Journal of Applied Statistics

Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Causal inferences from many experiments

Wendy K. Tam Cho

To cite this article: Wendy K. Tam Cho (2016): Causal inferences from many experiments, Journal of Applied Statistics

To link to this article: http://dx.doi.org/10.1080/02664763.2016.1266468

Published online: 08 Dec 2016.

JOURNAL OF APPLIED STATISTICS, 2016
http://dx.doi.org/10.1080/02664763.2016.1266468

Taylor & Francis
Taylor & Francis Group

Causal inferences from many experiments

Wendy K. Tam Cho

Departments of Political Science and Statistics, and National Center for Supercomputing Applications, University of Illinois at Urbana–Champaign, Urbana, IL, USA

ABSTRACT

The underlying statistical concept that animates empirical strategies for extracting causal inferences from observational data is that observational data may be adjusted to resemble data that might have originated from a randomized experiment. This idea has driven the literature on matching methods. We explore an un-mined idea for making causal inferences with observational data – that any given observational study may contain a large number of indistinguishably balanced matched designs. We demonstrate how the absence of a unique best solution presents an opportunity for greater information retrieval in causal inference analysis based on the principle that many solutions teach us more about a given scientific hypothesis than a single study and improves our discernment with observational studies. The implementation can be achieved by integrating the statistical theories and models within a computational optimization framework that embodies the statistical foundations and reasoning.

ARTICLE HISTORY
Received 26 March 2016
Accepted 3 November 2016

KEYWORDS
Causal inference; optimization; subset selection; computational methods; operations research

1. Many experiments within an observational study

The core literature on making causal inferences from observational data rests upon the expectancy and possibility that experiments hide within observational data. When randomization in an experiment is successful, the treatment effect is isolated from potential confounders. Differences in response can then be interpreted as a treatment effect [13]. The hope for observational studies is that *if* one can organize or weight the observations in an observational study such that their configuration resembles data from a randomized experiment, then one may be able to make the leap from associational inferences to causal inferences [20]. This line of reasoning has animated much of the work on the statistical adjustment of observational data.

A number of methods for seeking the latent experiment in observational data have been proposed [9,19,29]. These methods are not identical in implementation or outcome, but they share a common goal – to identify the observational design that is 'closest to the experiment'. They also share a common process that first defines a distance metric for assessing how 'close' two observations are to one another and then collects units into homogeneous

CONTACT Wendy K. Tam Cho ✉ wendycho@illinois.edu Departments of Political Science and Statistics, and National Center for Supercomputing Applications, University of Illinois at Urbana–Champaign, 420 David Kinley Hall, 1407 W. Gregory Dr., Urbana, IL61801, USA

sets or pairs. These approaches emulate the latent block- or pair-randomized experimental framework and provide a single matched solution. [1]

The 'one solution' from these observational studies is akin to running 'one experiment'. Certainly, conducting just one experiment is common since experiments are often costly, with respect to both time and resources. If time and resource limitations are lifted, it is easy to see how one would have more confidence in a particular finding if it were replicated across many experiments. The treatment effect, after all, is a random variable. One experiment provides simply one realization of that random variable. For both experiments and observational studies, it would be helpful if we could interpret them in context – as one experimental realization among a host of possible experimental realizations. While we can all agree on this ideal, it is unclear *how* we can interpret experiments in context until many experiments have been conducted. Perhaps unsurprisingly, then, methods for making causal inferences from observational data seem to also hold tightly to the one experiment framework.

Our analysis takes a different turn with a focus on a research design that exploits the ability to create a large number of independent 'as-if-randomized' designs. We show how a single observational study can potentially provide many matched subsets and thus more information that is useful for interpreting treatment effect estimates. Our method is not a panacea for the overarching issues that plague inferring causality from observational data, namely the selection on observables assumption. The current implementation of the algorithm can also be improved, especially with respect to the generation of more independent designs. However, we offer a novel perspective and method that utilizes more information, and as a result, offers a way forward for incorporating additional information for causal inferences models.[2]

2. Expanding beyond the one experiment framework

Partly because a single experiment requires significant time and resources, we tend to view the outcome of any single experiment in the best light possible – as an unbiased estimate of the true effect. We leave it to subsequent scholars to advance science by validating our findings under similar conditions. That is, even though the p-value for the treatment effect in a randomized experiment might be based on the thought experiment of repeating the random assignment, the idea as articulated by Fisher [8] is that one can advance science only through an accumulation of evidence from different yet complementary studies. Practical matters tend to lead researchers toward under-emphasizing repetition as part of the scientific enterprise, while also over-emphasizing the benefits of randomized assignment in a single study. We know, however, that estimates may differ across different models and analyses even when the data remain constant.

2.1. The variation embodied within many experiments

A randomized experiment can be repeated. If it were repeated, the next randomization process would yield empirically different units and a potentially different estimate of the treatment effect, highlighting that treatment effect is a random variable. That variation is embodied within the many potential experiments is plain – it is induced by the randomization process itself. Consider a simulation of 1000 randomized experiments, using data from the Current Population Study (CPS). We randomly draw 370 individuals from the data set, and then subsequently randomly place each in the treatment or control group. The chosen group size mimics the LaLonde NSW experiment [15]. Our outcome variable is real earnings in 1978 (RE78). We do not treat these chosen units in any sense. That is, there is no treatment effect because there is no treatment. So, the true difference in the outcome variable, RE78, between the treatment and control groups in our simulations is zero, in expectation.

The distribution of the estimated average treatment effects across our 1000 experiments is shown in Figure 1. For the CPS data, the mean of this distribution is approximately $6.90, which is very close, and given variability in the simulation, essentially identical to the true treatment effect of zero. The range of the treatment effect, however, is quite large (−$3470, $3042), despite the truly randomized experiment, reflecting the noisy nature of the outcome variable. The randomized experiment provides us with an unbiased estimate, but the noisy outcome variable ensured that we also have a large variance estimate of the average treatment effect.

When the estimated treatment effect is far from the true treatment effect, the standard error is not large enough to ensure that the result will be correctly identified as statistically insignificant. In our simulation, the SE remained in a fairly small range (942.1, 1075.0), essentially conveying the same uncertainty for the associated estimate, which is *assumed*

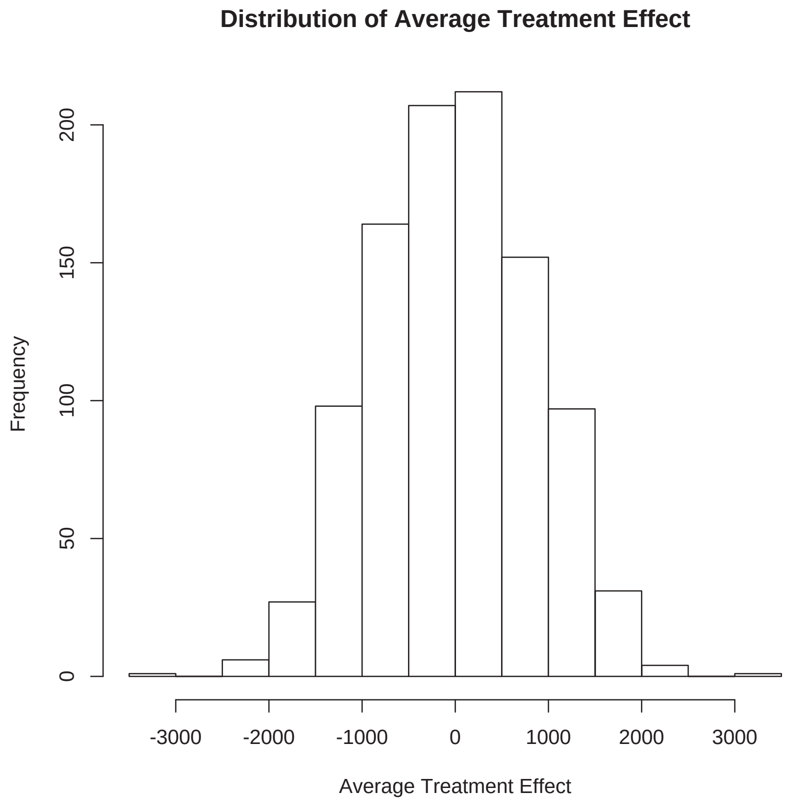

Figure 1. Average treatment effect across 1000 simulated experiments.

to be correct. When one calculates the SE, that calculation includes the sample size and an estimate of the population SD but does not incorporate information about how far the estimated treatment effect might be from the unknown true treatment effect. Indeed, the SD of the sampling distribution is 1037.68, so the various SE estimates are quite good, generally.

2.2. The relationship between balance and the ATE

For each of our single simulated experiments, we expect randomization to result in balanced treatment and control covariate distributions that are roughly equivalent. Figure 2 provides an illustration of the balance achieved across our 1000 simulated experiments. In the plot on the left, balance, b, is a summary measure of balance across the eight covariates. In particular, it is the sum of theKolmogorov–Smirnov statistic (comparing the treatment and control covariate distributions), the absolute value of the t-test for the difference in means, and absolute value of the difference between 1 and the ratio of the variances,

$$b = \sum_{j=1}^{8} KS_j + |t_j| + \left|1 - \frac{\sigma_{tj}}{\sigma_{cj}}\right|. \tag{1}$$

With perfect balance and no measurement error, both the Kolmogorov–Smirnov statistic and the difference of means would be zero, and the ratio of the variances would be one. So, the closer our balance measure is to zero, the more balanced our covariate sets are by this particular measure.

The least balanced experiment in our simulation was the 947th experimental iteration (shown in the lower right area of the plot), where the balance was about 16.36 (values shown in Table 1), and the average treatment effect was −$2167.14, far from the true value of 0. One might be tempted to believe then that when a randomized experiment results in poor balance, the estimate of the treatment effect will likewise be poor. Of course, this

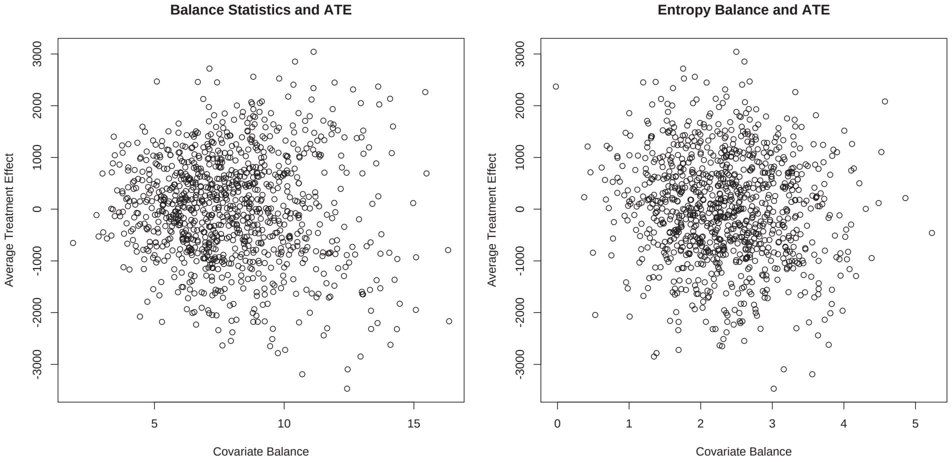

Figure 2. Average treatment effect across 1000 simulated experiments.

Table 1. Covariate balance for experimental iterations 947 and 693.

Experimental iteration 947, $\tau = -\$2167.14$, $b = 16.36$

Covariate	Kolmogorov–Smirnov	*p*-Value	Difference of means	*p*-Value	Variance ratio
1	0.093	0.401	−1.020	0.309	1.110
2	0.081	0.579	1.936	0.054	0.739
3	0.049	0.979	1.647	0.100	1.645
4	0.008	1.000	0.366	0.715	1.174
5	0.124	0.117	−2.606	0.010	1.259
6	0.077	0.636	−1.761	0.079	0.791
7	0.127	0.100	−2.190	0.029	1.105
8	0.152	0.028	−2.230	0.026	1.128

Experimental iteration 693, $\tau = -\$2077.06$, $b = 4.44$

Covariate	Kolmogorov–Smirnov	*p*-Value	Difference of means	*p*-Value	Variance ratio
1	0.047	0.986	0.156	0.876	1.132
2	0.055	0.944	0.400	0.690	0.787
3	0.005	1.000	−0.186	0.853	0.944
4	0.014	1.000	0.533	0.594	1.225
5	0.058	0.920	−1.232	0.219	1.138
6	0.007	1.000	−0.143	0.887	0.989
7	0.083	0.545	−0.212	0.832	1.126
8	0.051	0.969	−0.276	0.783	1.082

may occur and, generally, better balanced experiments exhibit less variation, but as a rule, it is neither correct nor guaranteed. Consider the results in the lower left or upper left area of the plot. These data points represent simulations where the balance was good, but the ATE was not good. In the 693rd simulation, for instance, the treatment effect estimate was −\$2077.06 while the balance value was 4.44. The balance is quite good for every covariate *yet* the estimate of the treatment effect, τ, is far from zero.

Since our measures of balance involve only marginal distributions, one might rightly wonder whether $\hat{\tau}$ is off in experimental iteration 947 because of imbalance in the joint distributions. This is a reasonable and testable hypothesis. One way in which we can examine the difference in joint or higher order distributions is through the Kullback–Leibler Information Criterion (KLIC). KLIC is a measure of the similarity of a probability distribution, p, to another probability distribution, q.

$$\mathrm{KLIC}(p, q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right). \tag{2}$$

In our data, p arises from the control group while q is defined by the treatment group covariates. In particular, we create an entropy covariate balance measure that combines the Kullback–Leibler divergence measure for each covariate individually as well as including the KLIC entropy measure for all 28 of the two-way joint covariate distributions. The results are shown in the right-hand side in Figure 2. We can see that $\hat{\tau}$ may be far from the true treatment effect even when the marginal and joint distributions are quite similar. In summary, balance among the covariates can be quite good while $\hat{\tau}$ is not particularly close to the true treatment effect *even in a truly randomized experiment*.

We know that when experiments are run well, we are not likely to obtain estimates on the tail of the distribution. All the same, it is also clear that given one single estimate, we have no idea where that estimate lies in the true underlying distribution. Given the inability to repeat the experiment, we assume that the estimate is a good estimate of the true ATE. If one could easily replicate and run an experiment repeatedly, one would surely choose to do so as this option is plainly superior to running an experiment once. In a larger sense, this is how all science proceeds. For any single study, no matter how well conducted the research may be, if no other researcher ever identifies similar results, that single study will be discounted as an anomaly, on the tail of the distribution.

While these results and statements are not earth-shattering, the insights inherent in them have nonetheless been neglected in the designs of methods for making causal inferences from observational data. Consider, for instance that the statistical methods do not incorporate the notion that there is not one valid experiment but many valid experiments with different control individuals that yield essentially the same level of covariate balance. Instead, the methods identify one 'best' matched group without regard to a possible myriad of other sets of individuals that could have constituted another valid experiment. They do not identify other matched groups that other equally valid experiments may embody nor do they contextualize the solution they identify. One has no sense of how much better this 'best' identified group is from the next best group or whether there are other matched groups that correspond to an equally valid experiment. Possibly there are other equally or better matched groups that yield conflicting estimates of the average treatment effect. In short, these methods place undue confidence in the identification of one matched group. More ideally, a single experiment should be regarded as a *single* experiment, a single realization of a random variable. To understand this single value well, it needs to be properly contextualized.

3. Computational modeling to capture statistical reasoning

For an experiment, it is difficult to garner the resources to repeat an experiment many times. Time and resource constraints, however, are alleviated when one wishes to make causal inferences from observational data since no experiment is actually conducted. Instead, the question becomes: if it is possible to statistically adjust once, is it possible to modify the process so that we are able to extract insight into the distribution of experimental outcomes?

The Balance Optimization Subset Selection (BOSS) method for making causal inferences from observational data [2] allows us to achieve these goals by incorporating a computational optimization model that captures the statistical reasoning we have outlined. This method seeks to identify the solution with the best balanced covariates, but in the process, finds many other solutions in the solution space that are also consistent with randomization and outputs a host of solutions that satisfy a given criterion. It is akin to other causal inference methods in that it identifies control groups that are close to the treatment group. It is different, however, in its fundamental design because it identifies and saves all control groups that are above some standard of randomization.

3.1. Mimicking completely randomized experimental framework

BOSS also embodies important departures from the general design of matching methods. First, while other methods attempt to match individual treatment and control units to one another in paired unit sets, BOSS examines *subsets* of the control group and identifies those subsets that achieve an optimal level of covariate balance between the treatment group and the control group in the aggregate.[3] BOSS seeks to post-process observational data so that they resemble a randomized control trial. By shifting the focus from matching *individual units* to the overall balance in *treatment and control groups* as a whole, BOSS reframes the causal inference problem from a matching problem to a subset selection problem. For BOSS, given the treatment group, T, and a control pool, C, the goal is to find $S_t \in T$, a subset of the treatment pool and $S_c \in C$, a subset of the control pool, so that a measure of balance, $b(S_t, S_c)$, is maximized.

For the traditional matching problem, one seeks to identify individual treatment units, $t \in T$, and individual control units, $c \in C$, so that a defined distance between two units, $\delta(c, t)$, is minimized. One may formulate this as an optimization problem [18]. Akin to the personnel assignment problem [14], the objective is to identify matched pairs so that the total distance between all matched pairs is minimized,

$$\begin{aligned} \underset{a}{\text{minimize}} \quad & \sum_{t \in T} \sum_{c \in C} \delta(c, t)\, a_{tc} \\ \text{subject to} \quad & \sum_{c \in C} a_{tc} = 1, \quad \forall t, \end{aligned} \tag{3}$$

where a_{tc} is 1 if treated unit t and control unit c are matched and 0 otherwise. The constraint (3) indicates that the matches are 1–1 and include every treated unit. This formulation can be easily changed so that matches are 1–k and/or not all treated units are included. Neither modification is significant in the basic formulation of the problem. Once the optimization is completed and the matched pairs are identified, the matched controls units are placed in S_c, the treated units are placed in S_t, and then the balance, $b(S_c, S_t)$ is computed. The hope is that the identified set of matched pairs results in sufficient covariate balance between the identified control and treatment sets. This assessment is made after the optimization is complete by computing statistics related to the empirical covariate distributions from the identified treatment and control units.

With BOSS, the objective function directly minimizes the imbalance between the treatment and control subsets. The process does not involve computing any distances or similarity measures between individual units. Computing these distances is not necessary for obtaining covariate balance and restricts the design to the pair-randomized framework. Our objective function is

$$\begin{aligned} \underset{(S_c, S_t)}{\text{minimize}} \quad & \sum_{i} w_i\, b_i(S_c, S_t) \\ \text{subject to} \quad & |S_c| = |S_t|, \end{aligned} \tag{4}$$

where w_i is a weight for the ith balance measure, b_i. Given some set of covariate balance measures, b_i, BOSS identifies the subsets S_c and S_t that minimizes imbalance, subject to equally sized treatment and control subgroups. The constraint (4) is flexible – the cardinalities of the subsets need not be the same. In the current algorithm, the cardinality of the control group is fixed by the user at the outset and, in our examples, is set to the cardinality of the treatment group. Once the BOSS optimization routine is complete, the subsets identified will be maximally balanced. Unlike traditional matching methods that hope to obtain balance by minimizing distance between matched pairs, the BOSS optimization routine directly balances the covariates.

By identifying entire treatment and control subsets that minimize differences in the empirical covariate distributions, BOSS differs from other approaches because it mimics a completely randomized experiment rather than a block-randomized (or pair-randomized) experiment that lies at the foundation of propensity score methods or Mahalanobis metric matching. To be sure, any of these frameworks approximates a valid experimental design that may yield covariate balance consistent with randomization. These approaches generally will not yield identical solutions but will rather yield different solutions with possibly indistinguishable balance. Just as there are many different valid research designs for experiments, there are many ways in which subsets of observational data may be extracted to approximate these varied research designs. Notably, the BOSS framework subsumes these experimental frameworks that are incorporated into existing matching procedures since block-randomized designs and pair-randomized designs are also identified by BOSS's computational modeling approach.

The computational/optimization framework of BOSS procedure allows one to extract information from many different subsets. While BOSS is searching the solution space for all control subsets that are consistent with a randomization framework, the best subset is recorded along with many of other subsets encountered in the optimization search for control and treatment groups that are statistically indistinguishable from those that might have arisen from a randomized experiment. The information from the many different as-if-randomized subsets allows us to gain a sense of the underlying distribution for the treatment effect, τ.

BOSS's shift from individual matching to subset selection highlights an interesting combinatorial aspect of both the matching methodologies and the subset selection methodology. In particular, for even moderately sized data sets, the set of possible 'solutions' is extremely large. For instance, if our control pool has 100 members, and we wish to choose a subset of size 20, there are $\binom{100}{20} = 5.359834 \times 10^{20}$ possibilities. Given the sheer size of the problem, finding the best or most balanced subset in this solution space proves challenging. Moreover, the solution landscape is not rugged. While the landscape is hilly, these peaks and valleys are not a rapid succession of precipices, but instead, a series of vast plateaus. These expansive plateaus manifest themselves throughout the landscape because many possible subsets of the data are similar to one another. It is readily evident that swapping several units for other units in a large data set should not alter the covariate balance much. Yet, even among subsets with substantially or completely different composition, many of these different subsets are equally consistent with a randomization process. Indeed, just as a randomized experiment can be conducted repeatedly, there may be many subsets of the observational data that are consistent with a randomization procedure.

While it would be ideal to conduct a series of randomized experiments, doing so is often not practical. Estimating causal effects from observational studies is understandably more complex and less reliable than estimating causal effects from a randomized experiment. Data limitations and the Selection on Observables assumption cannot be overcome. However, a heretofore overlooked point is that the extraction of so many different subsets of observational data consistent with randomization procedures can yield a tremendous advantage. Given some threshold dividing 'putatively randomized' from 'putatively not randomized' studies like a p-value on an omnibus balance test or a function of a collection of p-values, we can produce many different designs, all of which would prima facie qualify as a randomized experiment.

4. The LaLonde NSW data

We turn to the LaLonde data to demonstrate how integrating a computational approach with the statistical foundations of causal inference models can yield substantive insights. The LaLonde [15] data hail from an experiment designed to capture the effect from participation in a temporary employment program designed to help disadvantaged workers lacking basic job skills to move into the labor market by providing work experience and counseling in a sheltered environment. Qualified applicants to the training program were randomly assigned to treatment and control groups, creating a randomized job training experiment. The treatment group received the benefits of the NSW program while the control group did not. The NSW provided wages for the participants that could increase based on job performance. After the program period expired, the participants were forced to find regular employment. Earnings and demographic data from both the treatment and control group were collected every nine months. A number of scholars have pursued varied paths for causal analyses with the LaLonde data, raising a number of estimation issues, and presenting different substantive interpretations [3–7,10,15,25–28].

In our analysis of the LaLonde data, we use the Dehejia and Wahba [5] subsample for the treatment group, which includes pre-treatment income in 1974 as a covariate, and the individuals from the *Current Population Survey* (CPS) for the control pool. We do not use the control group from the LaLonde data.[4] The treatment group contains $m = 185$ individuals and the control pool contains $N - m = 15{,}992$ individuals. In this data set, there are eight covariates, $x_1, \ldots, x_8$. The purpose of the study was to discover how real earnings in 1978 (RE78) might have changed as a result of being a part of the NSW job training program.

In these data, each person i has two potential outcomes, one from assignment to the job training program, $Z_i = 1$, and one from exclusion from the program, $Z_i = 0$.[5] The potential outcomes, $y_{i,Z_i=1} \equiv y_{i1}$ and $y_{i,Z_i=0} \equiv y_{i0}$, represent the person's real earnings in 1978. If the two potential outcomes differ, $y_{i1} \neq y_{i0}$, then we say that the job training program had a causal effect for person i. The fundamental problem of causal inference is that it is impossible to observe the value of both y_{i1} and y_{i0}, because each subject was either exposed to the job training condition or was not. To gain some traction in this situation, Neyman [16] suggested reconceptualizing the framework to focus on the average causal effect across the treatment and control groups,

$$\tau = \frac{1}{N}\left(\sum_{i=1}^{N} Z_i y_{i1}\right) - \frac{1}{N}\left(\sum_{i=1}^{N} (1 - Z_i) y_{i0}\right), \tag{5}$$

so that either the potential outcome under treatment or under control, but not both, needs to be observed for each unit [12,22,23]. He showed that, across repeated randomizations of treatment, the observed difference of means between the treated and control groups is an unbiased estimator of the unobserved average treatment effect (ATE), i.e. $E(\hat{\tau}) = \tau$.

For the LaLonde data, BOSS identifies subsets of the control pool that achieve a particular level of balance with the treatment group for the eight covariates in the data set. Initially, we began with the balance measure (1) that we used in our earlier simulations. The subsets that emerged had reasonable balance, but we noticed that it was difficult to find enough blacks for the control subset. Since the balance on all of the variables is equally weighted and some variables are much simpler to balance, the algorithm tended to sacrifice balance on the black variable for very close balance on other variables. To encourage the optimization routine to obtain better balance on the substantively important black covariate, we weighted that variable more highly than the others. Note that the weighting here is simply a method for guiding the optimization routine toward portions of the solution space that are difficult to reach. The weights do not imply substantive changes in modeling the underlying phenomenon.

Another difficulty arises from the bi-modal nature of the real earnings variable distribution. For the large number of the observations where a job was not sustained, the real earnings was \$0, reflecting lack of employment. The information contained in the mean and Kolmogorov–Smirnov statistic are insufficiently nuanced to capture the distributional shape of these variables. Accordingly, we created a number of new variables intended to provide data points that were more expressive of the idiosyncratic nature of the particular covariates in the data set. Real earnings for 1974, real earnings for 1975, and the age variable were broken down into quintiles. For the real earnings variables, we also created indicator variables for when the value fell at the minimum or the maximum values of the distribution. Education was broken down into three levels. These additional variables provided additional guidance for identifying subsets where the distributions of the control and treatment variables were more closely aligned.

Our balance variable was

$$b = \sum_{j=1}^{26} w_j \left(KS_j + |t_j| + \left| 1 - \frac{\sigma_{tj}}{\sigma_{cj}} \right| \right), \tag{6}$$

where w_j is a weight for the jth variable. In our case, the weights for the black variable was 3 while the weight for the other 25 variables was 1. The weighting helped the algorithm to more successfully balance the distribution of blacks in the control sets without having an adverse effect on the balance for the other variables.

Once we identified our subsets, we computed the ATE as specified in Equation (5). A summary of the BOSS solution search using the LaLonde data is shown in Figure 3 while the summary statistics for the results are provided in Table 2. In the figure, the dark solid line shows the mean ATE for solutions that fall in a range of objective functions values. The lighter solid horizontal line displays where the experimental benchmark value lies. The lighter grayed area displays one standard deviation of these estimates while the darker shaded gray area shows two standard deviations for the estimates.[6] Despite the fairly large range and standard deviation of our estimated treatment values, both the range and the variance tend to become smaller as the objective function value improves. Though the range is fairly large, as we expected, the minimum value is well above zero, indicating a positive effect from the job training program.

Our lowest objective function value was 10.62.[7] This particular solution yielded an estimated treatment effect of \$1870.08 (\$129.33 from the experimental benchmark). In addition, we found three other solutions whose balance was within a hundredth of 10.62.

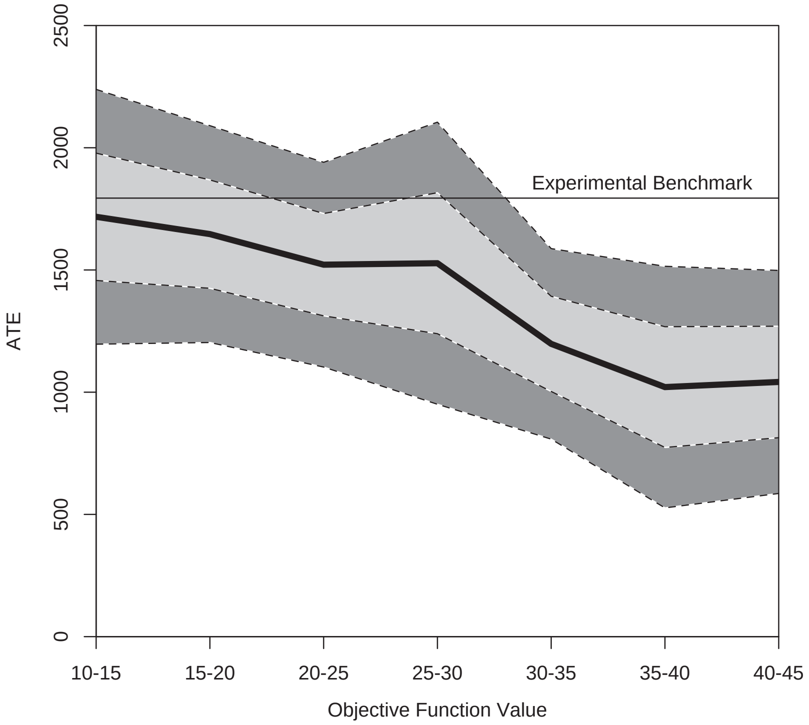

Figure 3. LaLonde data: average treatment effect by objective function range.

Table 2. LaLonde data: solutions sorted by objective function value.

		Treatment effect				Kolmogorov–Smirnov	
Objective function range	Observations	Mean	SD	Minimum	Maximum	Mean	SD
10.0–15.0	2578	1717.54	260.54	1211.40	2232.90	0.02	0.01
15.0–20.0	2607	1646.77	221.73	1030.04	2316.60	0.01	0.01
20.0–25.0	2595	1521.63	209.35	765.45	2014.36	0.01	0.02
25.0–30.0	2536	1527.65	288.47	970.47	2249.91	0.03	0.03
30.0–35.0	2727	1197.72	194.80	506.34	1779.61	0.07	0.06
35.0–40.0	2897	1020.89	247.05	414.86	1777.75	0.12	0.12
40.0–45.0	2598	1041.91	228.05	422.61	1695.36	0.08	0.07

Note: Control and treatment group sizes are constrained to be equal. Control groups do not contain any duplicate observations (i.e. individuals are chosen without replacement).

The ATEs for these solutions were in the range [1893.88,1914.97]. We also identified a solution with an ATE within a penny of the experimental benchmark. Its balance value was 15.80. Importantly, note that for these data, there was a good deal of variance in the treatment effect estimate even among the set of solutions with the best objective values between 10 and 15. In this set, we identified 2578 subsets that together had a mean of $1717.54 and a standard deviation of $260.54. The experimental benchmark is well within a standard deviation of our solutions in this objective function range. Our emphasis, of course, is not on the single experimental benchmark estimate. Instead, our point here is that there is much more information revealed through the set of solutions within a well-balanced range of objective functions as these all ostensibly represent valid experimental designs. In the balance range between 10 and 15, we see more uncertainty about the estimated treatment effect than is unearthed by any one solution, including the solution associated with our lowest objective value. In the experimental simulation that we presented in Section 2.1, 137 out of the 1000 simulations yielded an objective function value that exceeded our minimum value of 10.62. Our best balanced subset, despite being somewhat on the outskirts compared to our simulated experimental data, still comfortably resembles a subset that might have arisen from a statistically valid randomized experiment. At the balance value obtained, this solution is also the subset that embodies the best balance achieved with these data from a procedure that chooses a subset with or without replacement [4–7,25–27].[8]

5. Discussion and conclusion

The causal inference literature strongly favors the outcome with the 'best balance'. However, as we clearly see, the best balance does not necessarily yield the estimate for the average treatment effect that is closest to the true treatment effect. In our simulation, all of the subsets generated were the result of a randomization process. Accordingly, while there are many distinct solutions, they, *together*, yield collective information regarding the phenomenon in question. It is the mean that is an unbiased estimate of the true treatment effect. Any one estimate, no matter how well balanced the control and treatment group are, may not be close to the true treatment effect.

It is also important to note that many levels of balance are indistinguishable. In our simulated experiments, each outcome was the result of a randomization process. The levels of balance in each experiment, as we would expect, are all different, but consistent with a randomization process. Even for the experiment that, in isolation, yields a p-value that implies it may not have arisen from a randomization process, needs to be understood in context. Randomization sometimes produces odd results. If the p-value is .05, then that result may be odd, but we do expect to regularly see that result 1 out of every 20 times that the experiment is conducted. For the other 95%, while both the p-values and the balance differed, they are essentially indistinguishable in the sense that they are all consistent with randomization at the .05-level. We ought to view the experiments as a collective rather than choosing the best balanced experiment and considering it only in isolation, which is especially true when the outcome variable is noisy.

Since randomization is the standard for experiments, any solution that satisfies some randomization threshold should be included in this set. The reported ATE estimate should then be the mean ATE of this set of solutions. The SE may be calculated as the SD of this set of ATEs. It is difficult in the framework of the extant causal inference models to extract this type of information. It is much simpler to achieve these goals within the framework of a computational model. An advantage of our approach is that we realize and utilize more information than other matching methods. A linear model would also use information from more data, but a linear model permits only associational inferences from the unbalanced data set. Our approach first creates balanced data sets and then uses them in a framework that permits causal inferences.

The LaLonde data have been highlighted in many different causal inference analyses. Some of this debate has centered around achieving sufficient covariate balance with matching methods. We have identified the best subsets to date, but more importantly, have identified a large number of subsets that are consistent with randomization, giving us confidence that the NSW program had a positive effect on future earnings.

We have presented an overarching research design framework via design-based approaches that can be incorporated into causal inference analyses. While we relied on BOSS as our method for identifying experiments that are latent in an observational study, our framework is general and not limited to the BOSS methodology. The ideas translate to many matching methods, including designs identified by methods that mimic only block-randomized designs. BOSS automates the process by saving all solutions beyond a specific balance threshold.

It is certainly possible that, despite our best efforts, our analysis does not capture the underlying truth. In an observational study, we must always consider the possibility that unobserved covariates confound the analysis. Estimates from statistical models can certainly be affected when the underlying data are problematic. The methods and design we have presented do not differ from other statistical methods in this regard: statistical modeling cannot solve data woes. Sensitivity analyses that consider the potential impact of unmeasured confounders are still important [1,11,17,21,24]. If these data problems do not exist, however, our methodology provides interesting and new information for making causal inferences.

Indeed, there are multiple avenues for future research. First, the computational approach identifies a large-scale optimization instance for which there is ample room for the development of appropriate and efficient heuristic algorithms. Second, we do not purport that the set of solutions identified consists of independent sets. If they were, we would be able to gain additional leverage toward the identification of the true underlying distribution of the treatment effect. Toward this end, a research direction may be to refine the computational algorithm to limit the extraction of solutions to those that exceed a threshold of independence from previously identified solutions. Another option would be to quantify the new information in each set and then to weight the solutions accordingly. Either advance would add to the idea advanced by this paper – multiple as-if-randomized solutions are better than a single as-if-randomized solution.

Plainly, our approach is computationally intensive. In the course of our research, we have discovered many computational challenges presented by the very large number of essentially equivalent experiment-like designs that are discoverable within an observational study. The computational burden is non-trivial. At the same time, we enthusiastically welcome the challenges because they illuminate opportunities to increase our understanding of how to obtain causal inferences. We know with certainty that computing power is on the rise (along with easier parallelization and hardware advances), so enumerating and extracting insights from permutations of many solutions become faster and simpler every day. We embrace this sign of the times and present an analysis of the Lalonde data to illustrate the virtues of integrating statistical modeling into a computational approach.

Notes

1. An exception is Zubizarreta's design, which is more flexible and optimizes directly on particular balance measures. All of the designs are the same, however, in that the implementations of these approaches provide *one solution* to whichever version of the matching problem is posed.

2. We also note that there are situations that are more amenable to causal inference analyses than others. These do not change with our method. For instance, matching methods and our method are more ideal when the control population is much larger than the treatment group size.
3. Other matching technologies match treatment and control units (not subsets) first then assess the success of the matching later by the level of balance achieved. Without knowing how all matching methods perform, it is difficult to assess if balance is good or 'good enough' because the baseline or optimal level of balance in a particular data set is unknown. In BOSS, the goal is optimal balance, not 'good balance'. The optimal level of balance is the baseline or standard for assessing any particular balance level.
4. Note that we are not using randomized experimental data in our analysis, but using only the treatment group from the LaLonde data. We use the LaLonde data because it has been widely used in the literature. This provides a comparison for our model vis-à-vis other models as well as a benchmark estimate from the original data.
5. We presume that there is no spillover between individuals (i.e. we make the stable unit treatment value assumption (SUTVA).
6. We can see from the figure that as balance improves, one standard deviation around the estimate of the treatment effect includes the experimental benchmark, $\hat{\tau} = \$1794$. To be sure, we do not know the true value of the treatment effect in this instance. We refer to the experimental benchmark here simply to offer some guidance, without certitude, and with sufficient wariness.
7. The objective function, meant to measure covariate balance is flexible in the BOSS technology. A researcher can define balance in any way. The BOSS routine will seek to optimize whatever balance measure is given to it. Our particular formulation for the LaLonde data is specified in Equation (1).
8. It is possible that we have not identified the best subsets. Certainly, the optimization procedure can still be and is still being refined.

Disclosure statement

No potential conflict of interest was reported by the author.

References

[1] B.A. Brumback, M.A. Hernan, S.J.P.A. Haneuse, and J.M. Robins, *Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures*, Stat. Med. 23 (2004), pp. 749–767.

[2] W.K.T. Cho, J.J. Sauppe, A.G. Nikolaev, S.H. Jacobson, and E.C. Sewell, *An optimization approach for making causal inferences*, Statist. Neerland. 67 (2013), pp. 211–226.

[3] K.A. Couch, *New evidence on the long-term effects of employment training programs*, J. Labor Econ. 10 (1992), pp. 380–388.

[4] R. Dehejia, *Practical propensity score matching: A reply to Smith and Todd*, J. Economet. 125 (2005), pp. 355–364.

[5] R.H. Dehejia and S. Wahba, *Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs*, J. Amer. Stat. Assoc. 94 (1999), pp. 1053–1062.

[6] R.H. Dehejia and S. Wahba, *Propensity score matching methods for nonexperimental causal studies*, Rev. Econ. Stat. 84 (2002), pp. 151–161.

[7] A. Diamond and J.S. Sekhon, *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies*, Rev. Econ. Stat. 95 (2013), pp. 932–945.

[8] R.A. Fisher, *Design of Experiments*, Hafner, New York, 1935.

[9] B.B. Hansen, *Full matching in an observational study of coaching for the SAT*, J. Amer. Stat. Assoc. 99 (2004), pp. 609–618.

[10] J.J. Heckman and V.J. Hotz, *Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training*, J. Amer. Stat. Assoc. 84 (1989), pp. 862–874.

[11] J.J. Heckman, H. Ichimura, J. Smith, and P. Todd, *Characterizing selection bias using experimental data*, Econometrica 66 (1998), pp. 1017–1098.
[12] P.W. Holland, *Statistics and causal inference*, J. Amer. Stat. Assoc. 81 (1986), pp. 945–960.
[13] D.R. Kinder and T.R. Palfrey, *On behalf of an experimental political science*, in *Experimental Foundations of Political Science*, University of Michigan Press, Ann Arbor, 1993, pp. 1–39.
[14] H.W. Kuhn, *The Hungarian method for the assignment problem*, Naval Res. Logist Q. 2 (1955), pp. 83–97.
[15] R. LaLonde, *Evaluating the econometric evaluations of training programs with experimental data*, Amer. Econ. Rev. 76 (1986), pp. 604–20.
[16] J. Neyman, *On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1923)*, Stat. Sci. 5 (1923 [1990]), pp. 463–480. reprint. Transl. by Dabrowska and Speed.
[17] J.M. Robins, *Association, causation, and marginal structural models*, Synthese 121 (1999), pp. 151–179.
[18] P.R. Rosenbaum, *Optimal matching for observational studies*, J. Amer. Stat. Assoc. 84 (1989), pp. 1024–1032.
[19] P.R. Rosenbaum, *A characterization of optimal designs for observational studies*, J. R. Stat. Soc. 53 (1991), pp. 597–610.
[20] P.R. Rosenbaum, *Choice as an alternative to control in observational studies (with discussion)*, Stat. Sci. 14 (1999), pp. 259–304.
[21] P.R. Rosenbaum, *Observational Studies*, 2nd ed., Springer, New York, 2002.
[22] D.B. Rubin, *Estimating causal effects of treatments in randomized and nonrandomized studies*, J. Educ. Psychol. 66 (1974), pp. 688–701.
[23] D.B. Rubin, *Bayesian inference for causal effects: The role of randomization*, Ann. Stat. 6 (1978), pp. 34–58.
[24] C. Shen, X. Li, L. Li, and M.C. Were, *Sensitivity analysis for causal inference using inverse probability weighting*, Biometr. J. 53 (2011), pp. 822–837.
[25] J.A. Smith and P.E. Todd, *Reconciling conflicting evidence on the performance of propensity score matching methods*, AEA Papers Proc. 91 (2001), pp. 112–118.
[26] J.A. Smith and P.E. Todd, *Does matching overcome LaLonde's critique of nonexperimental estimators?* J. Econometr. 125 (2005), pp. 305–353.
[27] J. Smith and P. Todd, *Rejoinder*, J. Econometr. 125 (2005), pp. 365–375.
[28] Z. Zhao, *Matching estimators and the data from the national supported work demonstration again*, IZA Discussion Papers, No. 2375, 2006.
[29] J.R. Zubizarreta, *Using mixed integer programming for matching in an observational study of kidney failure after surgery*, J. Amer. Stat. Assoc. 107 (2012), pp. 1360–1371.