Matching for DR and EE Impacts

Seth Wayland, Opinion Dynamics, Oakland CA

ABSTRACT

A central problem for impact evaluation is how to minimize bias and model dependence in impact estimates. Precise estimates of a potentially-biased impact estimate can be misleading when the size of bias and model dependence is unknown or unreported. How do we know if the reported impact estimates are accurate? Designed experiments, where the participants are randomly assigned to the treatment or comparison group, offer the reassurance that "over repeated experiments" the control group will be equivalent to the treatment group. However, we only have one instance of the experiment. What if the control and treatment are not balanced? What if the program was not a designed experiment? We can use models to correct for the differences, we can use matching to improve balance and reduce model dependence, or both.

Standard approaches to demand response and energy efficiency program impact analysis use regression models to adjust for the differences between the treatment and comparison groups that are not caused by the program. These methods perform well when the comparison group is equivalent to the treatment group so that the amount of model adjustment is small, the comparison group and treatment group have similar covariate distributions, and there are no outliers in the control or treatment groups. Using matching as non-parametric preprocessing can help reduce uncontrolled variation, identify outliers, help balance covariate distributions, and reduce model dependence and bias. For these reasons, we propose adding a matching step into the impact evaluation process.

Introduction

Impact evaluation is the science of comparing what would have happened in the absence of a program to what actually happened. Our central problem is the fundamental problem of causal inference: for any given unit, we cannot simultaneously observe both the unit's response when treated and it's response when not treated. Paraphrasing Rubin (1974), for a particular unit, the causal effect of a treatment at time t is the difference between what would have happened at time t if the unit was exposed to the treatment and what would have happened at time t if the unit was not exposed to the treatment.

This paper discusses the implications of some of the choices we make as part of that comparison. We focus on the choice of treatment and comparison groups, and how to choose those groups to help minimize bias and variation in the impact estimates. This discussion applies to both observational quasi-experiments which may have selection bias and randomized experiments, which, over repeated experiments, do not. There are many other issues in causal analysis that we do not cover here. See e.g. (Pearl and others 2009) (Lam 2009) (Keele 2014) (Stuart 2010) for much more.

Causal inference using the Rubin Causal Model (Rubin 1974) requires two main assumptions:

- Stable unit treatment value assumption (SUTVA)---The outcomes of any unit (individual) are not affected by the treatment assignment of any other units (individuals). An example where we see a violation of SUTVA is spillover, where comparison group members may e.g. learn about energy saving behaviors by talking to their treatment group neighbors. In the case of the home energy reports used in behavioral programs, the effect of spillover is likely to be so small that it can be safely ignored.

- Ignorable treatment assignment---For every unit, it must be possible that that unit could have been assigned to either treatment or comparison group. Further, that treatment assignment is independent of the outcome, given the covariates. This is sometimes called "unconfounded" or "no hidden bias."

We can help insure ignorability through the combination of preprocessing by matching, followed by modeling. Both help to correct for the dependence of the outcome on covariates other than the treatment assignment. Modeling directly adjusts the estimate of the treatment effect by applying a statistical model to the outcome, making adjustments to the treatment estimate according to the values of covariates. Matching works less directly, by balancing the distribution of covariates between the treated and comparison groups. Balanced groups will only be different from one another on covariates not included in the matching process, and then only inasmuch as those covariates are uncorrelated with the covariates used in matching.

Matching improves the balance of the distribution of all covariates included in the matching. For this reason, we want to include all covariates that are known to be (or could be) related to both treatment assignment and outcomes (Stuart 2010). However, we should not include any covariates that may have been affected by the treatment.

When estimating treatment effects, we select from among a number of possible effects to measure. This selection is dependent on the design and the effect of interest. A few of these are:

- Intention to Treat (ITT), which measures the effect on all units assigned to treatment, whether or not they complete treatment.
- Average Treatment Effect (ATE), which measures the difference in outcomes for the treated group versus the comparison group.
- Average effect of Treatment on the Treated (ATT), which is an estimate of the difference between the outcomes for the treated group, and what those outcomes would have been if they were not treated.
- Sample ATT (SATT), which measures ATT among only those actually in the treatment group during the experiment.
- Feasible Sample ATT (FSATT), which measures ATT among only those actually in the treatment group during the experiment who have good matches in the comparison group.

Established matching methods require evaluators to iterate through: evaluation of the equivalence of the comparison group to the treatment group (balance), followed by the application of a matching method designed to adjust sample size based on a set measure of balance, or a method designed to improve balance based on a chosen sample size. This operation is difficult enough that it is nearly impossible to simultaneously optimize the balance and sample-size. There have been recent improvements to this process, namely a method for finding the balance/sample-size frontier without iteration, recently proposed by King, Lucas, and Nielsen (2015).

Discussion

EE and DR impact evaluations often use matching when the design is quasi-experimental. It can also be useful to perform matching as a part of the modeling done for experimental designs. Consider that a standard EE impact evaluation for a behavioral home energy report program with an RCT design uses a fixed-effect or lagged dependent variable model to help correct for household-specific differences in energy usage and weather response when calculating impacts. The RCT experimental design only requires that we use the difference in mean usage between the treatment and control groups to estimate impacts, but we use the model to improve the impact estimates and shrink the standard error of the estimates. We suggest using matching in addition to the model, because matching can reduce bias introduced by outliers

of poor balance between the treatment and control group. Thus, we propose matching as preprocessing to reduce bias and model dependence in impact estimates.

Matching can improve balance in pre-treatment covariates through selective removal of observations (King et al. 2011). The purpose is to reduce model-dependence and bias in estimates of treatment effects. We want to reduce model dependence because some methods, such as linear regression, can increase bias when the true relationship is non-linear and there are differences in covariate means and variance between the control and treatment groups (Stuart 2010).

We cannot directly control the bias-variance trade-off while matching. According to best practice, we do not include the outcome variable during matching, to insure that we do not cause selection bias. We can optimize on balance-sample size, which is an analogue to bias-variance. Balance is similar to bias because balance, along with variable importance, determine bias (King, Lucas, and Nielsen 2015). Sample size is one of the determinants of variance, so controlling the balance and sample size indirectly controls the bias-variance.

With most matching methods, we choose the sample size in advance (e.g. for 1:1 nearest neighbor matching, the comparison group is *a priori* set to the treatment group size) and we iteratively try to improve balance, sometimes adjusting sample size along the way. Coarsened exact matching improves on this by setting an upper limit to imbalance and selects the best treatment and control group size to achieve that balance. Even better, the matching frontier method selects the best of all possible combinations of treatment and comparison group for every possible sample size, allowing us to select the best possible balance and sample size given the data we have.

As evaluators, we often think of matching as just a way to help alleviate opt-in bias, but in both randomized controlled trials (RCT) and observational quasi-experiments, matching can decrease model dependence and variation when the treatment group is not well balanced to the control group. Inclusion of treatment units that do not have reasonable matches among the available control units yields (unacceptably) high model dependence (Iacus, King, and Porro 2011a). Think of matching as non-parametric preprocessing for causal inference that can be used to identify subsets of the data where impact estimates can be made with reduced model dependence (King, Lucas, and Nielsen 2015). Then report impact estimates (SATT) for the N treated units as a weighted combination of two groups, one where there is a well-matched[1] comparison group (estimate feasible sample average treatment effect on the treated (FSATT)) with N_f treatment units and another having poor matches (estimate non-feasible sample average treatment effect on the treated (NFSATT)) with N_{nf} treatment units. Reporting impact estimates in this way helps to explicate the amount of model dependence in the analysis.

$$\text{SATT} = \frac{\text{FSATT} \cdot \text{N}_\text{f} + \text{NFSATT} \cdot \text{N}_\text{nf}}{\text{N}}$$

Whether the energy efficiency or demand response program is a randomized experiment or an observational quasi-experiment, information about the level of model dependence is valuable when assessing the internal validity of impact estimates.

Steps in matching, from Stuart (2010):

1. Select a distance measure to determine "closeness", whether one unit is a good match for another.
2. Select and implement a matching method, using your definition of "closeness".
3. Assess the balance of the resulting treatment and comparison group. If necessary, return to 1 or 2 and iterate until treatment and comparison groups are well-matched with the largest possible sample size.
4. Use matched data to perform analysis and estimation of the treatment effect.

Important considerations:

- Choice of treatment effect ITT, ATE, ATT, SATT, FSATT

1 We discuss one way to categorize the treatment group into feasible and non-feasible in the Measuring Balance section below.

- Choice of variables to match on -- critical for ignorability
- Choice of distance metric
- Choice of matching method
- Balance checking

Distance Metrics

Exact

Exact matching guarantees that the treatment group is perfectly matched to the comparison group on pre-treatment covariates, creating a dataset with perfect balance. To perform exact matching, first discretize all continuous covariates, then match a treatment unit to a comparison unit only if the levels of all covariates are exactly the same.

The distance metric for exact matching is:

$$D(X_i, X_j) = \begin{cases} 0 \; if \; X_i = X_j \\ \infty \; otherwise \end{cases}$$

Propensity Score

The propensity score distance metric uses a model (usually logistic regression, but alternatively boosted CART, random forests, or generalized boosted models) to calculate a score for how likely each unit is to be in the treatment group. Propensity score matching is widely used in the social sciences and sometimes appears in EE or DR impact analysis.

A recent paper (King and Nielsen 2015) strongly argues against using propensity score distance for matching. Their primary argument is that propensity score distance matching substantially underperforms other matching methods, often actually increasing imbalance and model dependence.

The scores are used to calculate distance between pairs of units. There is a wide variety of ways to calculate this distance (Ho et al. 2011), but the simplest distance metric is:

$$\mathrm{D}(\mathrm{X_i}, \mathrm{X_j}) = |\mathrm{e_i} - \mathrm{e_j}|$$

Mahalanobis Distance

The Mahalanobis distance metric directly calculates a weighted distance between a pair of points in p-dimensional space, where p is the number of covariates used for matching, and the distance is weighted by the inverse of the covariance matrix. Mahalanobis distance suffers from the curse of dimensionality, so adding more covariates can be counterproductive. Stuart (2010) suggests that performance drops off with more than 8 covariates, and when the covariates are not normally distributed.

Matching on the sum of squared differences (Euclidian distance) is a special case of Mahalanobis distance matching with $\mathrm{S} = \mathrm{I}$. Where I is the identity matrix. Glinsmann and Provencher (2013) use Euclidian distance matching with replacement on 12 months of pre-treatment usage in their matching example. Euclidian distance is appropriate when all covariates are on the same scale and have similar range and variance. If one of the covariates has a substantially larger range than the others, the matching will only increase balance in that one dimension.

The Mahalanobis distance metric is:

$$\mathrm{D}(\mathrm{X_i}, \mathrm{X_j}) = \sqrt{(\mathrm{X_i} - \mathrm{X_j})\mathrm{S}^{-1}(\mathrm{X_i} - \mathrm{X_j})}$$

Matching Methods

We highlight a few of the many available matching methods, describing only those we have seen used in evaluation or that we think are especially interesting. For a more complete overview of matching methods, see Stuart (2010).

Exact

Exact matching is the ideal, ensuring that there is no model dependence in the matched units for the covariates included in matching. Unfortunately, exact matching is rarely possible in practice when there are more than a few covariates with few discrete levels, as the probability of finding an exact match quickly decreases as we add more covariates and more levels per covariate. In the rare situation where there are only a few covariates, exact matching may be a good approach. Exact matching is the simplest type of matching, but not applicable for most EE or DR impact analysis since we match on many continuous covariates.

K Nearest Neighbors

K nearest neighbors selects the k closest (defined by the choice of distance metric) comparison units to each treatment unit. This method is straightforward and easy to implement. The comparison group matches for each treated unit are selected in a "greedy" algorithm that moves through the treated group in order and selects the nearest comparison group unit. It is also possible to use an "optimal" algorithm that minimizes a global distance measure such as the average Mahalanobis imbalance or the standardized difference in means. A serious problem with k nearest neighbors, whether using a "greedy" or "optimal" algorithm, is the possibility of very poor matches remaining in the dataset, since the number of treatment and control units is fixed in advance of the matching.

Coarsened Exact

In contrast to k nearest neighbors, Coarsened exact matching fixes an upper limit to balance in advance, and adjusts the size of the comparison (and optionally the treatment) group (Iacus, King, and Porro 2011a). Coarsening combines levels of factor covariates and discretizes continuous covariates, and then applies exact matching. The units are then matched when one or more treatment units and one or more comparison units match exactly. Treatment units without matching comparison units are put into the non-overlap set.

Matching Frontier

A new option is the matching frontier method from King, Lucas, and Nielsen (2015), which calculates the balance and sample size for all possible combinations of treatment and comparison units, and returns the set of units with the best possible balance for every possible sample size. The set of assumptions and constraints are fairly reasonable. The authors of King, Lucas, and Nielsen (2015) have made an R (R Core Team 2015) software package available (King, Lucas, and Nielsen 2014) that calculates the frontier. The software is still nascent, and when it is more mature, it will be interesting to explore the usefulness of this matching method.

Measuring Balance

Measuring balance and assessing treatment-comparison overlap are key, often overlooked steps in the matching process. It is often useful to use two or even all three of the following balance measures to assess the matching, especially when trying more than one matching method. In the example below, we also describe the use of space diagrams to visually assess the comparative effectiveness of different matching methods.

Difference in Means

The most common way to measure covariate balance while matching is to look at the difference in means between the treatment and comparison group for each of the covariates. Stuart (2010) recommends calculating the "standardized difference in means" for each covariate two-way interactions and squares. The standardized difference in means is the difference in means between the treatment and comparison groups, divided by the standard deviation of the covariate in the full treatment group: $\frac{\overline{X}_t - \overline{X}_c}{\sigma_t}$.

Average Mahalanobis Imbalance

For continuous distance metrics, we can use the Average Mahalanobis Imbalance (AMI) metric (King et al. 2011), which is just the mean Mahalanobis distance between all matched pairs:

$$D = \text{mean}_i[D(X_i, X_{j(i)})]$$

We can identify the overlap and non-overlap sets using AMI by defining the non-overlap set as the set of treatment units that are not chosen as a match by any control unit (King, Lucas, and Nielsen 2015). Then run the model analysis separately on the overlap (yielding FSATT) and non-overlap sets (yielding NFSATT) and combine the results according to the definition of SATT from above.

Median L1 Distance

For discrete distance metrics, median L1 distance is the distance between the multivariate histograms of the treatment and comparison groups. Iacus, King, and Porro (2011b) interpret L1: "If the two empirical distributions are completely separated, then $L1 = 1$; if the distributions exactly coincide, then $L1 = 0$. In all other cases, $L1 \in (0,1)$. If say $L1 = 0.6$, then 40% of the area under the two histograms overlap." The median L1 distance is:

$$L1(H) = \frac{1}{2}\sum (l_1 \cdots l_k) \in H | f_{l_1 \cdots l_k} - g_{l_1 \cdots l_k}|$$

The non-overlap set is then all treated units in bins with no comparison units. Estimation of the SATT can then proceed as described in the AMI section above.

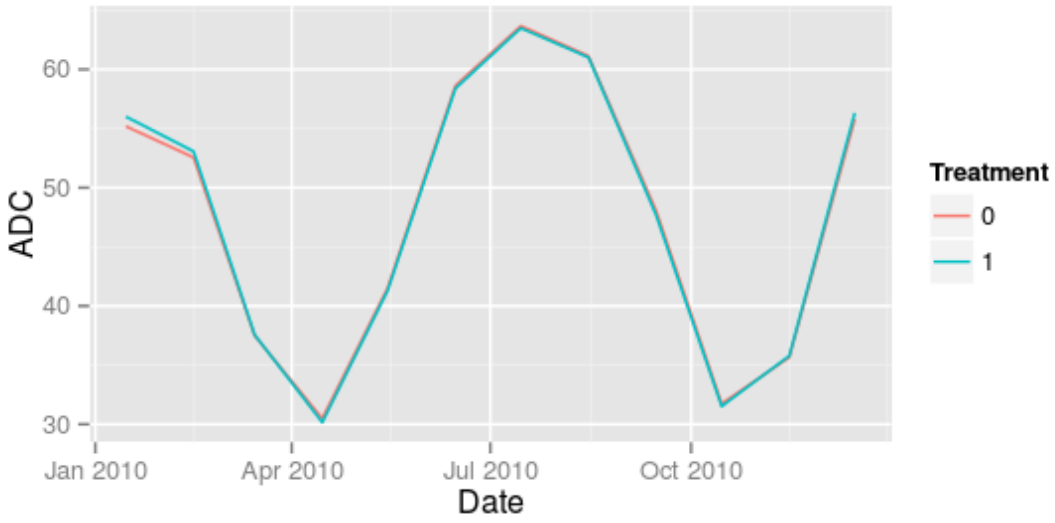
Example

We use a simplified dataset from a behavioral home energy report program where the program was designed as a randomized experiment. We filtered the data to include only 12 months of pre-period data before the start of the program and 12 months of post-period data for the second year of the program for 10,000 treated households and 10,000 control households. This allows us to simplify estimation of the savings impact of the reports, since we don't have to account for the roll out period. For estimates of program savings, we use a simple model of average daily consumption (ADC) that includes a household specific intercept and adjustments for heating and cooling degree days:

$$adc_t = \beta_0 + \beta_i + \beta_{treat} \cdot treat \cdot post + \beta_{HDD} \cdot HDD + \beta_{CDD} \cdot CDD + \varepsilon_t$$

Without Matching

We start with a dataset with an equal number of comparison and treatment group households. Checking balance, we find a median L1 distance of 0.56 and average mean difference of 0.03 kWh/day. For month to month comparison, we plot the monthly mean for the comparison group and the treatment group for the 12 months in the pre-treatment period.



The comparison and treatment groups appear well balanced. It seems that the randomization to comparison and treatment group was quite effective. We will still investigate the effectiveness of matching to see what we can learn. Our first estimate of savings attributable to the program using the full dataset is 4.9%.

K Nearest Neighbors

We use 1:1 matching with two different distance measures to look at how these two distance measures affect the balance, sample size and savings estimates. We match with replacement, which allows us to drop comparison group households that do not have a good match in the treatment group while retaining all treatment group households. The savings estimates are SATT estimates.

First, with a propensity score distance metric, we match the 10,000 treatment group households with 5,580 comparison households. The median L1 distance is now 0.90, and the average mean difference is -0.22, so according to these metrics, the balance is actually worse than in the original dataset. The savings estimate for this propensity score matched dataset is 4.8%, so the estimate moved very little.

Second, using the Mahalanobis distance metric, we match the 10,000 treatment group households with 5,762 comparison households. The median L1 distance is now 0.66, and the average mean difference is now 2.9! What happened here to make the balance so much worse? We are using Mahalanobis distance matching on 12 covariates (monthly usage for 12 months prior to the program start), more than the recommended maximum of 8 from Stuart (2010), so this matching is suffering from the curse of dimensionality, which makes the balance almost 100 times worse when measured as average difference of covariate means. The savings estimate is now 5.0%.

It is important to realize that k-nearest neighbors matching is not guaranteed to improve balance, so it is extremely important to check balance before and after matching to see if the balance has improved. If it has not, it is better to either try a different matching method, such as coarsened exact that is more likely to improve the balance, or use the full, unmatched dataset if it is sufficiently balanced.

Coarsened Exact Matching

Coarsened exact matching allows us to gain some more insight into which treatment households are well matched, and which are not. Here, we will measure the FSATT and the NFSATT, combining them into a SATT. For the feasible sample, coarsened exact matching yields 9,408 treated households and 9,355 comparison households. The median L1 distance for this group is 0.09 and the average mean difference is 0.37. The estimated FSATT savings of 4.3% is lower than the original SATT of 4.8%.

Combining the FSATT with the NFSATT, which is 9.6%, yields a SATT of 4.6% which is just lower than the SATT estimate from the full dataset.

The insight that we get from using CEM is that the 592 unmatched treatment households appear to have substantially more savings than the rest of the treated households, but this increased savings is highly model dependent because these treatment households do not match well with the comparison group, so we rely on the model to reduce bias. The 9,408 matched treatment households in the FSATT estimate have less model dependence, so we model them separately from the unmatched treatment households. For this reason, we would prefer to report the 4.6% SATT savings from the weighted combination of FSATT and NFSATT as a better representation of the overall program savings, rather than the 4.8% from the full dataset.

Conclusion

Matching, when performed carefully, substantially improves our understanding of our treatment effect estimates (impact estimates), and can substantially reduce bias in those estimates. However, when it is performed without balance checks and careful consideration of matching method and distance metric, it can increase bias and variation in causal estimates.

This example only covers one possible case, where the treatment and comparison groups are well balanced as part of an RCT. Matching can actually be more valuable in other situations, for randomization failure when the comparison and treatment groups are not balanced, or for quasi-experimental designs where the comparison group must be selected from a larger pool of non-treated units. In these more extreme cases, the matching method and balance checks are even more important because the bias of the impact estimates is primarily due to the selected comparison group.

References

Glinsmann, Bethany, and Bill Provencher. 2013. "I Can't Use a Randomized Controlled Trial - NOW WHAT? Comparison of Methods for Assessing Impacts from Opt-in Behavioral Programs." In *2013 International Energy Program Evaluation Conference, Chicago*.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software* 42 (8): 1–28. http://www.jstatsoft.org/v42/i08/.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011a. "Causal Inference Without Balance Checking: Coarsened Exact Matching." *Political Analysis*. http://gking.harvard.edu/files/gking/files/political_analysis-2011-iacus-pan_mpr013.pdf.

———. 2011b. "Multivariate Matching Methods That Are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106 (2011): 345–61. http://gking.harvard.edu/files/gking/files/cem_jasa.pdf.

Keele, Luke. 2014. "The Statistics of Causal Inference: The View from Political Methodology." http://www.personal.psu.edu/ljk20/causal.pdf.

King, Gary, and Richard Nielsen. 2015. "Why Propensity Scores Should Not Be Used for Matching." http://gking.harvard.edu/files/gking/files/psnot.pdf.

King, Gary, Christopher Lucas, and Richard Nielsen. 2014. *MatchingFrontier: R Package for Calculating the Balance-Sample Size Frontier*. http://gking.harvard.edu/publications/matchingfrontier-R-Package-Calculating-Balance-Sample-Size-Frontier.

———. 2015. “The Balance-Sample Size Frontier in Matching Methods for Causal Inference.” http://gking.harvard.edu/files/gking/files/frontier_2.pdf.

King, Gary, Richard Nielsen, Carter Coberley, James E. Pope, and Aaron Wells. 2011. “Comparative Effectiveness of Matching Methods for Causal Inference.” http://gking.harvard.edu/files/gking/files/psparadox.pdf.

Lam, Patrick. 2009. “Causal Inference.” http://www.people.fas.harvard.edu/~plam/teaching/methods/causal/causal_print.pdf.

Pearl, Judea, and others. 2009. “Causal Inference in Statistics: An Overview.” *Statistics Surveys* 3. The author, under a Creative Commons Attribution License: 96–146. http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66 (5). American Psychological Association: 688. http://www.biostat.jhsph.edu/~dscharf/Causal/rubin.journ.psych.ed.pdf.

Stuart, Elizabeth A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science*. http://biostat.jhsph.edu/~estuart/Stuart10.StatSci.pdf.