An overview of Propensity Score Analysis including illustrations*

Bob Pruzek, University at Albany (Biostatistics)

Abstract

Propensity Score Analysis (PSA) was introduced by Rosenbaum & Rubin (Biometrika, 1983). Since then PSA has become one of the most studied and frequently used new methods in statistics. Hundreds of papers have been published, covering philosophy, statistical theory and a wide variety of applications (health, medicine, behavioral sciences, biology, education, economics, etc). Here I focus on the background and logical foundations of PSA, as well as the key questions and basic forms of the analytic method. Two data sets are used to illustrate how the method works, how results might effectively be interpreted, and what some of the key ancillary questions or issues are; several novel graphical methods are demonstrated using freely available software in the form of R.

**Talk on 4/28/10 to Albany Chapter of the American Statistical Association*

PSA owes its central rationale to the logic that underpins the analysis of true experiments. In true experiments, units are ***randomly allocated*** to the (two) treatment groups at the outset of an experiment, that is, before the treatments begin. In the words of Fisher, randomization is the 'reasoned basis for causal inference' in experiments. Its role is to ensure that units allocated to each treatment group do not differ systematically from one another on ***any*** covariate. Randomization supports causal interpretations: If the one group scores systematically higher than the other, then, thanks to randomized allocation of units to treatments, this finding can (with qualifications*) be attributed to the treatments, and not other factors. **Three caveats are in order: 1. Randomization can go awry in practice, particularly when samples are not large;*
2. Much depends on the details of how experiments were run; &
3. To say that 'treatments caused the differences' is not to say that one knows what feature(s) of the treatments had the noted effects. We study the 'effects of causes,' not 'causes of effects'.

Observational studies entail comparison of groups that were not formed using randomization; this means that observational studies carry with them a greater likelihood for ***Selection Bias*** *(SB)*. SB refers to systematic covariate differences between groups being compared, differences that can confound attempts to interpret treatment differences when they are found. SB is the central problem that propensity score analysis aims to reduce, if not eliminate (usually -- but not always -- in the context of observational studies).

Three people have written key articles and books that underpin propensity score methods: William Cochran, his student Donald Rubin, and then his student, Paul Rosenbaum. A review of one of Cochran's reports, done 40 years ago is worth brief examination.

Cochran (1968) studied death rates of smokers and non-smokers. It had been found, when using unstratified data, that death rates for smokers and non-smokers were nearly identical (evidence that many smokers and manufacturers of tobacco products found greatly to their liking). But Cochran decided to sort both smokers and non-smokers by age. Following age-based stratification he re-calculated death rates, only to find that they were on average 40 - 50% higher for smokers than non-smokers -- and this was for very large samples. Results of this kind represent early versions of what now can be seen as propensity score analysis (a term that gave nearly a million hits in a recent Google search!).

Note that when there is only one confounding variable (such as age, in Cochran's case) in an observational study then mere stratification (of subgrouping) on that variable is likely to work well when comparing two (or more) treatments with one another. But this prospect is most unrealistic; in general practice numerous covariates can confound interpretations, and for many years analysts found it most difficult to account for confounding effects. The key breakthrough came when Rosenbaum and Rubin (1983) showed how to produce a single variable, a propensity score, whose use could greatly simplify treatment comparisons in observational studies. They noted that conditions may exist where treatment assignment Z (binary) is independent of potential outcomes Y_0 & Y_1 , conditional on observed baseline covariates, X. That is, $(Y(1), Y(0)) \perp\!\!\!\perp Z|X$, if $0 < P(Z=1|X) < 1$. This condition was defined as ***strong ignorability,*** which essentially means that all covariates that effect treatment assignment are included in X.

These authors then went on to define the ***propensity score e(X) (a scalar function of X) as the probability of treatment assignment, conditional on observed baseline covariates***:

$e(X) = e_i = Pr(Z_i = 1 \mid X_i)$.

They then demonstrated that the propensity score is a ***balancing score***, meaning that, conditional on the propensity score, the distribution of measured baseline covariates is similar between treated & untreated (or treatment and control) subjects. This means that $(Y(1), Y(0)) \perp Z | e(X)$, an analog of the preceding expression. In effect we see that e(X) summarizes the Information in X. Again, they assume that strong ignorability holds.

In practice, the preceding leads to an interest in estimating the (scalar) propensity score from the (vector) of (appropriately chosen) covariates, say X, so that comparisons of treatment and control response score distributions can be made, conditional on an estimated propensity score. The most common method for estimating e(X) entails use of logistic regression (LR).

In practice, there are two main stages or Phases of a propensity score analysis. In Phase I, pre-treatment covariates are used to construct a single variable, a propensity score, that summarizes key differences among units (or respondents) with respect to the two* treatments being compared.

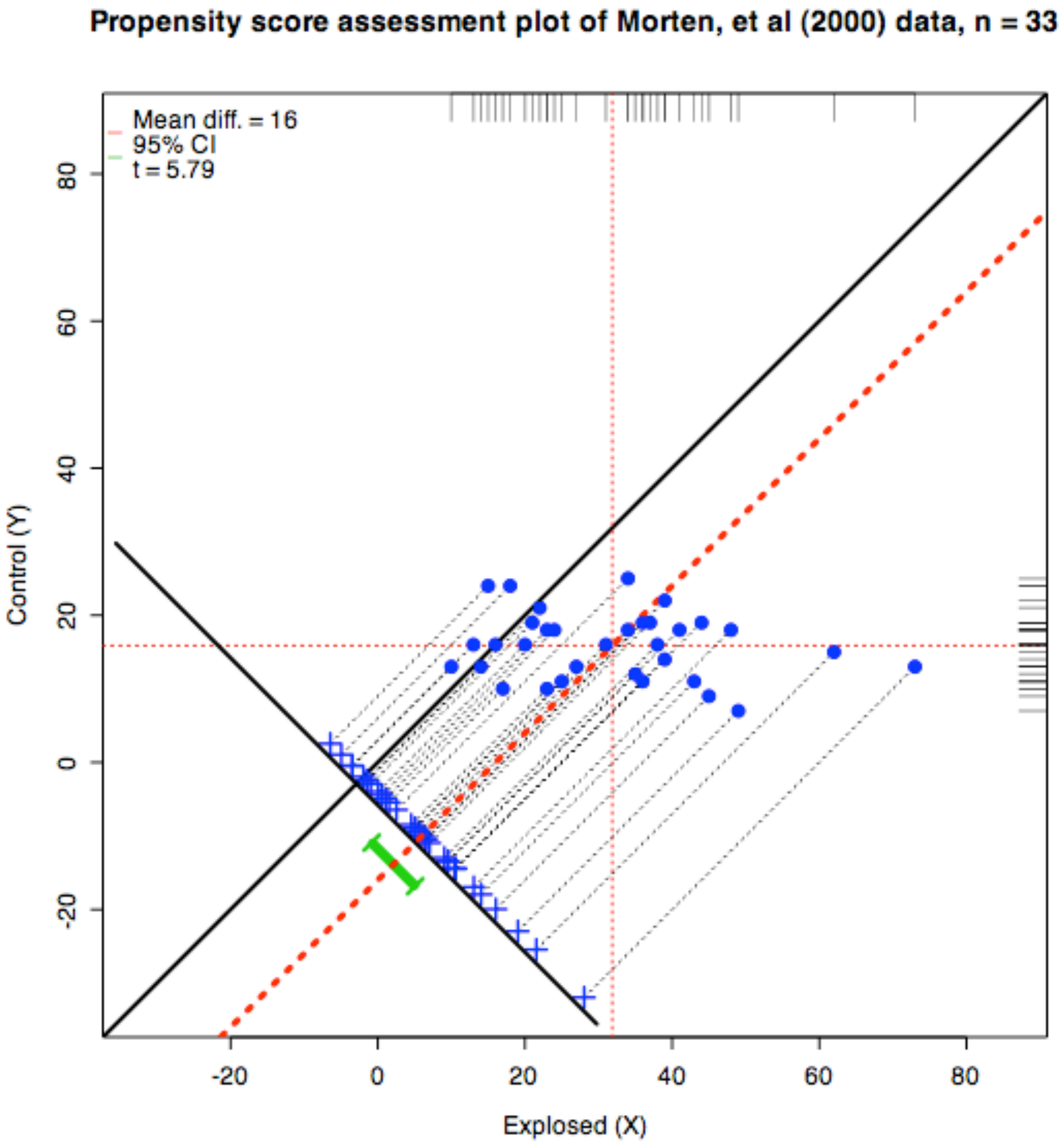
These P-scores are then used in Phase II in two main ways: Units in the treatment and control groups are either matched or stratified (sorted), and then ***the two groups are compared on one or more outcome measures, conditional on these propensity scores***. For matching, the usual approach begins from selection of a treated unit (individual) and tries to match that unit with a control unit with a similar propensity score; in the case of stratification, responses of units are compared within propensity-based strata. Both methods are illustrated below.

*Except for some recent work, nearly all PSA's to date have focused on two group comparisons.

Data shown in the next slide* derive from an observational study by Morten, *et. al* (1982, *Amer. Jour. Epidemiology*, p. 549 ff); this entails a relatively simple form of propensity score analysis.

Children of parents who had worked in a factory where lead was used in making batteries were ***matched by age and neighborhood*** with children whose parents *did not* work in lead-related industries. Whole blood was assessed for lead content to provide responses. Results shown compare Exposed with Control Children in what can be seen as a ***paired samples design***. Conventional dependent sample analysis shows that the (95%) C.I. for the population mean difference is far from zero. The mean of the difference scores is 5.78, and the results support the interpretation that parents' lead-related occupations tend to influence how much lead is found in their children's blood.

- This plot is called a *Propensity Score Assessment Plot* and is produced by the function **`granova.ds`** in R package **`granova`** (Pruzek & Helmreich, 2007). Note that the heavy black line on the diagonal corresponds to X = Y, so if X > Y its point lies below the identity diagonal. Parallel projections to the lower left line segment show the ***distribution of difference scores corresponding to the pairs***; the red dashed line shows the average difference score, and the green line segment shows the 95% C.I.


Propensity score assessment plot of Morten, et al (2000) data, n = 33
Mean diff. = 16
95% CI
t = 5.79
Control (Y)
-20
0
20
40
60
80
-20
0
20
40
60
80
Explosed (X)

The graphic shows more, however. Note the *wide dispersion of lead measurements* for Exposed children in comparison with their Control counterparts. A follow-up to the main study showed that parental hygiene differed largely across the battery-factory parents, and the variation in hygiene in large measure served to account for the dispersion of their children's lead measurements (a finding made possible because of the authors' close attention to detail in their data collection). Although it is not certain that Control & Exposed children did not differ in other ways (than age and neighborhood of residence) these data seem rather persuasive in showing that working in a lead-based battery factory puts the workers' children at major risk for high levels of blood lead, except when personal hygiene of the worker was 'generally satisfactory'.

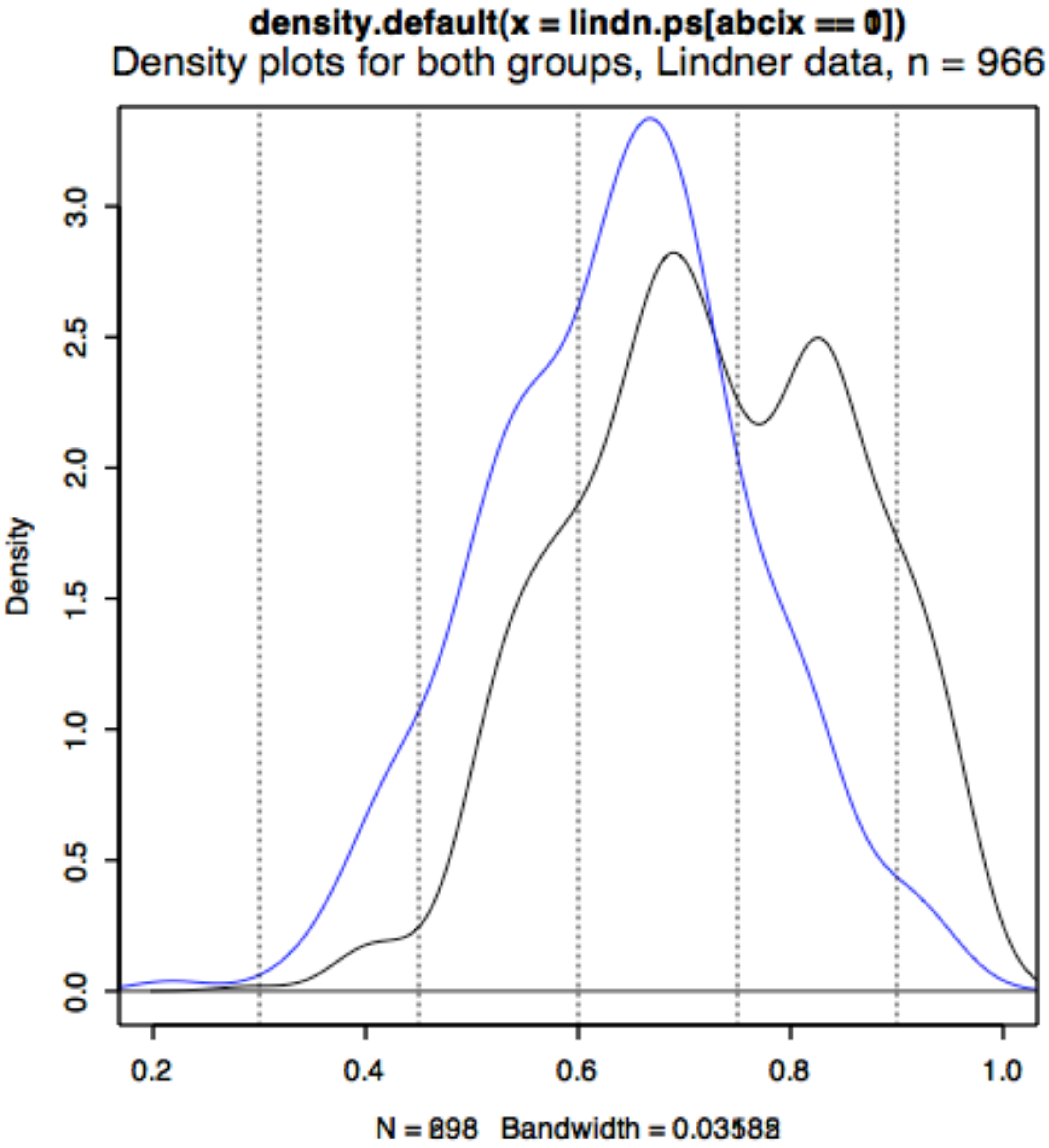
Rosenbaum (2002), who discusses this example in detail, uses a ***sensitivity analysis*** to show that the hidden bias would have to be extreme to explain away differences this large. Sensitivity analyses can be essential to a wrap-up of a PSA study, but they are often not completed. In summary, these observational data provide useful evidence to support ***causal conclusions*** about the specified treatments.

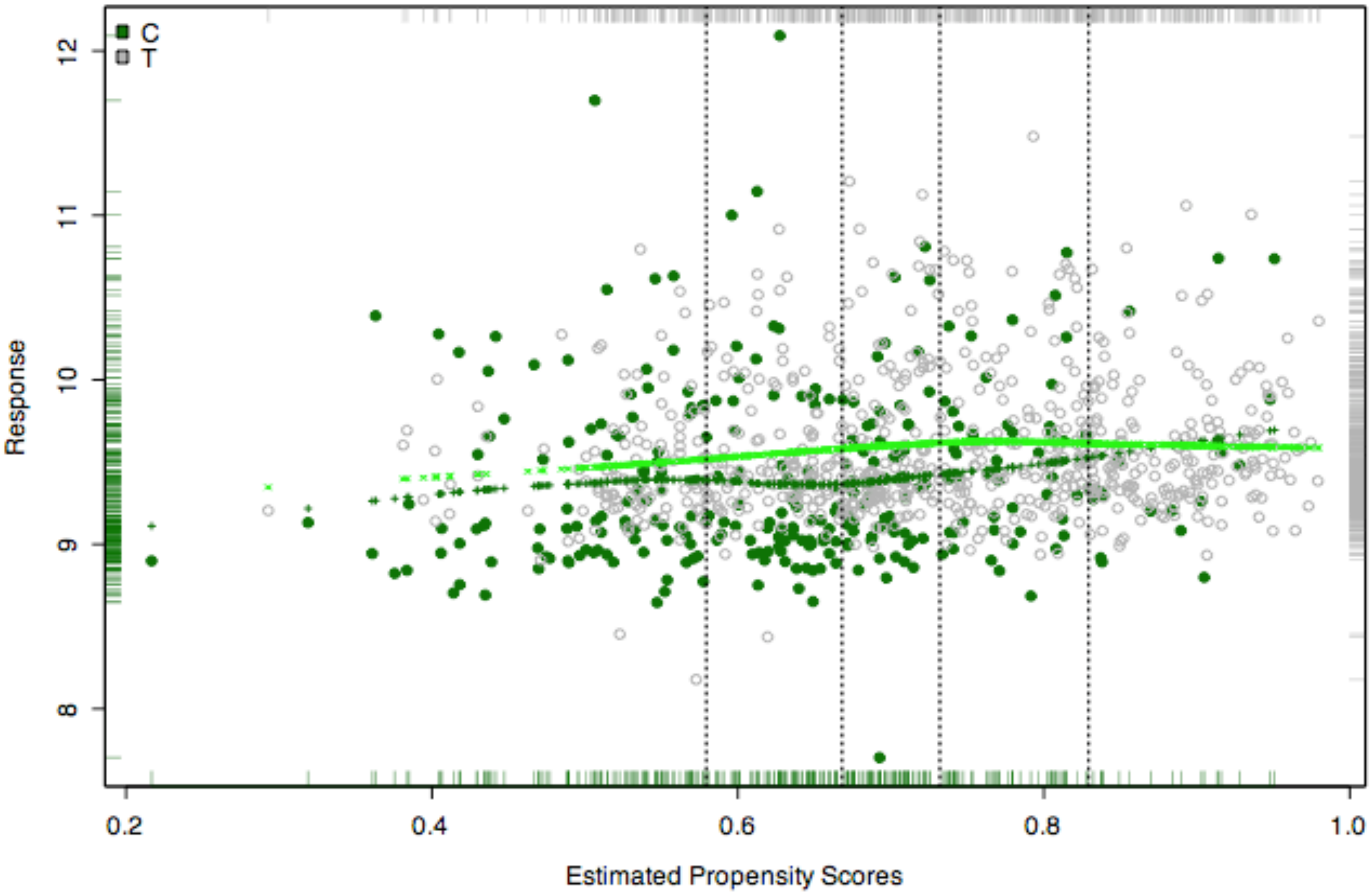
Consider next, estimation of propensity scores for a medical study using data on 996 initial Percutaneous Coronary Interventions (PCIs) performed in 1997 at the Lindner Center, Christ Hospital, Cincinnati. The goal was to assess the effect of a drug ‘abcix’.

Description: Data from an observational study of 996 patients receiving a PCI at Ohio Heart Health in 1997 and followed for at least 6 months by the staff of the Lindner Center. This is a landmark dataset in the literature on propensity score adjustment for treatment selection bias due to recent practice of evidence based medicine; patients receiving abciximab tended to be more severely diseased than those who did not receive a cascade blocker.

The binary variable abcix indicates control (0) or treatment (1). Logistic regression was used initially to estimate propensity scores.

The next two slides show two graphics. In the first, (densities) for both of the Lindner treatment groups, where the Counts were 298 for Control (0), and 698 for Treatment. Five strata are also identified by vertical lines in the plot. The loess plot, the second figure, is particularly informative, as it shows the (non-linear) regression lines for predicting costs (log metric) of the two treatments. Interpretation of the loess graphic will be provided after it's presentation. Because the cost variable was strongly positively skewed, the response variable analysis was done in log metric, i.e. log(cost).


density.default(x = lindn.ps[abcix == 0])
Density plots for both groups, Lindner data, n = 966
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8
1.0
N = 898   Bandwidth = 0.03588

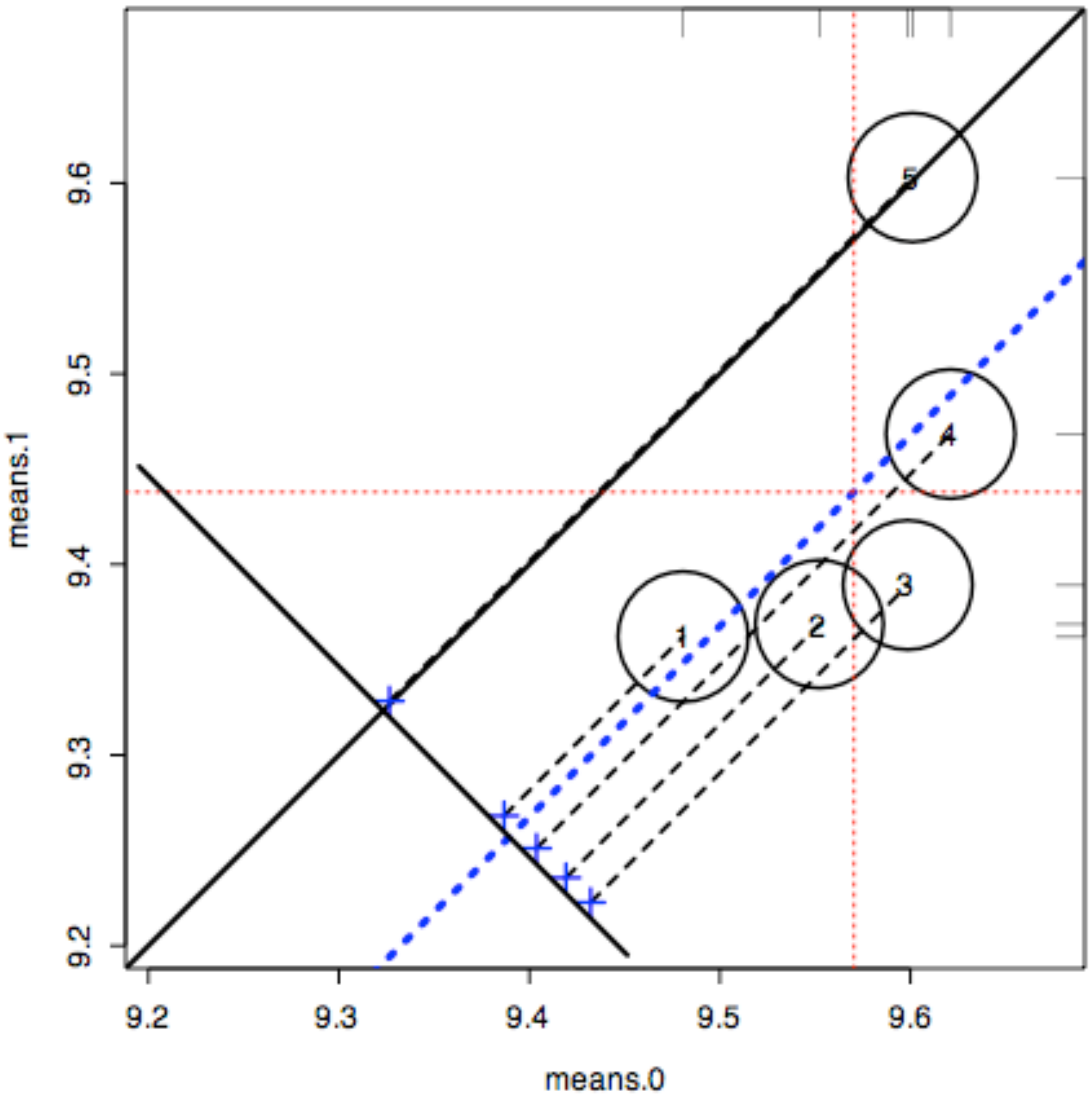


Estimated Propensity Scores

The preceding graphic shows all 996 data points colored to correspond with the treatment/control designations. Note that the loess (non-linear) regression curve for abcix = 1 lies above the curve for the control group across nearly the full range of the propensity scores. This indicates that costs of treatments (this being a proxy for health problems) for abcix were almost universally higher than for their control counterparts after adjusting for all covariate differences using LR-based P-scores; 95% C.I.=(.05,.21) in Log metric. Summary results for strata (noted by vertical lines in the preceding plot) are shown in the table below, as well as in the graphic on the next page. (Alas, the x- and y-axis labels are reversed; they should be x: mean.1, y: mean.0.)

	counts.0	counts.1	means.0	means.1	diff.means
1	97	107	9.36	9.48	0.12
2	73	122	9.37	9.55	0.18
3	62	138	9.39	9.60	0.21
4	48	151	9.47	9.62	0.15
5	18	180	9.60	9.60	0.00

Propensity score assessment plot, Lindner data, LR-based strata, n = 966



Summary for PSA of Lindner PCI Data (only parts of which were seen here)

(The lindner data are available in the USPS & the PSAgraphics (R) packages.)

Before adjustment (n=966: 298 control & 698 treatment) (not shown) 5 out of 7 covariates, including ejecfrac, Ves1pro, acutemi, daibetic and stent showed notable mean differences between treatment and control.

For the LR-based method, the loess regression plot shows the full range of adjusted effect results; this response variable analysis produced a 95% CI that did not span zero, showing significantly higher overall costs for abcix after PS adjustment. Next, five strata were defined. Treatment effects differed somewhat across strata, but the effects again showed greater costs overall for abcix than for control after adjustments. Strata 1 - 4 showed the strongest treatment effects while stratum 5 showed a change of sign in the mean log(cost) difference, but essentially a null treatment effect.

For the classification tree stratification method (not shown), 6 strata were found and concordance with the preceding treatment effects across propensity score levels were observed. The log(cost) analysis yielded a confidence interval that did not to span zero showing that there were statistical differences in costs for the treatment and control groups, results that were consistent with those found using LR.

The graphical-visualization functions (in R package PSAgraphics) for PSA provided helpful images of the sizes and direction of treatment effects, and clarified the differences found for the different PSA methods. In general, supposing that the available covariates accounted for most of the selection bias, the results might be taken to imply causal effects of increased abcix costs compared to the control after P-score adjustments.

The foregoing are only two examples. Others are easy to imagine: Comparing two behavioral patterns with one another, two diets or two exercise plans; or two food supplement schedules.

In many situations it is either unethical or impractical to use randomization to allocate individuals to treatment and control groups (or simply to the treatments). In such cases, given that an appropriate or ‘reasonable’ set of covariates can be observed for all units, propensity score methods can facilitate comparison of treatments in a way that removes the notable effects of selection bias.

The key requirement in such observational studies is the selection of covariates that are responsible for most of the selection bias, and the use effective numerical or graphical (statistical) methods to make comparisons with respect to appropriate outcome measures.