Propensity Score Analysis in R: A Software Review

Bryan Keller & Elizabeth Tipton

Teachers College, Columbia University

Author Note

We thank the authors of the R packages reviewed herein for making their software open source. Furthermore, we thank Beth Ann Griffin, James Helmreich, James Honaker, Gary King, Jasjeet Sekhon, and Elizabeth Stuart for helpful discussion, and the Software Review Editor for helpful comments.

Abstract

In this article we review four software packages for implementing propensity score analysis in R: **Matching**, **MatchIt**, **PSAgraphics**, and **twang**. After briefly discussing essential elements for propensity score analysis, we apply each package to a data set from the Early Childhood Longitudinal Study in order to estimate the average effect of elementary school special education services on math achievement in fifth grade. In the context of this real data example, we evaluate documentation and support resources, built-in quantitative and graphical diagnostic features, and methods available for estimating a causal effect. We conclude by making some recommendations aimed at helping researchers decide which package to turn to based upon their familiarity with propensity score methods, programming in R, and the type of analysis being conducted.

Propensity Score Analysis in R: A Software Review

Introduction

Propensity score methods are commonly used to estimate the causal impact of a treatment or intervention when random assignment is not possible. The methods are particularly useful in situations in which it is unethical or impossible to randomly assign an intervention (e.g., student retention, Hong & Raudenbush, 2006; special education services, Morgan, Frisco, Farkas, & Hibel, 2010; smoking, Rubin, 2001; mode of child delivery, Li, Kleinman, & Gillman, 2014). If the treatment had been assigned to participants randomly, then the treated and untreated participants would be identically distributed on all pretreatment covariates, both observed and unobserved, on average (what is referred to as *balance*). In the absence of random assignment, however, those receiving the treatment can differ in systematic ways from those not receiving the treatment (what is referred to as *confounding*). Propensity scores provide a method for adjusting for these systematic differences, allowing for an unbiased treatment effect to be estimated.

Over the past ten years, the use of propensity score methods in the social and educational sciences has grown rapidly. The methodologies are now commonly used to answer questions regarding the effectiveness of programs, as well as for more novel uses, such as verifying assumptions in multiple imputation (van Buuren, 2012), equating tests (Longford, 2015) and making generalizations from experiments (Tipton, 2014). At least in part, this useage has increased in response to the advent of new statistical software implementing the methods, including packages in R, and macros in Stata, SAS, and SPSS. As many have noted (see Stuart, 2010), the gap between propensity score theory and application can be large, and understanding the differences between these statistical methods and programs can be difficult.

In this article, we provide a review of software for implementing propensity score analysis. We assume that the reader is familiar with key concepts in the propensity score literature; readers interested in a broad overview of these methods should see Stuart (2010)

and Steiner and Cook (2013). We limit our scope to full-featured packages currently available in R (R Core Team, 2015). We focus here on comparing the features of the programs, highlighting the types of output that is available, as well as places where a new user may be confused. Those interested in implementation of a particular package should see the associated tutorial paper or manual, which we cite in each section. We review packages in R both since the software is free and since it is typically the program in which new statistical methods are first made available. We define full-featured to mean that the package includes functionality to (a) assess the adequacy of propensity score estimates through balance plots or summary statistics and (b) use the propensity score estimates to estimate an average causal effect and its standard error. The four packages we review are **Matching** (Sekhon, 2011), **MatchIt** (Ho, Imai, King, & Stuart, 2011), **PSAgraphics** (Helmreich & Pruzek, 2009), and **twang** (Ridgeway, McCaffrey, Morral, Burgette, & Griffin, 2015).

In this review, we begin by introducing an example which we use throughout. We then sequentially review features of each package, with a focus on (a) documentation and support resources with an eye for any challenges a new R user might face, (b) quantitative overlap and balance diagnostics, (c) graphical displays for overlap and balance, and (d) methods for conditioning on the estimated propensity scores to estimate an average causal effect and its standard error. In particular, we examine (b) through (d) in the context of a real data example. Finally, we conclude the review with a comparison of packages aimed at helping researchers decide which package to turn to based upon their familiarity with propensity score methods, programming in R, and the type of analysis being conducted.

The Data

Following Morgan et al. (2010), we use data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K), to examine the effect of student participation in special education in 2002 on math and reading scores in 2004. Thirty five pretreatment covariates were selected by Morgan et al. (2010) based on theory and

previous results in the literature that linked them to special education placement. After we deleted cases without complete covariate information 7362 cases remained, 429 of which were associated with students who had received special education services. A dichotomous variable indicating if a student lived with neither biological mother nor biological father was eliminated because after listwise deletion there were no positive cases.

A child was defined as a recipient of special education services based on his or her special education status gathered from school administrative records from the spring of 2002; for our analysis, the outcome of interest is the ECLS-K Revised IRT scaled math achievement test score. The variables used in the analyses are described in Table 1. For each variable, we include the means and standard deviations in the treatment and control groups, as well as two numerical summaries typically used to assess balance, the standardized mean difference (d) and the variance ratio (r), defined as follows,

$$d = \frac{\bar{X}_T - \bar{X}_C}{s_{\text{pooled}}}, r = \frac{s_T^2}{s_c^2}$$

where X_T and X_C are the sample means and s_T^2 and s_C^2 are the sample variances of the treated and comparison groups, respectively, and $s_{\text{pooled}} = \sqrt{\frac{(N_T-1)s_T^2+(N_C-1)s_C^2}{N_T+N_C-2}}$ is the pooled standard deviation across groups, where N_T and N_C are the treated and comparison group sample sizes, respectively[1]. The goal of propensity score analysis is to reduce these imbalances.

Data Analysis

In this section we use each software package to conduct a propensity score analysis to estimate the average treatment effect on the treated (ATT) of exposure to special education services on mathematics achievement. Note, however, that the analyses we carry out are illustrative; resultant estimates should not be interpreted as robust estimates of the

[1]Standardized differences for measuring balance may be calculated differently for different causal estimands. For example, when estimating the ATT, some prefer to standardize by the sample standard deviation of the treatment group alone.

Table 1

Variable Names, Descriptions, and Unadjusted Standardized Mean Differences (d) and Variance Ratios (r) by Special Education Status (T = Received Services, C = Did Not Receive Services)

Variable Name	Description of Variable	Values	d	r
DEMOGRAPHIC				
GENDER	Male	0, 1	0.38	0.88
WKWHITE	White	0, 1	0.17	0.79
WKSESL	Socioeconomic Status	[-4.8, 2.8]	-0.29	0.89
ACADEMIC				
RIRT	Kindergarten Reading Score	[23.17, 139.36]	-0.65	0.53
MIRT	Kindergarten Math Score	[11.9, 99.0]	-0.71	0.77
S2KPUPRI	Public School	0, 1	0.44	0.25
P1EXPECT	Parental Expectations	Integers 1–6	-0.32	1.22
P1FIRKDG	First-Time Kindergartener	0, 1	-0.41	3.26
P1AGEENT	Child's Age at K Entry (Months)	[54, 79]	0.08	1.08
apprchT1	Approaches to Learning Rating	Integers 1–4	-0.70	1.20
P1HSEVER	Attended Head Start	0, 1	0.19	1.42
chg14	Ever Changed Schools	0, 1	0.02	1.09
SCHOOL COMPOSITION				
avg_RIRT	Reading IRT	[27.9, 80.0]	-0.23	0.79
avg_MIRT	Math IRT	[16.1, 66.1]	-0.18	0.82
avg_SES	SES	[-2.2, 2.5]	-0.16	0.88
avg_apprchT1	Approaches to Learning	[1.5, 4.0]	-0.14	0.80
S2KMINOR	Percent Minority Students	Integers 1–5	-0.20	0.77
FAMILY CONTEXT				
P1FSTAMP	Received Food Stamps	0, 1	0.12	1.26
ONEPARENT	One-Parent Family	0, 1	0.13	1.22
STEPPARENT	Stepparent Family	0, 1	0.05	1.19
P1NUMSIB	Number of Siblings	[0, 10]	0.16	1.17
P1HMAFB	Mother's Age at First Birth	Years [12, 45]	-0.26	1.00
WKCAREPK	Nonparental Pre-K Child Care	0, 1	-0.07	1.14
HEALTH				
P1EARLY	Number of Days Premature	[0, 112]	0.19	2.05
wt_ounces	Birth Weight (Ounces)	[17, 214]	-0.11	1.24
C1FMOTOR	Fine Motor Skills	Integers 0–9	-0.63	1.27
C1GMOTOR	Gross Motor Skills	Integers 0–8	-0.43	1.54
PARENT RATING OF CHILD				
P1HSCALE	Overall Health	Integers 1–5	0.12	1.17
P1SADLON	Sad/Lonely	Integers 1–4	0.10	1.32
P1IMPULS	Impulsive	Integers 1–4	0.41	1.55
P1ATTENI	Attentive	Integers 1–4	0.72	1.45
P1SOLVE	Problem Solving	Integers 1–4	0.68	1.55
PSPRONOU	Verbal Communication	Integers 1–4	0.86	1.51
P1DISABL	Child has Disability	0, 1	0.82	2.38
OUTCOME VARIABLE				
C6R4MSCL	Fifth Grade Math Score	[50.9, 170.7]	-0.77	1.40

causal effect. We discuss the following steps in conducting a propensity score analysis.

1. **Estimation of propensity scores.** We use built-in functionality for propensity score estimation, if provided; otherwise we use a logistic regression through the `glm` function in R with one main effect for each covariate.

2. **Assessment of overlap.** In a propensity score analysis, it is important for every unit to have a non-zero probability of receiving both the treatment and control. This means the distributions of propensity scores in the treatment and control groups must share common support. Here we look for functionality to detect and handle lack of common support in the estimated propensity score distributions. We also report on plots that can be used to visually assess common support.

3. **Assessment of balance.** When the treatment and comparison groups differ in distribution on important confounding covariates, as found in the example, the naive treatment effect estimate is biased. The goal of the propensity score analysis is therefore to reduce this bias by improving covariate balance. The standardized mean difference and the variance ratio are two common measures of balance, though others also exist, including distributional measures like the Kolmogorov-Smirnov distance. For each package, we report on balance statistics produced and the manner in which they are reported. In addition, we focus on balance plots, which can provide an overall sense of covariate balance before and after conditioning on estimated propensity scores.

4. **Estimation of the average causal effect and its standard error.** In the propensity score literature, there are three estimands that can be easily defined: the average treatment effect (ATE), the average treatment on the treated (ATT), and the average treatment on the control (ATC). Each of these answer different questions and users should think in advance about which is appropriate for their analysis. In the example used here we focus on estimation of the ATT. Throughout, however, we note

both which estimands are available in each package, as well as methods available for conditioning on the estimated propensity score to get an estimate of the treatment impact and whether standard error estimation takes into account variability due to the estimation of the propensity scores themselves.

Matching

As the name suggests, the **Matching** package is centered around the implementation of *matching* methods for causal inference. In addition to matching based on propensity scores, the package also features Mahalanobis distance, inverse variance, and genetic matching. The implementation of genetic matching for causal inference, which was proposed by Diamond and Sekhon (2010) and algorithmically developed by Mebane, Jr and Sekhon (2011), is a unique contribution made by this package. However, because genetic matching is not a propensity score method we do not attend to it in this review.

The primary resource for the package is a *Journal of Statistical Software (JSS)* paper (Sekhon, 2011), which features a summary of the theoretical background for matching methods and several examples with code. New users interested in propensity score analysis will likely be most interested in Section 3.1, entitled "Propensity score matching example", which uses data from LaLonde's (1986) job training experiment to demonstrate propensity score matching. The example is well-written and the code is easy to follow, with helpful guidance for new R users. For example, on p. 9 the author notes that after running several commands no output has been produced yet; he then describes how to coax R into displaying output. This aspect of the object-oriented nature of R may be surprising to those used to programs such as SPSS, which produce copious output after every command.

1. Estimation of propensity scores. The package does not include built-in methods for estimating propensity scores; rather, it relies upon existing generalized linear modeling machinery in R. Thus, our first step in analyzing the ECLS-K data was to estimate propensity scores using a logistic regression model with one main effect for each covariate.

2. Assessment of overlap. A straightforward way to limit analysis to areas of common support on the estimated propensity score when using a matching method is through the use of a propensity score caliper; this is the approach recommended in the user manual (Sekhon, 2013), though neither "overlap" nor "common support" are mentioned in the JSS paper. The number of units discarded due to lack of a suitable match within the caliper is recorded (for our analysis the number was 18 based on a caliper size of 0.1 pooled standard deviations). No plots of common support are available.

3. Assessment of covariate balance. Two functions may be used in tandem to produce balance statistics. First, `Match` is run to create matches based on the propensity score; second, `MatchBalance` is run on the output from the `Match` function along with a formula that includes the variables for which balance statistics are desired. The printed output from the `MatchBalance` function includes a variety of useful balance statistics for the baseline data and for the data after matching. The printed output is stacked vertically for each covariate; for the 34 covariates in our analysis the function printed about 15 pages of output. If the output is saved as a named object, it is stored as a list with three nested layers (values nested within covariates nested within list). Because of the list format some data-wrangling is necessary to produce a table based on balance statistics. Output for RIRT, a continuous variable representing pretest reading scores, is shown below. Output for dichotomous predictors is identical except that Kolmogorov-Smirnov tests are not reported.

```
***** (V4) RIRT *****
                       Before Matching     After Matching
mean treatment........     35.242          35.429
mean control..........     42.607          36.052
std mean diff.........     -87.59           -7.34

mean raw eQQ diff.....     7.4525           3.601
med  raw eQQ diff.....       6.52           1.545
max  raw eQQ diff.....     31.845           39.29
```

```
mean eCDF diff........     0.2175       0.053429
med  eCDF diff........    0.22951       0.031347
max  eCDF diff........    0.39192        0.17758

var ratio (Tr/Co).....    0.53001         1.4112
T-test p-value........ < 2.22e-16        0.14528
KS Naive p-value...... < 2.22e-16     < 2.22e-16
KS Statistic..........    0.39192        0.17758
```

R's `qqplot` function may be used to create quantile-quantile (QQ) plots of individual covariates separated by treated and comparison groups to illustrate balance before and after propensity score matching; see Figure 1 for an example. No balance plots are implemented through the package.

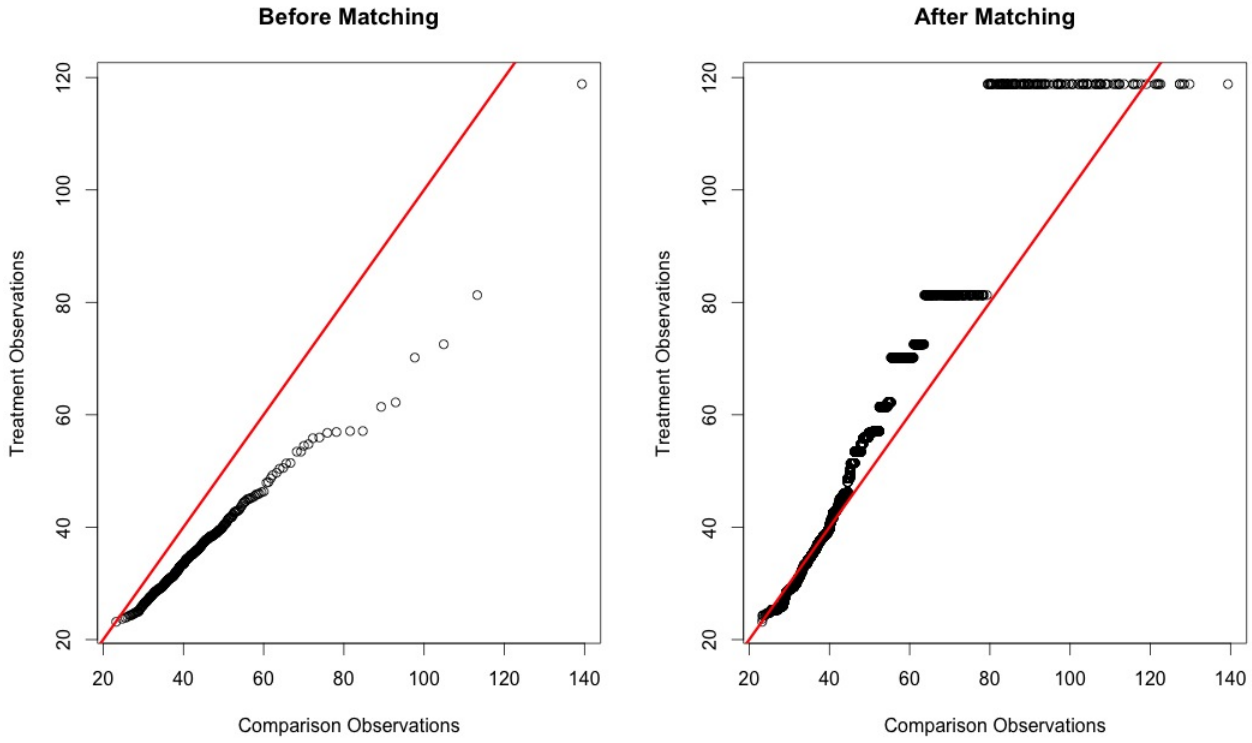

Figure 1. Empirical QQ Plots of Reading IRT Score Pretest Before and After PS Matching

4. Estimation of the ATT and its standard error. The treatment effect

estimate and its standard error are calculated through the `Match` function. By default **Matching** reports the Abadie and Imbens (2006) standard error (AI SE) that accounts for the uncertainty of the matching procedure, although the usual model-based standard error is also reported if requested. The unadjusted difference (T - C) was -19.2 points. After one-to-one nearest neighbor propensity score matching with replacement (default settings) with a caliper of 0.1 pooled standard deviations, we found that special education had a significant and negative effect on mathematics achievement of -5.14 points (AI SE = 1.624, p-value = 0.002).

MatchIt

MatchIt was created as a companion to a paper by Ho, Imai, King, and Stuart (2007) to facilitate matching techniques for causal inference in R; an overview and tutorial for the package are provided in Ho et al. (2011). In the framework outlined by Ho et al. (2007), matching techniques are advocated as part of doubly robust estimation. An estimator that models both the selection mechanism (e.g., through propensity score matching) and the outcome (e.g., through a regression analysis) is referred to as *doubly robust* if correct specification of at least one of the models implies the estimator is consistent (Bang & Robins, 2005). Thus, the authors refer to propensity score matching as "preprocessing" because they have in mind that the next step will be to apply a parametric linear or generalized linear regression analysis to estimate the average causal effect and its standard error.

Of course, this type of estimation is possible in other packages as well. For example, propensity score weights produced by package **twang** can be incorporated into a weighted least squares regression of the outcome variable on the covariates. However, the dual-modeling approach is central to the Ho et al. (2007) paper and, thus, the language permeates the documentation for the **MatchIt** package.

1. Propensity Score Estimation. The default option for estimation of propensity scores is logistic regression, however, many more options are available. Any of ten link

functions for generalized linear models, ten link functions for generalized additive models, neural networks, or classification trees may be used within the package to estimate propensity scores. Following Ho et al. (2011) we used logistic regression for propensity score estimation.

2. Assessment of Overlap. The package offers two options for labeling cases as non-overlapping; the first is based on the estimated propensity score distributions and the second is based on on the convex hull. The *convex hull* of a set of points in a covariate space is defined as the smallest convex subset of the covariate space that contains all the points; a set is *convex* if for any points A and B in the set, the straight line from A to B is also in the set. King and Zeng (2006) proposed the convex hull for assessing common support and their approach is implemented in **MatchIt**. Whether the propensity score or convex hull are used to label cases as non-overlapping notwithstanding, several options for the deletion of non-overlapping cases are available: (a) remove cases in the comparison group that are outside the support of the treatment group, (b) remove cases in the treatment group that are outside the support of the comparison group, and (c) a combination of (a) and (b). The `matchit` function is the main workhorse for the package and the function includes an argument `discard` that triggers the assessment and treatment of non-overlapping data.

The `plot` command with argument `type = "jitter"` produces a jittered scatterplot of propensity score distributions by treatment status, before and after matching; see Figure 2. The plot allows the user to point and click to identify particular points, which is useful for identifying non-overlapping units. Alternatively, the same information can be displayed with histograms by specifying `type = "hist"`.

3. Assessment of Covariate Balance. Nearest neighbor matching is the default option for the `matchit` function, however, optimal matching, full matching, genetic matching and subclassification are also available. **MatchIt** calls package **optmatch** (Hansen & Klopfer, 2006) for optimal matching and full matching and calls package **Matching** (Sekhon, 2011) for genetic matching. The `summary` function is used to display

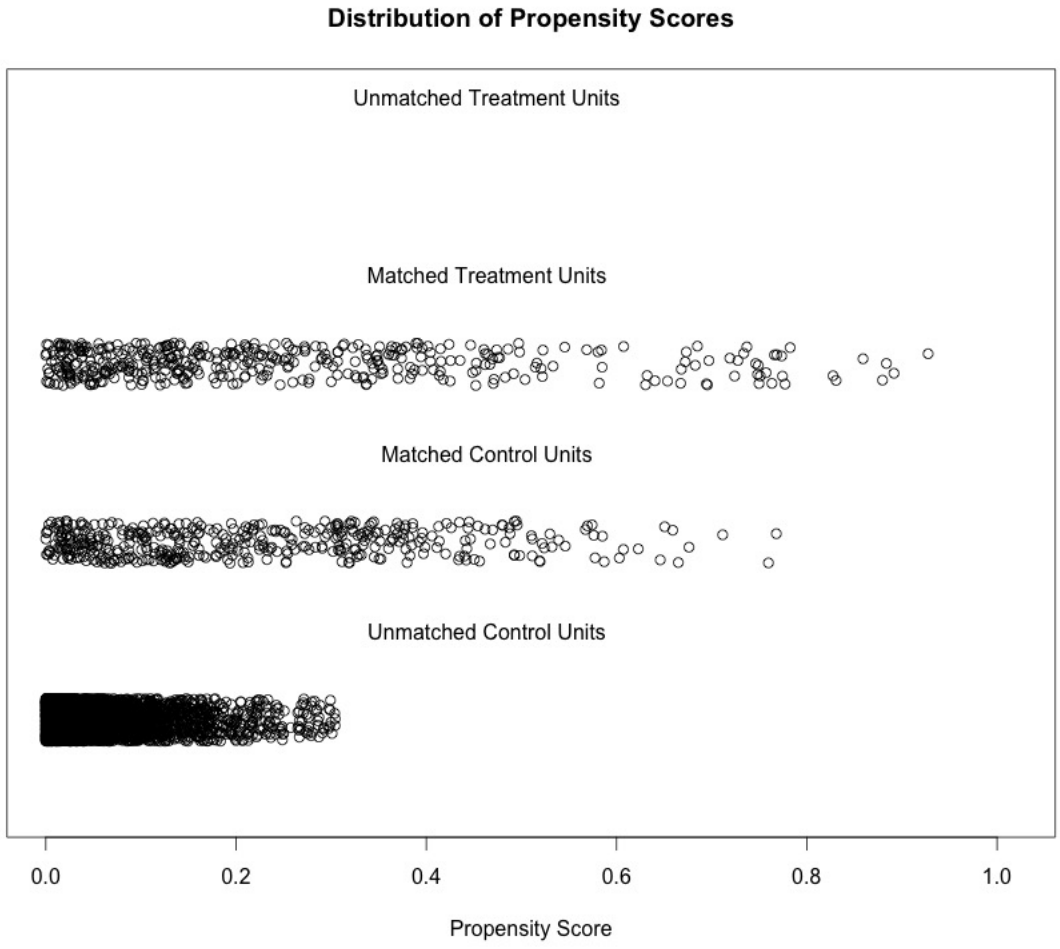

Figure 2. Jitter Plot of Propensity Score Distributions Produced by Package **MatchIt**

balance statistics before and after matching. The first six rows of the after-matching balance tables are displayed below. Note that the first row, labeled "distance", represents balance for the distance measure used for matching; here the distance is based on the estimated propensity score, so the distance row shows the balance on the estimated propensity score itself.

```
Summary of balance for matched data:
```

	Means Treated	Means Control	SD Control	Mean Diff	eQQ Med	eQQ Mean	eQQ Max
distance	0.2433	0.2171	0.1693	0.0262	0.0001	0.0263	0.2070
GENDER	0.6737	0.6713	0.4703	0.0023	0.0000	0.0023	1.0000
WKWHITE	0.8182	0.8112	0.3918	0.0070	0.0000	0.0070	1.0000

```
WKSESL          -0.0408      -0.0333     0.7474   -0.0076  0.0300   0.0392  0.5600
RIRT            35.2418      35.9584     7.3288   -0.7165  1.2450   1.4648 12.7600
MIRT            26.7245      27.6646     7.8351   -0.9401  1.3550   1.4397 20.2550
```

The output also displays the percent improvement due to matching for balance statistics and reports sample sizes and number of units discarded (not shown here). By default, `summary` does not report *standardized* mean differences, only the actual mean difference; standardized mean differences may be produced by adding the argument `standardized = TRUE` to the function call. One other argument worth mentioning is the `interactions` option. Setting this argument equal to `TRUE` triggers the reporting of balance on all two-way interactions. For our example with 34 covariates plus the propensity score itself there are $\binom{35}{2} = 595$ two-way interactions, 35 first-order terms, and 35 quadratic terms, for a total of 665 rows in the balance table. This option facilities the examination of balance on second-order terms, a recommended practice in assessing the adequacy of propensity score estimates (Hill, 2008; Lee, Lessler, & Stuart, 2009; Keller, Kim, & Steiner, 2015).

The default action of the `plot` function is to produce QQ plots of treatment vs comparison units before and after matching for each variable. The QQ plot for the kindergarten reading score (RIRT) variable is displayed in Figure 3.

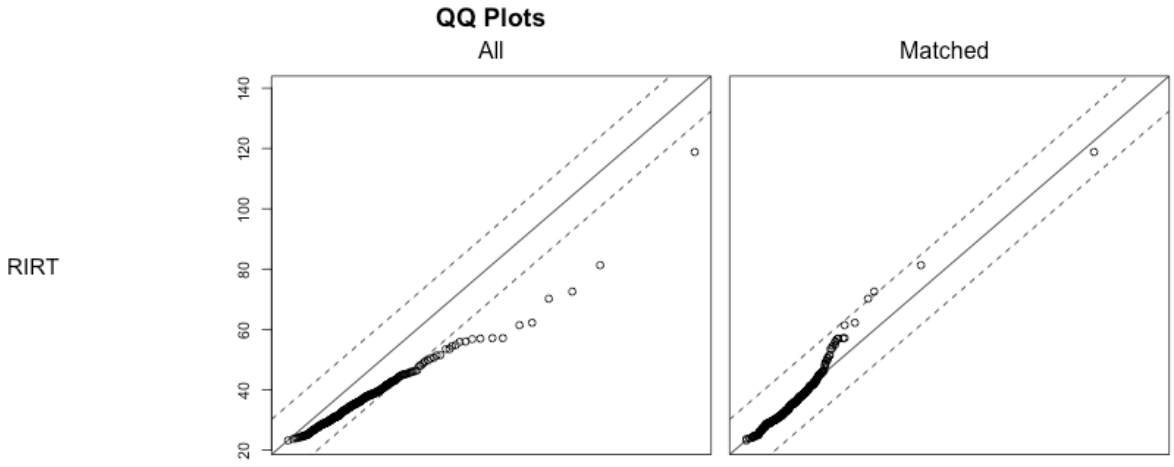

Figure 3. QQ Plots of Treated vs Comparison Units Before and After Nearest Neighbor Matching for the Kindergarten Reading Score (RIRT)

Estimation of the ATT and its standard error. MatchIt does not include options for estimating causal effects or standard errors. Ho et al. (2011) provide several examples estimating causal effects using package **Zelig** (Owen, Imai, King, & Lau, 2013) with data that has been preprocessed by **MatchIt**, though they note that any software for parametric statistical analyses may be used to analyze data that has been preprocessed by **MatchIt**. Neither package **MatchIt** nor package **Zelig** offer options for estimating adjusted standard errors that account for uncertainty due to the estimation of propensity scores.

For those unfamiliar with package **Zelig**, the examples given in Ho et al. (2011, pp. 12-13) may be difficult to follow because they focus on Monte Carlo simulation to estimate the causal effect. Furthermore, examples demonstrating estimation of the ATT and ATE produce errors when run in the current version (4.2-1) of **Zelig**, though this functionality will likely be restored in a forthcoming update (James Honaker, personal communcation, September 5, 2015). Although simulation is featured in the examples, it is nevertheless possible to estimate the ATT through maximum likelihood-based regression in **Zelig**. With our data, after preprocessing in **MatchIt** via nearest neighbor matching where only comparison group cases were dropped, the ATT was estimated to be -5.04 (SE = 1.23, p-value < 0.001).

PSAgraphics

The **PSAgraphics** package (Helmreich & Pruzek, 2009) is aimed at providing graphics for balance checking and to help visualization and interpretation of results. The focus of the package is on stratification; neither matching nor weighting is supported.

1. Estimation of propensity scores. Though propensity scores are not estimated within the package itself, Helmreich and Pruzek (2009) provide examples using logistic regression and recursive partitioning via package `rpart` (Therneau, Atkinson, & Ripley, 2015). We used logistic regression with one term for each covariate to estimate propensity scores for our analysis.

2. Assessment of overlap. The `loess.psa` function produces a scatterplot of the outcome on the vertical axis against the propensity score on the horizontal axis, with separate nonparametric loess regressions by treatment group; see Figure 4. Helmreich and Pruzek (2009) suggest examining rug plots produced at the top and bottom margins of the `loess.psa` plot to assess propensity score overlap.

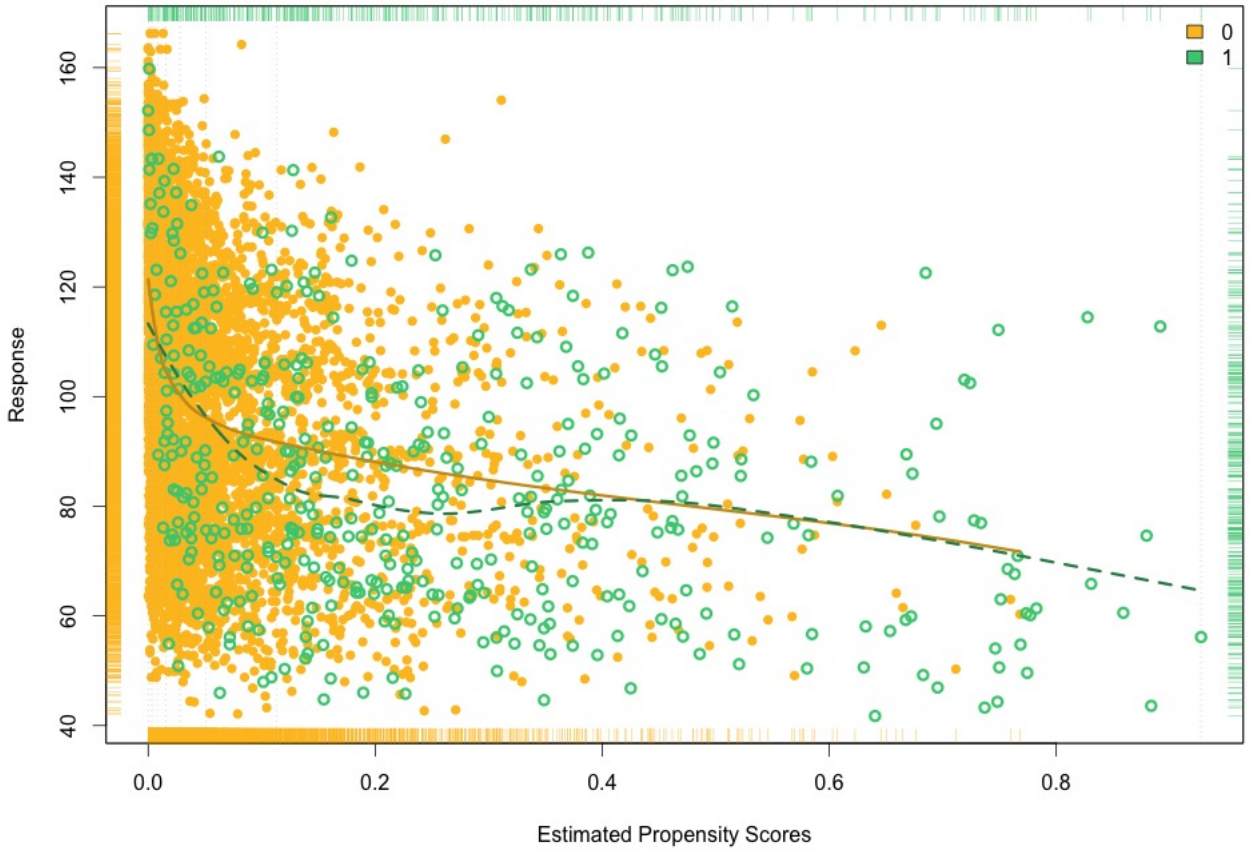

Figure 4. Plot Produced by the `loess.psa` Function in Package PSAgraphics; Open Circles and the Dashed Line Represent Cases that Received Special Education Services, Filled Circles and the Solid Line Represent Comparison Group Cases

3. Assessment of covariate balance. With propensity score stratification, covariate balance must be examined within each individual stratum. Lack of balance suggests more strata may be necessary; however, the trade-off between bias and variance limits the number of strata because as the number of strata increases, the within-strata

sample sizes and, hence, the stability of the estimates of treatment effect, decreases. Thus, with large sample sizes, it makes sense to use more strata. With smaller sample sizes, regression estimation within-strata may be used to reduce residual bias due to an insufficient number of strata (Lunceford & Davidian, 2004). In **PSAgraphics**, side-by-side boxplots for each strata are produced for continuous covariates; see Figure 5 for an example with the continuous WKSESL variable. Side-by-side barplots showing proportions within-strata are produced for categorical covariates (not shown).

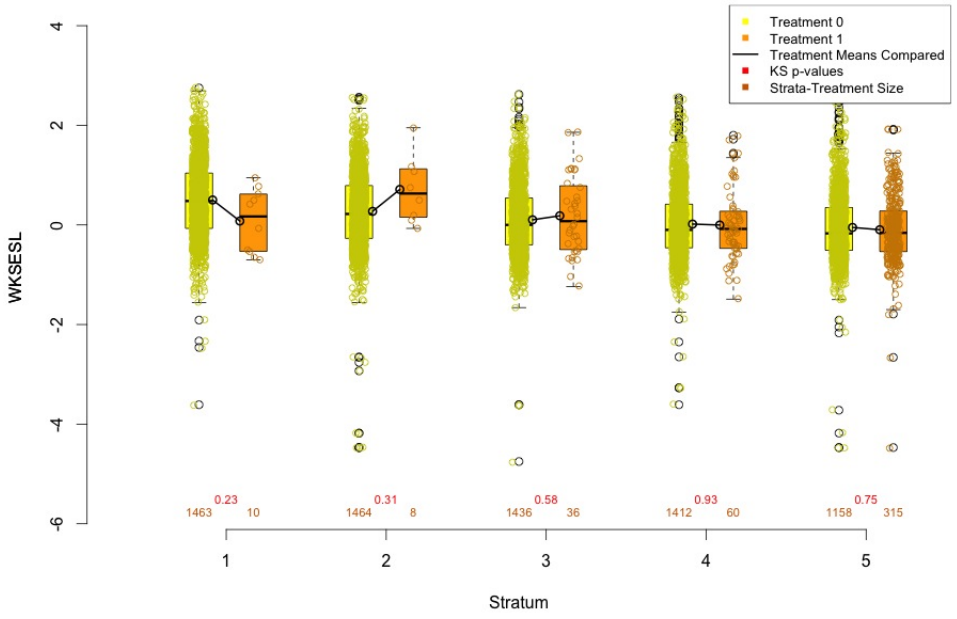

Figure 5. Boxplots Produced Within Strata by `box.psa` for Continuous Variable WKSESL

Though the emphasis is on the production of graphics to display balance, the plotting functions in **PSAgraphics** also produce stratum-specific p-values based on the Kolmogorov-Smirnov test for continuous covariates and Fisher's exact test for categorical covariates; standardized mean differences are not produced.

Estimation of the ATT and its standard error. The `circ.psa` function creates a plot that includes both stratum-specific and overall estimates of the average treatment effect. The estimates of the ATE, its standard error, and 95% confidence interval are

reported as output; adjusted standard errors are not available. For our analysis, the within-strata sample sizes and means were reported as follows.

```
$summary.strata
   n.0 n.1  means.0  means.1
1 1463  10 140.8739 156.2850
2 1464   8 133.4403 142.5263
3 1436  36 127.3028 121.0078
4 1412  60 121.9288 118.5260
5 1158 315 114.2722 103.4142
```

The `loess.psa` function produces a scatterplot of the data with points differentiated by treatment group and displays separate nonparametric loess regression fits for the treatment and comparison cases, respectively (see Figure 4). `loess.psa` also estimates the ATE by using a stratification approach based on a number of bins or cut points set by the user. The difference here is that the ATE is estimated based on the predicted values from the loess fits within each stratum; this is a within-strata regression estimation approach. Both functions `circ.psa` and `loess.psa` produce plots and generate estimates only for the overall ATE; the ATT is not supported. Because our goal for data analysis was to estimate the ATT with each package, we used the stratum weights for the treated cases, reported above in column "n.1", to re-weight the means in order to provide an estimate of the ATT; formulas for ATT weighting with stratification are provided, for example, in Schafer & Kang, 2008, p. 292. We estimated the ATT based on the stratification scheme above to be -8.45 (SE = 1.26, $p < 0.001$).

twang

Package **twang**, short for Toolkit for Weighting and Analysis of Nonequivalent Groups, implements propensity score-based weighting approaches for the estimation of average treatment effects. The main supporting document for the package is a tutorial

paper available through the Comprehensive R Archive Network (CRAN) (Ridgeway et al., 2015), and there is a website devoted to the package at www.rand.org/statistics/twang. The paper is easy to read and begins with an example using data from LaLonde's (1986) job training experiment (the same data used in the **Matching** documentation).

1. Estimation of propensity scores. The development of **twang** was motivated by the use of generalized boosted modeling for propensity score estimation, as described in McCaffrey, Ridgeway, and Morral (2004). Thus, while it is possible to conduct analyses using propensity scores estimated by logistic regression or other methods in **twang**, we use GBM here because it is a focal point of the package.

Two advantages to using GBM relative to logistic regression are (a) it automatically handles interactions and other nonlinear relationships between the response variable (in this case, the logit of the propensity score) and the predictors, which frees the analyst from the potentially time consuming iterative process of balance checking and model re-specification, and (b) the algorithm produces a quantitative ranking of relative influence for each covariate; see Figure 6. Disadvantages include (a) increased computational time[2] and (b) some of the time saved on model specification will be spent learning about and selecting optimal values for the various tuning parameters for GBM.

2. Assessment of overlap. A number of plots can be produced within **twang** based on the estimated propensity scores, including side-by-side boxplots of propensity scores by treatment group, shown in the left panel of Figure 7. Although boxplots do allow for visual assessment of overlap at the lower and upper extremes, they obscure any gaps in common support that occur in the interior of the distributions.

3. Assessment of Covariate Balance. Balance statistics are produced by calling the `bal.table` function. When GBM is used to estimate propensity scores, `bal.table` produces a table of balance statistics at baseline (i.e., before propensity score weights are applied) and a table of balance statistics after propensity score weights are applied for each

[2]For our analysis (7362 cases on 34 covariates), the estimation of propensity scores using GBM with default values for tuning parameters took a little less than nine minutes on a 2014 iMac running at 3.2GHz with 8GB RAM.

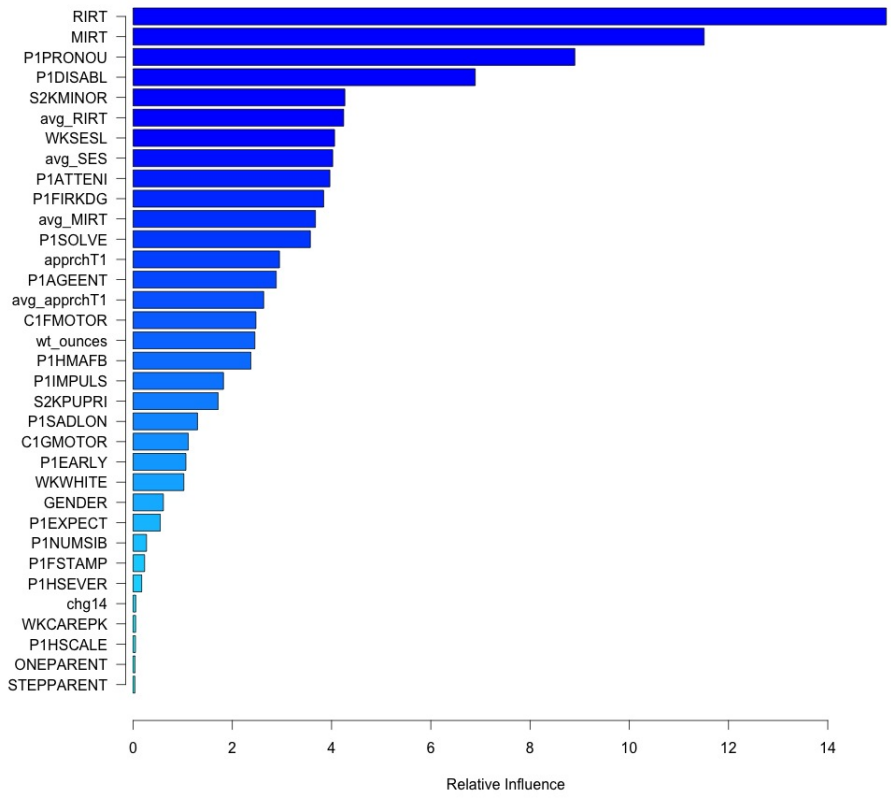

Figure 6. Relative Influence of Covariates on the Estimated Propensity Score

stopping method specified when running the GBM algorithm. Three tables were produced for our example: baseline, es.mean.ATT, and ks.mean.ATT. The latter two denote that the number of iterations was selected that minimized the standardized mean difference or the mean of the Kolmogorov-Smirnov statistics, respectively. For our example, this iteration number happened to be the same for both methods, so plots and tables are identical. We display the first five rows of the es.mean.ATT table below. The output is succinct and the table formatting makes it relatively easy to export to word processing programs such as LaTeXor Microsoft Word.

```
$es.mean.ATT

              tx.mn  tx.sd   ct.mn  ct.sd std.eff.sz   stat     p    ks ks.pval
GENDER        0.674  0.469   0.617  0.486      0.121   1.951 0.051 0.057   0.322
```

WKWHITE	0.818	0.386	0.791	0.406	0.070	1.241	0.215	0.027	0.984
WKSESL	-0.041	0.718	-0.003	0.745	-0.053	-0.814	0.416	0.038	0.797
RIRT	35.242	8.409	36.513	9.484	-0.151	-2.722	0.006	0.090	0.022
MIRT	26.725	8.779	27.904	8.821	-0.134	-2.420	0.016	0.070	0.128

We display two plots for balance assessment, though more are available. The right-hand panel of Figure 7 shows the change in absolute value of standardized mean difference for each covariate before and after applying propensity score weights. Bold lines indicate an increase in standardized mean difference, lighter lines indicate a decrease; filled circles indicate a statistically significant difference, open circles a non-significant difference. Other balance plots based on p-values from Kolmogorov-Smirnov tests and t tests are available, though not shown here.

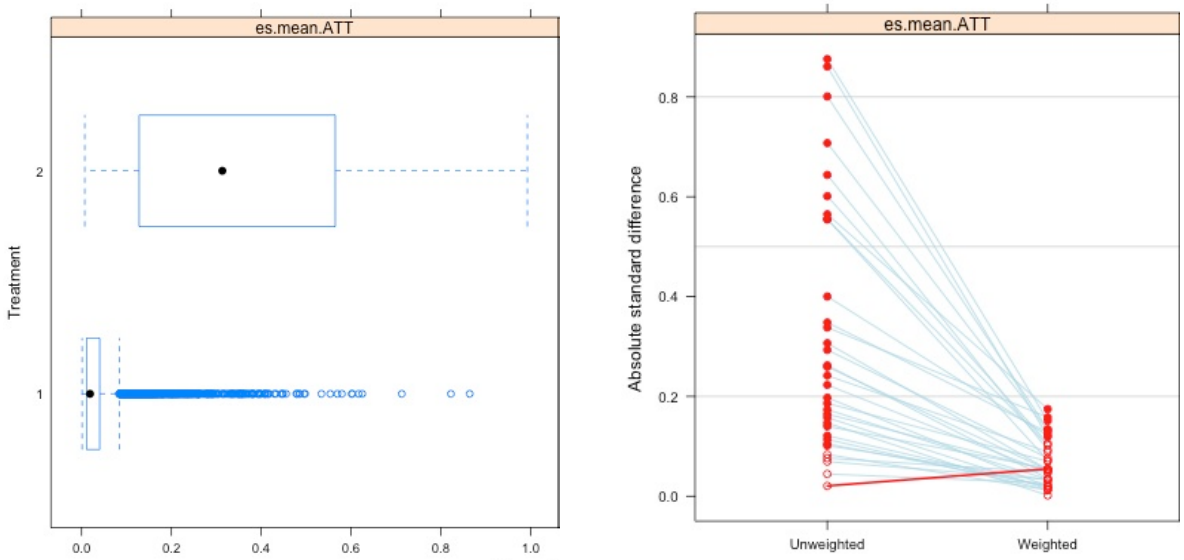

Figure 7. Left Panel: Boxplots of Estimated Propensity Scores by Treatment Group; Left and Right Panels Differ as to the Stopping Criteria Used for the GBM Algorithm; Right Panel: Absolute Values of Standardized Mean Differences Before and After Weighting

The Reading IRT pretest score (RIRT) was the only variable with a significant p-value on the Kolmogorov-Smirnov test after PS weights were applied; RIRT was also the most influential covariate in estimating the propensity score according to Figure 6. Because reading is also an important predictor of math achievement, the outcome variable, it is

arguably one of the most important covariates in terms of bias reduction potential. At baseline, the special education group scored about 88% of a pooled standard deviation lower than the comparison group. After propensity score weighting the difference was reduced to 15% of a pooled standard deviation. However, given the importance of the reading pretest variable, a careful analyst might not be satisfied.

This situation highlights one of the complications that can be encountered when using data mining methods to estimate propensity scores; it is not always clear how to proceed to make targeted changes to improve upon unsatisfactory balance. With a logistic model we might consider adding transformations of, or interactions with, the reading variable to improve upon balance. Such model-based changes do not, however, translate directly to GBM because the interaction depth has already accounted for up to three-way (by default) interactions.

Estimation of the ATT and its standard error. There are no options for estimation of causal effects built into **twang**. Instead, the tutorial paper provides an example of how to use the **survey** package (Lumley, 2014), which accounts for weights in the estimation of standard errors, to estimate treatment a treatment effect and its standard error. For our analysis, the ATT was estimated to be -7.29 (SE = 1.69, p-value < 0.001).

The use of GBM to estimate propensity scores here would necessitate a computational approach such as the bootstrap or jackknife to estimate standard errors that account for uncertainty due to propensity score estimation; these approaches are not implemented in the package. While not covered here, **twang** also includes relatively new functionality for causal effect estimation in the presence of multiple (i.e., more than two) treatments (McCaffrey et al., 2013). Research in this area is new and ongoing, so users should expect future packages and updates in this area.

Recommendations

In this section, we make recommendations to both users and those creating and maintaining packages. For users, the answer to the question “Which package should I

use?" depends largely on two aspects: (a) level of expertise with respect to R and (b) preference for propensity score estimation and application techniques and plots that may or may not be offered by each package. Regarding the second point, we encourage readers to consult Table 2, which compares the essential features of each of the packages we review herein. With respect to the first point, we make some recommendations.

New R users will likely prefer a package that offers defaults that provide them with the variety of information required for each of the four steps. Unfortunately, as Table 2 highlights, none of the packages excel in all four areas. For example, only two packages – **Matching** and **PSAgraphics** – provide a fully integrated outcome analysis, whereas the other two packages direct users to outside packages for analysis (i.e., to **Zelig** for **MatchIt**, and to **survey** for **twang**). However, **PSAgraphics** only provides an option for estimating the ATE, not the ATT, and **Matching** does not provide convenient output for covariate balance or adequate defaults for assessing common support.

For new users, a package with easy to use defaults for statistical and graphical balance assessment may be more useful. Here the **twang** and **MatchIt** packages excel. Both provide balance output in tabular form and in terms of standardized mean differences for each covariate. Both also provide useful visualization methods, including methods for assessing common support (e.g., parallel boxplots in **twang**; jittered scatterplots and histograms in **MatchIt**). In terms of assessing balance, **MatchIt** provides diagnostic plots for each covariate (i.e., Q-Q plots) and has an option for an overall summary plot like that produced by **twang**, shown in the right panel of Figure 7, while **twang** provides only overall summary plots of balance. In comparison, **PSAgraphics** offers several plotting options too, though these are only available for a stratification approach to conditioning on the estimated propensity score. Furthermore, **PSAgraphics** does not produce a dedicated plot for assessing common support. Rug plots, which can be used for examining overlap, are displayed in the upper (treated cases) and lower (comparison cases) margins of plots produced by the `loess.psa` function (see Figure 4); however, calls to `loess.psa` also produce output including the estimated ATE and its standard error, which is at odds with

Table 2: Comparison of features of the packages under review

Category	Matching	MatchIt	PSAgraphics	twang
(1) Propensity score estimation				
Logistic PS estimation	N	Y	N	N
Other PS estimation	N	CT[†]; GAM[‡]; NN[¶]	N	GBM[§]
(2) Overlap /Common Support				
Method	Caliper	Min/Max; CH[§§]	N	N
Option to discart T/C/Both	Only Both	Y	N	N
(3a) Conditioning				
Nearest neighbor	Y	Y	N	N
Subclassification	N	Y	Y	N
Full Matching	N	Y+; optmatch	N	N
Optimal Matching	N	Y+; optmatch	N	N
Weighting (IPW)	N	Y	N	Y
Genetic Matching	Y	Y+; matching	N	N
(3b) Balance Assessment (Statistics)				
Concise table	N*	Y	N	Y
Standardized mean differences	Y	Y	N	Y
Variance ratios	Y	N	N	N
Q-Q comparisons	Y	Y	N	N
Distributional comparisons	Y	Y	N	Y
Statistical Tests	KS[¢]; t[¢¢]	N	KS[¢]; FET[¶¶]	KS[¢]; t[¢¢]
Variables not in model	Y	Y	N	N
(3c) Balance Assessment (Graphical)				
PS distribution	N	Y	N	Y
Covariate distributions	N	Y	Y	Y
Other	N	N	N	Y**
(4) Outcomes				
User calls other package	N	Zelig	N	survey
Estimands	ATE, ATE, ATC	ATE, ATT, ATC	ATE	ATE, ATT
Unconditional SE	Abadie-Imbens	N	N	N

*Notes: † CT = classification trees; ‡ GAM = generalized additive models; ¶ NN = neural networks; § GBM = generalized boosted models; §§ CH = convex hull; Y+ = package calls another package (specified after ";"); * tables of covariate balance produced, but separately, as a long list; ¢ KS = Kolmogorov-Smirnov test; ¢¢ t = t test; ¶¶ FET = Fisher's exact test for categorical variables; ** plots indicating influence of each covariate on the PS, and balance by iteration.*

the goal of separating design from analysis in quasi-experimental research (cf. Rubin, 2007). **Matching** does not produce plots at all as a default, making it perhaps the least friendly towards new users.

For advanced users who are more comfortable manipulating objects and making graphics in R, the **Matching** and **twang** packages may be of particular interest, as both of these packages provide users with advanced methods (i.e., genetic matching for **Matching** and GBM for **twang**). The packages also offer other benefits, including several options for standard errors (i.e., **Matching**) and methods for assessing covariate influence (i.e., **twang**).

These recommendations can also be read as recommendations to those creating and maintaining packages. While R is being used by more and more analysts in a variety of fields, the ability to use its advanced features is not as common. For this reason, a package developed in R with good defaults covering all four of the steps used in an analysis will reach a larger variety of users than a package with a more limited scope. For those not interested in developing a full set of tools, they may wish to then design their package to easily integrate with other existing packages, either by being called by those other packages or by calling these other packages themselves. The key focus should be on providing a broad set of defaults, including easy to read tables and simple standard graphics. In the end, the ideal - which is currently not available - is for there to be a package that a new user can easily use from the beginning of the analysis (i.e., propensity score estimation) until the end (i.e., estimation of ATT or ATE and its standard error).

Other Resources

This paper has focused on reviewing packages in R that use propensity score based methods to estimate causal effects. Not reviewed here are packages or macros in other software environments; those interested should see Elizabeth Stuart's website, which contains several resources http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html. Notably, the current

set of tools includes nearly as wide a variety of options in Stata, with fewer in SAS, and only one (with very limited capabilities) in SPSS.

While not covered here, *sensitivity analysis* is a method for assessing how statistical inferences based on propensity score analysis would change in the presence of hidden biases of different sizes (see Rosenbaum, 2002). Sensitivity analysis may be condcuted in R with the **rbounds** package (Keele, 2014) and the `sensitivity` function in package **twang**.

Finally, in the area of causal effect estimation, there is also a growing body of research that focuses on the maximization of covariate balance directly, without necessarily requiring propensity score estimation. Those interested may want to explore packages in R focusing on mixed integer programming (see **mipmatch**; Zubizarreta, 2012 and **designmatch**; Zubizarreta & Kilcioglu, 2016); modified support vector machinery (see **SVMMatch**; Ratkovic, 2015); coarsened exact matching (see **cem**; Iacus, King, & Porro, 2015); covariate balancing propensity scores (see **CBPS**; Fong, Ratkovic, Hazlett, & Imai, 2015); or genetic matching (see **Matching**; Sekhon, 2011).

References

Abadie, A., & Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*, 235–267.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–972.

Diamond, A., & Sekhon, J. S. (2010, November 14). *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies.* Retrieved from `http://sekhon.berkeley.edu/papers/GenMatch.pdf` (Unpublished manuscript)

Fong, C., Ratkovic, M., Hazlett, C., & Imai, K. (2015). Cbps: Covariate balancing propensity score [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=CBPS` (R package version 0.10)

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*, 609–627.

Helmreich, J. E., & Pruzek, R. M. (2009). PSAgraphics: An r package to support propensity score analysis. *Journal of Statistical Software*, *29*.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, statistics in medicine. *Statistics in Medicine*, *27*, 2055–2061.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

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. Retrieved from `http://www.jstatsoft.org/v42/i08/`

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case

study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*, 901–910.

Iacus, S. M., King, G., & Porro, G. (2015). cem: Coarsened exact matching [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=cem` (R package version 1.1.17)

Keele, L. J. (2014). rbounds: Perform rosenbaum bounds sensitivity tests for matched and unmatched data. [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=rbounds` (R package version 2.1)

Keller, B., Kim, J.-S., & Steiner, P. M. (2015). Neural networks for propensity score estimation: Simulation results and recommendations. In L. A. van der Ark, D. M. Bolt, S.-M. Chow, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology research.* New York, NY: Springer.

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*, 131–159.

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

Lee, B. K., Lessler, J., & Stuart, E. A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346.

Li, L., Kleinman, K., & Gillman, M. W. (2014). A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity. *Journal of Developmental Origins of Health and Disease*, *5*, 435–447.

Longford, N. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics*, *40*, 227–253.

Lumley, T. (2014). survey: analysis of complex survey samples [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=survey` (R package version 3.30)

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

Medicine, *23*, 2937–2960.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multinomial treatments using generalized boosted models. *Statistics in Medicine*, *32*, 3388–3414.

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.

Mebane, Jr, W. R., & Sekhon, J. S. (2011). Genetic optimization using derivatives: The `rgenoud` package for r. *Journal of Statistical Software*, *42*, 1–26.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*, 236–254.

Owen, M., Imai, K., King, G., & Lau, O. (2013). Zelig: Everyone's statistical software [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=Zelig` (R package version 4.2-1)

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ratkovic, M. (2015). Svmmatch: Causal effect estimation and diagnostics with support vector machines [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=SVMMatch` (R package version 1.1)

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2015, January 30). *Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the* `twang` *package.* Unpublished manuscript. Retrieved from `http://cran.r-project.org/web/packages/twang/vignettes/twang.pdf`

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169–188.

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*, 20–36.

Schafer, J., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, *42*, 1–52.

Sekhon, J. S. (2013). Matching: Multivariate and propensity score matching with balance optimization [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=Matching` (R package version 4.8-3.4)

Steiner, P. M., & Cook, T. D. (2013). Matching and propensity scores. In T. Little (Ed.), *Oxford handbook of quantitative methods.* Oxford: Oxford University Press.

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

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=rpart` (R package version 4.1-9)

Tipton, E. (2014). How generalizable is your experiment? an index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*, 478–501.

van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL: Chapman & Hall/CRC.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, *2012*, 1360–1371.

Zubizarreta, J. R., & Kilcioglu, C. (2016). designmatch: Construction of optimally matched samples for randomized experiments and observational studies that are

balanced by design [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=designmatch (R package version 0.1.1)