & Imbens, G. (2018). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447-456. https://doi.org/10.1080/07350015.2017.1366909

↩

- 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. https://doi.org/10.1093/pan/mpr025 ↩
- Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609–618. https://doi.org/10.1198/016214504000000647 ↩
- Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. https://doi.org/10.1198/106186006X137047 ↩
- Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. CRC Press. ↩
- Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), Article mpr013. https://doi.org/10.1093/pan/mpr013

↩

- Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243–263. https://doi.org/10.1111/rssb.12027

↩

- Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373–419. ↩
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. ↩
- Newhouse, J. P. (1993). *Free for all?: Lessons from the rand health insurance experiment*. Harvard University Press.

↩

- Peters, C. C. (1941). A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, *34*(8), 606–612. https://doi.org/10.1080/00220671.1941.10881036 ↩

- Pimentel, S. D., Kelz, R. R., Silber, J. H., & Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, *110*(510), 515–527. https://doi.org/10.1080/01621459.2014.997879 ↩
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.2307/2290910 ↩
- Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, *22*(4), 544–559. https://doi.org/10.1214/07-STS227D ↩
- Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394. https://doi.org/10.1080/01621459.1987.10478441 ↩
- Rosenbaum, P. R. (1987a). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*(1), 13–26. https://doi.org/10.2307/2336017

  ↩
- Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032. https://doi.org/10.1080/01621459.1989.10478868

  ↩
- Rosenbaum, P. R. (2002). *Observational studies*. Hardcover; Springer. ↩
- Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In Brian S. Everitt & David C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1809–1814). Wiley. https://doi.org/10.1002/9781118445112.stat06358

  ↩
- Rosenbaum, P. R. (2010). *Design of observational studies*. Springer. ↩
- Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, *7*, 143–176. https://doi.org/10.1146/annurev-statistics-031219-041058 ↩
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41 ↩
- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903 ↩
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739 ↩
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214%2F09-STS313 ↩

- VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine*, *167*(4), 268–274. https://doi.org/10.7326/m16-2607 ↩
- Yu, R., & Rosenbaum, P. R. (2022). Graded matching for large observational studies. *Journal of Computational and Graphical Statistics*, *31*(4), 1406–1415. https://doi.org/10.1080/10618600.2022.2058001 ↩
- Yu, R., Silber, J. H., & Rosenbaum, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statistical Science*, *35*(3), 338–355. https://doi.org/10.1214/19-STS699 ↩
- Zhao, Q., Small, D. S., & Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: (Statistical Methodology)*, *81*(4), 735–761. https://doi.org/10.1111/rssb.12327 ↩
- Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, *107*(500), 1360–1371. https://doi.org/10.1080/01621459.2012.703874 ↩
- 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. https://doi.org/10.1080/01621459.2015.1023805 ↩
- Zubizarreta, J. R., Paredes, R. D., & Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Annals of Applied Statistics*, *8*(1), 204–231. https://doi.org/10.1214/13-AOAS713 ↩