Summary and discussion of "The central role of the propensity score in observational studies for causal effects"

Statistics Journal Club, 36-825

Jessica Chemali and Michael Vespe

1 Summary

1.1 Background and context

Observational studies draw inferences about the possible effect of a treatment on subjects, where the assignment of subjects into a treated group versus a control group is outside the control of the investigator. Let $r_1(x)$ be the response when a unit with covariates x receives the treatment ($z = 1$), and $r_0(x)$ be the response when that unit does not receive the treatment ($z = 0$). Then one is interested in inferring the effect size:

$$\mathbb{E}[r_1(x) - r_0(x)] = \int_{\mathcal{X}} (r_1(x) - r_0(x))p(x)dx, \tag{1}$$

where $\mathcal{X}$ is the space of covariates, and $p(x)$ is the distribution from which both treatment and control units are drawn.

An unbiased estimatemate of the effect size is the sample average:

$$\hat{\mathbb{E}}[r_1(x) - r_0(x)] = \frac{1}{N}\sum_{i=1}^{N}(r_1(x_i) - r_0(x_i)),$$

where $\{x_i\}_{i=1}^N$ are drawn from $p(x)$. In observational studies, however, one has no control over the distribution of treatment and control groups, and they typically differ systematically (i.e $p(x|z = 1) \neq p(x|z = 0)$). Consequently, the sample average becomes a biased, unusable, estimate for the treatment effect.

The goal of the paper we discuss here is to present methods to match the units in the treatment and control groups in such a way that sample averages can be used to obtain an unbiased estimate of the effect size. The key is to use balancing scores, defined below, conditioned on which the covariates become independent of the treatment received.

1.2 Theoretical notions

We start with some notation and definitions, reflecting what is used by Rosenbaum and Rubin. Let $z_i \in \{0, 1\}$ be the indicator variable that unit i is assigned to the treatment group. Let x_i be a vector of covariates observed for unit i.

A **balancing score** is any function $b(x)$ such that $x \perp z \mid b(x)$, that is, *conditional on $b(x)$, the distribution of x is independent of z.*

The **propensity score** $e(x)$ is defined by Rosenbaum and Rubin to be

$$e(x) = \mathbb{P}(z = 1|x)$$

that is, the probability of a unit with covariate values x receiving the treatment. (Hence the term "propensity.")

Theorem 1 asserts that the propensity score $e(x)$ is a balancing score. Theorem 2 asserts that any $b(x)$ is a balancing score if and only if the propensity score can be expressed as a function of that $b(x)$, i.e., $e(x) = f(b(x))$. Rosenbaum and Rubin prove Theorem 2, noting that Theorem 1 is just the special case of Theorem 1 wherein f is the identity function.

We will say that a treatment assignment is **strongly ignorable** given a vector v if

$$(r_1, r_0) \perp z| \ v, \text{ and}$$
$$0 < \mathbb{P}(z = 1|v) < 1, \text{ for all v.}$$

When a treatment assignment is strongly ignorable given covariates x, we will just say it is strongly ignorable.

In plain words, strong ignorability given v means that (a) all the possible confounding phenomena (in the sense that they influence both r and z) are measured in v, so that conditioning on v removes the direct dependence between r and z; and (b) that there is a nonzero probability of a unit with covariates v receiving either treatment.

Theorem 3 asserts that if a treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score $b(x)$; thus, in particular, the treatment assignment is strongly ignorable given $e(x)$.

Combining these ideas, as the authors do in Theorem 4, gives a justification for the idea of *propensity score matching.* For a given propensity score $e(x)$, suppose that we randomly sample two units from the entire population, one of which is a treatment unit and the other of which is a control unit. This is called a *matched pair.* Strongly ignorable treatment assignment implies that

$$\mathbb{E}[r_1|e(x), z = 1] - \mathbb{E}[r_0|e(x), z = 0] = \mathbb{E}[r_1|e(x)] - \mathbb{E}[r_0|e(x)] = \mathbb{E}[r_1 - r_0|e(x)]$$

Then, by the law of iterated expectations,

$$\mathbb{E}_{e(x)}\Big[\mathbb{E}[r_1|e(x), z = 1] - \mathbb{E}[r_0|e(x), z = 0]\Big] = \mathbb{E}_{e(x)}\Big[\mathbb{E}[r_1 - r_0|e(x)]\Big] = \mathbb{E}[r_1 - r_0]$$

Hence, the mean of the differences between the effects in matched pairs is an unbiased estimator for the treatment effect.

1.3 Modeling the propensity score

If we knew $e(x)$ for each unit, the preceding suggests a scheme wherein each treatment unit is matched with a control unit with the same propensity score. In practice $e(x)$ is not known and will have to be modeled. A natural model would be the logit model, where we model the log-odds that $z = 1$ conditional on x.

Suppose we assume that the covariates follow a polynomial exponential family distribution within the treatment and control groups, that is,

$$p(x|z=t) = h(x)\exp\left(P_t(x)\right) \text{ for t} = 0,\ 1$$

where P_t are polynomials of degree k. Then,

$$\begin{aligned}
\log\frac{e(x)}{1-e(x)} = \log\frac{p(z=1|x)}{p(z=0|x)} &= \log\frac{p(x|z=1)p(z=1)}{p(x|z=0)p(z=0)} \\
&= \log\left(\frac{p(z=1)}{p(z=0)}\right) + \log p(x|z=1) - \log p(x|z=0) \\
&= \log\left(\frac{p(z=1)}{p(z=0)}\right) + P_1(x) - P_0(x) \\
&= \log\left(\frac{p(z=1)}{1-p(z=1)}\right) + Q(x)
\end{aligned}$$

which is also a polynomial of degree k in x.

It is easy to imagine a setting where there are many covariates (and perhaps interactions between covariates). Indeed, in the example given in Rosenbaum and Rubin, "covariates and interactions among covariates were selected for the model using a stepwise procedure." Thus, the authors allow that some variable selection procedure may be necessary to properly model the propensity to receive treatment.

1.4 Applications

1.4.1 Pair matching on propensity scores

Matching is a method to draw samples from a large reservoir of controls to produce a population of controls that is similar to that of the treatment group. Matching is appealing because it results in intuitive comparisons between control and treated groups; it produces an estimate of the effect size that has lower variance than in the case where random samples are used; models adjusted on matched samples are more robust than ones based on random samples; and it allows for controlling of confounding variables better than multivariate matching methods can.

Theorem 2 shows that it is sufficient to match exactly the controls on any balancing score to get the same distribution of covariates for treated and control units. Moreover, if treatment is strongly ignorable, then the sample average would be an unbiased estimate of the effect size.

In practice, however, matching is not exact and approximations are used instead. Multivariate matching methods are said to be equal percent bias reducing if the bias in each coordinate is reduced by the same percentage. In the case where the relationship between (r_1, r_0) and x is suspected to be approximately linear, then matching methods that are equal percent bias reducing will reduce the bias of the effect size estimate in comparison to random samples. Theorem 6 of the paper shows that propensity matching is such a method.

1.4.2 Subclassification on propensity scores

In subclassification, control and treated groups are divided based on x into subclasses. Although it offers an intuitive comparison platform for the results, a major problem is that as the number of strata required to divide the space grows exponentially in the number of covariates.

Theorem 2 and strong ignorability show that if the space is divided into subclasses such that the balancing score, $b(x)$, is constant for each subclass and that there exists at least one treatment unit and at least one control unit in each subclass, then the sample average is unbiased at that $b(x)$, and the weighted average across all subclasses is an unbiased estimate of the effect size.

In practice the balancing score will not be constant in every strata, and there will be residual bias due to x. Theorem 7 of the paper shows that direct adjustment based on a balancing score $b = b(x)$ will reduce bias in each coordinate of x providing that the adjustment reduces the bias in b.

1.4.3 Covariance adjustment on propensity scores

Analysis of covariance is an important method for reducing bias in observational studies. After fitting r_1 and r_0 using a linear regression model, a plot of residuals $r_{ki} - \hat{r}_{ki}$ versus the discriminant of x_i is usually helpful in identifying nonlinear or nonparallel response surfaces as well as extrapolations that might distort the effect size estimate.

The paper argues that the plot against the propensity score $e(x)$ is more appropriate in the general case because if treatment assignement is strongly ignorable, then at every $e(x)$ the expected difference in response is an unbiased average treatment effect at $e(x)$. This property holds for balancing scores, but not necessarily for other function of x such as the discriminant. The paper argues that examples where covariance adjustment was seen to perform poorly are exactly places where the discriminant was not a monotone function of the propensity score.

1.5 Simulations

1.5.1 Illustrating matching on estimated propensity score

1. Matching

   We draw the covariates $\{x_1^i\}_{i=1}^{50}$ of the treatment group units from a normal bivariate distribution with mean $\mu_1 = [1, 1]$ and covariance matrix $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. Similarly, we draw the covariates $\{x_i^0\}_{i=1}^{150}$ of the control group units from a normal bivariate distribution with mean $\mu_0 = [0, 0]$ and same covariance Σ. We fit a logistic regression model using all the data and estimate $\hat{e}(x)$. Using this estimate of propensity, we match every treatment unit to its closest control unit from the reservoir. An illustration of the matching is given in Figure 1.

2. Estimating effect size

   We model r_0 as univariate normal distribution with mean $\beta x[2]$ and standard deviation 0.25, where $x[2]$ denotes the second covariate. r_1 is modeled as a univariate

normal distribution with mean $3 + \beta x[2]$ and standard deviation 0.25. Hence, the true treatment effect size is equal to 3.

Without matching on propensity score, the simple mean difference estimate of the effect size will be contaminated by the fact that both r and z depend on $x[2]$. Indeed, we find the simple difference in mean r between treated and control groups to be equal to 5.16, compared to a mean difference of 3.11 between pairs matched on the basis of propensity score. The latter estimate is much closer to the true effect size.

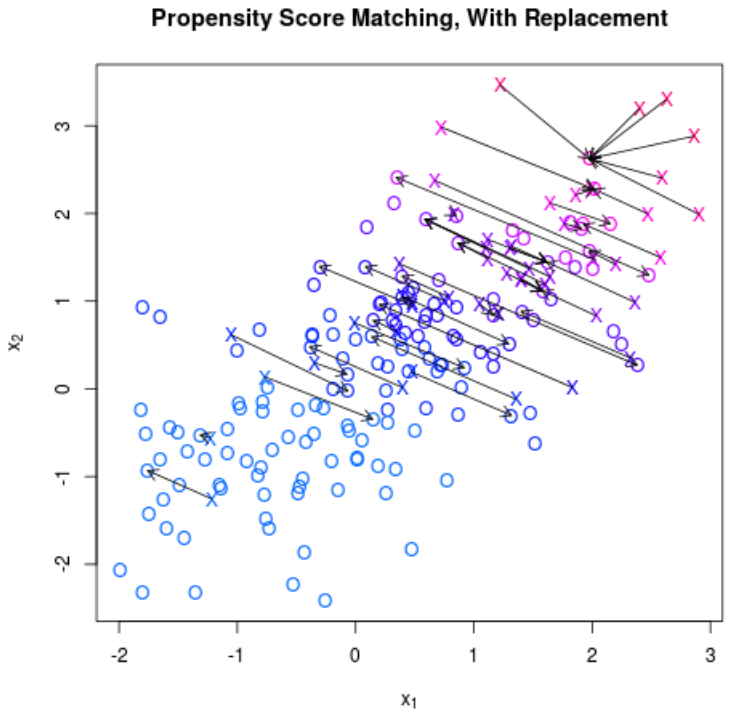

Figure 1: Results of matching based on the propensity score. The cross indicates a treated unit, and the circle indicates a control unit. The propensity score is indicated by the coloring scheme, where light blue indicates lowest propensity and purple indicates highest propensity to receive treatment. Finally the arrows indicate which units are matched based on their propensity scores.

1.5.2 Illustrating the problem of variable selection

In this simulation, courtesy of Rob Tibshirani, the propensity to receive treatment is a sparse combination of a large number of covariates. Four methods are used to evaluate the effect size with the results displayed in Figure 2.

- Taking the simple difference in treatment and control means, without matching

- Using the true propensity score
- Using an estimated propensity score based on a logistic regression on all covariates
- Using an estimated propensity score based on a logistic regression with forward stepwise selection of the covariates

As illustrated in the figure, the propensity score matching without variable selection isn't much better than no matching at all. The best estimate was the result of using the true propensity score, but this is not possible in practice. Furthermore, Larry Wasserman pointed out that it is technically incorrect to use the true propensity score, even if you knew it; that you should still use an estimate from the data, the MLE being the best one can do.

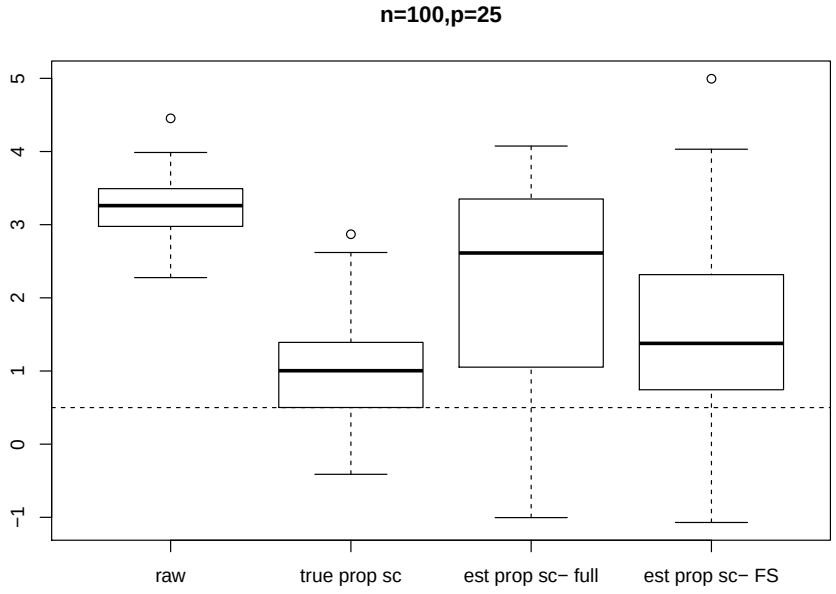

Figure 2: Different methods of estimating effect size. From left to right: no matching (simple difference in means); matching via true propensity score (usually unavailable); matching via propensity score estimated using all covariates; matching via propensity score estimated using covariates chosen via forward stepwise selection.

2 Discussion

2.1 Propensity scoring versus regression

Larry Wasserman made the point that regression is simpler, more standard and can achieve the same goal as propensity scoring. Propensity scoring boils down to assuming a logit relationship between the treatment and the covariates. Regression assumes a linear relationship between the response and both the treatment and covariates.

Ryan Tibshirani was arguing that the fact that you're modeling the response r in terms of x and z might be harder than first modeling z as a function of x and then estimating the effect size. But in t-test, there's an implicit assumption of linearity. So it might be the case that the two models are actually operating under the same assumptions. An interesting experiment to perform would be to simulate a model where z and x interact and see whether propensity scoring or regression perform differently.

Jared Murray said that researchers today don't use regression or propensity scoring on their own but instead use both in doubly robust methods (For further reading see [2]). He also made the point that if there are some regions of the covariate space that have only treatment units or only control units, then it doesn't even make sense to estimate a causal effect there because those units will be matched with units from the other group that aren't sufficiently similar.

2.2 Variable selection

The issue of variable selection was raised, due in part to Rob Tibshirani's simulation exercise.

1. What happens when $z|x$ depends on a few covariates (sparse relationship) but we put in a lot of covariates in the estimation of the propensity score? The intuition is that it will overfit the data, that the scores will be really close to 0 (control) and 1 (treatment), and matching will be fruitless.

2. What happens when $z|x$ depends on a lot of variables (dense relationship)? Then the model will typically be poor because sample size will be too small to estimate the model parameters accurately.

References

[1] Rosenbaum, Paul R., and Donald B. Rubin. *The central role of the propensity score in observational studies for causal effects.* Biometrika, 70.1 (1983): 41-55.

[2] Kang, Joseph D.Y. and Joseph L. Schafer. *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.* Statistical Science, 22.4 (2007): 523-539.