Paper 314-2012

PROPENSITY SCORE ANALYSIS AND ASSESSMENT OF PROPENSITY SCORE APPROACHES USING SAS® PROCEDURES

Rheta E. Lanehart, Patricia Rodriguez de Gil, Eun Sook Kim, Aarti P. Bellara, Jeffrey D. Kromrey, and Reginald S. Lee, University of South Florida

ABSTRACT

Propensity score analysis is frequently used to reduce the potential bias in estimated effects obtained from observational studies. Appropriate implementation of propensity score adjustments is a multi-step process presenting many alternatives for researchers in terms of estimation and conditioning methods. Further, evaluation of the sample data after conditioning on the propensity score informs researchers about threats to the validity of the adjustments obtained from such an analysis. This paper describes the steps required for a propensity score analysis, and presents SAS code that can be used to implement each step.

INTRODUCTION

Causality and the identification of causal relationships are often sought across various disciplines. In order to verify whether X causes Y, the following must hold true: (a) X precedes Y, (b) X is related to Y, and (c) no plausible alternative explanations for Y exist other than X (Shadish, Cook, & Campbell, 2002). Experiments estimate causal inferences using a counterfactual model, which is simply the difference between what did happen after an individual received a treatment versus what would have happened if the same individual did not receive the treatment (Campbell & Stanley, 1963; Holland, 1986; Shadish et al., 2002).

A precise estimate of the treatment effect could be estimated if a unit was assigned to treatment and control condition at the same time (Holland, 1986; Rubin, 1974, 1978). However, it is impossible to assign a unit to both conditions (e.g. treatment and control) in order to yield both outcomes- the “Fundamental Problem of Causal Inference" (Holland, 1986, p. 947; Rubin, 1978). Therefore, only one outcome is observed for each unit, the outcome related to the condition the unit was assigned, while the other outcome is missing. The missing outcome is considered the counterfactual, which is why causal relationships cannot be precisely identified, only estimated.

The potential outcomes framework, also known as Rubin's Causal Model (RCM) (Holland, 1986) provides a framework of the conceptualization of causal inference. Units, treatments, and potential outcomes are the three components to RCM (Shadish, 2010). In the simplest application of this model, there are two possible treatments (e.g. treatment and control) and each individual, i, has a potential outcome for each condition: that is, $Y_i(0)$ for control and $Y_i(1)$ for treatment. For each individual, the treatment effect, τ_1, is defined as the difference between the two outcomes:

$$\tau_1 = Y_i(1) - Y_i(0)$$

However, as previously mentioned, the fundamental problem does not allow τ_1 to be calculated for each individual. Therefore, RCM presents a statistical solution to estimate the average treatment effect (ATE) for the sample, based on the counterfactual, or the expected value of the differences in outcomes. Consider an experiment with two treatment levels, t (treatment) and c (control), where Z=1 when treatment is administered to an individual and Z=0 when the individual receives the control, and Y represents the outcome variable of interest. In the counterfactual model, each individual will have one observed outcome Y, dependent upon Z, and the counterfactual will be missing. The counterfactual model estimating the treatment effect for a unit i is given by

$$Y_i = Z_i\, Y_i(1) + (1 - Z_i)\, Y_i(0)$$

Observed outcomes for each condition can be averaged, as these averages from the sample also correspond to the population average. The ATE of the population, U can be estimated from sample, u outcomes. The two distributions of observed outcomes Y_{z1} and Y_{z0} are formed by separate individuals and these distributions represent hypothetical distributions of the population, had all individuals received treatment and control respectively (Lunceford & Davidian, 2004). Consequently, the differences between the aggregated outcomes represent the ATE.

$$ATE = \bar{Y}_{z1}(u) - \bar{Y}_{z2}(u)$$

In some instances, the average treatment effect of those treated (ATT) is of interest rather than the ATT. Here, observed outcomes for those in treatment group should be compared to the observed outcomes for those individuals who may have been eligible to receive the control, but instead did not. This provides information regarding the treatment’s effect on those who participated, rather than a population average effect. The formula to estimate the ATT is the same as the ATE; however, the units are representative of the population that received the treatment.

One important aspect of RCM, is that treatment effects are estimated using information observed from different units, that is, the units in the treatment group (Z=1) are not the same as the units in the control group (Z=0). Therefore, selection of units, and assignment to treatment to draw causal inferences requires consideration of statistical design elements (Holland, 1986; Rubin, 2007, 2008). RCM, and causal inference in general, rely on several assumptions, specifically, the assumption of strongly ignorable treatment assignment and the stable unit treatment value assumption (SUTVA).

The assumption of strongly ignorable treatment assignment is satisfied when alternative explanations have been accounted for and there is no hidden bias in treatment effects. This assumption refers to the mechanism or process used to assign individuals to conditions and requires the assignment to condition be independent and not associated with the outcome or other factors. When satisfied, causal inferences can be drawn for population *U* using the average observed outcomes for all *u* in *U* exposed to *t* and *c* only (Holland, 1986). While the strongly ignorable treatment assignment assumption focuses on the process of assigning units to conditions, SUTVA focuses on the relationships between the units. SUTVA is defined as an "a priori assumption that the value of Y for unit *u* when exposed to treatment *t* will be the same no matter what mechanism is used to assign treatment *t* to unit *u* and no matter what treatments the other units receive" (Rubin, 1986, p. 961). Simply, SUTVA assumes the outcomes from two individuals, irrespective of treatment assignment, are independent from one another. When experiments employ random assignment, both these assumptions are presumed satisfied, and the estimated treatment effects are considered accurate.

In conclusion, RCM is a statistical solution that "replaces the impossible-to-observe causal effect of *Z* on a specific unit, *i*, with the possible-to-estimate average causal effect of *Z* over a population of units, *U*" (Holland, 1986, p. 947) in order to identify causal relationships among variables.

ESTIMATION

To date, researchers have most frequently used binomial regression models, i.e., logistic regression or probit models, to estimate propensity scores (Austin, 2011; Shadish & Steiner, 2010). Other propensity score estimation methods include classification trees, bagging/boosting, neural networks, and recursive partitioning (Austin, 2011). A mandatory rule that applies, regardless of the estimation method, is that all covariates that were measured *before* treatment and that are related to the treatment and outcome should be included in the initial propensity score estimation model (Shadish & Steiner, 2010),The following SAS code uses PROC LOGISTIC to estimate the propensity scores for the treatment variable GROUP, predicted from three covariates (var1 – var3). The resulting propensity scores (PS_PRED) are stored in the SAS data set PS_P.

```
proc logistic  descending data = ps_est;
   title 'Propensity Score Estimation';
   model group = var1-var3/lackfit outroc = ps_r;
   output out= ps_p XBETA=ps_xb STDXBETA= ps_sdxb PREDICTED = ps_pred;
run;
```

TREATMENT EFFECTS

There are 2 estimands of interest in propensity score (PS) analysis: 1) the average treatment effect on the treated (ATT); and 2) the average treatment effect in the population (ATE) (see Stuart, 2010; Harder, Stuart, & Anthony, 2010). Choosing the appropriate causal estimand to report for research questions, in addition to including a propensity score conditioning method that will estimate the effect, is an important decision for researchers. According to Harder et al. (2010), any propensity score conditioning method can be combined with any propensity score estimation method.

Conditioning Methods

Conditioning methods commonly used to estimate the ATT include matching (1:1, 1:*k*, full), weighting by odds, and subclassification (Austin, 2011). Conditioning methods commonly used to estimate the ATE include full matching, subclassification, inverse probability of treatment weights (IPTW), propensity score weighting, ANCOVA including propensity score as a covariate, and ANCOVA without PS (Harder et al., 2010; Stuart, 2010; Steiner et al., 2010). In this section, four conditioning methods will be discussed: stratification, 1:1 matching, inverse weighting of propensity scores, and ANCOVA including propensity scores as a covariate.

Stratification

Stratification divides individuals into many groups (or subclasses) on the basis of their propensity score values (Rosenbaum & Rubin, 1984). It is similar to full matching but creates fewer groupings (Harder et al., 2010). The optimal number of strata depends on the sample size and the amount of overlap or common support between the treatment and control groups' propensity scores. However, five subclasses, purported to remove 90% of the bias due to measured confounders, have been used by the majority of propensity score studies (Thoemmes & Kim, 2010) based upon recommendations by Cochran (1968) and Rosenbaum and Rubin (1984).

The following SAS code uses PROC RANK to divide the sample into five equal strata based on the value of the propensity score (PR_PRED). The output data set (PS_STRATARANKS) contains all of the variables contained in the input data set (PS_STRATA) and includes an additional variable (PS_PRED_RANK) that indicates the stratum number for each observation.

```
proc rank data = ps_p out= ps_strataranks groups=5;
   var ps_pred;
   ranks ps_pred_rank;
run;
```

Treatment effects are determined for each subclass and averaged across strata using stratum-specific weights:

$$\hat{\mu}_{1i} - \hat{\mu}_{2i} = \frac{\sum w_i \left(\bar{X}_{1i} - \bar{X}_{2i} \right)}{\sum w_i}$$

where $w_i = \frac{1}{SE_i^2}$ (that is, the square of the standard error of the difference between means) and the variance of the estimated mean difference is given by $Var(\hat{\mu}_1 - \hat{\mu}_2) = \frac{1}{\sum w_i}$

The following SAS code sorts the data by stratum, then calls PROC TTEST to compute the difference between the group means within each stratum as well as the standard error of this within-stratum difference. Using ODS, the mean differences and standard errors are output to the SAS data set STRATA_OUT. Within the data step WEIGHTS, the stratum-specific weights are calculated and each mean difference is subsequently weighted. The weighted means and the weights are summed using PROC MEANS (with an OUTPUT statement) and the data step TOTAL2 computes the overall estimated mean difference and its standard error.

```
proc sort data = ps_strataranks;
      by ps_pred_rank;

proc ttest;
      by ps_pred_rank;
      class group;
      var outcome;
      ods output statistics = strata_out;

data weights;
      set strata_out;
      if class = 'Diff (1-2)';
      wt_i = 1/(StdErr**2);
      wt_diff = wt_i*Mean;

proc means noprint data = weights;
      var wt_i wt_diff;
      output out = total sum = sum_wt sum_diff;

data total2;
      set total;
      Mean_diff = sum_diff/sum_wt;
      SE_Diff = SQRT(1/sum_wt);

proc print data = total2;

run;
```

Matching

The goal of matching is to obtain similar groups of treatment and control subjects by matching individual observations on their propensity scores. One of the most common matching methods used in propensity score analysis is 1:1 matching (Thoemmes & Kim, 2010) which forms pairs of treated and control subjects. Nearest neighbor (NN) or

greedy matching selects a control unit for each treated unit based on the smallest distance from that treated unit in PS. The selection process can be done without replacement, i.e., subjects are not returned to the sample after being pair-matched, therefore many of the subjects in the dataset are discarded, reducing power and generalization. Another problem associated with NN matching without replacement is that the final estimates depend on the order in which the observations are matched; therefore it is important to randomly order the sample before matching. A method used to improve the quality of paired matches is to specify a caliper width (that is, a maximum allowable difference between propensity scores if two units are allowed to be matched). Although it is difficult to know beforehand the optimal choice of caliper width, some researchers (Rosenbuam & Rubin, 1985; Austin, 2011) have recommended using a caliper width that is equal to 0.2 of the standard deviation of the logit of the propensity score, i.e., $0.2\sqrt{(\sigma^2_1 + \sigma^2_2)/2}$. Caliper width can also be designated by simply assigning a value, for example, 0.1.

Several SAS macros are available for propensity score matching (see, for example, Coca-Perraillon, 2007). However, simple 1:1 matching may be implemented with the following SAS code. This code sorts the observations into random order within each group, then transposes the data file to obtain separate data sets with treatment and control group observation identification numbers (id) and propensity scores (ps_pred). These data sets are merged in the SAS data set ALL, and matching is implemented using a series of explicitly subscripted arrays. This algorithm attempts to match each observation in the treatment group with a single observation in the control group (1:1 matching without replacement). If no control observations are available within the range specified by the caliper for a treatment observation, then no matched pair is created. The observation identification numbers for the matched pairs are stored in the SAS data set MATCHES and are printed for review.

```
data one;
 set ps_p;
 ranvar = ranuni(0);

proc sort data = one;
 by group ranvar;

proc transpose data = one out = data1;
 by group;

data id_t (rename=(COL1-COL5 = tid1-tid5));
       * Note: N of columns is number of obs in treatment group;
 set data1; if group = 1 and _NAME_ = 'id';
data ps_t (rename=(COL1-COL5 = tps1-tps5));
 set data1; if group = 1 and _NAME_ = 'ps_pred';
data id_c (rename=(COL1-COL8 = cid1-cid8));
       * Note: N of columns is number of obs in control group;
 set data1; if group = 0 and _NAME_ = 'id';
data ps_c (rename=(COL1-COL8 = cps1-cps8));
 set data1; if group = 0 and _NAME_ = 'ps_pred';

data all;
 merge id_t ps_t id_c ps_c;
 caliper = .10; * Note: caliper for matching is specified here;
 array treat_id {*} tid1-tid5;
 array ctl_id {*} cid1-cid8;
 array treat_p {*} tps1-tps5;
 array ctl_p {*} cps1-cps8;
 array used_i {*} used1 - used8;
 array matched_t {*} m_tid1-m_tid5;
 array matched_c {*} m_cid1-m_cid5;
 match_N = 0;
 do i = 1 to 5;
       min_diff = 1;
       best_match = 0;
       do j = 1 to 8;
              if used_i[j] = . then do;
                     if ABS(treat_p[i] - ctl_p[j]) < caliper then do;
                            if ABS(treat_p[i] - ctl_p[j]) < min_diff then do;
                                   min_diff = ABS(treat_p[i] - ctl_p[j]);
                                   best_match = j;
                            end;
                     end;
              end;
       end;
       if best_match > 0 then do;
```

```
            match_N = match_N + 1;
            used_i[best_match] = 1;
            matched_t[match_N] = treat_id[i];
            matched_c[match_N] = ctl_id[best_match];
      end;
 end;

data matches;
 set all;
 array matched_t {*} m_tid1-m_tid5;
 array matched_c {*} m_cid1-m_cid5;
 do match = 1 to match_N;
      Treatment_IDN = matched_t[match];
      Control_IDN = matched_c[match];
      output;
 end;
 keep match treatment_idn control_idn;
proc print;
      var match treatment_idn control_idn;
      title 'Matched Observations in Treatment and Control Groups';
run;
```

After the matched pairs have been identified, the difference between the outcome means in the two groups can be tested using PROC TTEST. It is important to specify a correlated-means *t*-test (rather than an independent-means *t*-test) because the matched pairs impose a correlated structure to these data. Use of the PAIRED statement with PROC TTEST will request the appropriate version of this inferential procedure.

Inverse Probability of Treatment Weights

In IPTW, individuals are weighted by the inverse probability of receiving the treatment that they actually received. Treated individuals receive an IPTW equal to $1/p_i$ and control individuals receive a weight equal to $1/(1-p_i)$ (Harder et al., 2010). The weights are then used in a weighted least squares (WLS) regression model along with other predictor covariates. The IPTW method is inclusive of all subjects in a study, therefore no loss of sample occurs as in other conditioning methods, i.e., matching, stratification. A drawback of the IPTW method is the possibility of extreme propensity scores that can result in very large weights that can bias the treatment effect estimates (Austin, 2011; Shadish & Steiner, 2010). Bias from extreme weights can be adjusted using a stabilization technique (Harder et al., 2010; Robins et al., 2000) which multiplies the treatment and comparison weights by a constant or by using a trimming technique (Harder et al., 2010) that trims the stabilized weights within a specified range.

The following SAS code creates an IPTW weight variable (PS_WEIGHT) as described above. The mean weight is computed using PROC MEANS with an OUTPUT statement. This mean weight is used in the data step PS_WEIGHT2 to normalize these weights.

```
data ps_weight;
set ps_p;
if group = 1 then ps_weight = 1/ps_pred;
else  ps_weight = 1/(1-ps_pred);
run;

proc means noprint data = ps_weight;
 var ps_weight;
 output out = q mean = mn_wt;
run;

data ps_weight2;
 if _n_ = 1 then set q;
 retain mn_wt;
 set ps_weight;
 wt2 = ps_weight/mn_wt; * Normalized weight;
run;
```

The weights are used to estimate the treatment effect via weighted least squares. The estimated treatment effect may be obtained using PROC GLM or PROC REG (both are illustrated below). Note the use of the WEIGHT statement

with these PROCs to obtain weighted least squares estimates.

```
proc glm data = ps_weight2;
       class group;
       model outcome = group / ss3 solution;
       weight wt2;
       means group;
       run;

proc reg data = ps_weight2;
       model outcome = group;
       weight wt2;
       run;
```

ANCOVA Using the Propensity Score as a Covariate

This estimation method uses the propensity score and treatment status to predict the potential outcome (Austin, 2011; Shadish & Steiner, 2010). If the outcome is continuous, then an OLS regression model is selected and the treatment effect is estimated by the adjusted difference in means. If the outcome is dichotomous, a logistic regression model is used and the treatment effect is estimated by the adjusted odds ratio (Austin, 2011). Since there are no subjects discarded because of non-overlap, generalizability is maintained, assuming that the relationship between the propensity score and the outcome has been correctly specified.

This analysis is illustrated below using PROC GLM. The outcome variable is modeled as a function of both the propensity score and the treatment group. Requesting the Type III sum-of-squares analysis (using the SS3 option on the MODEL statement) ensures that the group difference is tested after adjusting for the propensity score. The LSMEANS statement provides the adjusted means for the two groups.

```
proc glm data = ps_p;
       class group;
       model outcome = group ps_pred / ss3;
       lsmeans group;
run;
```

EVALUATION

The quality of the propensity scores estimated in a sample are evaluated using two types of comparisons: comparing the distributions of the PS across the two groups and comparing the distributions of each covariate across the two groups.

COMMON SUPPORT

The comparison of PS distributions is typically referred to as the evaluation of common support. Ideally, propensity score distributions would overlap entirely (see Figure 1) indicating that observations from both groups are available across the range of the PS. Figure 2 illustrates samples with substantial areas of non-overlap. Group 0 provides no observations with large PS and Group 1 provides no observations with small PS. Such a comparison indicates that comparison of the two groups on the outcome variable will be compromised, even after conditioning on the propensity scores. Trimming the samples by discarding cases in the region of non-overlap appears to restore the interval validity of subsequent group comparisons, at a cost of generalizability (we cannot estimate the treatment effect for people with very high or very low PS).

Methods for evaluating common support include graphical displays and comparison of summary statistics from each distribution. A formal assessment of the equivalence of the two distributions (e.g., K-S test) may be an appropriate tool as well. The side-by-side boxplots illustrated here can be obtained using PROC BOXPLOT. The data are first sorted by group and common options used with PROC BOXPLOTS are implanted in this code.

```
proc sort data = ps_p;
       by group;

proc boxplot data=ps_p;
       symbol width = 2;
       plot ps_pred*group /
              cboxes=black
              cframe = white
              idsymbol = circle
              idcolor = black
              font='times new roman' height=3.5
```

```
            boxwidth=6
            boxstyle=schematic
            waxis = 2
run;
```

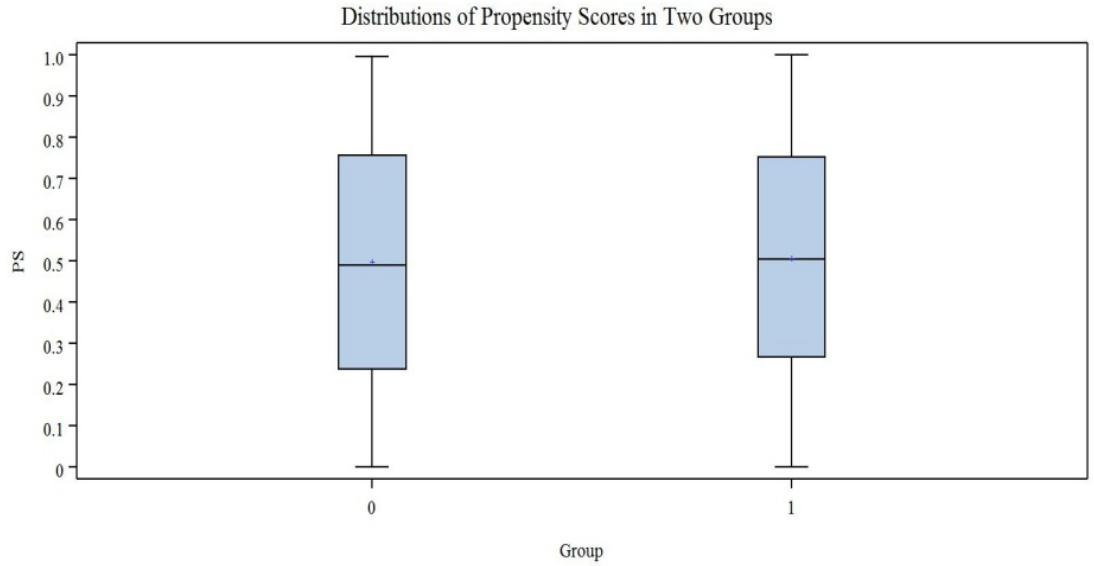

Figure 1. Distributions of Propensity Scores with Ideal Common Support

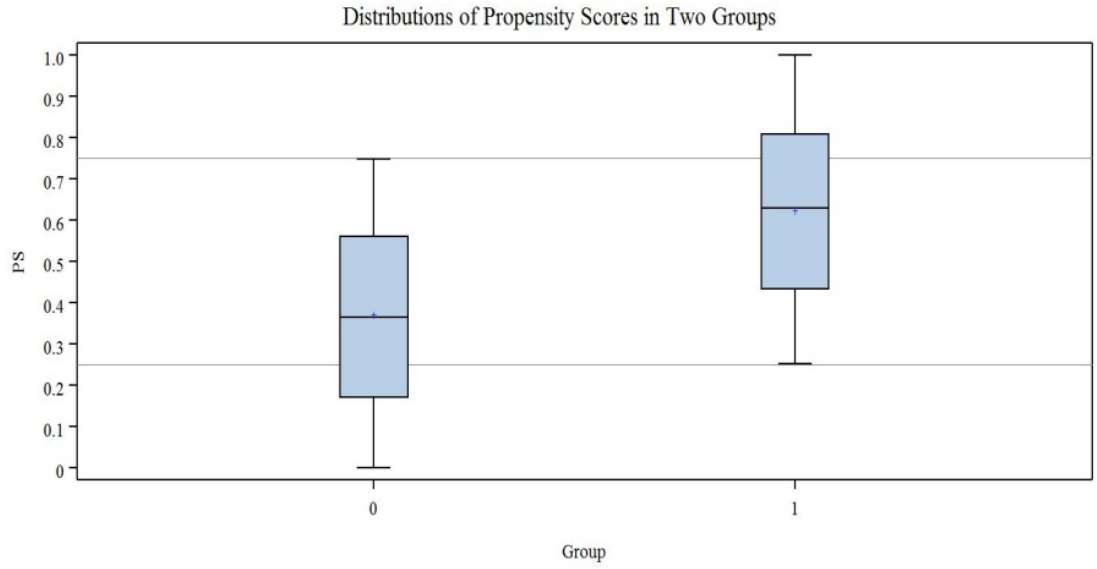

Figure 2. Distributions of Propensity Scores with Substantial Areas Lacking Common Support

BALANCE

The comparison of the distribution of each covariate across the two groups is referred to as evaluation of balance. Unlike the evaluation of common support, assessment of balance takes place after conditioning on the PS. In ideal circumstances, after conditioning the two groups would be very similar on all covariables.

The assessment of balance usually takes the form of computing a standardized mean difference between the two groups and standardized differences that are less than some arbitrary criterion indicate balance on these covariates. This technique is sometimes advocated for dichotomous as well as continuous covariates (although many researchers are more comfortable with using odds ratios for dichotomous variables). In addition, hypothesis testing is sometimes used to evaluate balance, but most researchers do not advocate this. When hypothesis testing is applied, researchers should take into account the probability pyramiding that naturally occurs when multiple hypotheses are

tested (if fewer than 5% of the null hypotheses are not rejected, we have evidence of balance).

Such assessments are focused on the similarity of means in the two groups, but a comparison of variances should also be undertaken. In addition, a comparison of distribution shapes or a multivariate approach to comparing two groups (Mahalanbois distance as a multivariate version of the standardized mean difference) may be appropriate (but these are rarely seen in published applications).

The evaluation of balance should reflect the type of conditioning that is planned for the evaluation of treatment effects (e.g., if stratification is planned, then evaluate balance within each stratum; if weighting is planned, then use a weighted analysis to assess balance). Evidence of imbalance suggests that the propensity score adjustment will not be sufficient to remove the effects of the covariates that are not balanced. The typical recommendation in such a circumstance is to return to the statistical model that gave rise to the PS. A re-estimation of PS using a more elaborate model (with interaction effects or polynomial terms) may provide sample propensity scores that will balance the groups.

An elegant approach to comparing the distributions of continuous covariates across the two groups is to use PROC NPAR1WAY, as demonstrated below. This procedure provides a graphical overlay of the cumulative distributions for the two groups. The code below illustrates the use of ODS to create an RTF file containing the output from PROC NPAR1WAY.

```
ods graphics on;
ods html file = 'balance_covar_edfplots.rtf';
proc npar1way d edf  plots=edfplot data=ps_p scores=data ;
  class group;
  var var2;*example continuous covariate;
run;
ods graphics off;
ods html close;
```

SENSITIVITY ANALYSIS

The legitimacy of propensity score analysis is based on the assumption of strongly ignorable treatment assignment which assumes all relevant covariates are employed in the treatment assignment and the bias due to the unmeasured covariates is ignorable (Thoemmes & Kim, 2011). However, testing this assumption is empirically impossible without the access to the unmeasured covariates. Alternatively, through sensitivity analysis researchers explore how sustainable the treatment effect is with the potential effect of unmeasured covariates. If the estimated treatment effect is sensitive to the presence of unmeasured covariates, or in other words, the estimated treatment effect is possibly washed away with the unmeasured covariates, the treatment effect may be due to the bias of unobserved covariates rather than a true effect. On the other hand, if a considerable magnitude of unobserved covariate effect is not likely to mitigate the treatment effect, researchers gain confidence on the treatment effect as an unbiased estimate.

Re-estimation of Treatment Effects

Given the unmeasured covariates (U), the treatment effect can be re-estimated as follows:

$$\delta^{*} = \delta - \gamma(\mathrm{E}[U_1] - \mathrm{E}[U_0])$$

where δ is the treatment effect after controlling for the observed covariates, γ(E[U1] - E[U0]) is the effect of unobserved covariates, and δ* is the adjusted treatment effect. That is, the adjusted treatment effect (δ*) can be obtained by removing the hidden bias due to unmeasured covariates (γ(E[U1] - E[U0])) from the estimated treatment effect (δ).

Sensitivity analysis is a type of what-if analysis because the effects of unmeasured covariates with two sensitivity parameters (γ and E[U_1] - E[U_0]) are not empirically estimable. The proxy of sensitivity parameters can be obtained either from the observed data or from theory and literature (Li, Shen, Wu, & Li, 2011), which requires researchers' substantive knowledge on the research field and thoughtful inspection on the observed data. For illustration, we adopt Hong's (2004) approach when matching is used for conditioning. Hong identified a covariate with the largest standardized coefficient in predicting the treatment effect assuming the impact of unobserved covariates does not exceed that of any observed covariate. The corresponding unstandardized coefficient and the unstandardized mean difference between treatment and control groups of the covariate are employed as a proxy of γ and E[U_1] - E[U_0], respectively.

The SAS code presented below illustrates the use of PROC REG in combination with PROC SQL to identify the covariate with largest standardized partial regression weight when all covariates are used to predict the outcome variable. The standardized mean difference between treatment and control groups for this variable is then obtained. Finally, these statistics are used to estimate the treatment effect in the presence of a hypothetical unmeasured covariate with these attributes.

```
proc reg data=ps_est;
       model outcome = var1 var2 var3 /stb ;
       ODS output ParameterEstimates= regcoeff ;

 run ;

* Find the largest standardized regression coefficient;

proc sql;
       create table mxcoeff as
       select Variable as mxvar, Estimate as mxest, StandardizedEst as
        mxstdest
       from regcoeff
       having StandardizedEst = max(StandardizedEst);

quit;

data _null_; set mxcoeff;
       call symput('mxvar', mxvar);

 run;

/*************************************************************************/
* Compute the standardized and unstandardized mean differences of observed
covariates. See Faries, Haro, Leon, and Obenchain (2010) for the macro;
* Suppose that "diff" is the data file of the standardized and unstandardized mean
differences;
* Obtain the unstandardized mean difference of the identified variable with the
largest standardized coefficient ;
/*************************************************************************/

data diff_mx ;set diff;
       where label ="&mxvar";run;

* "Product" is the effect of the unobserved covariates;

data mxcoeff_merged ;
       merge mxcoeff diff_mx;
       product = mxest*d_un ;
run;
```

Evaluation of Hidden Bias on the Treatment Effect

Reestimating the treatment effect (δ^*) by adding or subtracting the product of the identified sensitivity parameters and reconstructing confidence intervals of the adjusted treatment effect researchers evaluate the impact of the hidden bias on the treatment effect. When the impact is trivial not altering the statistical inference, the treatment effect can be considered as an unbiased estimate with confidence.

```
/*************************************************************************/
* Extract statistics from the dependent samples t-test which estimates the
treatment effect after controlling for observed covariates (i.e., matching);
/*************************************************************************/

ods output "Statistics" = stats
           "T-Tests" = ttests ;
proc ttest data= matched;
       paired outcomeC*outcomeT;
 run;
ods output close;

/*************************************************************************/
* Reestimate the treatment effect considering the hidden bias due to
unmeasured covariates (product) and reconstruct the 95% confidence interval ;
/*************************************************************************/

data sensitivity ;
       merge mxcoeff_merged stats;
       treateff_adj_lower = Mean - product;
       treateff_adj_upper = Mean + product;
       CL_adj_lower = treateff_adj_upper - (1.96*StdErr) ;
       CL_adj_upper = treateff_adj_upper + (1.96*StdErr) ;
       keep N Mean LowerCLMean UpperCLMean StdErr treateff_adj_upper CL_adj_lower
            CL_adj_upper treateff_adj_lower;
```

```
proc print data = sensitivity;
       var N Mean LowerCLMean UpperCLMean StdErr treateff_adj_upper CL_adj_lower
           CL_adj_upper treateff_adj_lower;
run;
```

CONCLUSION

Propensity score analysis plays an important role in making causal inferences for observational studies (Rosembaum & Rubin, 1983) and it is regarded as an effective covariate-balancing strategy over other matching techniques for obtaining unbiased estimates of treatment effects for causal inferences (Bai, 2011). However, potential selection bias is still possible and researchers should put special care when conducting PS analysis to avoid making unwarranted causal claims. SAS provides the flexibility to implement a variety of procedures that may be used in propensity score analyses. The aim of this paper was to illustrate some of these SAS procedures and to offer practical guidelines related to the treatment of data required when conducting propensity score analysis, from covariate selection, estimation of the propensity score, and its application in conditioning methods, to evaluating the effectiveness of the PS for balancing treatment and control groups, before estimating the effects of treatment.

It is important for researchers to clearly and explicitly report each step of such an analysis, from the selection of covariates through the final estimation of treatment effects. Details on the procedures followed to assess balance and common support, as well as the results of such assessments, allow readers to identify the credence with which causal claims may be founded. The methods by which propensity scores were estimated, and the conditioning methods employed should be presented as thoroughly and transparently as possible to allow other researchers to replicate these methods in future research. Finally, clearly presented sensitivity analyses provide critical information related to the potential of unmeasured covariates to alter conclusions about estimated treatment effects.

REFERENCES

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46(1)*, 399-424.

Bai, H. (2011). Using propensity score analysis for making causal claims in research articles. *Educational Psychology Review, 23*(2), 273-278.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: RandMcNally.

Coca-Perraillon, M. (2007). Local and global optimal propensity score matching. *SAS Global Forum*, Statistics and Data Analysis, 1-9.

Cochran, W. G. (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295-313.

Faries, D. E., Haro, J. M., Leon, A. C., & Obenchain R. L. (2010). *Analysis of observational health care data using SAS*. Cary, NC: SAS Institute.

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

Harding, D. (2003). Counterfactual models of neighborhood effects: The effect of neighborhood poverty on dropping out and teenage pregnancy. *American Journal of Sociology, 109*(3), pp. 676-719. Retrieved from http://www.jstor.org/stable/10.1086/379217

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945-970.

Hong, G. (2004). *Causal inference for multi-level observational data with application to kindergarten retention.* (University of Michigan, University of Michigan). *ProQuest Dissertations and Theses,* Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305181900?accountid=7082

Li, L., Shen, C., Wu, A. C., & Li, X. (2011). Propensity score-based sensitivity analysis method for uncontrolled confounding. *American Journal of Epidemiology, 174*(3), 345-353. doi:10.1093/aje/kwr096

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score estimation of causal treatment effects: A comparative study. *Statistics in Medicine 23*(19), 2937-2960.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550-560.

Rosenbaum, P. R. (1995). Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association, 90*(432), pp. 1424-1431. Retrieved from http://www.jstor.org/stable/2291534

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society.Series B (Methodological), 45*(2), pp. 212-218. Retrieved from http://www.jstor.org/stable/2345524

Rosenbaum, P. R. & Rubin, D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516-524.

Rosenbaum, P. R. & Rubin, D. B. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*, 33-38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology,* 66, 688-701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annuals of Statistics*, 6, 34-58.

Rubin, D. B. (1986). Statistics and causal inference: Comment: What ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961-962.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20-30.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808-840.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity score: Relating theory to practice. *Biometrics, 52,* 249-264.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods* 15(1), 3-17.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Belmont, CA: Wadsworth, Cengage Learning.

Shadish, W. R. & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews, 10(1)***,** 19-26.

Steiner, P. M., Cook, T. D., Shadish, W. R. & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15(3)***,** 250-267.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1-21.

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

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Reginald Lee
Enterprise: University of South Florida
Address: 4202 E Fowler Avenue EDU105
City, State ZIP: Tampa, FL 33617
Work Phone: 813.974.6457
Fax: 813.974.5132
E-mail: rlee@usf.edu
Web: