Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 13, June 2015 ISSN 1531-7714

A Practical Guide for Using Propensity Score Weighting in R

Antonio Olmos & Priyalatha Govindasamy
University of Denver

Propensity score weighting is one of the techniques used in controlling for selection biases in non-experimental studies. Propensity scores can be used as weights to account for selection assignment differences between treatment and comparison groups. One of the advantages of this approach is that all the individuals in the study can be used for the outcomes evaluation. In this paper, we demonstrate how to conduct propensity score weighting using R. The purpose is to provide a step-by-step guide to propensity score weighting implementation for practitioners. In addition to strengths, some limitations of propensity score weighting are discussed.

The use of propensity scores is becoming part of the evaluation landscape (Guo & Fraser, 2015). Rosenbaum and Rubin (1983) introduced the concept of propensity score analysis to address selection bias when random assignment is not feasible. As defined by Rosenbaum and Rubin, a propensity score is the conditional probability of assignment to a treatment condition given a set of observed covariates: e = p(z=i|X). When propensity scores are used, the resulting groups will have similar characteristics to those created through random assignment. Most of the applications related to propensity scores are in matching (Thoemmes & Kim, 2011). Recently, Randolph, Fable, Manuel, and Balloun (2014) in this journal, described in detail how to conduct propensity score matching using R.

A potential drawback of propensity scores when used for matching is that a very large number of subjects may be needed, especially in the control group (Guo & Fraser, 2015). And, dependent on the specific matching technique, the use of a caliper, and the number of subjects being matched to every subject in the control group (1 or more), a large number of those subjects in the control group may not be used (see Randolph et. al, 2014 for more information about propensity score matching). Given that in evaluation settings, data collection is costly for both treatment and control subjects, techniques that may be able to use all the subjects in the study pool should be preferred to techniques that discard substantial amounts of data. Propensity scores can also be used as weights in a linear model such as regression or ANOVA, so all the subjects in the control and treatment group can be used for this application.

This article will illustrate how to use propensity scores as weights in a weighted regression using R. Program evaluators can benefit tremendously from the ability to use propensity scores to create treatment and control groups that are matched in every way except for the intervention. This is especially appealing when this ability to match individuals will not mean sacrificing individuals who cannot be matched. In that sense, using propensity scores as weights represents a very powerful combination.

Using Propensity Scores as weights in a weighted regression

The idea behind the use of propensity scores as weights is to control the influence of participants by weighting their responses based on their propensity scores (also known as reweighting, McCaffrey, Ridgeway & Morrall, 2004). The key of this analysis is the creation of weights based on propensity scores.

Thus, one advantage compared to matching is that all the individuals in the sample are used (Guo & Fraser, 2015) rather than only matched cases. The relevance of cases in the analysis however can be down-weighted dependent on their propensity scores. A second advantage of this approach is that most statistical software allows the use of weights in linear models such as regression, ANOVA, or multivariate analysis (Green, 2013). Therefore, its implementation may be easier for users who may not be familiar with R or Stata. Finally, when using propensity scores as weights, several treatment effects can be estimated. Most social scientists are familiar with the so-called **A**verage **T**reatment **E**ffect (or **ATE**), which is the difference in the outcome variable between the average score for the individuals in the treatment group and the individuals in the control group. In this paper we will present the estimation of weights using the ATE model only.

Introduction to R and the required packages

Brief introduction to R

In this article, we use R (R Core Team, 2014) to demonstrate the implementation and use of propensity scores as weights in a regression model. R is becoming an important resource in the program evaluation community because it is very powerful, it is continuously updated and maintained by the top statisticians in the world, and it is open-source software and so it is free. There are several interfaces that can be used to run the software. The authors use R-studio (RStudio, 2015). The R software requires different packages, just like other statistical programs require specific routines for specific purposes (i.e., proc's in SAS, or modules in SPSS). The user can download those packages from the Internet and install them, and it is recommended to run updates on all the different components about every two months. There are a very large number of packages intended to run almost every imaginable analysis. Although readers are encouraged to search for resources in the R project for statistical computing website (http://www.r-project.org/), Beaujen (2013) in this journal provides a very good introduction to R using Factor analysis.

For this article, there are four packages that need to be installed: 1) Generalized boosted regression models (gbm; Ridgeway, 2015), 2) procedures for psychological, psychometrics and personality research (psych; Revelle, 2015), 3) regression modeling strategies (rms; Harrell, 2015), and 4) nonparametric preprocessing for parametric causal inference (MatchIt; Ho, Imai, King, & Stuart, 2011).

Description of the data

For this demonstration, we use one of the datasets that is included in several packages used to compute propensity scores. The datafile name is "lalonde", and the version used in this paper is included in the package MatchIt. This datafile is widely used to illustrate the use of propensity score packages, and was developed by Lalonde (1986) to demonstrate the impact of a retraining program (National Supported Work Demonstration). The data file included in MatchIt contains 614 observations, with 185 in the treatment group and 429 in the control group. The outcome variable is re78, which is the income for individuals in both groups during 1978.

Steps in conducting propensity score weighting

In order to conduct an analysis involving propensity scores, the authors follow a very specific set of steps that include:

1. Outcome analysis without the use of propensity scores
2. Balance analysis prior to the implementation of propensity scores
3. Propensity score estimation
4. Weight estimation using propensity scores
5. Balance analysis after implementing propensity scores
6. Outcomes analysis using propensity scores in a weighted regression

Before the steps are detailed, readers should be aware that methodologists speak about two models when using propensity scores: 1) a **selection model**, which is intended to estimate the effect of selection bias on the treatment variable, and 2) an **outcome model**, which is intended to explore the effect of the treatment variable (and other potential covariates) on the outcome variable. Adequate selection models are critical for the success of propensity scores, whether as part of matching or as weights. Selection models should always include variables believed to have an impact on the selection process, and be based on a deep understanding of the literature related to the specific topic under study. Guo & Fraser (2015) among others have demonstrated that misspecification of the

selection model can have serious consequences in their effectiveness for controlling selection bias.

```
# Regression analysis before introducing weights

>model0 <- lm(re78 ~ treat + black + hispan + married, data = lalonde)
>summary(model0)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6292.9      566.7  11.104  < 2e-16 ***
## treat         1294.1      820.5   1.577 0.115266
## black        -1939.4      806.9  -2.404 0.016531 *
## hispan        -369.4      971.4  -0.380 0.703877
## married       2217.4      644.0   3.443 0.000614 ***


## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7360 on 609 degrees of freedom
## Multiple R-squared:  0.03579,    Adjusted R-squared:  0.02945
## F-statistic: 5.651 on 4 and 609 DF,  p-value: 0.0001803
```

Figure 1. Regression analysis before introducing propensity score as weights. Note for all the figures related to R code: # = comment, ## = output

1. Outcome analysis without the use of propensity scores

In this step, we run an outcome analysis without the use of propensity scores. This analysis is helpful to gauge what might have been the result of the outcome analysis had we not used propensity scores to control for potential selection biases associated with group assignment. Figure 1 below presents the result of this analysis.

In this example, the outcome variable is re78, the treatment variable is treat, and all the other variables are covariates that might affect the outcome. As can be observed in Figure 1, there is no treatment effect (the variable treat is not statistically significant), but the black and married covariates are statistically significant at $p < 0.05$.

2. Balance analysis prior to the implementation of propensity scores

This step is intended to assess the degree of bias between the groups **before** the propensity score is incorporated in the analysis. This step also helps to assess the degree of improvement after propensity scores are included. Typical analyses include statistical comparisons between the covariate (as DV) and the treatment variable (as IV). Figure 2 below is an example of the code used to assess statistical balance for both a continuous (re74) and a categorical variable (nodegree) using linear regression.

Figure 2 shows an imbalance for both the continuous and categorical variables included in the pre-analysis between the treatment and control groups (variable "treat"), and highlights the importance of conducting balance adjustments. Both variables show some serious discrepancies between groups, and if it is assumed that if these variables are important for group selection, then it is very important to try to correct this imbalance. Decisions like this require not only statistical knowledge but also knowledge about the field.

```
#--- re74 (a continuous variable)
>bef.re74.ATE <- lm(re74 ~ treat, data=lalonde)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5619.2      303.1  18.540  < 2e-16 ***
## treat        -3523.7      552.2  -6.381 3.46e-10 ***



#--- nodegree (a dichotomous variable)
>bef.nodegree.ATE <- glm(nodegree ~ treat, data=lalonde, family = binomial
())
summary(bef.nodegree.ATE)


## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.39189    0.09842   3.982 6.84e-05 ***
## treat        0.49433    0.18931   2.611  0.00902 **
```

Figure 2. Pre-analysis assessment of the balance for a continuous (re74) and a categorical (nodegree) variable

3. Propensity score estimation

The next step is the estimation of the propensity scores that will be used as weights in the analysis. In this paper, two techniques used to estimate propensity scores are illustrated : a) Logistic Regression, b) Generalized Boosted Models

a) Logistic regression

Logistic regression is the technique that is generally associated with propensity scores (Austin, 2011). Logistic regression is used to determine the probability of membership in the treatment or control group, given the specific set of selection variables included. Figure 3 includes the code used to: 1) estimate propensity scores using logistic regression, 2) convert the model results into predicted values that can be used as weights, and 3) bind those scores to the original data file:

The selection model presented in Figure 3 assumes that the variables age, educ, nodegree, re74, and re75 have an influence on the assignment to either the control or the treatment group. Figure 3 also shows

that both age and re75 are not statistically significant. There is some dispute in the field regarding the variables to be included in the final selection model. Authors such as Caliendo and Kopeinig (2008) suggest the addition of all the variables affecting selection, regardless of their statistical significance. This is an area where more research is needed.

```
#---Using selected covariates to estimate Propensity score
>ps<- glm(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, fami
ly =binomial())
>summary(ps)

## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.694e+00  7.989e-01  -3.372 0.000746 ***
## age          2.464e-03  1.025e-02   0.240 0.810019
## educ         1.569e-01  5.299e-02   2.962 0.003059 **
## nodegree     8.502e-01  2.813e-01   3.023 0.002503 **
## re74        -1.225e-04  2.576e-05  -4.756 1.98e-06 ***
## re75         2.574e-05  3.955e-05   0.651 0.515252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


#---Attaching propensity score to the lalonde dataset
>lalonde$psvalue <- predict(ps, type = "response")
```

Figure 3. Code to estimate propensity scores using logistic regression, create predicted values, and bind the scores to the original file.

Note: the subcommand type = "response" saves the predicted value from the logistic regression so it can be used as the propensity score.

b) Generalized Boosted Model

Models developed using logistic regression may not produce the best propensity scores . In those cases, it is suggested to try to modify the model by adding polynomial or interaction terms (Dehejia & Wahba, 1999). However, there is little guidance that can be used to determine what specific combination of terms may produce the best model. Furthermore, there is some evidence that suggests that logistic regression can be sensitive to the functional form of the relationship between the set of variables in the model and the treatment variable (McCaffrey, Ridgeway, and Morral, 2004). McCaffrey et. al. have suggested an alternative approach for the development of propensity scores by using Generalized Boosted Models. Generalized Boosted Models are based on decision trees that create a complex model by combining multiple simple models using iterative algorithms. Through this iterative process, these models will then include interactions and polynomial terms that will produce a better model without external guidance. A potential drawback of this approach for some might be the fact that the specific characteristics of the model (i.e., the variables as well as variable polynomials or interaction terms) cannot be identified. However, for the purposes of fitting the model, this may not be an important disadvantage. Readers are encouraged to go to the original sources for more information about the technique. Figure 4 includes the code needed to run a generalized boosted model.

```
#---Estimate propensity score using GBM----------------------------------
>library("gbm")

>gps<- gbm(treat ~ age + educ + nodegree + re74 + re75, distribution = "be
rnoulli", data=lalonde, n.trees=100, interaction.depth=4, train.fraction=
0.8, shrinkage=0.0005)
>summary(gps)
```

Figure 4. Code to run a Generalized Boosted Model

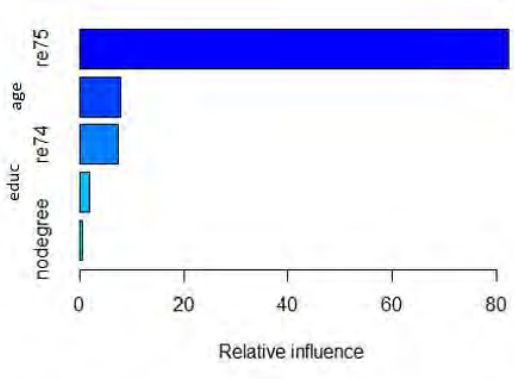
It is important to be aware that the specific fit is dependent on parameters such as the number of trees (n.trees), the interaction depth (interaction.depth), the fraction of the data used for training (train.fraction), and the threshold to stop the iterations (shrinkage). The parameters included in this example (except for the number of trees) are the defaults recommended by the developers (Ridgeway, 2015). Table 1 contains the function name and its meaning according to the gbm package (Ridgeway, 2015).

Table 1. Functions and their meanings in gbm

Function name	Function meaning
n.trees	Maximum number of iterations that gbm will perform
interaction.depth	It controls for the level of interactions allowed in gbm
shrinkage	Regulates the amount of smoothness of the resulting model
train.function	Portion of the sample used to compute out-of-sample estimates of the loss function

Figure 5 includes part of the output that indicates the relative influence of the variables in the final model, as well as the code to bind the predicted scores to the datafile. The figure shows that the variable re75 had the greatest influence in the boosted model, followed by age and re74. However, compared to the output from the logistic regression (Figure 3), it can be noticed that in this case, other than the relative influence, very little is known about the model. Readers are directed to Ridgeway, McCaffrey, Morral, Griffin & Burgette (2013) for more information about fit interpretation. It is difficult at this point to determine what approach

may be more effective for the estimation of propensity scores. Some evidence seems to suggest that generalized boosted models may outperform logistic regression (McCaffrey et al, 2004), but this is an area that needs more research. Readers are encouraged to describe in detail the procedure they followed for their development of propensity scores, as well as to scan the literature for new developments in this area.




```
##                var    rel.inf
## re75          re75 82.3429390
## age            age  7.9577202
## re74          re74  7.3572500
## educ          educ  1.9110088
## nodegree nodegree  0.4310819

#--- Attaching propensity score to the lalonde dataset
>lalonde$gpsvalue <- predict(gps, type = "response")
```

Figure 5. Relative influence of the variables in the generalized boosted model and code to bind the predicted scores to the datafile.

Note: the subcommand type = "response" saves the predicted value from the Generalized Boosted Model so it can be used as the propensity score.

4. Weight estimation using propensity scores

In order to use propensity scores in a weighted regression, the propensity scores ($\hat{e}(x)$) need to be transformed so they can be used as weights in a linear regression. The weight estimates for the **ATE** are estimated as follows: for the individuals in the treatment group:

$$\omega = \frac{1}{\hat{e}(x)} \quad (1)$$

And for the individuals in the control group:

$$\omega = \frac{1}{1 - \hat{e}(x)} \quad (2)$$

As explained earlier, every statistical software package that runs regression has routines for weighted regression. Figures 6 and 7 shows the code needed to transform the propensity scores according to equations 1 and 2. These equations can be used with propensity scores calculated using either logistic regression (Figure 6) or Generalized Boosted Models (Figure 7). The code used to compute and save weights to the dataset follows.

```
a. Estimation and storing weights using propensity score estimated using
logistic regression

#---Weights for ATE: We define the weights using :
## for the treatment group: 1/lalonde$psvalue
## for the control group: 1/(1-lalonde$psvalue)

#---Attaching weight to the lalonde dataset
>lalonde$weight.ATE <- ifelse(lalonde$treat == 1, 1/lalonde$psvalue, 1/(1-
lalonde$psvalue))
```

Figure 6. Code to transform the propensity scores to weights estimated using logistic regression.

```
b. Estimation and storing weights using propensity score estimated using
Generalized Boosted Model

# Weights for ATE: We define the weights using :
## for the treatment group: 1/lalonde$gpsvalue
## for the control group: 1/(1-lalonde$gpsvalue)


>lalonde$gweight.ATE <- ifelse(lalonde$treat == 1, 1/lalonde$gpsvalue, 1/
(1-lalonde$gpsvalue))
```

Figure 7. Code to transform the propensity scores to weights estimated using the Generalized Boosted Model

5. Balance analysis after implementing propensity scores

The ultimate purpose of using propensity scores is to balance the treatment/comparison groups on the observed covariates. This purpose does not change when using the propensity scores as weights in a weighted regression. To assess the success of the propensity scores as weights in a weighted regression for removing selection bias, a new set of tests to check the balance should be performed. Figures 8 and 9 present the results of tests similar to those presented in Figure 2. However, now weighted linear and generalized linear regressions are performed using the computed propensity scores as weights. Figure 8 presents the balance assessment for a continuous (re74) and a categorical (nodegree) variable using propensity scores weights computed using logistic regression. Figure 9 presents the balance for the same variables using propensity scores weights using the Generalized Boosted Model.

Figures 8 and 9 show that to run the balance test after creating the propensity scores the subcommand "**weights** = (**weight**.**ATE**)" for logistic regression, or

"(**weight.gATE**)" for the Generalized Boosted Model need to be added to the regression model. Figures 8 and 9 also show that although there was some improvement in the balance, there might still be some bias, since the coefficient for the variable treat is statistically significant.

```
## 1. Propensity score weights: Logistic regression

#--- re74 (a continuous variable)----------------------------------------
>aft.re74.ATE <- lm(re74 ~ treat, data=data, weights = (weight.ATE))
>summary(aft.re74.ATE)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4616.8      574.7   8.034 4.88e-15 ***
## treat         3910.1      787.2   4.967 8.82e-07 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#--- nodegree (a dichotomous variable)-----------------------------------
>aft.nodegree.ATE <- glm(nodegree ~ treat, data=data, family = binomial(),
 weights = (weight.ATE))
>summary(aft.nodegree.ATE)

## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.51160    0.08353   6.125 9.08e-10 ***
## treat       -0.32914    0.11294  -2.914  0.00357 **
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8. Balance checking using propensity score weights computed using logistic regression

```
## 2. Propensity score weights: Generalized Boosted Model

#--- re74 (a continuous variable)----------------------------------------
>aft.re74.gATE <- lm(re74 ~ treat, data=lalonde, weights = (gweight.ATE))
>summary(aft.re74.gATE)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5618.6      320.6  17.523  < 2e-16 ***
## treat        -3522.0      497.0  -7.087 3.8e-12 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#--- nodegree (a dichotomous variable)-----------------------------------
>aft.nodegree.gATE <- glm(nodegree ~ treat, data=lalonde, family = binomia
l(), weights = (gweight.ATE))
>summary(aft.nodegree.gATE)

## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3918     0.0777   5.043 4.57e-07 ***
## treat         0.4943     0.1261   3.921 8.81e-05 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9. Balance checking using propensity score weights computed using the Generalized Boosted Model

6. Outcomes analysis using propensity scores as weights in a weighted regression

The final step in the analysis is to run the outcomes model using the propensity scores as weights. In this example the outcomes model includes re78 as the outcome variable and treat, black, hispanic and married as independent variables. Figures 10 and 11 present the code to run a weighted regression to estimate ATE scores using weights computed through logistic regression (Figure 10) or the Generalized Boosted Model (Figure 11).

```
#---1.Weighted regression analysis using propensity score weights from
logistic regression

>model.ATE <- lm(re78 ~ treat + black + hispan + married, data = lalonde,
weights=(weight.ATE))
>summary(model.ATE)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2645.8      769.6   3.438 0.000626 ***
## treat         3684.0     1003.8   3.670 0.000264 ***
## black         1068.6     1076.7   0.992 0.321372
## hispan        -232.6     1389.7  -0.167 0.867133
## married       8018.0      787.4  10.183  < 2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10. Weighted regression using propensity score weights computed through logistic regression

```
#---2. Weighted regression analysis using propensity score as weights from
 Generalized Boosted Model.

>model.gATE <- lm(re78 ~ treat + black + hispan + married, data = lalonde,
 weights=(gweight.ATE))
>summary(model.gATE)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6318.6      601.8  10.499  < 2e-16 ***
## treat         1241.8      801.4   1.549  0.12179
## black        -1892.0      831.0  -2.277  0.02315 *
## hispan        -423.6     1032.0  -0.410  0.68161
## married       2163.1      665.2   3.252  0.00121 **
```

Figure 11. Weighted regression using propensity score weights computed through the Generalized Boosted Model

A comparison of the results of the analysis presented in Figures 10 and 11 to the results of the analysis presented in Figure 1 shows that the effect of the treatment was statistically significant after the groups are balanced using propensity scores. However, this was only the case when the propensity scores were estimated using logistic regression (Figure 10). There was no statistically significant treatment effect after balancing the groups using propensity scores estimated using the Generalized Boosted Model (Figure 11). The discrepancy in the results is a clear indication that the model fit achieved by both techniques is different. It also highlights the fact that more research is needed to determine which technique may yield results that are more accurate. To date, the authors are not aware of any formal comparison between the goodness of fit for these two techniques. Readers are encouraged to search the literature for new research that may help clarify this difference.

Conclusion

Propensity score analysis is a technique that has proven useful for evaluators trying to assess the outcomes of quasi-experiments and observational studies. A drawback associated with propensity scores

matching is the fact that, dependent on the specific matching technique used, there may be a need for a very large number of individuals in the control group. Furthermore, many of those individuals may not be used in the end, because they are not matched to individuals in the treatment group. This can be a major problem for evaluators trying to match individuals, because the sample sizes associated with local evaluations are not large enough to afford such waste. Similarly, treated subjects will be excluded if a match cannot be found in what is known as bias due to incomplete matching (Austin, 2014). Losing subjects in the treatment group threatens the internal consistency and generalizability of the treatment effect (Austin, 2014). Therefore, the ability to use all the individuals in the control and treatment groups for the outcomes analysis is an advantage of propensity score weighting. Not only are all the individuals included in the analysis, but statistical power to detect the treatment effect is maintained (Stone & Tang, 2013).

Similarly, using propensity scores as weights can be implemented in statistical techniques such as regression, ANOVA or any other multivariate technique that accepts weights. Glynn and Quinn (2010) propose a slight correction to the original propensity-scores-as-weights model, which makes propensity score weighting doubly robust. That is, as long as either the selection- or the outcomes-model is correct, the estimates will be satisfactory. Finally, propensity score weighting can incorporate time-dependent covariates and works with censored data (Curtis, Hammill, Eisenstein, Kramer & Anstrom, 2007; Xu et al., 2010). Readers are directed to the sources for more information about the benefits of using propensity scores as weights.

However, propensity score weighting has some limitations. In particular, it is very sensitive to the misspecification of the propensity score model (Freedman & Berk, 2008). When propensity score weights are estimated from misspecified models, they can exert a negative effect (Harder, Stuart & Anthony, 2010), and will result in biased treatment effect estimates (Stone & Tang, 2013). This limitation highlights the importance of a thoughtful specification of the selection model for the successful use of the propensity score weighting technique.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46, 399–424. DOI: 10.1080/00273171.2011.568786

Austin, P. C. (2014). Double propensity score adjustment: A solution to design bias or bias due to incomplete matching. *Statistical Methods in Medical Research*, 0(0), 1-22. DOI: 10.1177/0962280214543508

Beaujean, A. A. (2013). Factor analysis using R. *Practical Assessment, Research & Evaluation*, 18(4), 1-11. Retrieved from http://pareonline.net/getvn.asp?v=18&n=4

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72. DOI: 10.1111/j.1467-6419.2007.00527.x

Curtis, L. H., Hammill, B. G., Eisenstein, E. L., Kramer, J. M., & Anstrom, K. J. (2007). Using inverse probability weighted estimators in comparative effectiveness analyses with observational databases. *Medical Care*, 45(10), 103-107.

Dehejia, R.H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062. DOI:10.2307/2669919

Freedman, D. A., & Berk, R. A. (2008). Weighting regression by propensity scores. *Evaluation Review*, 32, 392-409.

Glynn, A. N., & Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, 18(1), 36-56. doi: 10.1093/pan/mpp036

Green, S. (2013). *Assessing sensitivity of early head start study findings to manipulated randomization threats* (Unpublished doctoral dissertation). University of Northern Colorado, Colorado.

Guo, S. & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3), 234-249.

Harrell, F. E. (2015). rms: Regression modeling strategies. R package version 4.3.1

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8), 1-28.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. American Economic Review, 76, 604-620.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9, 403-425.

R Core Team (2014). R: A language and environment for statistical computing. (3.0.3) [Computer software]. Vienna, Austria: Foundation for Statistical Computing.

RStudio (2012). RStudio: Integrated development environment for R (Version 0.96.122) [Computer software]. Boston, MA. Retrieved May 20, 2012

Randolph, J.J., Falbe, K., Manuel, A. K., & Balloun, J. L. (2014). A step by step guide to propensity score matching in R. *Practical Assessment, Research & Evaluation*, 19(18). Retrieved from http://pareonline.net/getvn.asp?v=19&n=18

Revelle, W. (2015). psych: Procedures for psychological, psychometrics and personality research. R package version 1.5.4

Ridgeway, G. (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2013*). twang: toolkit for weighting and analysis of nonequivalent groups*. Retrieved April 30, 2015, from http://cran.r-project.org/

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

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research & Evaluation*, 18(13), 1-12. Retrieved from http://pareonline.net/getvn.asp?v=18&n=13

Thoemmes, F., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, 46(1), 90-118.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, 13(2), 273-277.

Note:

The complete printout for the example described in the paper can be downloaded from http://pareonline.net/sup/v20n13sup.pdf.

Citation:

Olmos, A., & Govindasamy, P. (2015). A Practical Guide for Using Propensity Score Weighting in **R**. *Practical Assessment, Research & Evaluation*, 20(13). Available online: http://pareonline.net/getvn.asp?v=20&n=13

Corresponding Author:

Antonio Olmos
Associate Professor
Research Methods and Statistics program
Morgridge College of Education
University of Denver

email: Antonio.olmos [at] du.edu