An introduction to propensity score analysis

Bob Pruzek (with assistance of J. Helmreich*)

1. Introduction.

1.1 *Comparison of Treatments*

The principal topic of this document is the comparison of treatments, a matter of major concern to applied scientists in virtually all fields. My specific aim is to focus on *comparisons of treatments in the context of observational studies*, studies in which the investigator has not controlled treatment assignments. After some discussion of *true* (designed) experiments (where random assignments to treatments are central) I examine comparisons of treatments for which the entity or individual (can be said to have) *selected* his or her treatment. The emphasis on observational studies recognizes that human sciences in particular require well-researched methods that can support sound conclusions even when true experiments are impractical or unethical.

In general, however difficult the analysis of designed experiments can become, the analysis of treatment comparisons in the context of observational studies is typically even more difficult or more complicated. Although our broad aim is to examine treatment comparisons in observational contexts so as to be relevant to a wide range of applied sciences, my background, experiences and space considerations will of necessity limit illustrations to a relatively small domain of scientific application.

The focus here is on scientific applications where theory and prior research will typically have informed definitions of treatments; but there is no reason why treatments must be confined to science. We trust the reader will actively consider various alternatives to our examples of treatment comparisons, to aid understanding and generalizations about treatment comparison methods. A key aim is to help readers enter the literature of propensity score analysis with a realistic conception of the basic issues, and the corresponding complication, that bear on the conduct of meaningful and interpretable treatment comparisons of many kinds. That literature has grown enormously, both its theory and applications, in the quarter century since Rosenbaum and Rubin (1983) first introduced the main ideas of PSA.

1.2 *Treatments*

Anyone who aims to study methods for treatment comparisons would do well to ask a number of basic questions at the outset. *First, what exactly are treatments?* Or, how might one best think about them? Speaking broadly, almost any action, behavior, substance or condition can serve to define a treatment. When you take aspirin for your headache, massage sore muscles, or buy orthotics to insert in your shoes you are choosing treatments to ameliorate problems. When medical researchers develop vaccines to prevent disease, drugs to treat illness or methods to aid victims of injury they may be said to be designing treatments. When biologists study the effects of clear cutting forests, or educational psychologists evaluate how class size affects learning, they are studying the consequences of treatments. In short, treatments are ubiquitous, they can take virtually any form. But it is treatment comparisons that are often the essence of applied scientific study. What are the consequences of taking one kind of drug versus another if the aim is to reduce headaches? Does selective cutting of timber differ notably from clear cutting in its effects on an endangered species habitat? Does one instructional method facilitate learning more than another? Comparison of various treatment options is at the heart of many scientific endeavors. We seek clear definitions of treatments, but also of responses; but we recognize that in practice, especially in observational studies, an investigator will generally induce what treatments entail based on what has been observed – usually, the behavior of individuals (not necessarily persons). Furthermore, outcomes of interest will often extend to (unintended) side effects, not just (most) valued outcomes.

Ideally, treatments should be (reasonably) well-defined. This is particularly important as an to aid repetition, or replication. Carefully constructed definitions of treatments are especially important when those treatments are complex, or multifaceted. This is often the case with medical and behavioral therapies, or instructional methods. In true experiments it is *not* sufficient that conceptual definitions of treatments are clear; it is also necessary that their *implementations to be sound* and consistent. For example, consider the investigation of a novel instructional method. What specifically is the instructional method? To what (standard) method(s) will it be compared? How are these methods implemented? Will it be group instruction (what size groups?), or will it be individualized (and how long will it last?)? In what context (ability, age or grade level, school, etc.) with the methods be applied? Among other things, this means that treatments share common elements. What is called a treatment for one entity or unit should be common to others. For example, instructional methods should not change (much) across teachers or schools in which they are used – unless the teachers themselves constitute alternative treatments. Treatments need not be simple, nor one-dimensional, although interpretations of results of treatment comparisons may well be difficult when treatments are themselves overly complicated. For observational studies, treatments may well vary more across individuals than for true experiments because the individuals themselves, by their behaviors, their particular choices, define what investigators define treatments to be.

A particular treatment usually takes place over a well-defined period of time so that it has a clearly determined beginning and ending. This stipulation, that treatments generally have a temporal aspect, helps ensure that treatments are well defined, and is most helpful to the investigator when it is time to evaluate treatment effects. It is also helpful in knowing how to define or select important factors ('covariates') to measure that will serve an especially critical role when comparing treatments in observational contexts. Covariates are any variables that measure or characterize individuals or entities at the *outset* of treatment comparisons. Unless treatments have a discernable beginning, it can be difficult to identify covariates that will not have been affected by treatments. Treatments may also be longitudinal, with several parts or features that occur over some specified time period. When assessment of longitudinal treatment effects is attempted, however, it may be difficult to know what aspects of such treatments are being evaluated.

1.3 *Response Variables*

Empirical studies that compare treatments inevitably entail identification of one or more *response variables* or criteria that provide a basis for assessment of treatment effects. In the comparison of two drug therapies to treat AIDS a response variable might take the form of 'healthiness' ratings with scores supplied by attending physicians. Other possible criteria might be laboratory measures of immune response, T-cell counts, or a binary criterion such as survived or not. If the treatments consisted of two methods for teaching reading, response variables could take the form of measures of oral reading skill or reading comprehension. Response variable selection – or development – can become a major component of treatment comparisons, and to be done most effectively this step requires scientific subject-matter knowledge, or specialized skills and information about measurement.

Response measures in research studies must possess some degree of *reliability or dependability*. If one measures a response repeatedly, the values obtained should be comparable, and reasonably stable when the thing being measured itself does not change. Similarly, persuasive research comparisons require response variables to have *defensible validity*, which is to say the measure captures the construct or feature that s/he wants to measure. (The same may go for covariates, although reliability is arguably more important than validity for these variables.) Response variables should also have *central relevance* to the comparisons of interest. Is it reasonable to believe that the chosen response measures will be accepted by a critical audience as reflecting treatment effects, were they to be found? And how many response variables should be examined for particular treatment comparisons? One variable will often be insufficient to cover the intended effects of treatments. Sometimes it may be desirable to obtain measures at multiple time points following the specific period over which the treatments 'ran.'

When treatments are compared, response variables to be examined should often consider *side effects*; these are often termed 'secondary outcomes' of treatments. Side effects may be unintended effects, but they often take on major importance. However important the principal or targeted effects of treatments are, *side effects can never be wholly ignored,* and in some areas, especially medical sciences, side effects can overwhelm primary effects. A little reflection on drugs that had to be withdrawn from the commercial market will show that when drug use is found to be associated with untoward effects these results may undermine the intended purpose of the drug. Often, such effects are not anticipated, but guidelines have been produced to help reduce these problems. The reader has only to recall the devastating results caused by thalidomide, where many infants were born with missing limbs, to realize how devastating side effects can be. (Thalidomide was marketed in the late 1950's to reduce morning sickness among pregnant women.)

1.4 *Observational Units*

The *units* (sometimes referred to as experimental units) to which treatments are applied or administered are often human beings. In the social and psychological sciences, units are often called 'subjects;' but in the physical sciences, the term 'material' is often used to reference units. In any case observational units need not be single organisms, but may be aggregates such as classrooms or families. Alternatively, units may be rodents or blood samples or virtually any entity to which treatments may be applied. Particularly when units involve human beings, questions of ethics or morality arise in administrations of treatments. Such treatments must be properly screened for safety, and the avoidance of undue risk. The latter point is central to most applications where the units are human subjects and the treatments are related to either public health or education. There are often 'At risk' issues to be considered, matters that concern the ethicality of using human subjects in an experiment. Indeed, it is often because experiments cannot be done ethically that we will often look to observational studies.

1.5 *Covariates*

Various *preexisting conditions* or *specific characteristics* of individuals will often have a profound impact on response measures. In any situation, be it observational or experimental, it is necessary to carefully list and try to measure key *covariates,* that is, those that may have a material effect on interpretations of differences between treatment groups with respect to the outcome measures. Cholesterol levels, blood pressure, triglycerides, etc., have demonstrable correlations with heart disease, so a comparison of treatments for heart disease should certainly take these into account. The careful definition of treatment regimes, time periods, response variables and units all aid in making sound scientific judgments about what covariates to measure and include in a given analysis. In nearly any study that is undertaken it will be valuable to try to produce a reasonably comprehensive list of such covariates, as well as details about how such measures can be (or were) obtained, and whether the measures are likely to be of high quality (with, for example, few missing values).

In the case of a randomized experiment, units should generally be placed into similar groups, or *blocks*, based on relevant covariates. Ideally, each block of units is similar with respect to all major covariates; then, units will be (randomly) assigned into specific treatment groups within each block. Random assignment has the virtue of making the treatment groups being compared 'relatively similar' with respect to all covariates, whether or not they have been observed; more on this later. This allows more precise estimation of treatment effects, and blocking also *facilitates studies of interactions* between covariates and treatments. (In practice, blocking is used less often than would be ideal, which leads not only to relatively inefficient studies, but also in failures to discover interactions when they exist.)

In the context of observational studies, the data are generally taken from one or more 'naturally existing populations' where blocking can only be induced after the fact. Since units self-select for treatment, treatment groups being compared are *generally not balanced* with respect to any covariates, and this is especially problematic when what may be called *critical or major covariates* are not balanced. We might also speak of *imbalance*, a term that refers to *covariate distributions being different in the treatment and control groups*. The lack of balance associated with observational studies generally leads to what is called *selection bias*. Selection bias is central to propensity score studies because it is the *reduction of selection bias* that is the key motivating feature that led to the development of propensity score methodology.

2. Comparing treatments in observational studies

2.1 *Causal Inferences*

Few specialists would disagree with the statement that when feasible, the scientifically most informative comparison of two or more treatments occurs when units have been *randomly assigned* to treatments. Treatment comparisons based on random assignments of units to treatments are called *true experiments*. Over the past century numerous experimental design principles and methods have been developed to underpin sound, efficient and meaningful comparisons of treatments; moreover, true experiments have been central to scientific progress in many fields.

A key virtue of true experiments is that they can be especially helpful to underpin *causal inferences*. (Some caveats are in order here, but they can wait.) When the investigator randomly assigns units to treatment groups, then the groups are likely to be broadly similar with respect to all characteristics, both known and unknown, that may affect the measured response. Suppose two treatments have been compared following random assignment, and measurements of units taken; suppose further that responses are found to be *notably higher* for one treatment than the other. Since the treatment groups had been randomly defined, this finding can (generally) be defended as (strong?) evidence that the response difference was due to treatment effects. While in practice things are never quite this simple, it remains true that randomization is a revere ed principle in experimental science. For helpful scholarly reviews of causality, see Cox (1992) or Pearl (2000); also, see Holland (1986 & 1988).

Despite the central importance of random assignment in applied science, true experiments play a relatively small role in many sub-disciplines of science. The focus in these pages is to describe and illustrate what are deemed to be especially promising methods for comparing treatments in situations where true experiments are impossible or unfeasible. That is, the main interest here is comparing treatments in the context of *observational studies*.

The particular focus here is on a class of methods called propensity score analysis (PSA), first described by Rosenbaum and Rubin (1983). Since that notable publication, scores of articles have been published on aspects of PSA, and a second edition of Rosenbaum's definitive text *Observational Studies* has now been published (Rosenbaum, 2002). Many applications of PSA have appeared in the medical sciences, but the same cannot be said for the social, behavioral or educational sciences where relatively few PSA studies have been reported. Much of the current literature of PSA is highly technical, requiring graduate level knowledge of statistics for its comprehension. However, the goal here is to show that propensity score methods are highly intuitive, and relatively easy to use and comprehend, regardless of one's technical or mathematical background.

We see propensity score analysis as a most promising methodology that can aid sound causal inferences in observational studies, particularly in the human sciences. As PSA methodology becomes more widely used and understood, it is likely that it will be refined or even standardized according to developed needs. Although this book aims at a basic introduction to PSA, we also hope to provide guidance to the literature and some future directions of PSA research for the student who aims to go beyond these pages. The following example should help define and explain basic terminology, as well as highlight the questions and problems inherent in the analysis of observational data.

2.2 *A PSA example: Charter Schools*

Several states have recently implemented programs to encourage charter schools. Based in reform efforts, charter schools are usually autonomous public schools run by teachers, parents, and/or community organizations. A key argument often advanced to justify charter schools is that they can reduce bureaucracy and thereby improve efficiency. Charter schools are given autonomy and deregulated by the state in exchange for a time-limited contract for student achievement. But the question of how well charter schools have 'worked' is rarely (if ever?) studied intensively, using scientifically sound research methods.

First, what exactly do we mean by a charter school, or an aggregation of such schools? For that matter, how would we characterize the public school(s) with which they are compared? And what evidence would we seek (say if we were parents who were considering placement of their children in such schools) that the charter schools had in fact *worked effectively*? Since the goals may be different in public and charter schools, careful thought is required about outcomes and how to measure them. For instance, we might like to answer questions such as, 'What are the typical achievements of charter school students, as compared with their public school counterparts?' Achievements in what subjects or for what outcomes? Or, better, 'How, and how much, do various achievements and behavioral characteristics of charter school students differ from those of their public school counterparts?'

Having proposed lines of inquiry such as these, we would need to determine exactly what treatments we wish to investigate, and over what time period we seek to focus our efforts. Given the nature of these questions, it seems unlikely that any time frame shorter than a year or two would be appropriate. Since most charter schools focus on the elementary grades, we might limit our attention to selected elementary student comparisons? At what grade level will we make measurements? If we settle on, say, the fourth grade, will we want to limit comparisons to students who have spent the three previous years at the same charter school? Covariate choices are inevitably bound up in such decisions. Once we decide what charter grades to consider, we will want to choose public schools for comparison that are generally comparable. Think about why this should be the case. Should we match each charter school with a geographically and socio-economically similar school? Or should we measure by county, by state? The reader might consider pros and cons of each choice in relation to the others; since we are talking here about such a study in the abstract, no decisions need be made on this and other practical matters.

As we move to settle the various treatment issues we shall need to determine how to go about measuring outcomes. What response variables are possible that will inform the original questions posed above? This may or may not be an easy task, but let us assume that appropriate outcome measures have been selected: perhaps standardized test scores on core subjects, surveys, attendance records etc. For example, suppose charter schools result in better attendance than public schools in a certain community. Does this fact alone provide a basis to conclude that charter schools generally induce better attendance than public schools? Certainly not: herein lies the difficulty in analyzing observational data. Random assignments of students to charter and public schools would rarely be feasible, thus it is wholly possible that students who chose to go to the charter school were more motivated and had better patterns of attendance even before they began charter schools. Nor would a simple difference in any other measure be conclusive evidence for or against charter schools. Outcome measures are inherently influenced by differences in the types of students who self-select (or whose parents select) the school. This means that the problem of *selection bias* is likely to interfere with one's interpretation of data bearing on school performance. Propensity score analysis seeks to adjust for selection bias based on identified and measured ancillary, collateral or concomitant information or *covariates.*

Because it is generally recognized that students (or parents) self-select for type of school, actual comparisons of charter schools to date tend to not be 'outcomes based,' but rather mostly anecdotal and subjective. A good example of this is the government document, 'The State of Charter Schools, 2000, Fourth Year Report.'[1] As is typical, this report offers no evidential basis for comparing achievements of charter and public students. Its authors use survey data to focus on descriptions of the numbers of charter schools, where they exist, how they have grown, and what purposes they are intended to serve. But the desired comparison data are typically incomplete, unfocused, and easily criticized as a basis for judgments about the relative virtues of these two kinds of schools. Indeed, despite the fact that parents are routinely making major decisions that entail choices between charter, and say, public schools, there seems to be no sound basis in empirical studies to date that permit informed choices, studies where selection bias has been taken into account in the comparisons.

Given that random assignments of students to charter or public schools is not possible, PSA methods are worthy of examination for their potential to adjust for differences between students in the two kinds of schools. That is, a well-conceived PSA can account for selection bias, and thus provide a basis for sound comparison of these two kinds of schools, despite the unfeasibility of random assignments of students to

[1] Available at http://www.ed.gov/PDFDocs/4yrrpt.pdf.

schools. In what follows we shall sketch the logic of PSA in the context of comparing charter and public schools.

2.3 *Propensity Score Analysis of Charter School Data (draft section)*

The initial goal of PSA methods for comparing charter and public schools would be to observe as many covariates (just think, predictor variables) as possible that might account for major or critical differences between students who enter these two kinds of schools. That is, the aim would be to be able to distinguish between students (parents) who have chosen charter schools and students who have chosen public schools. Inevitably some covariates would work better than others to ***distinguish*** between charter and public school students. The initial aim would be to learn how students differed in these kinds of schools; any and all variables that serve to predict the distinction between the two kinds of schools might be a candidate covariate in this context. Student-level covariates of major interest might be (standardized) test scores for English-Language arts, reading comprehension, vocabulary, word attack; various components of mathematics achievement or skill, as well as science achievement scores. Additional student-level covariates to discriminate between charter and public school students could include measures of the numbers of books in the child's home, parents' education, family socioeconomic status, attendance history, and measures of motivation to learn. The more knowledge one has about the differences between the kinds of students that enter public and charter schools in a particular community the better equipped s/he would be to choose effective covariates.

Supposing that observations were made for a comprehensive selection of covariates, PSA methods entail using covariate information to *model or predict* a (binary) 'treatment assignment' variable, in this case, whether a student is in a charter or a public school. That is, we wish to model the *propensity*, or probability, given the covariates, of a student attending the charter school (say). The propensity can be modeled in any of several ways. One method is to use logistic regression analysis (see wikipedia) to generate a single new covariate (itself a function of many covariates) that would serve to predict the binary treatment variable. Such a derived variable, based solely on covariates, is called an *estimated propensity score*, as each value (a score in the range zero to one) indicates the 'propensity' of a single student to have chosen charter over public schools. The complement, one minus the estimated propensity score, indicates the propensity to have chosen the other type of school. Modeling (the probability of) treatment assignment constitutes the *first phase* of a PSA, which generally has two main parts, or phases.

Given a reasonable selection of covariates, students with similar propensity scores can usually be assumed to be comparable to one another; that is, within each (narrow) band of propensity scores, covariate distributions should tend to be similar when comparing treatment and control groups. These narrow bands are called propensity strata. This is exactly what randomization does in the context of a true experiment – indeed, to seek evidence of balance in covariate distributions is one of the main aspects of a comprehensive propensity score analysis. (Ideally, blocking considerations should be considered in a PSA applications – even though this is difficult, and therefore rare.)

The *second phase* of a PSA is to sort or *rank students* according to their estimated propensity scores, and then stratify students on these scores. Within strata (perhaps as few as five) propensity scores and so covariate distributions should be similar. That these distributions are similar within strata needs to be checked carefully. A standard approach is then to make comparisons within each PS stratum for *any* chosen outcome measure(s). Assuming a continuous measure, the response variable difference is found within each stratum and then averaged over the strata. If the strata have different sizes, the sizes determine the weights to be applied to the mean differences. This mean response variable difference between charter and public school is the basic comparison to be made for each response.

The key point in this context is that within any narrow band of propensity scores students will be comparable with respect to any covariate that distinguishes between these two kinds of schools because the bands have been derived by effectively taking student differences into account. A particular PSA application might focus on discriminating between treatments within a particular state, city or community. Students with similar propensity scores will thus not only be similar to one another with respect to the covariates, they will of course be comparable within a particular state, city or community. If we then compare other groups of students with similar but different from the first group of student's propensities, then when we average over the size of the groups we can get a measure of the overall effect size of the charter versus traditional school. Put another way, *PSA entails comparison of (subgroups of) likes with likes.*

In practice, the counterfactual approach to treatment comparisons is not feasible; but the CF approach does help to define the problem conceptually, or theoretically. Since students can only be in one place at a time, they can only have an outcome for one of the treatments. So while we cannot measure the outcomes of an individual student at both the charter and traditional school, we can measure the outcomes of students with similar propensities – i.e. propensity scores – some of whom attended the charter school, and some of whom attended the traditional school.

This example illustrates the essential ideas of PSA. Several other documents of interest can be found at propensityscoreanalysis.pbwiki.com . Check it out.