Propensity Score Methods for Longitudinal Data Analyses: General background, rationale and illustrations*

Bob Pruzek, University at Albany SUNY

Summary

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 appeared covering philosophy, statistical theory and a wide variety of applications (especially in health science & medicine). I focus on background and logical foundations of PSA, then present & discuss graphics that illustrate various PSA methods; lastly, I describe with examples how conventional PSA methodology can be extended to accommodate longitudinal data analysis. It is noted that while longitudinal PSA often often entails notable complications, special advantages can accrue to LDA-PSA if attention is given to certain aspects of observational study design.

*Talk for INTEGRATIVE ANALYSIS OF LONGITUDINAL STUDIES OF AGING Conference, Victoria, BC June 2010

PSA is based on the same *logic that underpins analyses of true experiments*. In true experiments, units are ***randomly allocated*** to (two) treatment groups at the outset of study, that is, before the treatments begin. Randomization, in the words of R. A. Fisher, provides the 'reasoned basis for causal inference' in experiments. Randomization ensures that units in the two treatment groups do not differ systematically on ***any*** covariate which is why this operation supports *causal interpretations*: When one group scores notably higher than another on ultimate response variable(s), this can (with qualifications*) be attributed to treatment differences; random assignment tends to make alternative explanations implausible.

**Three caveats, at least, are in order: 1. Randomization can go awry in practice, particularly when samples are not large;*

2. Much depends on details of how experiments are run; &

3. To say that "treatments caused differences" is not to say that one knows what feature(s) of the treatments had the noted effects. Statisticians generally study 'effects of causes,' not 'causes of effects'.

Observational studies entail comparison of groups not formed using randomization; units are said to "select their own treatments." This means that observational studies give rise to a greater likelihood for ***Selection Bias*** (SB). *SB refers to systematic covariate differences between groups differences that can confound attempts to interpret response variable differences.* 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). This tends to be facilitated if one conceptualizes each observational study as having arisen from a (complex) randomized experiment.

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 studies, done 40 years ago is worth brief examination.

Cochran (1968) compared death rates of smokers and non-smokers. It had been found, 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). Cochran decided to reanalyze the data after **stratifying** both smokers & non-smokers **by age** before computing death rates. After age-based stratification he re-calculated death rates. This led to the finding that death rates among smokers were on average 40 - 50% higher than for non-smokers! Moreover, this was for very large samples. Results of this kind represent early versions of what now can be seen as propensity score analysis. The advent of modern PSA methods helps investigators adjust for multiple covariates, not just one as in Cochran's case.

When there are many potential confounding variables in an observational study then direct stratification is unwieldy because the number of 'cells' associated with the crossing of covariates is often huge; also missing values will be found in many cells. Nevertheless numerous covariates can be expected to confound interpretations. For many years analysts found it especially 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, the use of which 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, $(Y(1), Y(0)) \perp Z|X$, if $0 < P(Z=1|X) < 1$. This condition was defined as ***strong ignorability*** – which essentially means that all covariates that affect treatment assignment are included in X.

Reference to 'potential outcomes' invokes counterfactual logic.

These authors defined 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 for the treated & untreated (or treatment and control) subjects. Therefore $(Y(1), Y(0)) \perp Z \mid e(X)$, an analog of the preceding expression. In effect e(X) summarizes the information in X. Rosenbaum and Rubin rely strongly on the assumption of strong ignorability.

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).

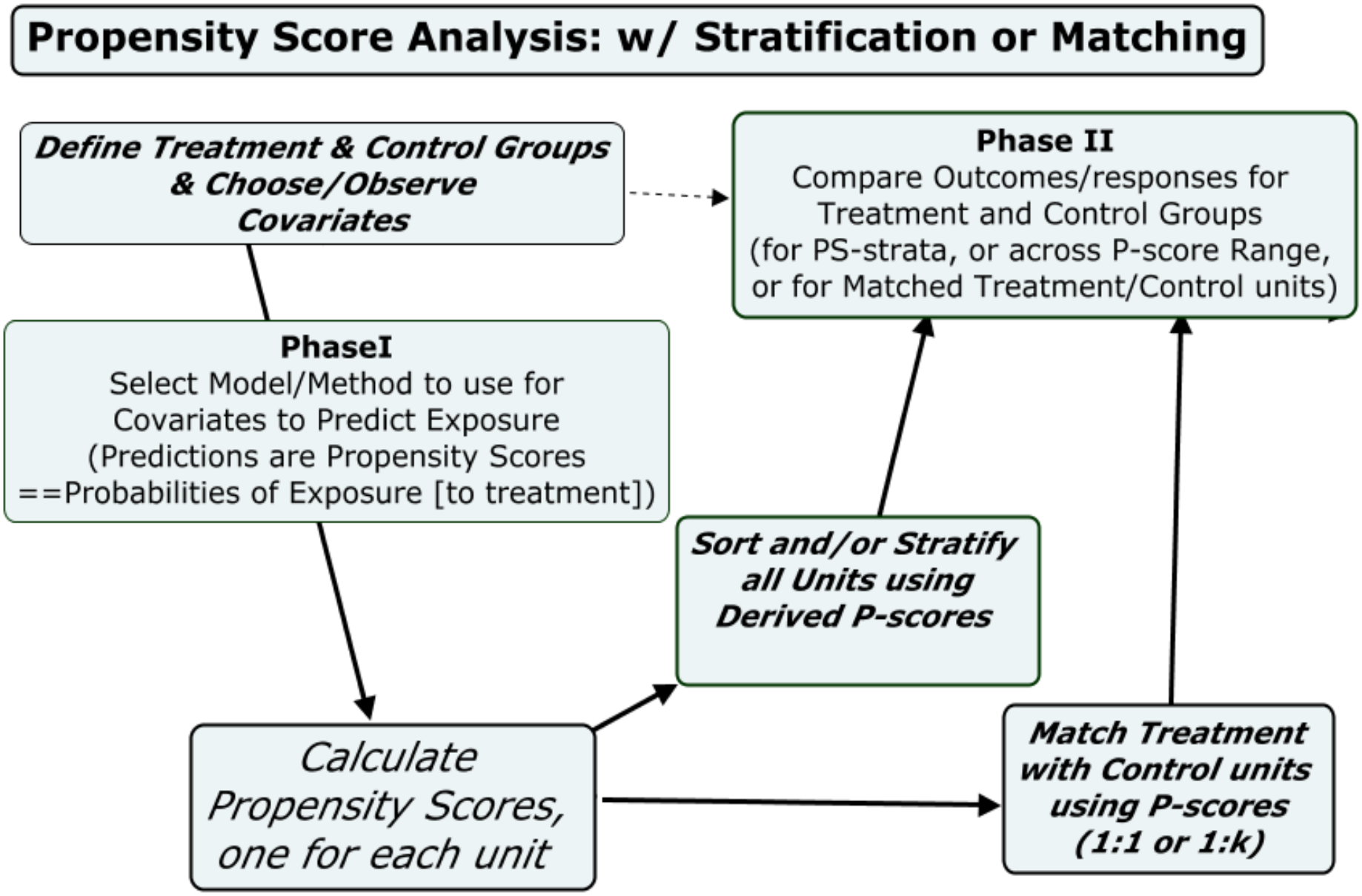
In practice, there are two main Phases of a propensity score analysis. In Phase I, pre-treatment covariates are used to construct a scalar variable, a propensity score, that summarizes key differences among units (or respondents) with respect to the two* treatments being compared. Generally the fitted values produced in *logistic regression* are taken as estimates of propensity scores, the e(X)'s. $e(x) = 1/(1 + e^{-\{\text{linear function of covariates}\}})$.

These e(X)'s are then used in Phase II in either of two ways: units in the treatment and control groups are either matched or stratified (sorted); then ***the two groups are compared on one or more outcome measures, conditional on the matches or strata***. For matching, an algorithm or rule is used to match individuals in the T & C groups whose P-scores are "reasonably close" to one another; numerous methods are available. With stratification responses of units in the two groups are compared within propensity-based strata. Both methods are illustrated below.

*Except for recent work, nearly all PSA's to date have focused on two groups. See my wiki: propensityscoreanalysis.pbworks.com

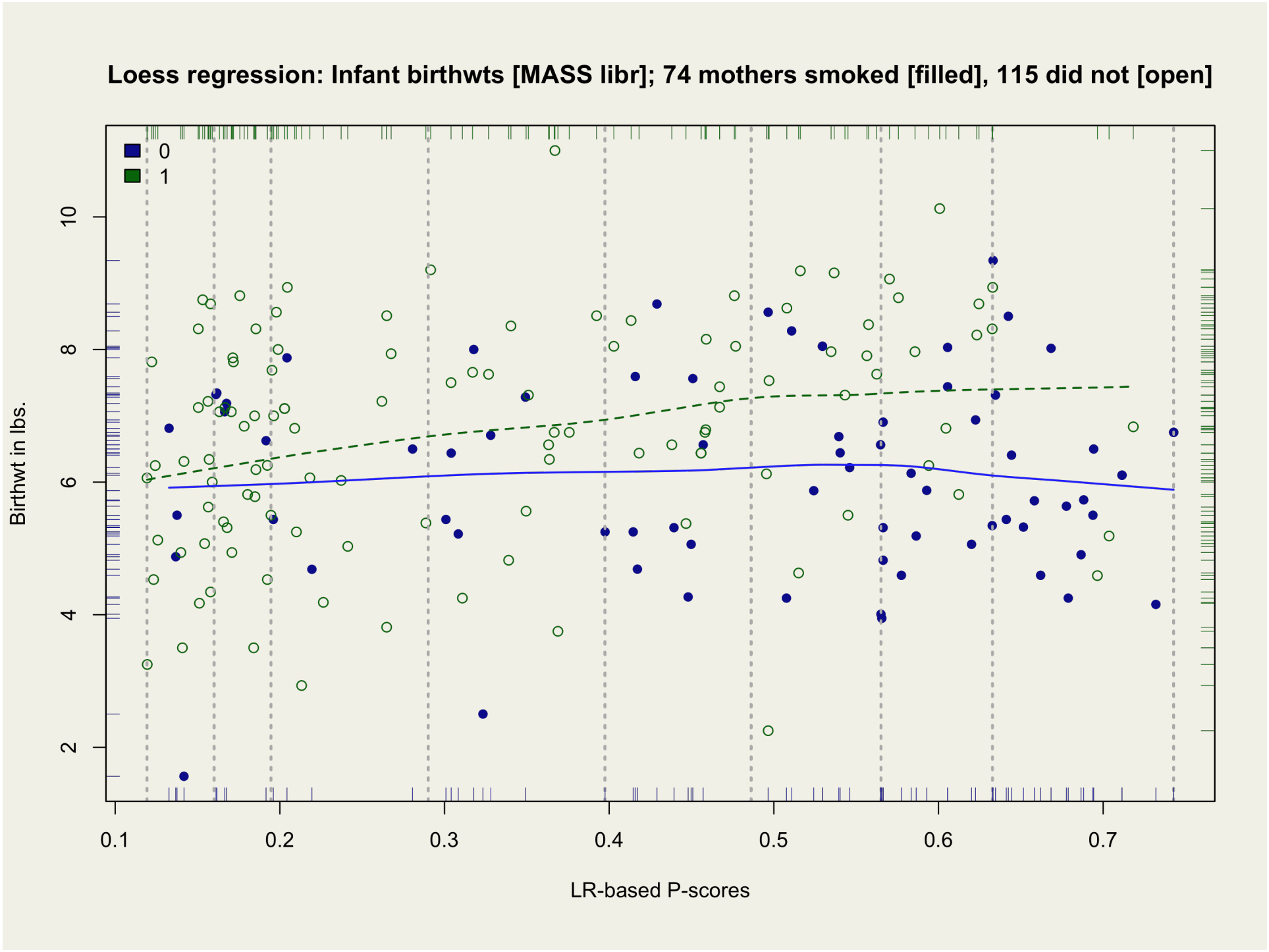
The following slide exhibits a flow chart showing how propensity score analysis proceeds when comparing two groups (to be read counterclockwise from the NW corner).

- Covariate selection is central. Once the T & C groups have been defined, the key problem is to decide *what covariates should be balanced* re: T & C comparison. Theory and prior evidence come into play. Use of all relevant covariates is advised; they should relate to the ultimate response variable, as well as the T vs. C distinction
- Logistic regression modeling should consider *main effects as well as interactions* (based on substantive relevance, and empirics)
- Once propensity scores have been calculated, it is helpful to demonstrate *overlap of P-score distributions* for the T & C groups
- Either or both, *matching and stratification, are generally used for analyses*; the estimands, however, differ in the two cases (ATT, ATE).
- Outcomes are readily compared across the range of P-scores; see the loess graphic that follows. For matched data either dependent or independent sample statistical methods and graphics may be used.


Propensity Score Analysis: w/ Stratification or Matching
Define Treatment & Control Groups & Choose/Observe Covariates
Phase II
Compare Outcomes/responses for Treatment and Control Groups (for PS-strata, or across P-score Range, or for Matched Treatment/Control units)
PhaseI
Select Model/Method to use for Covariates to Predict Exposure (Predictions are Propensity Scores ==Probabilities of Exposure [to treatment])
Sort and/or Stratify all Units using Derived P-scores
Calculate Propensity Scores, one for each unit
Match Treatment with Control units using P-scores (1:1 or 1:k)

The next slide illustrates a Phase II analysis, where *loess regression w*as used to compare infant birth weights of mothers who smoked (treatment group) with mothers who did not. Birth weights (in lbs.) are plotted (vertical) against LR-derived propensity scores (horizontal) for n = 189 infants. Two loess regression lines (dashed and solid) are shown, for infants of smoking (darkened points) and non-smoking (open circles) mothers. Vertical dashed lines depict eight quantile-based strata; effects are assessed within strata (and then averaged).

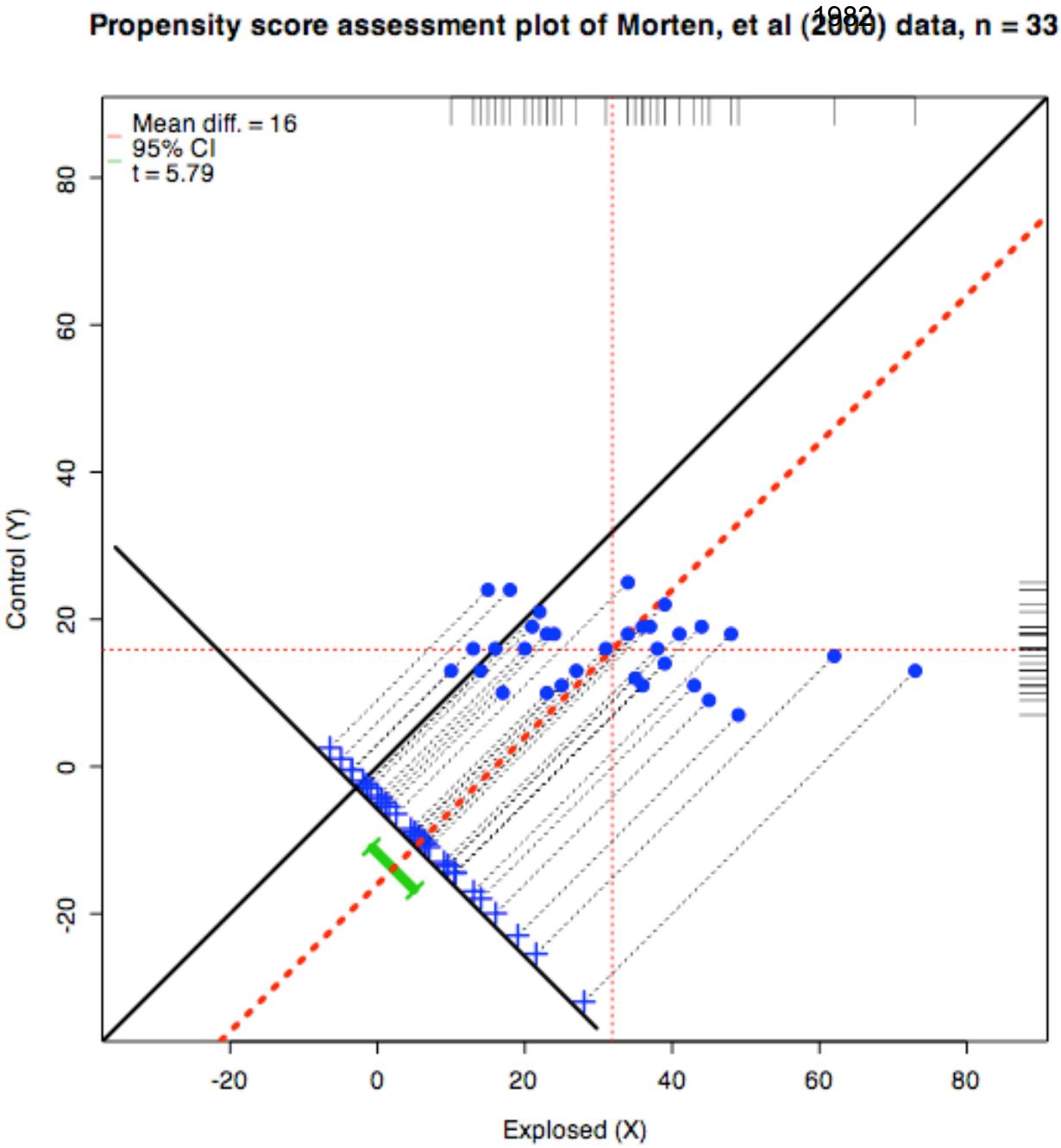
In this case, after adjusting for covariate effects using P-scores, it is seen that birth weights are notably lower for infants (whose mothers smoked) than for controls. (Notably, *overlap* of the two P-score distributions provided reasonable 'support' for the comparison and all covariates were reasonably balanced across the eight P-score strata.) To complete the illustration note that the Average Treatment Effect was .84 lbs., and the 95% CI yields the limits (0.30, 1.38) – failing to span zero. (The graphic is based on function **`loess.psa`** from the **`PSAgraphics`** package (R).)


Loess regression: Infant birthwts [MASS libr]; 74 mothers smoked [filled], 115 did not [open]
0
1
Birthwt in lbs.
2
4
6
8
10
0.1
0.2
0.3
0.4
0.5
0.6
0.7
LR-based P-scores

The next slide illustrates matching* in an observational study by Morten, *et. al* (1982, *Amer. Jour. Epidemiology*, p. 549 ff) that entailed an especially 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 neighbor-hood*** with children whose parents *did not* work in lead-based industries. Whole blood was assessed for lead content to provide responses. Results shown compare blood of Exposed with that of 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 (see line segment, lower left). The mean difference score is 5.78; results support the conclusion that a parents' lead-related occupation can 'cause' lead to be found in their children's blood.

*Using function **`granova.ds`** in package **`granova`** (R). The heavy black line on diagonal corresponds to X = Y, so if X > Y its point lies below the identity line. Parallel projections to 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)
Explosed (X)
-20
0
20
40
60
80

A graphic allows one to go beyond a numerical summary. In this case note the *wide dispersion of lead measurements* for exposed children in comparison with their control counterparts. A follow-up showed that parental hygiene differed largely across the battery-factory parents, and the variation in hygiene accounted in large measure for dispersion of their children's lead measurements (a finding made possible because of Morton's close attention to detail in initial data collection). Although Control & Exposed children may differ in other ways (than age and neighborhood of residence) these data seem persuasive in showing that lead-based battery factory work puts children at risk for high levels of blood lead -- except when personal hygiene of the worker is effective.

Rosenbaum (2002), who discusses this example in detail, uses a ***sensitivity analysis*** to show that the hidden bias would have to be substantial to explain away a difference this large. Sensitivity analyses can be essential to a wrap-up of a PSA study.

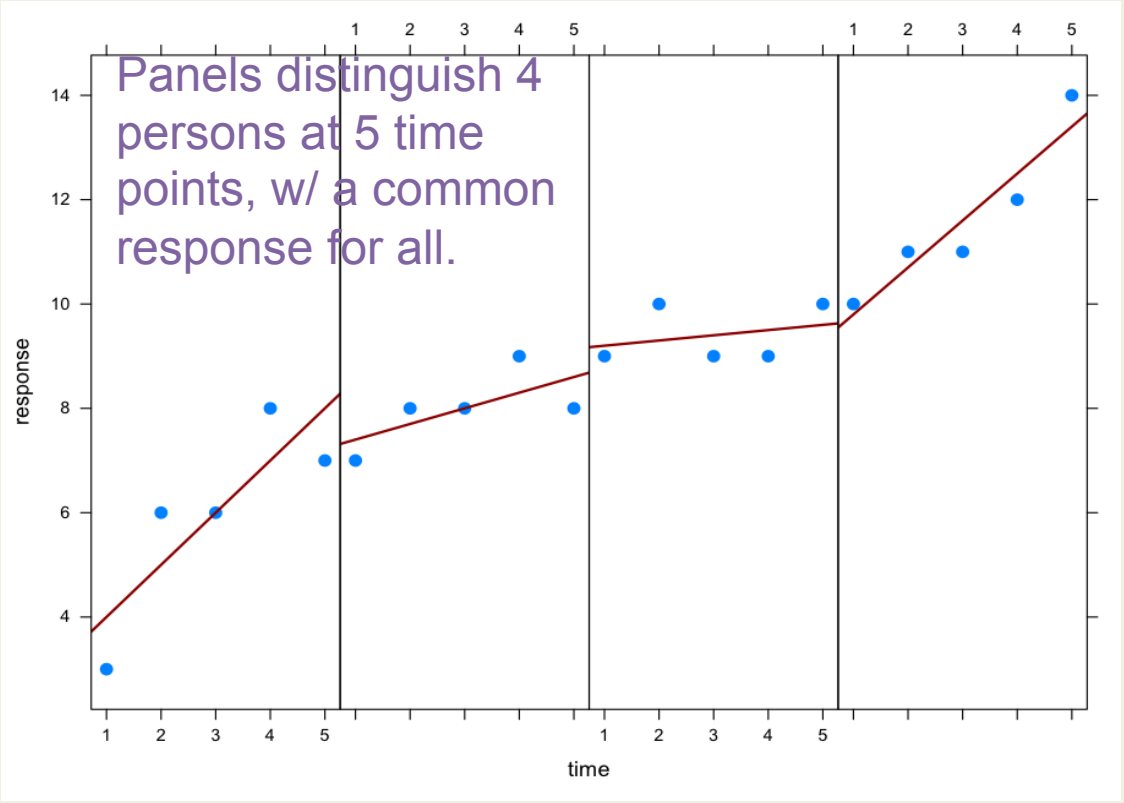
In summary, these observational data appear to provide valuable evidence to support ***causal conclusions*** *re: the hypothesis.*

The basic ideas of PSA have been simplified in order to focus on key principles and methods central to modern-day propensity score applications. Recent PSA investigations have begun to move beyond comparison of two treatments, to compare three or more (however, most authors assume an underlying continuum, e.g., dose-response groups). Multilevel methods for PSA have begun to be published, as have methods for studying mediation; the role of stratification has also begun to see attention. A few studies have been aimed at missing data imputation methods, including multiple imputation. Pearl (2010), in particular, has formalized basic ideas to help bridge the gap between mainstream PSA methods & structural/graphical modeling.

To date, only a handful of authors seem to have addressed the central issue of this conference, *viz*., ***analysis of longitudinal data*** -- in particular, P-score methodology to compare treated & control groups for observational data after adjusting for confounding covariates.

In what follows, I use a basic illustration to show how preceding methodology can be extended to deal with longitudinal data comparisons. The next slide sets the stage for how this might be done.

Longitudinal data are shown for four individuals, where slopes & intercepts are readily discerned for 4 individuals for 5 time waves. For data like these, where T & C groups could start *prior to time 1*, and key covariate data are available for all units, P-scores could be generated to *to assess T vs. C effects. Statistics that describe profiles could be used as responses in PSA.* A key is to find statistics *sufficient to characterize* trajectories (regardless of the # of data waves). In this way, LDA versions of PSA may be straightforwardly generalized, moving from univariate to multivariate PSA.

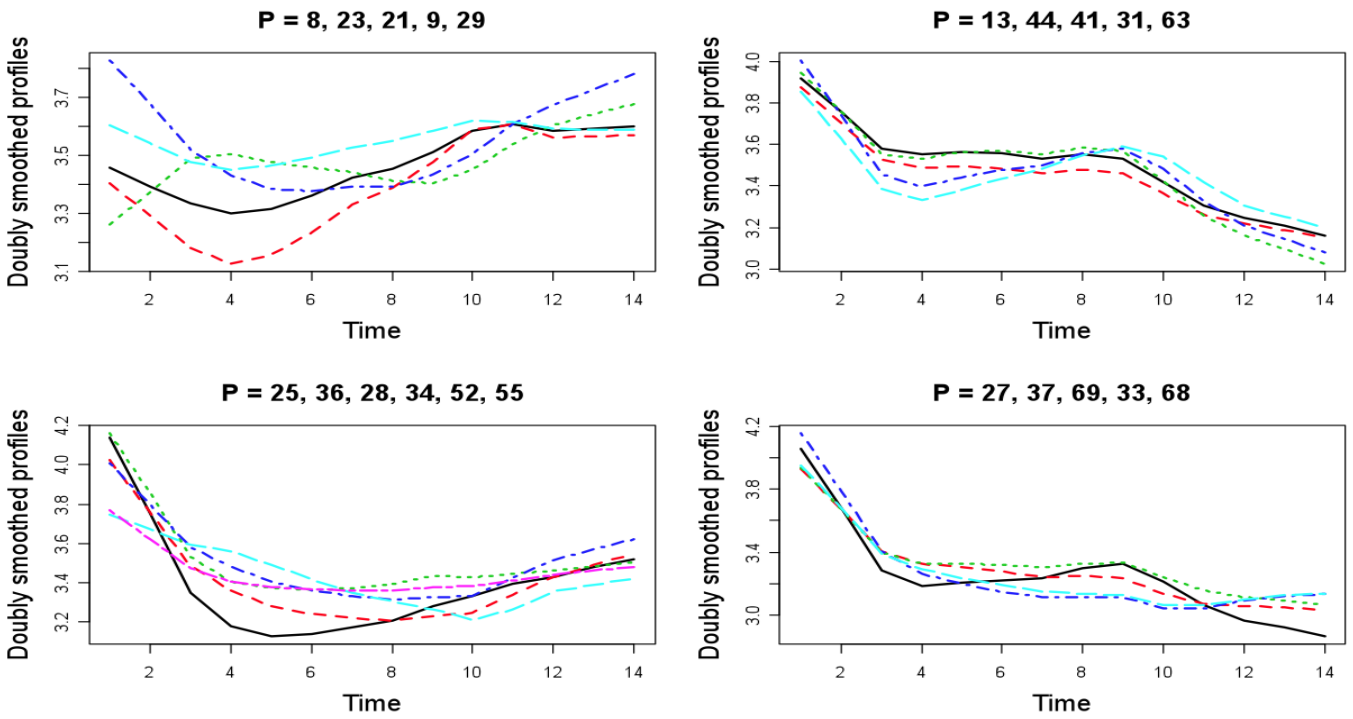


Assuming straight line regression for all panels, two statistics are sufficient regardless of the # of waves; moreover, many fitted (smoothed) curves might entail few (often no more than 3 or 4) statistics to characterize time trends for assessments of treatment effects. In such cases PSA can be generalized to (low dimensional) multivariate analysis to support observational LD analyses. *Smoothing is the key*; let us consider this topic next.

Given the preceding approaches for extending PSA to longitudinal data for observational studies, consider several further points:

1. A wide variety of so-called growth models are available to characterize longitudinal profiles; much recent work in this field has aimed at developing generalizations that extend the reach of models;
2. Some authors have focused on *smoothing profiles* -- in two distinctive ways: **a**. *smoothing individual profiles* by taking advantage of dependencies among adjacent or closely related observations in profiles, and **b**. smoothing by capitalizing on similarities among profiles for individuals; such smoothing entails 'borrowing strength' from mutually related profiles. Double smoothing may also be employed.
3. Those who model as in 1., are often advised to smooth initially when individual observations are subject to 'considerable noise'.
4. Model-based predictions (or fitted versions) of initial profiles, or at least their smoothed counterparts, are likely to be better targets for (PSA) studies than would be initial (raw-data) profiles.
5. As in the case of simpler forms of PSA, sensitivity analyses will generally be advisable. (Rosenbaum, in his two books, considers this topic closely.)
6. A great deal of work on PSA remains to be done for longitudinal problems and there are many opportunities for analysis in this area.

These four panels illustrate possibilities. They correspond to subsets of profiles, each for five animals, as clustered (see below). In particular, three principal components were derived from initially smoothed profiles that were in turn used to get *doubly smoothed profiles* computed as linear combinations of the PCs. Each profile can be fully described using 4 coefficients: intercept, & three PC regression coefficients. Clusters were based on these coefficients (in R).



Given appropriate covariate data for each animal, these might be used for observational study comparison of animals whose diets differed from one another; i.e. using constructed P-scores. (All initial responses were measures of *protein in milk over five weeks.* These data are part of the Milk dataset in the **nlme** package; they are real.)

As seen here, smoothing can work especially well for some data. Exploratory approaches (based on underlying components or latent variables) may often permit creation of smoothed versions of either original or pre- smoothed profiles. LDA versions of PSA may readily follow.

Although time may not permit discussion, it may be useful to make some further observations pertaining to mainstream IALSA interests. Consider an example of an *observational comparison of two groups (as suggested by S. Hofer).*

"Does engagement in intellectually challenging tasks, exercise, [or] social networks help to maintain cognitive functioning in later life?" There have been a number of analyses of longitudinal observational studies and experimental studies of this topic. There is some evidence to suggest that physical activity enhances cognitive performance. Suppose we revisit such a question using a modern PSA approach.

Let us limit attention to one measure of cognitive functioning and a clearly defined treatment (that could be a combination, but might be limited to one behavior, say engagement in exercise). Given a clear distinction between two groups, one of which will **not** have exercised (self-report?), and one of which will (at some defined level of rigor and regularity), we might aim to adjust for (all) relevant covariate differences. This is the hardest part, one that might ideally be done using a *prospective approach*, where one could have the luxury of naming in advance all covariates that seem likely to confound interpretations of T vs. C effects. Archival data might also be used with the proviso that such data rarely contain all variables that ultimately matter (think of the critics). If cognitive functioning scores are available for individuals *before and after* commencement of exercise, initial covariate scores might be used in the construction of P-scores.

It is almost inevitable in practice that some covariate and longitudinal data will have gone missing. This is one key reason imputation methods have garnered so much interest. (But the underlying theory that supports PSA methods, á la Rubin, is strongly based on counterfactual logic where one of two potential outcomes will always be missing). A special advantage for some longitudinal data sets is that missing LD values can be more reliably estimated than counterparts outside the longitudinal framework. This means that multiple LD imputations, and many products of analysis, may vary less than is typical.

An additional concern is that responses (such as cognitive function scores) are likely to have been obtained at different times for different individuals; i.e., different spacings and different numbers of times as well. Smoothing can often help with such problems, in which case the ultimate data used in the PSA can begin from statistics that characterize smoothed profiles, not original data. (This step may also help ameliorate problems induced by measurement errors.) Naturally, imputations can be done in different ways, perhaps using different imputation models; and the same goes for smoothing. When multiple analyses of the same data are employed one will want to learn how much results vary across methods. Further value may come from use of mixed models in analysis.

Ultimately, longitudinal PS analysis of such data may lead to fairly strong conclusions about “treatment effects”, conditional on the extent to which covariates are *strongly ignorable,* and the extent to which results do not depend heavily on the particular methods used for analysis. Design is likely to be central.