Applications and graphics for propensity score analysis*

Robert M. Pruzek

State University of New York at Albany

J.E. Helmreich

Marist College

July 30, 2004

*Requests for reprints and information concerning software should be addressed to R.M. Pruzek, Department of Educational and Counseling Psychology, 1400 Washington Aven., State University of New York at Albany, Albany, N.Y. 12222.

Applications and graphics for propensity score analysis

Abstract

Methods for propensity score analysis (PSA) originated with Rosenbaum and Rubin (1983), as vehicles to sharpen and clarify treatment group comparisons in observational studies. Although highly recommended by many statisticians, and applied often in medical sciences, PSA has seen relatively few applications in the social and behavioral sciences. This paper aims to facilitate sound PSA applications in psychological and other social sciences, and to emphasize the role visualization can play in such contexts. Numerous references to the expanding PSA literature are also provided.

Applications and graphics for propensity score analysis

Introduction

It has become a broadly shared opinion among those who have studied propensity score analysis (PSA) that it can sharpen and clarify analyses of treatment effects in many applied sciences, particularly in contexts where randomization is not feasible. The key reference for PSA is Rosenbaum and Rubin (1983), but by now several hundred articles, including tutorials (D'Agostino, 1998), and a major book (Rosenbaum, 2002) have focused on this topic. Despite its promise, however, PSA has rarely been applied in most psychological, social and educational sciences; moreover, students in these fields seldom learn even the rudiments of this new class of methods. The relative lack of PSA applications means that many opportunities are missed since social, behavioral and educational program evaluations routinely use observational data to compare treatments. For example, clinical therapies and behavioral regimens are often compared outside of randomized settings. Also, intact treatment groups are frequently compared, such as schools or classrooms using various educational programs, curricula and instructional methods.

Our main aim in this paper is to use real data to introduce and demonstrate the central ideas of PSA, with a focus on graphics to aid understanding and interpretation of the methods. Some of the graphical methods that we illustrate are new; others are not novel, but nonetheless are rarely used. We discuss how and why PSA can in certain situations facilitate getting clear answers to key questions that often drive treatment comparisons, as well as how effectively to pursue questions that arise in analyses. We also compare and contrast two methods for estimating propensity scores, showing each

may lead to different insights into the data at hand. Before describing PSA methods, however, it may be useful to consider some elements that distinguish experiments from observational studies.

In the first section below, we introduce and discuss the two main phases of PSA. The second section discusses briefly the role PSA has in establishing causal relations, and compares it to some other standard methods for addressing covariate imbalances. In the third and main section of the paper we present two analyses of the same data set, using two different methods for estimating propensity scores, *viz*., logistic regression and classification trees. Various graphical techniques are introduced and discussed, especially as they yield insights into the data and contrast the two estimation techniques. The fourth section discusses implementation strategies and available software, especially for the S+ and R statistical packages. We conclude with a expanded discussion of issues raised in the previous sections. Throughout we give some indication of the available literature on the subject. Before describing PSA methods, however, it may be useful to consider some elements that distinguish experiments from observational studies.

True experiments, studies in which individuals are randomly assigned to treatments, have played a vital role in applied science; they are commonly regarded as the best methodology available for answering unambiguously whether treatments have causal effects on outcomes. Still, critics often denigrate experimental studies, particularly when treatments seem somewhat artificial, or do not conform to behavioral conventions. Indeed, experimental treatment comparisons often differ from their observational counterparts, based on matters such as whether participants choose to participate in treatments when assigned, how fully they carry out treatment responsibilities, or simply

because the experimental implementation of treatments differs in notable ways from its observational counterpart. Often it is observational situations that scientists want most to understand, as they provide a better basis for generalizations to real-world situations. Further, by choosing whether and how to participate in various treatments subjects can have a strong role in how such treatments are even defined.

In practice surely the most important point is that randomization is ethically or administratively unfeasible for comparing many treatments since it is difficult or unreasonable in many social and behavioral research settings for an investigator either to seek or gain permissions required to assign individuals to treatment groups. Conniffe, Gash and O'Connell (2000) note, "Application of the direct experimental approach in the economy and society is usually considered unpalatable, or even unethical, even when it would clearly provide the ideal comparison" (p. 283). In general, those who argue in favor of experiments based on random assignments weight internal validity more heavily than external validity; but even some analysts who most strongly endorse randomization concede that observational studies that do not use random assignment are likely to be essential (cf. Cook and Payne (2002)).

The Central Ideas of PSA

PSA generally entails two phases. First, covariates are selected and used to distinguish between two groups; one of these groups is usually called a 'treatment', the other a 'control.' Ideally, covariates at Phase I should distinguish between the two groups and have some relationship with the ultimate response variable(s). Based on modeling of the probability of being in the treatment group, Phase I analysis yields estimated

propensity scores, where (the unobservable) propensity score for an entity or individual is defined as the probability of being in the treatment group, conditioned on the covariate values for that individual. Typically, estimation of propensity scores has been based on logistic regression. However, methods such as classification trees and discriminant analysis may also work effectively.

In Phase II of PSA, individuals are sorted on the basis of their propensity scores into a relatively small number of strata. Within each derived stratum, treatment and control groups are compared using one or more outcome measures; usually a difference in mean response is generated for each stratum, however, medians, trimmed means, and various other summary measures may also be compared. A summary measure of treatment effects, a Direct Adjustment Estimator (DAE), can be computed as the average of treatment effects across strata. Done effectively, the first phase of PSA yields covariate distributions that are similar across treatment and control groups within each stratum; this covariate balance is central to effective studies that use propensity scores (Rosenbaum and Rubin, 1984).

Everything depends on whether the available or observed covariates account for fundamental differences between treatment and control groups – differences identified as selection bias. A PSA based on observational data can in principle serve nearly as well as a randomized experiment to infer causal effects. But practical reality is that observed covariates are likely to be at least somewhat inadequate for the task they are being asked to do. Consequently, observational studies, even if analyzed carefully with propensity score methods, generally provide at least somewhat weaker evidence than would a corresponding experimental study – if the true experiment could be conducted. To the

extent that most of the selection bias can be accounted for, PSA results may support causal inferences about treatment effects even when data are observational. This is a key argument for trying to find effective approaches to propensity score analyses, to make observational study results stronger, more like those of experimental studies. The use of PSA for the analysis of observational data puts a premium on collection of relatively comprehensive covariate scores for all respondents based on how effectively to account for differences between the treatment and control groups; this issue had rarely been understood, much less raised to prominence, before propensity score methodology came on the scene.

Generalizing a result of Cochran (1968), Rosenbaum and Rubin (1983) advise that five strata are usually sufficient to remove 90% of the selection bias. When logistic regression is used to estimate propensity scores, strata are usually constructed to be of equal size. When strata have differing sizes (as is often the case with propensity scores estimated using classification trees) then mean differences for strata are weighted according to relative sizes of strata to obtain the DAE. It often happens, however, that effects for treatment comparisons differ across strata. In this case a summary measure such as the DAE may be of less interest than effects for individual strata. A main advantage of PSA is that the analyst may be able to investigate how values of particular confounders play out for particular strata, and various possible implications for causal mechanisms vis-à-vis outcomes. This point will be elaborated below in the context of our example, but in general this issue will require attention of an analyst with subject matter and data-specific knowledge. Of course there are no assumption-free methods for inferring causation.

As noted, a key goal of PSA is to achieve covariate balance within strata. That is, once Phase I has been completed and propensity scores have been used to sort entities into subgroups that are relatively homogeneous in these scores, covariate distributions should differ only minimally between treatment and control groups within strata. To the extent that covariate distributions are similar on all relevant covariates, the treatment and control group comparison will be little affected by selection bias. The central idea for PSA is to use covariates first to distinguish the treatment and control groups from one another, and then to assess outcome effects within strata that have been defined using propensity scores. As will be seen below, graphical methods can help show the extent to which this and related objectives have been achieved.

The Role of PSA in Supporting Causal Inferences

In the best of situations PSA can be seen as providing support for causal inferences in observational studies, support that can be nearly as strong as that of the corresponding true experiment, were the latter to be ethical or realistic. But the term 'support' should not be construed to imply 'direct attribution.' Even true experiments entail assumptions that when violated can obviate causal inferences. For example, randomization may not have worked as advertised to balance all relevant covariates. Stratified random assignment can be an improvement, but one may not have used the most appropriate covariates for stratification. Further, the possibilities of interaction can never be ignored with impunity. Treatments can have observed effects that misinform when notable interactions between treatments and (uncontrolled) independent variables are hidden and therefore mask real effects. It is also possible that treatment or control

groups may themselves be poorly defined, or poorly executed, in which case results of experiments can mislead or deceive the experimenter. In fact, empirical results generally do not warrant direct, assumption-free attributions of causation in virtually any real-life scientific study.

Notwithstanding the difficulties that can arise even in true experiments, it is not unreasonable to argue that observational studies can provide support for causal inferences. It is the thesis of many scholars that this support will tend to be strongest in situations when propensity scores are used to adjust for selection bias in the comparison of treatments, and when certain other conditions are met (as discussed at several places on the pages that follow). In such cases the homogeneously grouped entities in propensity-defined strata are formed in a way that makes these subgroups comparable with respect to relevant covariates. When all strata under comparison are very nearly comparable with respect to what appear to be all of the most relevant covariates then the degree of support for a causal inference in an observational study can closely approach that of the experiment that might be envisioned to stand in place of the observational study (Rubin, 1997).

Although PSA is not the only class of methods that has been developed with the goal of removing selection bias, it seems to have withstood its challenges better than its competitors. Authors such as Winship and Morgan (1999) discuss these issues in considerable detail, with special reference to counterfactual logic that underpins PSA and certain other methods. But it appears that non-PSA methods with similar purposes [*esp*. Instrumental variable methods] are often found to be "frustratingly sensitive to the validity of … underlying assumptions (Obenchain and Melfi, 1997)." In

contrast, PSA results tend to be robust to methodological variations (cf. Conniffe, Gash, & O'Connell (2000)).

Conventional statistical methods used to adjust for covariate effects, notably the analysis of covariance (ANCOVA), are *not* directed at reducing selection bias but at adjusting for covariate imbalance when randomization has been used. It is generally accepted that the key goal of ANCOVA is to improve the efficiency of statistical inferences when there are treatment group differences on one or more covariates. While ANCOVA has often been used in analyses of observational data, it has been subject to serious criticisms in this context, particularly because of its strong assumptions. Most importantly, use of ANCOVA generally entails the assumption that there are no interactions between treatments and any of the covariates, a criticism that need not apply when PSA is used in observational contexts. Furthermore, PSA methods should be seen as not only allowing, but in fact encouraging use of as many covariates as seem to be needed to make adjustments. In contrast, ANCOVA methods become more and more likely to fail as covariates are added to a model.

PSA, as a vehicle for studying causal claims, can also be contrasted with structural equation models (SEMs). Many authors have contributed to the literature that speaks to the issue of inferring causality from association in SEMs and this literature is both broad and deep. Some authors, cf. Spirtes, Glymour and Scheines (1993), argue aggressively that causal inferences are feasible with SEMs, and they provide algorithms intended to assist toward this end; others, cf. Freedman (1997, 1999), dispute such arguments and hold that SEMs provide little basis for causal inferences. Authors such as Stone (1993) and Clogg and Haritou (1997) have thoughtfully traced the arguments as to

what assumptions are required if causal claims are to be supported in regression analyses, but the topic of causation is far too large for even a cursory review here. Pearl (1998, 2000) is well worth studying by those who aim to understand causality in the context of SEMs, but it is notable that SEMs generally do not aim to adjust for selection bias.

A PSA Illustration Based on Real Data

We present two analyses of a data set from the early 1960's concerning the relationship between maternal smoking during pregnancy and infant birthweight. [Data source: http://stat-www.berkeley.edu/users/statlabs/labs.html] The data were excerpted from a larger database of all births to women enrolled in the Kaiser health plan in California from 1960 to 1967. The original Berkeley Statlabs data set (cf. Nolan and Speed, 1999) contained 1237 cases, but some data cleaning, followed by elimination of infants who were not born full term, gestations less than 37 or more than 43 weeks, reduced the sample to 954 cases, all with complete records.[1]

We defined smoking mothers as the treatment group, with the control group identified as those mothers who did not smoke during pregnancy. The outcome measure of interest is the weight of the infant at birth. Confounding variables chosen for this analysis included the length of gestation, mother's height, weight, age, race, education level and marital status, as well as the number of her previous pregnancies; also, the age, race and education level of the father. Our aim is to provide a demonstration of various graphical techniques and discuss their consequences in the context of two different

[1] Recently, Rubin and colleagues have reexamined PSA in the context of applications where some data are missing data, noting that missingness should also be balanced when estimating propensity scores. Interestingly, we note that classification trees handle missingness more easily than do parametric methods such as logistic regression, something not pursued by Rubin.

approaches to the estimation of propensity scores: logistic regression and classification (regression) trees. We note that the mean birthweight for infants born to mothers who smoked during pregnancy was 3255 grams and for nonsmokers, 3509, leading to an (unadjusted or raw) difference of 254 grams. Consequently, the unadjusted estimate of the affect of smoking is to lower birthweights by about 9 ounces.

Logistic Regression Results: At the outset, several preliminary logistic models were estimated. Statistically unimportant predictors were eliminated and this led to a main effects logistic regression model incorporating the factors gestation, height, weight, education and race (Mexican, African American, Asian, mixed or White). Table 1 summarizes the results based on the logistic model.

Using this model, subjects were classified into five strata according to quintiles of the estimated propensity score distribution. The following table lists strata, numbers of non-smokers and smokers, as well as mean birthweights in grams for each group within strata; the final column contains the difference in mean birthweights for strata.

Because each stratum consists of approximately 20% of the sample, sizes of strata are 190 or 191, from the total of 954. Given the equally sized strata, the average of differences across strata equals 219.5 grams, the Direct Adjustment Estimator (DAE). These results show that after PSA adjustment for selection bias, using gestation length, height, weight, education and race of mother, smokers gave birth to infants who were about half a pound lighter than those born to nonsmoking mothers, a slightly smaller value than the unadjusted estimate noted above. For those not familiar with studies of neonatal effects, it may be important to note that birthweight is generally regarded as the

single best indicator of infant morbidity so that a result like this is useful evidence about the effects of maternal smoking.

A more nuanced analysis is shown in Figure 1 where birthweights are plotted against propensity scores; Obenchain (2002) seems to have been the first to have suggested this kind of plot, but we have never seen it published before. Open circles correspond to Nonsmokers; filled circles correspond to Smokers. Separate loess regression lines (cf. Cleveland, Grosse and Shyu (1992)) are shown for the groups of Smokers and Nonsmokers. These are non-parametric regression curves for the treatment and control groups. Although not computed here, the weighted average difference between the two loess lines might be generated as a summary estimate of the effect of smoking (DAE), having adjusted for selection bias with this logistic regression model.

Note that for most of the left side of this propensity score distribution the distance between the Nonsmoker and Smoker regression lines is about the same. However, at the upper levels of propensity scores, just past the middle of the PS distribution, estimated birthweights tend to decrease for both groups, where the downward trend is larger for Smokers than Nonsmokers. This leads to interesting questions about the covariate levels that, while indicative of a higher propensity to smoke, are also apparently indicative of a tendency for smaller birthweights irrespective of smoking status. The plot also shows cutoffs using vertical lines at the quintile points of the PS distribution; these lines distinguish strata for which counts and means are given in Table 2. It is evident that there are several Smokers and Nonsmokers in each stratum, but as can be most readily seen in Table 2, there are fewer Smokers than Nonsmokers in the left-most stratum and more

Smokers than Nonsmokers in the right-most stratum. This kind of imbalance in counts in strata is characteristic of propensity score strata constructed using covariates in PSA.

That the loess lines lie strictly above and below each other and do not cross is an indication of concordance of treatment effect across propensity score levels and thus covariate levels. In some applications the loess lines may cross to indicate values of the covariates, and subsets of units, where treatment effects are manifestly more complicated than seen here; naturally, details of context would be needed to know how effectively to follow up on such results.

In Figure 2 boxplots are used to compare covariate distributions within propensity score strata; however, only one figure is given here, that for gestation. These five strata again correspond to those discussed above for the logistic regression analysis. The boxplots are presented in pairs for the respective strata, where the boxes are colored differently for each stratum, for Nonsmokers on the left, Smokers on the right. The endpoints of each box correspond to the first and third quartile points for the corresponding distributions; the medians correspond to the horizontal line segments within each box. The so-called box whiskers identify the ends of each distribution, unless there are notable outliers, as seen in the second, fourth and fifth strata. Because knowledge of means generally adds to information in the form of medians, the short line segments that connect each of the comparable boxplots, for each stratum, are based on means of the corresponding distributions. Detailed examinations of such graphics may provide insights about treatment differences as related to covariate distributions.

Figure 3 shows a counterpart of Figure 2 where the covariate, education, is an ordered categorical variable. Each pair of barplots corresponds to Nonsmokers (left) and

Smokers (right) for a particular stratum. Again, the goal of PSA stratification is to achieve balance within strata for the covariates, which in the case of a categorical variable like education means to learn whether proportions in the respective education categories are similar within strata.

The boxplots and barplots of Figure 2 and 3 provide detailed visual indications of covariate balance within strata. If desired, these graphics can be used in conjunction with standard numerical tests for independence of the categorical variable and treatment variable or difference of means or medians in the continuous case.

In this example, before stratification on the propensity score, differences between Nonsmokers and Smokers were significant, and notable, for several covariates, including gestation, mother's and father's ages; also, the categorical variables for mother's and father's education and mother's race were significantly and strongly related to smoking. After stratification on the estimated propensity score, two sample t-tests for comparison of means by covariate and strata were conducted for the continuous variables; tests of independence within strata of categorical covariates and smoking were also conducted. With minor exceptions, Nonsmokers and Smoker distributions were approximately balanced for all strata, as there were no significant differences in means or deviations from independence. These results were the same whether the propensity scores were derived from the logistic model or from the classification tree (see below).

Note that covariate distributions may reasonably be compared not just for covariates used in the logistic regression, but also for covariates that were not part of in the logistic model to discriminate between treatment and control groups. Recall that the goal in PSA is to form strata for which all generally relevant covariates will be balanced

in each stratum, not just the covariates that happen to have been used in Phase I of the PSA. Inspection of the range of all such distributions, comparable to those shown here as Figures 2 and 3, shows that balance seems largely to have been achieved in this case.

Another graphical aid to interpretation of results is apparently novel. We name it the PSA Assessment Plot, seen in Figure 4. This is an enhanced scatterplot based on circles rather than points, each sized so their areas will appear proportional to sizes of strata; the coordinates for each stratum are summary measures for control and treatment groups respectively, in this case mean birthweights of Nonsmokers on the horizontal axis and Smokers on the vertical axis. In the example based on logistic regression, where cutoff points for the strata were based on PS quintiles, all circles have the same size. Note that if the means were the same in a stratum for Smokers and Nonsmokers then the corresponding circle lies on the identity diagonal, the solid line with intercept zero and a slope of unity. That all circles in Figure 4 lie on the same side of the identity line indicates a concordance in direction of effects across strata for these data, echoing that indication from the loess lines shown earlier. The heavy dashed line parallel to the identity diagonal corresponds to the weighted mean of effects across the five strata; that is, this dashed line shows the DAE. The distribution of effects across strata is shown by the crosses in the lower section of the plot, where effects, i.e. differences in means within strata, have been projected downward to a line segment perpendicular to the identity diagonal. In addition, the thin horizontal and vertical dashed lines locate the weighted means for birthweights by strata for Smokers (horizontal) and Nonsmokers (vertical). Each stratum is identified by its number inside the corresponding circle, so graphical results seen in Figure 4 can be compared to numerical counterparts in Table 2.

Figure 4 makes it apparent that results for strata 2 and 3 are similar; but while mean Nonsmoker birthweights for strata 1, 4, and 5 are similar, the same measure for Smokers in these strata varies notably. Though stratum 1, the subset with lowest propensity scores, is closest to the identity diagonal, it still shows the same direction of effect as the other strata. Stratum 5 yielded the strongest effect among the strata, a finding that recalls the loess plot above where the largest difference between Smokers and Nonsmokers was found for the highest propensity scores. It appears that the propensity score assessment plot yields insights that extend beyond those of the basic table depicted above; similarly the loess regression plot provides far more information than given in the count and mean summary table.

Classification Tree Results: Next we compare the analysis based on logistically estimated propensity scores with an analysis based on a classification tree. As we discuss below, this different estimation strategy here finds a similar DAE, yet yields potential insights different from the analysis above using logistically derived propensity scores. In fact, the two different derived propensity score estimates differ notably from one another, although each achieves reasonable covariate balance in our example.

A tree was used to partition the data set recursively, based on the same covariates available initially in the preceding logistic regression analysis. Figure 5 shows that the covariate that best split the Smoker and Nonsmoker groups into two subgroups was education, an ordered categorical variable, using the cutpoint 3.5. That is, the variable education was selected by the tree algorithm at the first level, where all mothers whose education level above the third went to the left-hand branch of the tree, the rest to the

right (see the website given above for details about variable codes). Two groups of size 429 and 525 respectively were formed initially; in the higher education group 32% were smokers, whereas in the lower education group 44% were smokers, a difference of 12%. The classification tree algorithm is such that any other partition, using either a different covariate or a different cutpoint for education, would have led to a split whose difference in proportions would be smaller than 12%. Recursive partitioning was continued, where again the goal was to use a covariate (Education being eligible again) to partition each of the education-based subgroups so that the maximum difference between proportions of smokers would be found. In this case the higher education group was split on gestation at 282.5 days. The lower education group cut at 32.5 years, based on father's age; 34% of the fathers under this age were found to be smokers, and since this subgroup was not split further it is called a terminal node or leaf. For older fathers, the tree was extended on the basis of mother's race and gestation, as shown in the tree graphic.

The final tree in Figure 5 used the covariates age, education, and race of mother, as well as age of the father and gestation to partition the full data set. One of the inherent benefits of such a tree is that the strata are determined in a natural manner by the data themselves; interactions of covariates with respect to the prediction of a two-category criterion are found automatically using a tree. Strata sizes are unconstrained, and the tree algorithm is non-parametric in the sense that the tree is unchanged when quantitative and ordered categorical covariates are reexpressed using (order preserving) transformations (such as logs or powers).

The counts for terminal leaves of the tree ranged from a high of 258 (stratum 4) to a low of 23 (stratum 4). Although judgment may be required in deciding how far to grow

a tree, there is reason, based on prior experience with logistic regression and the results of Rosenbaum and Rubin (1983), to aim for five to seven terminal leaves. Ultimately, one does not want so many leaves that many cells in the strata by treatment table are empty, nor so few that considerable selection bias remains.

Table 3, based on the classification tree, shows counterpart counts and means that had earlier been presented in Table 2 for logistic regression. Propensity scores in the case of a classification tree are derived by the Smoker/Nonsmoker counts within strata. For example, in the first stratum 5 of 23 mothers are smokers, yielding a propensity score of 5/23 = 0.217, where all individuals are assigned the same estimated propensity score in each stratum. Weighting the mean birthweight difference in each stratum by the proportion of subjects in that stratum yields a DAE of 206.4 grams. Thus both estimates of the direct effect indicate that the unadjusted difference overemphasized the negative effect of maternal smoking on birthweight.

In Figures 6 and 7 we present boxplot and barplot comparisons for these classification tree-derived strata to aid comparison with the earlier logistically-derived strata. Recall that gestation was significantly different for the two groups before stratification. Thus Figure 6 shows considerable differences across strata in the (conditional) distributions of gestation, but again, balance within strata has been fairly well achieved. Figure 7 generally shows approximate balance for mother's education levels across strata, but the first stratum indicates exceptionality. The latter result is probably not of great importance in this case, however, since the size of this stratum is small (23 cases). The question of whether 'sufficient balance' has been achieved remains open, as it nearly always will in applications. But note that even in randomized studies,

particularly when samples are not large, at least some covariate distributions are likely to become unbalanced by chance.

The tree-based PSA assessment plot is shown in Figure 8, where numerical summaries in Table 3 can be used in conjunction with this Figure. Again, the assessment plot shows a concordance in directions of effects across strata. Compared to the preceding logistic regression results, however, the overall spread of the strata is somewhat diminished, with the majority of birthweight means lining up with the mean effect across strata. Stratum 2 shows the smallest treatment effect, while stratum 1, which is the smallest, shows the largest effect of smoking. Note that sizes of circles differ for strata derived from the classification tree, in contrast with the logistic regression illustration above. In general there may be merit in further examination of the distinguishing characteristics of units in a given stratum with respect to the size of the treatment effect for that stratum.

Implementation

In our use of logistic regression in Phase I above, we have chosen not to use the fullest model, more for simplicity's sake than from need. Since the model is not to be used for effect estimation, over-fitting is generally not a serious issue. It is possible that the Phase I estimates of propensity scores are such that there is little overlap in propensity score distributions for the treatment and control groups. Larger samples can help ameliorate this problem, but it must be remembered that many observational data sets simply cannot provide an empirical basis for answering questions about treatment effects. A more parsimonious model in Phase I may sometimes be helpful, provided reasonable

covariate balance is maintained. A similar argument can be used regarding the question of how far to grow the classification tree; there need not be much disadvantage to having a more elaborate tree except that as more nodes are added it is more likely that counts for some leaves will be too small and too imbalanced to permit comparison of the treatment and control groups. The key issues have to do with matching, about which Rubin and Thomas (1996) provide especially useful information.

Statistical methods for both phases of PSA are in general within the main stream of statistical software so that conduct of a PSA may require little more than what can be found in any comprehensive package. Authors such as Conniffe, Gash and O'Connell (2000) provide relevant discussion of the numerical aspects, including, for example, equations showing how to estimate the standard error of a DAE.

Graphical methods are another matter, however, at least those of the kind illustrated in this article. We are aware of several attempts to develop graphics for PSA using a variety of software packages, but it would appear that the most successful ones were based on the S language, in particular the S+ and R packages. We have used R here because it is free (website: r-project.org) and has reached a relatively refined stage of development that lends itself to graphical displays of data. Not only are our PSA functions available without cost (website to be provided) but there are others as well, such as that of Obenchain (2004).

Discussion

An example has been used to illustrate propensity score analysis, and especially to highlight graphical methods as aids to interpretations. Several larger points should be

made about such propensity studies, and our example affords an opportunity to point up some relevant issues. First, a caveat: the data used for our illustration were archival from the 1960s, something to be taken into account in order to avoid unwarranted generalizations. For example, nearly 39% of this sample of pregnant women reported that they smoked to some extent during their pregnancies, a significantly larger percentage than currently seen in the U.S.

A deeper question in this and many observational studies has to do with what it means to be in the so-called 'treatment' and 'control' groups. In this case ancillary data were available showing how many cigarettes were smoked prior to, or during pregnancy. We made some rather arbitrary choices in our final decisions about whom we would call Smokers and Nonsmokers; the only reason for not providing details is the methodological emphasis of this article. Another critical issue is whether self-reported data (as smoking levels were here) are sufficient to adequately define treatments. Clearly behavioral records would be more satisfactory, but they were not available in this case, and would usually not be in our experience. Similar questions about the measurement properties of the measured covariates need also to be considered. Careful design of observational studies is at least as important as careful design in the context of true experiments (cf. Rosenbaum, 2001, 2002).

It should be recognized that many observational data sets will not provide a sound basis for propensity score analysis. This is particularly true when treatments or control groups cannot be clearly defined, when most needed covariates are not available or not soundly measured, or when the differences between the groups being compared are simply too great to have a basis for comparison. These points should not be seen as

criticisms of PSA itself, however, because lacking the desired data, no statistical methods may provide a sound basis for comparing groups, especially to support causal inferences. Some methods, especially those based on regression, may appear to 'work' when PSA may not. But this may be illusory; PSA is fundamentally self-critical while most other methods are not. See Rubin (1997) for a relevant discussion. PSA leads the analyst in directions that conventional methods do not because the key problem of observational group comparisons, selection bias, is addressed directly and because PSA need not entail any assumptions about particular functional relationships among variables.

Analysis of observational data using propensity scores can clarify, and therefore improve interpretations of group comparisons by removing selection bias. To the extent that covariates effectively account for important or relevant pre-treatment differences between groups (differences in shoe size would probably not be important), comparing groups whose propensity scores are similar will often sharpen the focus of comparisons, and may help support causal interpretations about treatment effects. Graphical methods were emphasized above as a means to assess key questions, especially that of covariate balance. The foregoing loess plot and corresponding covariate balance plots showed that detailed examinations of group differences can help to show where and to what extent particular covariate differences exist for specific strata. Recall that for one stratum, the fifth, the fitted lowess lines decreased for both Smoking and Nonsmoking groups, and the covariate balance plots showed how gestation differed when comparing this stratum from the others. This example provided strong hints about how particular covariate differences could further affect birthweights, irrespective of smoking status. Alternatively, these plots

can be seen as indicative of interactions between gestation and effects of smoking on birthweights.

Because one never observes *all* potentially relevant covariates the question is always open, at least to some extent, as to whether treatment effects in PSA reflect causality, or simply a failure to include key covariates. This is where the analyst's subject-matter knowledge comes in, as well as sensitivity analysis when feasible. Sensitivity analysis attempts to quantify the hidden bias, or strength of an unknown and unobserved confounding variable that would be needed to explain the observed DAE; see Rosenbaum (1991, 2002).

Both the loess regression plot in the case of logistic regression applications, and the PSA assessment plot, which applies for both logistic regression and classification trees, provide useful visual images of the size and direction of treatment effects. In the case of the assessment plot, it is clear that if all circles, one for each stratum, lie on the same side of the identity diagonal, this lends strength to the conclusion about treatment effects. When circles-as-strata in an assessment plot do not line up parallel to the identity diagonal this is prima facie evidence of interactions between covariates and treatments, since treatment differences are thus seen to depend on particular combinations of covariate score values or categories. The graphics of PSA may be incisive for the analyst in learning what investigative directions may be worth following. As always, of course, care is needed in order not to over-interpret data.

Conclusion

In the end a great deal depends on details such as how the data were collected, how the ostensible treatments were defined, whether at least most of the most desirable covariates were measured, and if so with not too much missingness; what outcome measures were used, and how effectively to analyze outcomes as well. Some of these and related issues can be usefully illuminated through careful use and interpretation of graphical displays. Technical details may matter in the conduct of PSA of course, but it has been satisfying to learn that PSA results have often proven robust to many methodological variations. Nevertheless, in view of the paucity of PSA studies in the psychological and social sciences to date, the situation in these fields is currently not clear as to what issues will turn out to be most central in future applications. Many additional references of relevance are available: a tutorial, D'Agostino (1998); details about what 'choice' means in an observational study, Rosenbaum (1999); an application, Perkins, et al (2000); and discussions of broad issues and technical details pertinent to many aspects of PSA, Rosenbaum (2002).

References

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992) Local regression models. Chapter 8 in Chambers, J.M. and T.J., Editors (1992) *Statistical models in S*. New York: Chapman & Hall.

Clogg, C.C. nd Haritou, A. (1997) The regression method of causal inference and a dilemma confronting this method. In V.R. McKim and S.P. Turner (Editors), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences,* pps 83-112, South Bend, IN: University of Notre Dame Press.

Cochran, W. G. (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 205-213.

Conniffe, D., Gash, V., & O'Connell, P. J. (2000) Evaluating state programmes: "Natural experiments" and propensity scores. *The Economic and Social Review, 31*, *4*, 283-308.

Cook, T. & Payne, M. (2002) Objecting to the objections to using random assignment in educational Research, Chapter 6 in Mosteller, F. & Boruch, R. (Editors) (2002) Evidence matters: randomized trials in education research. Washington, D.C.: Brookings Institution Press.

D'Agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265-2281.

Freedman, D.A. (1997) From association to causation via regression. In V.R. McKim and S.P. Turner (Editors), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences,* pps 113-161, South Bend, IN: University of Notre Dame Press.

Freedman, D.A. (1999). From association to causation: some remarks on the history of statistics. *Statistical Science*, *14,* 243–58.

Nolan, D. and Speed, T.P. (1999) Teaching Statistics Theory Through Applications. *American Statistician, 53, 4*, 370-376.

Obenchain, R.L. (2004) R functions for propensity score analysis. http://www.math.iupui.edu/~indyasa/bobodown.htm.

Obenchain, R.L. & Melfi, C.A. (1997). Propensity score and Heckman adjustments for treatment selection bias in database studies", *American Statistical Association 1997 Conference Proceedings, Biopharmacology Section*, 297-306.

Pearl, J. (1998) Graphs, causality, and structural equation models. *Sociological Methods and Research, 27*, 226-284.

Pearl, J. (2000) *Causality: models reasoning and inference*. Cambridge, UK: Cambridge University Press.

Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X.-H., & Murray, M. D. (2000) The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and drug safety, 9*, 93-101.

Rosenbaum, P. R. (1991) Discussing hidden bias in observational studies. *Annals of Internal Medicine, 155*, *11*, 901-905.

Rosenbaum, P.R. (1999) Choice as an alternative to control in observational studies (with discussion). *Statistical science, 14, 3*, 259-304.

Rosenbaum, P. R. (2001) Replicating effects and biases. *The American Statistician,* 55, 3, 223-227

Rosenbaum, P. R. (2002) *Observational Studies, 2nd ed.,* New York: Springer-Verlag.

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

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

Rubin, D. B. (1997) Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757-763.

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

Spirites, P., Glymour, C. and Scheines, R. (1993) *Causation, Prediction and Search*. New York, NY: Springer-Verlag.

Stone, R. (1993) The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society*, *Series B, 55*, 455–66.

Winship, C. and Morgan, S.L. (1999) The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659-706.

Tables and Figures

Table 1

Main effects logistic regression model

Smoking Predictor	β	SE	t	$p(>\|t\|)$
(Intercept)	3.238	2.575	1.257	.2086
Gestation	−0.023	0.006	**−3.448**	.0005 ***
Height	0.077	0.032	**2.376**	.0175 *
Weight	−0.011	0.004	**−2.747**	.0060 **
Education	−0.222	0.050	**−4.455**	.0000 ***
Race – *White vs:*				
Mexican	−1.054	0.487	**−2.161**	.0307 *
African American	−0.827	0.187	−0.928	.3535
Asian	−0.827	0.452	−1.831	.0671
Mixed	−2.483	1.037	**−2.394**	.0166 *
Unknown	−0.282	0.715	−0.394	.6936

*p<0.05, **p<0.01, ***p<0.001

Table 2

Count and Mean Summaries Based on Logistic Regression

Strata	Counts		Means		
Propensity Score Range	NonSmokers	Smokers	NonSmoker Birthweights	Smoker Birthweights	Difference
(.014,0.284]	150	40	3479	3421	58
(.284,0.356]	122	69	3573	3309	264
(.356,0.415]	124	66	3569	3319	250
(.415,0.485]	108	83	3451	3275	176
(.485,0.946]	80	111	3457	3108	349

Table 3

Propensity scores, counts and means for birthweight using classification tree strata

Propensity Score	Nonsmoker	Smoker	Nonsmoker Birthweight	Smoker Birthweight	Difference
.217	18	5	3548	3175	373
.228	132	39	3622	3540	82
.341	118	61	3580	3278	302
.376	161	97	3382	3170	212
.432	92	70	3596	3426	170
.602	64	97	3326	3090	236

Figure 1

Loess Regression of Birthweight on Propensity Score for Smokers and Nonsmokers

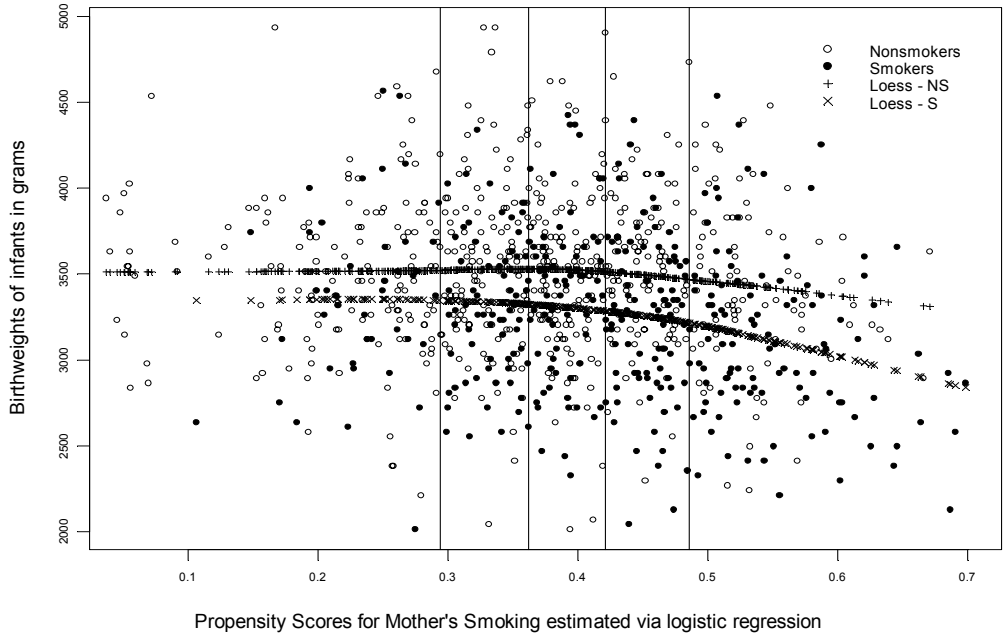

Figure 2

Boxplots to compare covariate distribution within strata: Gestation

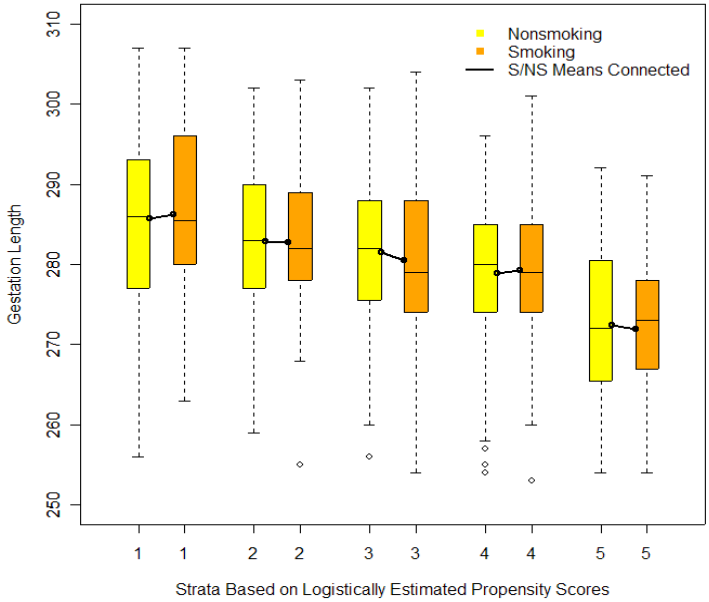

Figure 3

Barplots to compare categorical distributions across strata: Education

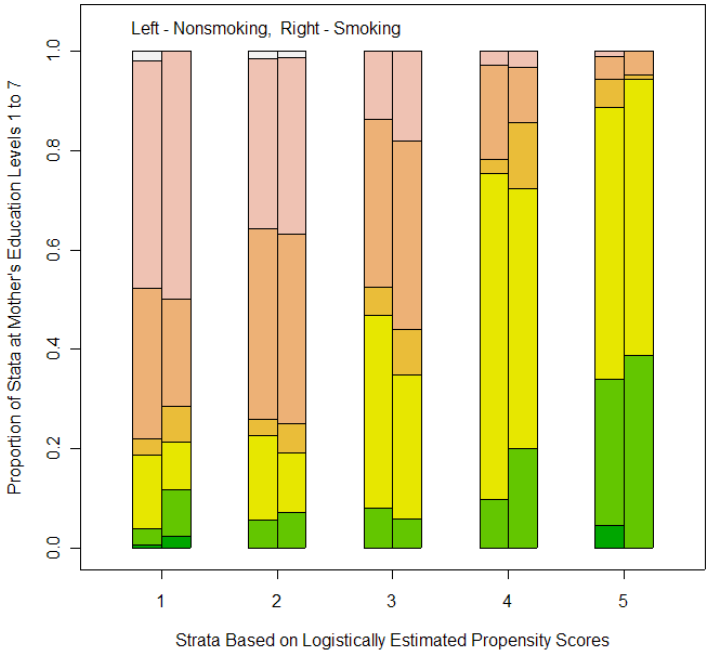

Figure 4

PSA assessment plot, strata from logistic regression

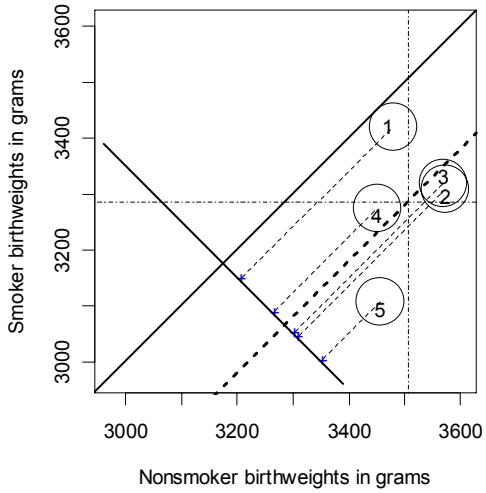

Figure 5

Classification Tree for Birthweight Data, predicting smoking status

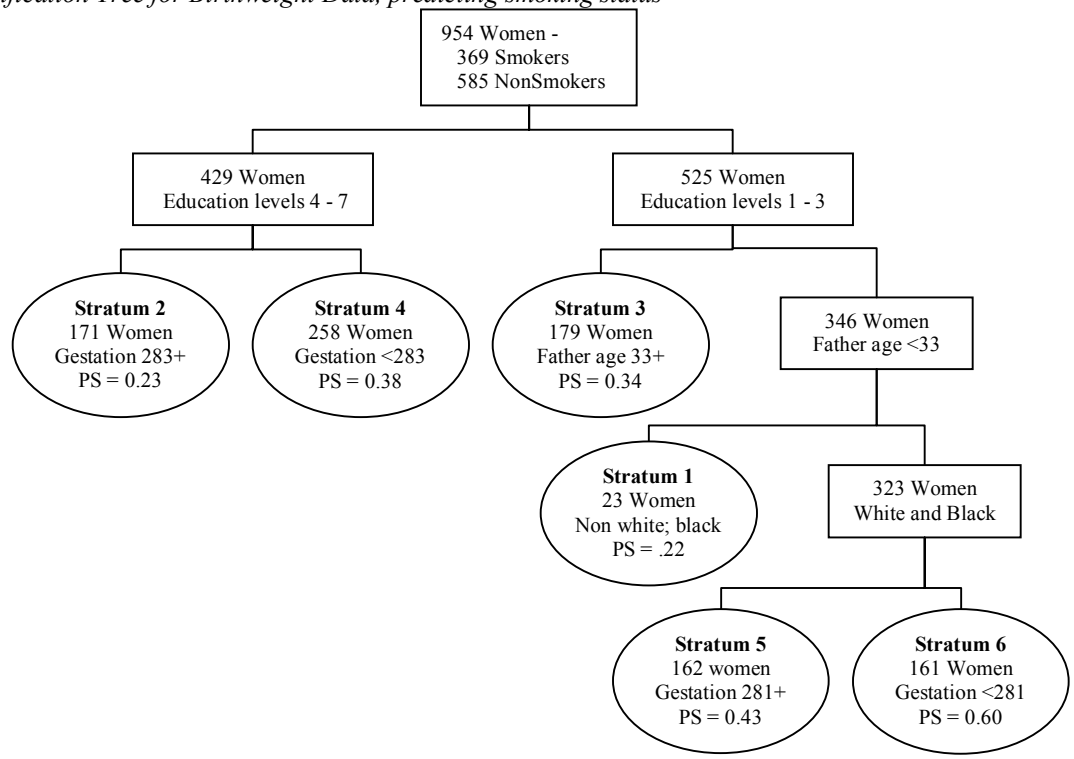

Figure 6

Boxplots to compare covariate distributions within strata: Gestation

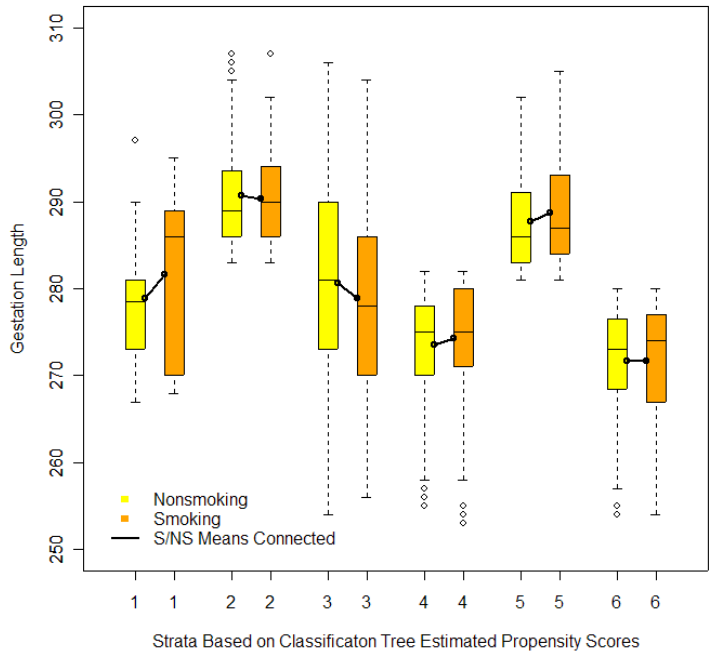

Figure 7

Barplots to compare categorical distributions across strata: Education

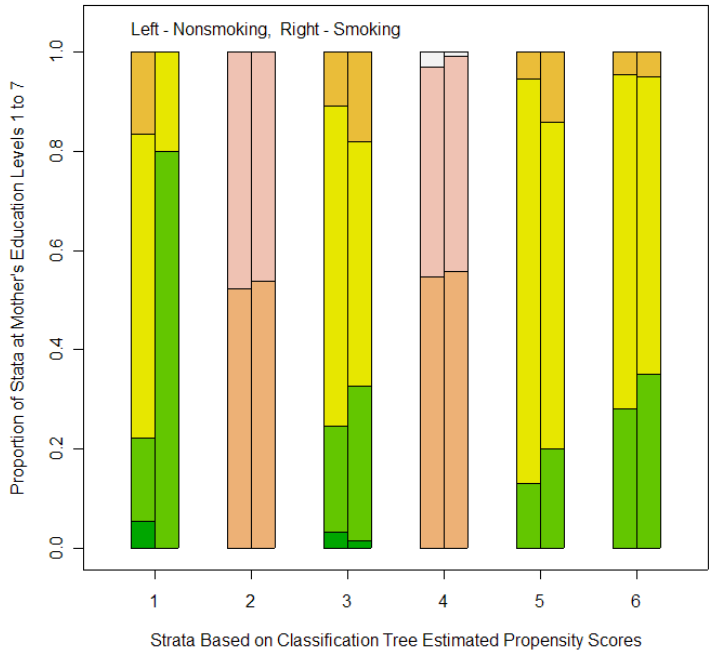

Figure 8

PSA assessment plot, strata from classification tree