Published in final edited form as:

Health Serv Outcomes Res Methodol. 2012 June ; 12(2-3): 104–118.

Bias and variance trade-offs when combining propensity score weighting and regression: with an application to HIV status and homeless men

Daniela Golinelli,

RAND, 1776 Main Street, Santa Monica, CA 90407, USA, daniela@rand.org

Greg Ridgeway,

RAND, 1776 Main Street, Santa Monica, CA 90407, USA, daniela@rand.org

Harmony Rhoades,

School of Social Work, University of Southern California, Los Angeles 90089, CA, USA,
hrhoades@usc.edu

Joan Tucker, and

RAND, 1776 Main Street, Santa Monica, CA 90407, USA, jtucker@rand.org

Suzanne Wenzel

School of Social Work, University of Southern California, Los Angeles 90089, CA, USA,
swenzel@usc.edu

Abstract

The quality of propensity scores is traditionally measured by assessing how well they make the distributions of covariates in the treatment and control groups match, which we refer to as “good balance”. Good balance guarantees less biased estimates of the treatment effect. However, the cost of achieving good balance is that the variance of the estimates increases due to a reduction in effective sample size, either through the introduction of propensity score weights or dropping cases when propensity score matching. In this paper, we investigate whether it is best to optimize the balance or to settle for a less than optimal balance and use double robust estimation to adjust for remaining differences. We compare treatment effect estimates from regression, propensity score weighting, and double robust estimation with varying levels of effort expended to achieve balance using data from a study about the differences in outcomes by HIV status in heterosexually active homeless men residing in Los Angeles. Because of how costly data collection efforts are for this population, it is important to find an alternative estimation method that does not reduce effective sample size as much as methods that aggressively aim to optimize balance. Results from a simulation study suggest that there are instances in which we can obtain more precise treatment effect estimates without increasing bias too much by using a combination of regression and propensity score weights that achieve a less than optimal balance. There is a bias-variance tradeoff at work in propensity score estimation; every step toward better balance usually means an increase in variance and at some point a marginal decrease in bias may not be worth the associated increase in variance.

Keywords

Propensity score; Double robust estimation; HIV status; Homeless men

1 Introduction

Propensity score (PS) methods (Rosenbaum and Rubin, 1983) have become the tool of choice when in need of removing bias due to observed confounders. The quality of the PS is usually measured with an assessment of the covariate balance between the two groups being compared. Achieving good balance is key for obtaining less biased estimates of the treatment effect. However, optimizing balance can be costly in terms of the estimator's precision.

Among the PS methods, weighting has recently received attention (Hirano and Imbens 2001). In particular, PS weights obtained with modern statistical learning algorithms have been shown to perform well in a variety of settings (McCaffrey et al, 2004, Lee et al, 2010). Propensity score weights can be highly variable (Kang and Schafer, 2007) and this variability directly affects the effective sample size (ESS). The variance of a weighted mean is σ^2/ESS where $\text{ESS} = n/(1+\text{Var}(w))$. If all of the observation weights are 1 then $\text{Var}(w)=0$ and the variance of the weighted mean is the standard σ^2/n . However, after applying propensity score weights $\text{Var}(w) > 0$, which decreases ESS and increases the variance of the weighted mean. Propensity score matching and stratification have a similar effect on the ESS.

Compared to propensity scoring, regression can produce treatment effect estimates with smaller variance, but potentially with more bias when the form of the regression model selected by the analyst neglects important nonlinear terms and interactions (Rubin 1973). In fact, if the imbalances for some of the covariates are large between the treated and control groups and such covariates are non-linearly associated with the outcome, linear regression will produce a biased estimate. However, regression can be effective in eliminating small imbalances. See Tsiatis et al (2000) for a review.

As a result, the analyst needs to weigh the risks of using regression, which is prone to bias but usually has smaller variance, and the risks of propensity scoring, which can have large variance but smaller bias. Naturally we would prefer a method that could mediate between these two in order to produce an estimator that is robust with small mean squared error (MSE). Double robust (DR) estimation (Bang and Robins, 2005) offers a potential solution. DR estimation combines regression with PS weights so that if either the propensity score model or the regression model is accurately specified then the estimator will be consistent. We further explore whether we can further reduce MSE by using DR estimation with PS weights that do not quite achieve optimal balance, accepting a small amount of bias in order to decrease variance. When the covariate distributions of the treatment and control groups are close, the regression step of DR estimation is less sensitive to the functional form and can eliminate the bias due to small covariate differences between the two groups (Tsiatis et al 2000).

This analysis was motivated by the desire to obtain an estimate of the difference in outcomes (such as risky behaviors, and service use) by HIV status among heterosexually active homeless men in the Central City East area of Los Angeles, often referred to as Skid Row. HIV remains a serious health problem in the U.S., particularly among vulnerable populations, such as homeless persons, where rates of HIV are estimated to be three times higher than those in the general population (National Alliance to End Homelessness 2006). Competing priorities associated with homelessness may also complicate health care and treatment for those living with HIV (Gelberg et al. 2000; Henry et al. 2008). Sexual risk behavior, drug and alcohol use, and service utilization have been found to be associated with HIV status in other populations, but to our knowledge little research has been done about this issue for the homeless male population. Understanding how these outcomes are related

to HIV status could inform the development and improvement of programs to deliver care to HIV-positive homeless men and reduce the further spread of HIV among these men and their sexual partners.

Obtaining cases from this population is expensive and as a result the sample size is not large. When optimizing the propensity scores to balance the covariates of the HIV positive and HIV negative groups, power and precision can degrade substantially. Some of this reduction in precision is necessary as some comparison cases simply do not resemble the exposed/treatment cases. However, perhaps being overly aggressive in aligning the covariate distributions eliminates more cases than needed. Therefore, we turned to methods that could better preserve the sample size without risking too much of an increase in bias.

2 Methods

2.1 Data

2.1.1 Participants—Participants in this study were 305 homeless men randomly sampled and interviewed in 13 meal programs in the Central City East area of Los Angeles. This area is home to the highest concentration of homeless persons in Los Angeles County. Men were eligible if they were at least age 18, could complete an interview in English, and had experienced homelessness in the past 12 months (i.e., stayed at least one night in a place like a shelter, abandoned building, voucher hotel, vehicle, or outdoors because they didn't have a home to stay in). As this sample was collected as part of a study of heterosexual risk behavior, all participants had had vaginal or anal sex with a female partner in the past 6 months. Of the 338 men who screened eligible for the study, 320 men were interviewed (18 refusals). Of these 320: 11 were partial completes/break-offs (when interviews could not be completed because the respondent had to leave suddenly, refused to finish the interview, or otherwise did not fully complete the interview process), and 4 were later found to be repeaters. The final sample size was 305, for a completion rate of 91% (305/334). Individual, computer-assisted face-to-face structured interviews were conducted by trained male interviewers. Men were paid \$30 for participation in the interview, which lasted on average 83 minutes. The research protocol was approved by the institutional review boards of the University of Southern California and the RAND Corporation.

2.1.2 Sample Design—To obtain a representative sample of heterosexually active homeless men from the Central City East area of LA, we implemented a probability sample of men recruited from meal lines in the area. The list of operating meal lines in Central City East was developed using existing directories of services for homeless individuals and performing interviews with service providers. Our final list contained 13 meal lines: five breakfasts, four lunches and four dinners offered by five different organizations. Each of the meal lines was extensively investigated to obtain an estimate of the average number of men served daily. This information was used to assign an overall quota of completes to each site, approximately proportional to the size of the meal line. We then drew a probability sample of homeless men from the 13 distinct meal lines. When the assigned quota could not be achieved in a single visit, the quota was divided approximately equally across the number of visits for each meal line. The interview team randomly selected potential recruits for screening by their position in line using random number tables generated by the statistician. Tables were generated such that the daily site quota could be achieved before the meal line was exhausted. Once the field director selected a potential recruit, an interviewer would wait for him to finish his meal before screening him.

The adopted sample design deviates from a proportionate-to-size stratified random sample because of changes in sampling rates during the fielding period, differential response rates of men across meal lines, and variability in how frequently men use meal lines. This last

factor means that some men are more likely to be included in the sample. We accounted for the differential frequency of using meal lines by asking respondents how often they had breakfast, lunch and dinner at a meal line in the Central City East area in the past 30 days, and how much of the past 6 months they had been homeless. This information was used to develop and implement sampling weights to correct for departures from a proportionate-to-size stratified random sample and potential bias due to differential inclusion probabilities (Elliott et al. 2006). All analyses described below use the derived sampling weights.

2.2 Measures

2.2.1 HIV status—We asked a series of questions to assess whether men were HIV positive. Men were asked the following questions in the reported order: 1. whether they had ever been tested for HIV; 2. whether the last time they got tested they found out the results; and 3. whether they had ever been told, or had reason to believe, that they have HIV or AIDS. We coded as HIV positive those men that answered yes to all three questions. In other words, to be coded as HIV positive the men had to have been tested, found out the results and hence still have reason to believe to be HIV positive or have AIDS. We also defined a less restrictive measure of self-reported HIV seropositivity, where every man who had reason to believe that he was HIV-positive, independent of whether he had ever been tested or had found out the test's result, was coded as HIV positive.

2.2.2 Independent variables—In the PS and regression models we included the following variables: age, race/ethnicity, education (less than high school, high school or equivalent, or more than high school) and whether the men had any children. We also included whether the homeless men had ever served in the military, the total months in their lifetime that they had experienced homelessness, and how they would describe their sexual orientation (heterosexual (straight): yes or no).

2.2.3 Outcome variables—In this study we were interested in several outcomes. We briefly describe below how these variables are defined.

Sexual risk behavior: It was measured for the prior six months, and included the total number of sex partners, the proportion of total partners who were 'casual' (vs. primary), and having traded sex for money or goods (Kennedy et al. 2010; Wenzel 2009). Respondents reported the total number of sex events in the prior six months, and the total number of events where a condom was used; a single measure of unprotected sex was created to indicate any reported sex event where a condom was not used.

Drug and alcohol use: Use of drugs during the past 6 months was assessed using separate questions for marijuana, crack, cocaine, prescription drugs, heroin, methamphetamine and 'other' substances. For each substance, the men were asked: "During the past six months, how often did you use <substance> (0 'not at all' to 9 'every day')?" An indicator was created for any use of 'hard drugs' (defined as heroin, crack, cocaine, methamphetamine, or hallucinogens), as well as individual indicators of any marijuana use and any misuse of prescription medication (defined as use of prescription medications "without a doctor's prescription, in larger amounts than prescribed, or for a longer period than prescribed"). Men were also asked how often they had a drink containing alcohol in the past 12 months (0: 'never/no drinking' to 5: '6+ times per week'), and a dichotomous measure was created indicating drinking, on average, 2-3 times per week or more (vs. less) over the past year. All substance use items have been previously vetted with a population of homeless women (Wenzel et al. 2009).

Service utilization: Men were asked about their use of alcohol or drug counseling, mental health counseling, legal assistance and medical or dental care service use during the past 30 days in the Central City East area (Tucker et al. 2011). Indicators were created for each type of service utilization. Men were also asked whether they currently had a regular place to stay.

2.3 Analytical procedures

2.3.1 The statistical problem—Our aim is to understand the differences between those reporting being HIV positive and those not reporting being HIV positive. However, the two groups differ on numerous features. For example, 51 percent of HIV positive subjects served in the military compared with 16 percent among those not HIV positive. Therefore, if we find a difference by HIV status we will be unsure as to whether the difference is due to HIV status or military status. We also found large differences in other subject features such as the race distribution, age, and education further clouding the comparison.

To narrow in on the differences in outcomes by HIV status we want to compare the HIV positive subjects to “similarly situated” HIV negative subjects. This is a simpler goal than estimating the causal effect of HIV status. If the subject features, $\mathbf{X}$, were sufficiently rich so that the subjects’ potential outcomes are independent of the actual HIV status conditional on $\mathbf{X}$, known as strong ignorability, then the estimated effects would be causal effects. We are not in a position to make that assumption in this example as there surely are unmeasured confounders (we do not have a rich personal history of each subject) and some of the subject features could have been affected by earlier HIV status (e.g. months of homelessness, military service). Rosenbaum (1984) contains a detailed discussion of the potential consequences of adjusting for characteristics that are affected by the exposure variable. Nonetheless, our analysis assesses the effect of “living with HIV” on the various outcomes of interest adjusted for several confounding features.

The methodology and simulations we describe in this manuscript are relevant to both analytical problems, those in which a strong ignorability assumption is plausible and cases in which it is not and the inferential goal is simpler. When strong ignorability claims do not hold, sensitivity analysis methods exist to assess whether a causal effect is likely nonetheless (see Rosenbaum, 2005 for an overview). For the case study presented here we were interested in a simpler inferential goal: estimating the difference in outcomes between the HIV positive cases and similarly-situated HIV negative cases. This is, from a policy point of view, a very relevant question: if there are still differences in outcomes by HIV status once the differences in age, race, education, number of months homeless and so on have been eliminated service providers and policy makers will have useful information on how to better develop programs to serve the HIV positive homeless community. Numerous methods, both modern and traditional, aim to adjust for observed differences between groups and can be used to obtain estimates of interest. We consider regression, propensity score weighting, and double robust estimation.

2.3.2 Regression—Regression or covariate adjustment is the traditional and commonly used approach to adjust for differences in the covariates between the treated and control groups. If the distribution of the covariates across the treatment and control group differ greatly, as is frequently the case in non-randomized studies, then treatment effects estimated from regression models are highly sensitive to the form of the regression model (Cochran and Rubin, 1973). Failure to include an important non-linear term or an important interaction term can produce large biases in the treatment effect estimate. This lack of robustness and the difficulty of diagnosing the quality of the regression model (in spite of

numerous regression diagnostic methods) have led researchers toward using methods that make the covariate distributions more similar.

2.3.3 Propensity score weighting—Propensity score weighting is an approach used to reweight treatment and control groups so that their covariate distributions match, manipulating the observational study so that the observed covariates resemble what we would expect from a randomized controlled trial. Let $p(\mathbf{x})$ be the propensity score, the probability that a subject with covariates $\mathbf{x}$ belongs to the treatment group. We estimate the propensity score using the observed data to produce $\hat{p}(\mathbf{x})$. Traditionally, $\hat{p}(\mathbf{x})$ comes from a logistic regression model. However, as described later in this section, we use boosted propensity score estimation. By assigning weights of 1 to the treatment group cases and weights $\hat{p}(\mathbf{x})/(1 - \hat{p}(\mathbf{x}))$ to the control group cases, the covariate distribution of the control group is reweighted to match the covariate distribution of the treatment group. The difference between the average outcome for the treatment group and the PS weighted outcome of the control group yields an estimate of the “average treatment effect on the treated” (ATT). Compared to regression, PS weighting is easier to diagnose (with a simple comparison of the weighted treatment and control group covariates), explain, and analyze.

There are several PS methods used to align the covariate distributions; they include matching, stratification, and weighting. Recently PS weighting has gained some popularity (Lunceford and Davidian, 2004). However, Kang and Schaffer (2007) showed that PS weighting can perform poorly when the propensity score weights are highly variable. The risk of poor performance is greatly reduced when using more flexible propensity score estimation methods (Ridgeway and McCaffrey, 2007). Lee, Lessler, and Stuart (2010) concluded that PS estimation using boosting (labeled “boosted CART” by Lee et al.) had “consistently superior performance”. They note “under conditions of both moderate non-additivity and moderate non-linearity, logistic regression had subpar performance, whereas ensemble methods provided substantially better bias reduction and more consistent 95 percent confidence interval coverage. The results suggest that ensemble methods, especially boosted CART, may be useful for propensity score weighting.”

Boosted PS estimation (McCaffrey et al, 2004) uses a flexible, nonparametric estimation technique that adaptively captures the functional form of the relationship between the subjects’ characteristics and the treatment indicator with less bias than traditional approaches such as logistic regression. Logistic regression models the propensity score as $p(\mathbf{x}) = 1/(1 + \exp(-\alpha' \mathbf{x}))$ and selects α to maximize the log-likelihood shown in (1) where t_i is the treatment indicator and λ is set to 0.

$$\sum_{i=1}^N t_i \alpha' \mathbf{x}_i - \log(1 + \exp(\alpha' \mathbf{x}_i)) - \lambda \sum_{j=1}^d |\alpha_j| \quad (1)$$

The term after λ is a penalty term that controls the size of the coefficients in the logistic regression model. While $\lambda=0$ makes the penalty term disappear to produce ordinary logistic regression, setting $\lambda=\infty$ tells the model not to tolerate any coefficient that is not zero. Forcing $\alpha_1, \dots, \alpha_j$ to 0 leaves only α_0 for the estimation, which produces propensity scores that are constant for every case (producing an analysis with no covariate adjustment). In between $\lambda=0$ and $\lambda=\infty$ lies a large collection of possible propensity score models.

For large values of λ , many of the estimated values of the α_j will be 0. If we slowly relax λ to be smaller and smaller most of the α_j s remain 0 and only the most important predictors of treatment assignment will have non-zero α_j s. Consequently we can include a large set of covariates, even if they are highly correlated, and the penalty term will control the size of

their coefficients. We include all piecewise constant functions of the $\mathbf{x}$'s and their interaction terms, limited to at most three-way interactions. This means that the covariates used in the propensity score model include indicator variables of the form $I(\text{age} < 25)$, $I(\text{age} < 26)$, $I(\text{age} < 27)$, $I(\text{age} < 25 \text{ and black})$, $I(\text{age} < 26 \text{ and black})$, and so on. In our example, including all of these transformations of our original seven covariates results in over 250,000 terms. However, the penalty term results in most of them having coefficients equal to 0 and only the terms most predictive of treatment assignment receive non-zero coefficients.

The estimation algorithm, the GBM package in R (Ridgeway 2007), is implemented so that we never need to store their coefficients or include them in a design matrix making computation feasible. The algorithm iteratively reduces λ and in doing so selects the next term to be included in the model. Thus, iteration 0 produces constant propensity scores (no adjustment) and subsequent iterations incrementally include new terms that improve the propensity score fit.

A process for selecting the right value of λ remains to be determined. A λ that is too large results in differences in the covariate distributions between the two groups causing bias in the treatment effect estimate. A λ that is too close to 0 produces highly variable propensity score weights resulting in treatment effect estimates with large standard errors. The analyst is faced with the common question of a bias-variance tradeoff. In practice, we select the λ that optimizes balance as measured by the largest Kolmogorov-Smirnov (KS) statistic of the covariates.

In our study we examine the effect of λ on the covariate balance, the precision of the treatment effect estimate, and on the treatment effect estimate. To measure covariate balance we compute the Kolmogorov-Smirnov statistic for each covariate and report the largest of them. To measure precision we compute the control group's effective sample size, the number of independent unweighted observations that would result in the equivalent level of

precision as the weighted sample achieves, computed as $ESS = \left(\sum_{i=1}^N w_i \right)^2 / \left(\sum_{i=1}^N w_i^2 \right)$. The treatment effect is computed as difference between the average of the treatment group outcome and the propensity score weighted control group outcome. We repeat this analysis for each of our study's outcomes.

Recall that propensity score weighting reweights the data so that the observed covariates resemble what we would expect from a randomized study. Even with a randomized study good analytical practice suggests using covariate adjustment for those covariates that might be predictive of the outcome in order to remove bias due to small differences between the treatment and control groups and to reduce residual variance. Combining propensity score estimation with regression can be equally effective and will produce estimates that are *doubly robust*, consistent if *either* the propensity score model is correct *or* the regression model is correct.

2.3.4 Double robust estimation—Double robust (DR) estimation methods (Bang and Robins 2005; Kang and Schafer 2007) can reduce the risk of bias due to remaining small differences between the treatment and control groups and can reduce the uncertainty in the treatment effect estimator by reducing the outcome model's residual variance.

To obtain a DR treatment effect estimate we fit weighted regression models (Gaussian for continuous outcomes and logistic regression for dichotomous outcomes,) using the propensity score weights as observation weights. Let $\hat{y}(t, \mathbf{x})$ be the predicted value of the outcome (on the probability scale for logistic regression models) from the weighted

regression model for a case with treatment group indicator t and with features $\mathbf{x}$. Then a DR treatment effect estimate is

$$\frac{\sum_{i=1}^N t_i (y_i - \widehat{y}(0, \mathbf{x}_i))}{\sum_{i=1}^N t_i} \quad (2)$$

The estimator essentially computes the average difference between each treatment case's outcome and what the regression model predicts would be the case's outcome had they been in the control group.

As mentioned above, obtaining good balance is critical for producing unbiased treatment effect estimates. However, optimizing balance can impact the efficiency of the treatment estimate. On the other hand, using only regression to adjust for imbalances, when they are large, can be risky if the regression model is not correctly specified. In this paper we want to investigate whether there could be a better solution in terms of the bias-variance trade-off that sits somewhat in between regression and DR estimation with PS weights that achieve optimal balance. In other words, we think that we might achieve better efficiency (i.e.: lower variance estimates) and not impact bias too much if we fit a PS model that does not achieve perfect balance and then use PS weighted regression to eliminate the small imbalances that might be left.

We first test the proposed strategy with a simulation study. We then investigate it when estimating the association between HIV status and outcomes. In the results section we estimate the association between HIV status and outcomes using regression, propensity score weighting and DR methods and we also show how the effective sample size, the association estimate and its standard error vary as we fit a PS model that gets closer and closer to achieving optimal balance. The proposed analyses and reported results should be seen primarily as illustrative of the choices that analysts of observational data can consider when selecting among estimation methods.

3 Simulation study

To test our estimation strategy we generated 1000 simulated datasets following the simulation framework described by Rubin (1973). In Rubin (1973) the data generating mechanism is given by

$$\begin{aligned} x_i &\sim N\left(-\frac{1}{2} + T_i, \sigma_{T_i}^2\right) \\ Y_i &= \exp\left(\frac{x_i}{s}\right) + e_i \\ e_i &\sim N(0, 1) \end{aligned}$$

where $i=1, \dots, 900$ and T_i is a 0/1 treatment indicator with $T_i=1$ for 300 of the cases. Note that there is no treatment effect in this simulation; T_i affects Y only through the covariate. s is a scale parameter that controls the degree of non-linearity; we consider three scenarios with $s=2$ (mild non-linearity) and three scenarios with $s=1$ (strong non-linearity). In scenarios 1 and 4: $\sigma_0^2 = \sigma_1^2 = 1$, in scenarios 2 and 5 $\sigma_0^2 > \sigma_1^2$ (with $\sigma_0^2=4/3$ and $\sigma_1^2=2/3$); lastly in scenarios 3 and 6 $\sigma_0^2 < \sigma_1^2$ (with $\sigma_0^2=2/3$ and $\sigma_1^2=4/3$). We estimated the treatment effect for each simulated data set using five different methods: simple difference of means (labeled unadjusted); simple linear regression (labeled OLS); propensity score weights that minimize the largest KS statistic (labeled PS); propensity score weighted linear regression (labeled DR); and lastly we determined whether there was a DR estimate obtained with PS weights that do not quite minimize the largest KS statistic that achieved a smaller mean squared error (MSE) than the MSE achieved by the DR estimator (labeled

DR2). More precisely, to obtain DR2, for each simulated data set we computed the PS weights at 40 iterations equally spaced between iteration one and the iteration corresponding to the PS weights that optimize balance. For each set of weights we obtained the treatment effect estimate using the DR method. For each scenario and each estimation method we compute the MSE, the squared bias, and variance of the 1000 treatment estimates.

The results of this simulation study are reported in Table 1. For the first four scenarios we find that there exists a set of sub-optimal PS weights, weights that do not minimize the largest KS statistic, that used in combination with linear regression produce a treatment effect estimator (DR2) with smaller MSE.

Scenario 1 is the most favorable to regression since the outcome Y is mildly non-linear in x and the variances of x for the treatment and control groups are equal. In fact in this scenario OLS has a smaller MSE than both PS and DR. However, there is a set of sub-optimal PS weights for which the DR2 estimator has smaller MSE than OLS. The gain in MSE is due to both a smaller bias and variance in this case. For scenarios 2-4, instead, the gain in MSE with DR2 is due to a smaller variance. Scenarios 2-4 represent what we expected: using PS weights that do not quite minimize the largest KS statistic together with linear regression should produce an estimator that is slightly more biased but more precise due to the gain in ESS.

We note here that Scenarios 5 and 6 are the least favorable to regression because of the stronger non-linearity and inequality of the variances of x for the two groups. However, in Scenario 5, the scenario where the distribution of x for the control group covariate overlaps with the distribution of x for the treatment group, PS outperforms all other methods in terms of MSE and in particular in terms of bias. Note that Scenario 5 is also the case in which the unadjusted estimator outperforms OLS.

4 Results

In this section we use regression, PS weights, and DR methods to estimate the difference in outcomes by HIV status. We also investigate whether it is possible to find a set of sub-optimal PS weights that combined with regression produce an estimate of the difference in outcomes that is more precise and does not differ greatly from the estimate obtained with optimal PS weights and the DR method.

Table 2 shows means and/or proportions of the demographic and background characteristics used to fit the propensity scores for (1) the HIV positive group; (2) the unweighted HIV negative group; and (3) the propensity score weighted HIV negative group. Before weighting, the HIV positive and negative groups differed on six of the seven variables included in the propensity score model. The HIV positive group was older, less likely to be White, more likely to be more educated, more likely to have ever served in the military, less likely to be heterosexual and had been homeless for a longer period of time. Using only regression to remove imbalances could be risky in this case given the large imbalances between the HIV positive and negative groups. From Table 2 we can see that, for example, HIV positive men on average are almost five years older than the HIV negative men and about 50% of them have served in the military versus 16% of the HIV negative group. After PS weighting, the two groups look more similar, however some differences still exist. In particular, the standardized mean difference is larger than 0.2 for education and for the rate of having ever served in the military. Table 3 shows the means/proportions for all the outcomes for the 1. HIV positive group; 2. HIV negative group unadjusted; 3. regression adjusted HIV negative group; 4. propensity score weighted HIV negative group; and 5. propensity score weighted and regression adjusted HIV negative group (DR estimate). From Table 3 we see that, in general, the estimates obtained via propensity score weighting and

the DR method are very close, with few exceptions (see for example mental health counseling and medical and dental care). On the other hand the estimates obtained with regression (column 3 of Table 3) are in many cases dissimilar to the PS weighted and DR estimates. For example, there are notable differences in the number of partners, any prescription drug use, any methamphetamine use, and so on.

The propensity score weights used to obtain the estimates in column 4 and 5 of Table 3 are optimal in that they minimized the largest Kolmogorov-Smirnov statistic and hence should ensure a good balance. To assess our idea of whether we might be able to obtain estimates of the HIV status effect that are not too biased and more efficient we computed the effective sample size (ESS) of the control group; the largest KS statistic; the treatment effect estimate and its standard error for each iteration of the propensity score model used for the PS weights. More precisely we computed the estimate of the difference in proportion of casual partners by HIV status and its standard error using the PS weighting and the DR method. The results of such simulation are reported in Figure 1. The top left plot of Figure 1 shows how the ESS decreases as the number of iterations in the propensity score model construction increases. The vertical line represents the iteration at which the largest KS statistic is minimized. The bottom left plot shows how the largest (or maximum) KS statistic varies as the number of iterations increases. Both plots show a sharp decrease within the first 2000 iterations. While the ESS keeps decreasing as the number of iterations increases until it reaches a plateau; the largest KS statistics starts increasing again after the optimal iteration. The top right plot shows the estimate of the difference in the proportion of casual partners obtained via PS weighting (solid line) and via the DR method (dashed line); while the bottom right plot shows the standard error of the estimate of the difference. At the first iteration the PS weighted estimate essentially corresponds to the simple unadjusted difference in means (difference between columns 1 and 2 in Table 3); while the DR estimate corresponds to the estimate we would obtain with the regression adjustment (difference between columns 1 and 3 in Table 3). For this particular outcome the PS weighted and DR estimates differ somewhat for the few first iterations and then they become very close to each other and stabilize to the value obtained with the optimal number of iterations. This last plot suggests that at least for this particular outcome we could have stopped after 2000 iterations and essentially obtained the same treatment effect estimate with a lower standard error. In fact the standard error for both estimates keeps increasing as the number of iterations increases. While the patterns observed in the right two plots of Figure 1 are outcome dependent; we observed a similar general pattern in most of the outcomes.

Overall, 7.4% of the sample was coded as HIV positive using the stricter measure of self-reported HIV sero-positivity; while 9.4% was HIV positive using the more lax definition. Nominally (i.e. without using the sampling weights) 24 of the 305 surveyed homeless men were coded as HIV positive, compared to 31 men who had reason to believe they were HIV positive. This rate is much higher than the general population rate, which is around 0.38%, and higher than the rate among men only of 0.59% (rates obtained from Centers for Disease Control and Prevention and U.S. Census Bureau. State and County QuickFacts: USA). Generally, before performing any adjustment, HIV positive homeless men were more likely to engage in sexual risk behaviors, with a notable exception: while HIV positive men were more likely to have traded sex and had a larger proportion of casual partners, they were less likely to engage in unprotected sex than HIV-negative men. All differences remain, but become smaller, after PS weighting or DR estimation. The largest difference remaining is in the proportion of casual partners (77% versus 57%). The same trend holds for alcohol and hard drug use. However, there are some striking differences by HIV status in the use of marijuana, prescription drugs and amphetamines. HIV positive homeless men seem to be much less likely to use marijuana, while they are much more likely to use prescription drugs

to get high and amphetamines than the HIV negative homeless men. These differences become even larger after PS weighting or DR estimation.

In terms of utilization of services, with few exceptions, we find small differences between the two groups. Though, HIV positive homeless men are more likely to access medical and dental care services. Interestingly, the HIV positive group is also more likely to report having a regular place to stay. Both differences remain large and significant after PS weighting and DR adjustment.

4 Conclusions

The study that prompted this research was a study of heterosexually active homeless men, a population for which data collection is extremely costly and time consuming. A standard approach to propensity score analysis would trim our control sample of 281 HIV negative cases down to an effective sample size of 54. However, propensity score methods focus on reducing bias by optimally balancing the covariate distributions. As with virtually all statistical methods, there is a bias-variance tradeoff at work. Allowing a regression step to resolve small remaining imbalances rather than aggressively pursuing balance may be a path toward robust, low variance, low bias treatment effect estimators.

In this paper we saw that the use of simple regression to adjust for large imbalances between the treated and untreated groups can produce biased estimates of the treatment effect if the regression model is not correctly specified. On the other, we saw that aiming to achieve optimal balance can have a large impact on the ESS, and hence reduce the precision of the treatment effect estimate. The advantage of using generalized boosted logistic regression to obtain propensity score weights is being able to examine a family of propensity score models that range from a model with no adjustment for covariates to a model that optimizes covariates balance. We found that along the path that takes the PS model from an unadjusted analysis to optimal balance it is possible to find a treatment effect estimate that is more precise and with approximately the same quality of covariate balance. Furthermore, DR methods can improve the precision of the treatment effect estimate and it is preferable when some covariates remain imbalanced after propensity score weighting. Our simulations suggest that, in several of the considered scenarios, we can find a DR estimator using propensity score weights that do not aggressively optimize balance that outperforms the DR estimator using the optimal PS weights. The reduction in MSE achieved by this new estimator ranged between 4 and 15 percent in the simulations.

This suggests that careful thinking and additional research is needed to assess the roles of propensity scoring and regression in computing treatment effects. Understanding these roles is particularly important in the context of costly data collection efforts. When a dataset is small and cases are expensive to enroll, we should not necessarily down-weight cases to pursue perfect balance between the treatment and control groups when precision and power may be degraded more than can be compensated by reductions in bias. Our assessment is that settling for slightly lower quality balance can result in a larger effective sample size and an overall lower MSE.

Acknowledgments

This research was supported by Grant R01HD059307 from the National Institute of Child Health and Human Development. We thank the men who shared their experiences, the service agencies in the Central City East area that collaborated with us, and the RAND Survey Research Group for assistance in data collection.

References

- Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005; 61:962–972. [PubMed: 16401269]
- Centers for Disease Control and Prevention. HIV Surveillance - United States, 1981-2008. Morbidity and Mortality Weekly Report (MMWR). 2011; 60:689–93.
- Cochran WG, Rubin DB. Controlling Bias in Observational Studies: A Review. *Sankhya, Series A*. 1973; 35(4):417–446.
- Elliott MN, Golinelli D, Hambarsoomian K, Perlman J, Wenzel S. Sampling with field burden constraints: An application to sheltered homeless and low income housed women. *Field Methods*. 2006; 18:43–58.
- Gelberg L, Andersen RM, Leake BD. The Behavioral Model for Vulnerable Populations: application to medical care use and outcomes for homeless people. *Health Serv Res*. 2000; 34:1273–302. [PubMed: 10654830]
- Henry R, Richardson JL, Stoyanoff S, Garcia GP, Dorey F, Iverson E, et al. HIV/AIDS health service utilization by people who have been homeless. *AIDS Behav*. 2008; 12:815–21. [PubMed: 17682939]
- Hirano K, Imbens GW. Estimation of casual effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv Outcomes Res Methodol*. 2001; 2:259–278.
- Kang JDY, Schafer JL. Demystifying double robustness, a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007; 22:523–539.
- Kennedy DP, Wenzel SL, Tucker JS, Green HD Jr, Golinelli D, Ryan GW, et al. Unprotected sex of homeless women living in Los Angeles county: an investigation of the multiple levels of risk. *AIDS Behav*. 2010; 14:960–73. [PubMed: 19876728]
- Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010; 29:337–346. [PubMed: 19960510]
- Lunceford JK, Davidian M. Stratification and weighting via propensity score in estimating of casual treatment effects, a comparative study. *Stat Med*. 2004; 23:2937–2960. [PubMed: 15351954]
- McCaffrey D, Ridgeway G, Morral A. Propensity Score Estimation with Boosted Regression for Evaluating Adolescent Substance Abuse Treatment. *Psychological Methods*. 2004; 9:403–425. [PubMed: 15598095]
- National Alliance to End Homelessness. Fact Sheet: Homelessness and HIV/AIDS. Washington, DC: National Alliance to End Homelessness; 2006.
- Ridgeway G. GBM 1.6-3 Package Manual. R Project. 2007
- Ridgeway, G.; McCaffrey, DF. Comment on ‘Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data’. In: Kang; Schafer, editors. *Statistical Science*. Vol. 22. 2007. p. 540-543.
- Rosenbaum PR. The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment *Journal of the Royal Statistical Society. Series A (General)*. 1984; 147(No. 5):656–666.
- Rosenbaum, Paul R. *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd; 2005. Sensitivity Analysis in Observational Studies.
- Rosenbaum PR, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rubin D. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*. 1973; 29:185–203.
- Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statist Med*. 2000; 25:1–10.
- Tucker JS, Wenzel SL, Golinelli D, Zhou A, Green HD Jr. Predictors of substance abuse treatment need and receipt among homeless women. *J Subst Abuse Treat*. 2011; 40:287–94. [PubMed: 21185682]
- U.S. Census Bureau. State and County QuickFacts: USA. Washington, DC: U.S. Census Bureau; 2011. Available from: <http://quickfacts.census.gov/qfd/states/00000.html>

- Wenzel SL. Heterosexual HIV risk behavior in homeless men. National Institute for Child Health and Human Development. 2009
- Wenzel SL, Green HD Jr, Tucker JS, Golinelli D, Kennedy DP, Ryan G, et al. The social context of homeless women's alcohol and drug use. *Drug Alcohol Depend.* 2009; 105:16–23. [PubMed: 19616904]

Figure 1.

ESS, largest Kolmogorov-Smirnov statistic, estimate of the difference in proportion of casual partners by HIV status, and its standard error as functions of the GBM iterations. The solid line is the propensity score estimate and the dashed line is the DR estimate. The vertical line in each plot indicates the iteration that produced the weights with the optimal balance as measured by the largest KS statistic.

Table 1

MSE, squared bias, and variance of five treatment effect estimators (all numbers are multiplied by 1000)

	Method	Scenario					
		Mild non-linearity, s=2			Strong non-linearity, s=1		
		1 $\sigma_0^2 = \sigma_1^2$	2 $\sigma_0^2 > \sigma_1^2$	3 $\sigma_0^2 < \sigma_1^2$	4 $\sigma_0^2 = \sigma_1^2$	5 $\sigma_0^2 > \sigma_1^2$	6 $\sigma_0^2 < \sigma_1^2$
MSE							
	Unadjusted	86.68	42.88	150.79	730.91	101.52	2089.84
	OLS	5.56	10.99	13.92	19.54	285.19	283.17
	PS	5.97	6.95	8.75	12.9	13.17	72.28
	DR	5.76	6.77	6.50	8.81	35.45	32.40
	DR2	5.53	6.68	6.45	7.60	35.45	32.40
Squared Bias							
	Unadjusted	79.97	36.59	143.09	684.11	75.21	2023.76
	OLS	0.07	5.93	8.27	2.04	257.41	268.33
	PS	0.27	0.67	2.37	2.70	0.05	56.58
	DR	0.02	0.62	0.33	0.01	22.93	21.76
	DR2	0.01	0.76	0.37	0.09	22.93	21.76
Variance							
	Unadjusted	6.71	6.30	7.69	34.32	26.32	66.08
	OLS	5.48	5.06	5.65	13.33	27.78	14.84
	PS	5.69	6.28	6.37	7.92	13.12	15.70
	DR	5.74	6.15	6.17	8.80	12.52	10.64
	DR2	5.51	5.92	6.07	7.51	12.52	10.64

Table 2

Demographic and background characteristics of HIV positive and negative homeless men

Subject characteristics		HIV+ (n=24)	HIV- (n=281)	HIV-PS Weighted (ESS n=54)
Age		49.9	45.22 *	49.19
Race:	African American	0.79	0.71	0.81
	White	0.06	0.12 *	0.08
	Latino	0.08	0.11	0.06
	Other	0.07	0.06	0.05
Education:	Less than high school	0.24	0.27	0.28
	High school	0.23	0.38 *	0.31 *
	More than high school	0.53	0.35 *	0.40 *
Any children		0.62	0.56	0.63
Served in the military, ever		0.51	0.16 *	0.38 *
Total months homeless		96.39	62.07 *	81.97
Heterosexual		0.64	0.94 *	0.74

* Indicates those variables for which the absolute standardized mean difference between the HIV positive and negative group is > 0.2. P-values are less relevant in this analysis because of the modest sample size and hence lack of power.

Table 3

Sexual risk behaviors, drug and alcohol use, and service utilization by HIV status.

Measures	HIV+	HIV-Unadjusted	HIV-Regression	HIV-PS Weighted	HIV-Double robust
Sexual risk behaviors					
Any trading sex	65.6	40.5 [*]	51.4	56.6	57.3
Any unprotected sex	53.4	62.9	67.8	59.6	61.0
Number of partners	4.91	3.57	4.90	3.92	4.03
% casual partners	0.77	0.50 ^{**}	0.57 [*]	0.57 [*]	0.55 ⁺
Drug and alcohol use					
Alcohol use	47.5	39.3	38.2	40.6	39.3
Any hard drug use	63.8	49.7	58.3	57.0	54.8
Any marijuana use	15.6	58.7 ^{**}	58.1 ^{**}	60.9 ^{**}	58.1 ^{**}
Any rx misuse	34.6	15.4 [*]	19.8 ⁺	10.4 ^{**}	12.7 [*]
Any methamphetamine use	23.1	9.8	12.3 ⁺	5.3 ^{**}	7.8
Service utilization/Housing					
Alcohol or drug counseling	37.8	21.4	22.3	29.0	26.1
Mental health counseling	26.4	26.3	31.1	30.1	35.7
Medical and dental care	69.9	31.7 ^{**}	40.1 [*]	42.4 [*]	37.7 [*]
Regular place to stay	41.1	11.6 ^{**}	11.7 ^{**}	12.4 ^{**}	10.9 ^{**}

^{**} $p < 0.01$,^{*} $p < 0.05$, and⁺ $p < 0.10$.