Some Reflections on Rosenbaum and Rubin’s Propensity Score Paper

Roderick J. Little

Observational Studies, Volume 9, Issue 1, 2023, pp. 69-75 (Article)

Published by University of Pennsylvania Press

DOI: https://doi.org/10.1353/obs.2023.0006

For additional information about this article
https://muse.jhu.edu/article/877888
Some Reflections on Rosenbaum and Rubin’s Propensity Score Paper

Roderick J. Little
rlittle@umich.edu
Department of Biostatistics
University of Michigan
Ann Arbor, MI USA

Abstract

Rosenbaum and Rubin’s paper is highly cited because the basic idea is simple and insightful, and it has applications to important practical problems in treatment comparisons with observational data, and selection bias and nonresponse in surveys. I discuss several issues related to the method, including use of the propensity score for weighting or prediction, and two robust methods that use the propensity score as a covariate and can be more efficient that weighting when the weights are highly variable, namely Penalized Spline of Propensity Prediction (PSPP) and Penalized Spline of Propensity for Treatment Comparisons (PENCOMP). Approaches to addressing highly variable weights are discussed, including omitting variables in the propensity model that are unrelated to outcomes, and redefining the estimand.

Keywords: confounding by indication, nonresponse modeling, penalized spline of propensity, robust causal inference

Introduction

I was a colleague of Paul Rosenbaum and Don Rubin when their paper (Rosenbaum and Rubin, 1983) (henceforth RR) was under review for Biometrika. In 1980, David Rosenbaum (no relation of Paul) was the Deputy Assistant Administrator for Radiation Programs at the Environmental Protection Agency in Washington, DC. A pioneer in monitoring levels of household radon and a person with expansive ideas, David had convinced Rubin to form a group in his program that was going to MSG (make statistics great) at EPA. At that time I was working at World Fertility Survey in London, but that project was starting to wind down, and I joined Rubin in Washington for the chance to work with the guru of missing data – I must admit I had the vaguest of ideas about what the job actually entailed. A number of us, including Paul Rosenbaum and myself, assembled in Crystal City in the fall of 1980. Ronald Reagan also became U.S. President at that time, and environmental protection was not high on his list of priorities. So Rubin’s statistics group was dead in the water, and his government statistics career essentially over before it had started. He spent a lot of time trying to find alternative employment for the team he had assembled. The job was a bust, but the group had a lot of fun together, and by a stroke of serendipity, Rubin and I started working on a book on missing data (Little and Rubin, 1987), which now has about the same number of google scholar citations as RR. So things worked out well for me in the end.

I like simplicity (Little, 2013), and one reason why RR is highly cited is that the basic idea is tantalizingly simple – a natural reaction is “why didn’t I think of that?” It is harder to generate simple, novel and useful ideas than elaborations of a complex statistical method that may receive limited use in practice.

The propensity score is very widely used in applications because it addresses three important applied problems – selection bias, nonresponse, and bias in the assignment of treatments, through estimating and adjusting for the propensity of these quantities. The general underlying framework is laid out in a key seminal paper (Rubin, 1978). Another useful feature of the propensity score is
that the balancing property of the propensity score is easily checked, by comparing distributions of observed variables within propensity score categories. RR concerned a relatively novel aspect of robust statistical modeling, because it focuses on limiting misspecification of how outcomes depend on a set of covariates, rather than on outliers and the distribution of errors, which was the main focus the robustness literature in the 1960s and 1970s (e.g. Andrews and Hampel (1975)). Thus, the standard approach to assignment bias due to observed confounders was multiple regression of the outcome on treatment indicators and other confounding covariates. This approach is vulnerable to misspecification of the regression model, in ways that may be hard to detect with a large number of predictors. By reducing a set of covariates to a single combination, namely the propensity, it allows the injection of robustness into the analysis, by modeling a flexible relationship between the propensity score and study outcomes. RR has had a big influence on my statistics research, along with two other papers by Rubin, his development of calibrated Bayes as an inferential philosophy (Rubin, 1984; Little, 2006) and his paper (Rubin, 1978) on the Bayesian justification of randomization. In this commentary I discuss some of my work in this area.

Survey Nonresponse: Response Propensity Weighting

RR played an important role in my research on survey nonresponse. A canonical problem is unit nonresponse in surveys, where data are available for respondents on a set of survey outcomes and on both respondents and nonrespondents for a set of variables. In some settings values of some components of $X$ is available for the whole population, but I don’t focus on this distinction here. In David et al. (1983) the ideas of RR were applied in the setting of survey nonresponse rather than treatment assignment. Let $R$ denote the response indicator, taking value 1 for respondents and 0 for nonrespondents. The propensity to respond is estimated by a regression of $R$ on $X$, using a model such as logistic regression appropriate for a binary outcome. The estimate of the probability of response is computed for each respondent, and the inverse of this probability becomes a nonresponse weight. In response propensity (RP) stratification, categories of the estimated response probabilities are formed, on the nonresponse weight is the inverse of the sample response rate within these categories. This modification allows very large weights to be avoided by judicious choice of categories.

Response propensity weighting can reduce bias due to $X$ when it is related to the survey variable, but when $X$ is predictive of response (so that the weights are variable), but the response propensity is weakly related to the survey outcome, then weighting reduces precision with no compensating reduction in bias (Little and Vartivarian, 2005). One approach to reducing unnecessary variability in the weights is to restrict the predictors in response propensity model to variables that are predictive of the main survey outcomes. For an application of this idea, see Morral et al. (2014). Another approach is to simply drop respondent cases with low estimated response propensities and hence high weights, which may have undue influence on survey estimates. This improves precision, but effectively restricts the estimand to a subpopulation of units that is more likely to respond. A further approach to limiting the effects of weight variability is discussed in the next section.

The response propensity can be thought of the combination of the X’s that is most different between respondents and nonrespondents. An interesting alternative is the combination of X’s that is most predictive of a survey outcome $Y_j$, which can be estimated by a regression of $Y_j$ on $X$ based on respondent data. In predictive mean (PM) stratification, adjustment cells are based on categorizing the predictive means. Mean squared error properties of RP and PM stratification for inferences about survey means are compared in Little (1986). Both approaches limit nonresponse bias, if missingness is missing at random (MAR) (Rubin, 1975). PM has the advantage that unlike RP it also controls the variance and hence the mean squared error of estimates of means. However, RP stratification has the practical advantage that it leads to a single set of weights, whereas the weights in PM stratification are different for each variable $Y_j$. 

70
Some Reflections on Rosenbaum and Rubin’s Propensity Score Paper

I think both the RP and PM dimensions are important for nonresponse adjustment. The response propensity is a key tool in bias adjustment, but the predictive mean brings in the relationship between the response propensity and particular survey outcomes. The PM combination of $X$’s plays a key role in proxy pattern-mixture analysis, a method for assessing sensitivity to deviations from MAR (Andridge and Little, 2011). Vartivarian and Little (2003) considers jointly stratifying on RP and PM.

Penalized Spline of Propensity Prediction

The standard approach to unit nonresponse adjustment uses the inverse of the estimated propensity as a weight. From my (calibrated) Bayesian prediction perspective, it is more natural to treat the propensity as a covariate in a model to predict survey nonrespondents. What are the pros and cons of weighting and prediction?

For a single categorical predictor, weighting by the inverse response rate in a cell is equivalent to regressing on the set of dummy variables for the cells. In other settings with continuous covariates the methods differ. In the nonresponse setting, suppose we estimate the propensity to respond by a regression of the response indicator on a set of covariates, and compare estimates of the mean of a survey variable $Y$ by (a) weighting respondents by the inverse of the estimated response propensity, or (b) predicting nonrespondent values of $Y$ by linear regression on the estimated propensity to respond. Both approaches adjust appropriately for nonresponse bias due to the covariates, if the relevant models are well specified. The weighting approach is a bit simpler, because the weights are the same for all survey variables $Y$. The prediction approach is more nuanced, because the regression coefficient of the estimated propensity takes into account the degree of association between the propensity and $Y$. In particular, as noted above, if the covariates $X$ are strong predictors of the propensity but the propensity is a weak predictor of $Y$, weighting or prediction is not needed to adjust for bias, and the weighted mean has higher mean squared error than the unweighted mean, that is, weighting makes the estimator worse (Little and Vartivarian, 2005). Prediction is more efficient than weighting, because when the association is weak, the small resulting coefficient of the regression on the propensity dampens the adjustment.

If the linear regression model on the propensity is misspecified, then it doesn’t completely eliminate bias in the estimated mean. In fact, the potential bias from misspecification is most serious when the method has the greatest potential, namely the predictors are strongly related to both nonresponse and the survey variable. This motivates Penalized Spline of Propensity Prediction (PSPP) (Little and An, 2004; Zhang and Little, 2009), which regresses the survey variable on the penalized spline of the estimated propensity, thus providing a more flexible relationship between the variables. Other types of splines could be fitted, but I like the penalized spline because, it has a Bayesian interpretation via a mixed model with random regression coefficients and is easily fit with readily available mixed-model software.

What if the propensity is a weak predictor of the survey variable, but a different combination of the covariates is a strong predictor? For example there are two covariates $X_1$ and $X_2$, the logit of the propensity is linear in $X_1 + X_2$, and the best predictor of $Y$ is $X_1 - X_2$. Weighting or prediction are both inefficient here (weighting more so than prediction). Weighting can be improved by augmented inverse-probability weighting (AIPW), which regresses on the covariates and applies weights to the residuals. PSPP simply adds strong predictors as additional covariates in the regression model for the survey outcome, along with the penalized spline of the propensity. Including all the predictors leads to issues with multicollinearity, so one of the predictors typically needs to be dropped.

PSPP has a potential advantage in efficiency over AIPW in this setting – if the added covariates in PSPP yield close to best predictors of $Y$, then the residuals are $Y$ are weakly related to covariates and hence to the propensity. Weighting the residuals by the inverse propensity is less efficient than PSPP. In simulations I have found PSPP to be similar to AIPW in terms of bias reduction, but
potentially more efficient, particularly when the weights in AIPW are very variable (Zhou et al., 2019, 2021).

An interesting feature of the PSPP model is that the regression coefficients in the propensity are estimated, and hence are subject to error. The Bayesian version of PSPP propagates this uncertainty by including a prior distribution for these unknown regression coefficients in the propensity model. An alternative approach which seems to work well is to multiply impute nonrespondent values of $Y$, with each set of imputations based on the propensity model applied to a bootstrap sample of the observations.

Another interesting feature of PSPP is that it has a double robustness (DR) property, in that it yields consistent estimates of the mean if either (A) the prediction model is correctly specified, or (B) the propensity model is correctly specified and (C) the penalized spline correctly captures the relationship between the survey variable and the propensity; misspecification of the regression on other covariates does not yield bias, because of the balancing property of the propensity score in RR. This DR property is not quite as strong as the well-known DR property for AIPW, which requires only (A) and (B). But arguably the additional condition (C) is rendered innocuous by the flexibility of the spline. The Bayesian version of PSPP thus has a form of double robustness without the need for weights, and it tends to have good confidence coverage under weak priors for the parameters. I like the conceptual simplicity of the method, and the lack of weights avoid the messy practical issues of how to deal with extreme weights than can lead to noisy estimates.

PSPP can also be used to handle probability sampling with unequal probabilities of selection, where the selection probabilities are known (Zheng and Little, 2003, 2005; Zangeneh and Little, 2015).

**PSPP for Treatment Comparisons – PENCOMP**

Penalized Spline of Propensity for Treatment Comparisons (PENCOMP) (Zhou et al., 2019) applies the same basic model as PSPP to handle bias in the allocation of treatments due to observed confounders. The Neyman/Rubin causal model with $T$ treatments defines the $T$ outcomes under each treatment, with the $T-1$ outcomes corresponding to the treatments not assigned being viewed as missing data. A model involving the spline of the estimated propensity to be allocated treatment $t$ and other covariates is then applied to predict the outcomes under treatments not assigned, and hence to estimate causal effects. A convenient implementation is to multiply impute the outcomes under treatments not assigned, and then base inference on multiple imputation combining rules (Rubin, 1987). For cross-sectional observational studies this yields a robust approach to estimating treatment effects.

PENCOMP can also handle confounding by indication, when treatments are assigned at more than one time points, and outcomes from treatments at an intermediate time point are used to determine treatment allocations at that time point. Simple regression methods do not apply because the intermediate outcomes are both outcomes (of initial treatment assignment) and confounders (of later treatment assignment.) The key idea is to apply the Neyman/Rubin causal model to define outcomes to alternative treatments under both intermediate and final outcomes (Frangakis and Rubin, 2002). With two treatments on one intermediate outcome, there are two possible outcomes for the intermediate time point, one actually observed corresponding to the treatment assigned, and four possible outcomes at the final time point, of which one is observed corresponding to the treatment combination actually assigned. So there are a lot of missing data! But multiply-imputing the missing data and applying MI combining rules provides valid and robust inferences that compare favorably with weight-based alternatives in simulation studies (Zhang and Little, 2009, 2011; Yang and Little, 2015).
Positivity and Propensity Imbalance Across Treatments

A key assumption of propensity methods is positivity – individuals have to have a positive probability of being assigned to all of the treatments being compared. In observational studies this is far from a minor assumption – for particular combinations of covariates, it is common for one or more of the compared treatments not to be assigned at all. If the estimated propensity to be assigned a particular treatment is zero, clearly a weight cannot be assigned. If the propensity model yields very low propensities, they receive very large weights that can lead to very inefficient weighted estimates.

PSPP or PENCOMP somewhat ameliorates the problem of variable weights for reasons discussed above, but cannot be said to resolve them. If the distribution of propensities differs substantially across two treatment groups, then the penalized spline model can still be fitted, but predictions involve extrapolation of the model into regions where there is little data, leaving results that are vulnerable to model misspecification. Thus, tools to reduce disparities in the propensity score distributions are important.

The task of PENCOMP is to adjust for observed confounders, where a confounder is a pre-treatment variable that is related to both treatment assignment and the outcome. Thus, variables that are (A) related to the outcome but not treatment assignment, or (B) related to treatment assignment but not to the outcome, are not true confounders. Adjusting for type (A) variables in the outcome model does not reduce bias but can increase precision, by reducing the residual variance. Adjusting for type (B) variables does not reduce bias but can reduce precision, particularly if included in a weighting adjustment. If PENCOMP is the method of adjustment, then including type (B) variables in the propensity model increases disparities in the propensity score distribution between treatment groups, which is still an undesirable effect. Simulations in Zhou et al. (2021) show substantial improvements in PENCOMP estimates if type (B) variables are removed from the propensity model, or down-weighted using shrinkage methods such as the LASSO.

Applying such methods after data collection is subject to criticisms of “data snooping”, and potentially violates the important principle of applying thinking behind randomized trials to the analysis of observational studies; see for example Rubin (2007). However, biasing treatment effects can be minimized by prespecifying methods in detail in a study protocol before data collection, as is done in randomized clinical trials, and assessing potential confounders using a model that does not include treatment indicators as covariates.

The usual estimand for causal comparison is the ATE, the average treatment effect for the entire population of interest. However, when the degree of overlap in the propensity score distributions between treatment groups is limited, it may make sense to restrict the inference to subpopulations that do not have extreme response propensities close to zero or one. For example, Gutman and Rubin (2013) proposed dropping units outside of the overlap region of estimated propensity scores between treatment groups. In fact, the precise nature of a restricted estimand is unclear when sample cases are excluded based on estimated propensities, which vary in repeated sampling. However, restricting the estimand by dropping cases with extreme weights seems more defensible here than in the case of nonresponse discussed in Section 2, because, as Imbens and Wooldridge (2009) argue, the focus on the ATE estimand can be unrealistic, and comparisons might usefully be restricted to subpopulations where all compared treatments have a reasonable chance of being assigned.

These considerations become much more significant in longitudinal studies involving treatment allocations at multiple time points. In the AIDS application described by Zhou et al. (2019), data are available at 16 time points, and either of two possible treatments could be assigned at each time point. Thus there are over 30,000 (2^{15}) possible treatment combinations, nearly all of which are not seen in the data! Providing simple and interpretable causal conclusions in such a setting requires careful thought and modeling.
References


