Commentary on Rubin and Rosenbaum Seminal 1983 Paper on Propensity Scores: From Then to Now

Usha Govindarajulu

Observational Studies, Volume 9, Issue 1, 2023, pp. 19-22 (Article)

Published by University of Pennsylvania Press

DOI: https://doi.org/10.1353/obs.2023.0000

For additional information about this article
https://muse.jhu.edu/article/877882
Commentary on Rubin and Rosenbaum Seminal 1983 Paper on Propensity Scores: From Then to Now

Usha Govindarajulu  usha.govindarajulu@mountsinai.org
Center for Biostatistics
Department of Population Health Science
Icahn School of Medicine, Mount Sinai
New York, NY 10029, USA

Abstract

Rubin and Rosenbaum (1983) wrote about the theory and application of “propensity scores” in their landmark paper. Since that time, the method has still been in use or adapted for use in various contexts. In this commentary, I discuss their original paper and the latest in terms of criticisms and defense of the use of some of the theory they proposed for propensity score matching. Although the commentary is not exhaustive, I try to highlight important aspects of their theory as well as points made later for and against some of their originally proposed theory.

Keywords: Propensity score matching, observational, treatment assignment, balancing score, subclassification, weighting

Rubin and Rosenbaum’s seminal 1983 paper on propensity scores for balancing observational non-randomized data due to potential imbalance between groups has been an incredibly useful method for improving the analyses of observational studies. Even though the article was published in 1983, I did not become aware of the use and existence of the method until well into the 2000s and for that matter, not even until 2011. This is also potentially because of it becoming more conventional in computing in that decade. Since this method was not something I had come upon until then, I just thought that the best I could do was adjust for confounding by covariate adjustment in regression models and/or carefully use variable selection. Ironically, I had been working with cardiologists on some research who were fairly statistically savvy and they were apt to use the propensity score method in their modeling. From then on, I started really thinking about how to use the method in other contexts and how it was developed. Even until recently, not everyone is aware of the method and/or how to employ it which is often a conundrum for them to understand how it works.

Prior to their article coming out in 1983, there was no well-structured or well described way of adjusting for potential imbalance in an observational study when comparing two main treatment groups or other groups. Rubin and Rosenbaum (1983) methodologically describe with pain-staking effort their theory for propensity scores to be used as they describe in three areas: matched sampling, subclassification, and covariance adjustment. The paper is very forward thinking and shows that the authors were light-years ahead in their thinking to problems often faced by analysts in observational studies. The strategy of their theory is methodologically thought out by the authors. It is also clear that the paper had long been in making due to prior publications in development of the theory from Rubin starting from 1973 up until this bombshell manuscript. Clearly Rubin had been thinking for well over a decade about the problems in observational studies, matching, and causal inference, and worked with Rosenbaum, who appeared to have connected with Rubin when they worked at the Educational Testing Service (ETS), on this seminal paper. It was also a great jump start for Rosenbaum’s career in this area.

©2023 Usha Govindarajulu.
Govindarajulu

They in effect proposed a balancing score, which was a function of the observed covariates, \( x \). They then defined the conditional probability of assignment to treatment one (where \( z \) denotes treatment), given the covariates.

\[
e(x) = \text{pr}(z = 1 \mid x) \tag{1}
\]

This essentially became their propensity score, because by definition, it the propensity towards being exposed to treatment 1 given the observed covariates and the propensity score in effect is “coarser” then the balancing score, which is finer. This can be estimated by logistic regression but also be estimated by a probit model as well as by a discriminant function. Taking from ideas about randomized trials, in which they differ from randomized trials, where \( z \) and response \((r_1, r_0)\) are known to be conditionally independent given \( x \) and every unit in the population has a chance of receiving the treatment, they then stated that the treatment assignment is strongly ignorable given a vector of covariates, \( v \), if

\[
(r_1, r_0) \mid z, v, 0 < \text{pr}(z = 1 \mid v) < 1 \tag{2}
\]

for all \( v \). They went on to conclude that when treatment assignment is strongly ignorable, when Eq 2 holds with \( v=x \), they can more simply say that treatment assignment is strongly ignorable. This last argument went on to serve as a foundation to their theory on the propensity score, especially as stated in their theory that if treatment assignment is strongly ignorable given \( x \), then it is strongly ignorable given any balancing score. Essentially they stated in theorem 4 in their paper that if treatment assignment is strongly ignorable, then the expected difference in observed responses to two treatments at \( b(x) \) would be equal to the average treatment effect at \( b(x) \) where:

\[
E\{ r_1 \mid b(x), z = 1 \} - E\{ r_0 \mid b(x), z = 0 \} = E\{ r_1 - r_0 \mid b(x) \} \tag{3}
\]

This was perhaps a strong assumption but also what allowed this method to work in practice. They employed this in the next several theorems or corollaries in the paper. Once they formalize this they go on to discuss three standard techniques for adjustment: matched sampling, subclassification, and covariance adjustment.

The section on matched studies in their paper was straightforward and probably the most simplistic of their approaches for using the propensity scoring for matching (PSM). Interestingly enough, they relegated the Mahalonobis metric matching to not be as useful in matching as propensity scores are, especially in cases of a rare binary variables where it is difficult to match treatment and control groups individually. Also, ironically, Rubin himself had produced an article about bias reduction using Mahalanobic metric matching (Rubin, 1983b). They later addressed it in their subsequent work after this paper. Furthermore, later on, Mahalonobis distance matching ends up being used together with propensity scoring for an alternative specification to matching and one can find it as an option in several of the major statistical softwares that have functions for PS matching available.

In their next section they discussed subclassification. This is also fascinating that they explored how to deal with data that had strata and how to balance the data amongst these as well, since as they pointed out some studies have experimental and control groups divided on a main predictor into subclasses or strata. A useful insight they had is that as the number of confounding variables increases that this leads to a dramatic increase in subclasses for the variables and then not all subclasses will contain both treated and control units. They then suggested that subclassification on the propensity score is a natural way to avoid the problem.

In their last section, the final discussion about covariance adjustment appeared to be more of a discussion then a set of theorems and proofs like the prior section, although some minor proofs are provided. At times the authors speak in a language such that usage of current methods at that time would appear to make sense in the contact, like linear discriminant. They align covariance adjustment with adjustment for the linear discriminant. The cases, they suggest, where the covariance
adjustment performs poorly is the case where the linear discriminant is not a monotone function of the propensity score. Generally, they state that if multivariate normality holds than this assumption is met, but this may not always be the case, like in the cases where the variances of the $x$ in both control and treated groups are not equal. As long as the covariance matrices are equal between groups, implying the discriminant is a monotone function of the propensity score, then they added that covariance adjustment will remove most of the expected squared bias in the cases considered by Rubin, for which they reference his paper (Rubin, 1979). Of their final note and important point is that one cannot rely on covariance adjustment to do well unless the linear discriminant is highly correlated with the propensity score. It would seem this major conclusion is under-appreciated in today’s usage of propensity scores and one would suspect, that most users are unaware of this.

Over the years though either, changes or adaptations to the propensity score method have been made or in some cases, or a claim to one of its deficiencies has been made. Gary King and Richard Nielsen, more recently in 2019, pointed out that one of the fundamental assumptions of the propensity score method of achieving balance could itself be fraught with errors and therefore that propensity scores should not be used directly for matching, but PSM however, could be used in other ways like PS matching in subclassification and PS weighting and, thus, still had some utility.

King and Nielsen (2019) assured that PSM uses exact matching as per its theorem. However, Rubin and Rosenbaum (1983) had stated that exact matches are almost impossible and they suggested methods that approximate it such as nearest neighbor matching. In terms of matching though, King and Nielsen (2019) argues that PSM through its algorithm of matching of carefully trimming in order to match observations will lead to reducing causal inference which he called the PSM paradox. Their claim to the inadequacies were that it approximates a randomized trial instead of a fully blocked randomized experiment. King and Nielsen (2019) went on to prove that two other matching designs, rather than the nearest neighbor with caliper matching in PSM, more closely approximate a fully randomized block design than PSM. These methods were the Mahalanobis Distance Matching (MDM) and Coarsened Exact Matching (CEM). King and Nielsen (2019) likened PSM achieving a completely randomized design to be equivalent to random pruning, which he had defined as a process of deleting observations in a dataset that are independent of treatment assignment and $x$, which can reduce information in the data and increase imbalance. Otherwise, as they claimed, PSM in its originally written context would increase imbalance, model dependence, and bias.

Therefore according to King and Nielsen (2019), PSM is helpful in data with least likely causal inference and most helpful in data where causal inferences are likely. PSM also would work better in larger datasets where it has more room to make matches and less room to cause imbalance. In effect they do not completely discard the theory but they do minimize its uses and applicability. They did not appear to provide any negative consequences for using MDM or CEM.

However, we are left wondering how realistic that is in general for achieving a fully randomized block design in an observational study. Of course it would be important for observational studies to achieve a fully blocked design but this is also less common realistically. As according to Guo, Fraser, and Chen (2020), it would be difficult to know what the construct for blocking would be. It is almost to say that a fully blocked design would be the best design but in general but it is almost impossible to obtain, while aiming for an observational study to be more like a randomized study seems like a more realistic assumption. Also, it would make sense that Rubin would think through this strategy.

Guo et al (2020) in their article have taken a more serious look into this debate which King and Nielsen (2019) had published and they came across some inconsistencies in their arguments. Guo et al (2020) did highlight that King and Nielsen (2019) appear to criticize mainly one component of Rubin and Rosenbaum’s paper (1983), which is the PSM matching, and more specifically, with the nearest neighbor caliper matching (NNCM) they proposed. The NNCM uses a summary balancing score while the MDM or CEM match on individual covariates. In Guo et al (2020), they used a Monte
Govindarajulu

Carlo simulation to compare the performance of many different matching methods on bias reduction and sample size retention. They found that NNCM performed no worse than the MDM or CEM in most simulations. In general, they found the best performing methods were PS subclassification, propensity score weighting, and matching estimators. Of course, each method could be chosen based on the specific analytic situation, but certainly the claims by King and Nielsen (2019) are debatable and not completely well founded. Certainly there is relevance in regards to model specification and knowing or understanding potential causal structures. Also, it does seem like PS subclassification, as originally proposed by Rubin and Rosenbaum (1983), is still fairly robust in most data applications as well as using propensity scores in weighting as an adjustment for confounding. King and Nielsen (2019) did not have an issue with these latter two methods. However, it is worthwhile for any one employing propensity scoring or other distance metrics for creating balance between groups in an observational study to go back to Rubin and Rosenbaum’s 1983 landmark paper to understand the essential theory and the assumptions behind propensity scores. Interestingly enough, King and Nielsen (2019) were coming from political science and Guo et al (2020) were coming from social science, which just goes to show how widespread the use of propensity scores in different fields have become. Therefore, the legacy of Rubin and Rosenbaum’s 1983 paper still carries on.

References:


