Tim Vigers

3/26/19

Consulting 2

Group Project

*Propensity Scores*

*Introduction*

Randomized controlled trials (RCTs) ensure by design that treatment groups have similar baseline characteristics, but these characteristics can differ significantly in observational studies (for example through participant self-selection). It is important to account for these differences before making inferences on treatment effect based on observational data, since one cannot compare outcomes directly as with a RCT. One common method for dealing with this issue is regression adjustment. [1] Propensity scores (PS), or the probability that a participant is assigned to a particular treatment group given their observed baseline characteristics, are simply another way to “remove bias due to all observed covariates.” [2]

There are a few reasons that PS are used in place of regression adjustment. First, PS allow for separation between the design and analysis phases of the study because “a matched, stratified, or weighted sample can be constructed without any reference to the outcome.” [1] Also, research has shown that PS are more effective than regression when the outcome under study is rare [3] and slightly better for estimating hazard and odds ratios. [1]

However, once a researcher has decided to use a PS approach rather than regression adjustment, there are several questions regarding its implementation. First, how should one estimate the PS (which baseline covariates should be considered)? Next, how should one evaluate this score estimate? Finally, what is the best algorithm for matching treatment groups based on PS, and how should one assess match quality? [4]

*Methods*

Score model

The most common method for estimating the PS is using a logistic model regressing treatment group on a set of baseline characteristics. [1] Lee et al. and Setoguchi et al. examined logistic regression alternatives such as classification and regression trees (CART), random forests, and other machine learning or data mining techniques. [5, 6] However, many of these potential replacements were evaluated using c-index, which “provides no information as to whether the PS model has been correctly specified.” [1]

In theory, matching participants by true PS will ensure that baseline covariates are equally distributed between groups (i.e. independent of treatment), so evaluation of the score model should assess to what extent this is achieved. One method is to look at the standardized difference, which converts difference in means to units of pooled standard deviation. It is therefore unaffected by sample size and allows for comparison between variables measured in different units, although there is no consensus on a useful threshold for this measure. [1]

If systematic differences in potential cofounders remain (i.e. the groups are unbalanced) after conditioning on PS, then the score model requires adjustment. Assuming a logistic model, this correction can be performed using standard model building techniques (addition of interactions, non-linear terms, etc.), although Austin strongly discourages significance testing as p values are confounded by sample size, and suggests looking at overall group balance instead. [1]

Matching

Assuming one is able to develop a reasonable model for PS, the next step is to determine the appropriate matching algorithm. The three most common methods are nearest neighbor (NN), caliper/radius, and stratification. Of these, the easiest and most common is NN, where each participant from the control group is matched with the participant in the treatment group with the closest PS (or vice versa). This can be done with or without replacement, and is not limited to 1:1 matches. [4]

Caliper matching is essentially the same as NN, but with the additional constraint that the pairs must be within a certain range (the caliper) of one another. This also ensures closer matching but can lead to reduced sample size, as unmatched participants must be excluded from the analysis. Radius matching is a variant of caliper matching which uses “not only the NN within each caliper but all of the comparison members within the caliper.” [4]

Stratification by PS was proposed by Rosenbaum and Rubin in their original paper on PS. Participants are ranked based on their score, and then the cohort is divided into several strata. [2] Inference can then be made within in each PS interval, then these results can be pooled to estimate an average treatment effect. As one might expect, there is some debate surrounding the appropriate number of strata, although 5 is a common rule of thumb. [1]

Logistic regression for score modeling can be performed in base SAS or R, and matching can be done in SAS, or in R using the MatchIt [7] or Optmatch [8] packages. See Austin (2014) for additional information on matching algorithms.

Results

PS are increasing in popularity, [9] and appear to be particularly useful in estimating causal effects where it’s difficult or impossible to design an ethical RCT. Anecdotally, it’s most commonly used in the behavioral sciences for this reason.

For example, Ye and Kaskutas used PS to examine the effectiveness of alcoholics anonymous (AA). There is a strong selection bias in studying AA, so the authors use multivariate logistic regression to develop PS “based on known confounders including motivation, problem severity, and prior help-seeking.” [10] PS obviously cannot account for unknown or unobserved confounders, but nor can techniques like linear regression adjustment. For this study, the authors used both matching and stratification of PS, both of which resulted in the same conclusions. This was by design because “there is no single optimal approach to apply the PS method,” but “the use of these two techniques suggests the robustness of the finding.” [10]

Recommendations

Although PS requires additional research to determine the best approach for various situations, there are a few useful rules of thumb that can help applied researchers use the technique. First, for now logistic regression on known and potential confounders appears to be the best approach for modeling scores. Machine learning techniques are exciting, but still require somewhat specialized knowledge and software, and have not been extensively studied in this area.

The Ye and Kaskutas approach to matching is also relatively easy and can protect against the variance/bias tradeoffs of different matching algorithms. If the conclusions about causal effect are the same regardless of how the participants are matched, it seems safe to feel reasonably confident in the results. However, if the matching algorithm significantly affects study conclusions, that would suggest that perhaps something is wrong with the score model, or that there are unmeasured confounders which may require a completely different statistical approach. References

1. Austin, P.C., *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.* Multivariate Behav Res, 2011. **46**(3): p. 399-424.

2. Rosenbaum, P.R. and D.B. Rubin, *The central role of the propensity score in observational studies for causal effects.* Biometrika, 1983. **70**(1): p. 41-55.

3. Braitman, L.E. and P.R. Rosenbaum, *Rare outcomes, common treatments: Analytic strategies using propensity scores.* Annals of Internal Medicine, 2002. **137**(8): p. 693-695.

4. Caliendo, M. and S. Kopeinig, *Some Practical Guidance for the Implementation of Propensity Score Matching.* Journal of Economic Surveys, 2008. **22**(1): p. 31-72.

5. Lee, B.K., J. Lessler, and E.A. Stuart, *Improving propensity score weighting using machine learning.* Stat Med, 2010. **29**(3): p. 337-46.

6. Setoguchi, S., et al., *Evaluating uses of data mining techniques in propensity score estimation: a simulation study.* Pharmacoepidemiol Drug Saf, 2008. **17**(6): p. 546-55.

7. Ho, D.E., et al., *MatchIt: Nonparametric Preprocessing for Parametric Causal Inference.* Journal of Statistical Software, 2011. **42**(8).

8. Hansen, B.B. and S.O. Klopfer, *Optimal full matching and related designs via network flows.* Journal of Computational and Graphical Statistics, 2006. **15**(3): p. 609-627.

9. Austin, P.C., *A comparison of 12 algorithms for matching on the propensity score.* Stat Med, 2014. **33**(6): p. 1057-69.

10. Ye, Y. and L.A. Kaskutas, *Using propensity scores to adjust for selection bias when assessing the effectiveness of Alcoholics Anonymous in observational studies.* Drug Alcohol Depend, 2009. **104**(1-2): p. 56-64.