

# Bias Reduction in Observational Studies via Prognosis Scores

Ben B. Hansen

Technical Report # 441

Statistics Department, University of Michigan

439 West Hall, Ann Arbor, Michigan 48109, U.S.A.

ben.b.hansen@umich.edu

June 2006

## **Abstract**

The propensity score collapses the covariates of an observational study into a single measure summarizing their joint association with treatment conditions. The prognosis score — a complement or alternative to the propensity score — summarizes covariates' association with potential responses. As with propensity scores, stratification on the prognosis score brings to uncontrolled studies a concrete and desirable form of balance, a balance that is more familiar as an objective of experimental control. In parallel with the propensity score, prognosis scores reduce the dimension of the covariate; yet causal inferences conditional on them are as valid as are inferences conditional only on the unreduced covariate. They suggest themselves in certain studies for which propensity score adjustment is infeasible. Other settings call for a combination of prognosis and propensity scores; as compared to propensity scores alone, the pairing reduces both the variance and bias of estimated treatment effects. An example and a simulation study confirm that the combination of the two better reduces bias, and is much more efficient, than propensity scoring alone.

Key words: covariate balance, matched sampling, matching, multivariate distance, propensity score, quasi-experiment, regression discontinuity, stratification, subclassification

# 1 Introduction

Following Neyman (1990) and others (Cox 1958; Rubin 1977), let us construe an intervention's effect in terms of potential outcomes, as the differences between subjects' potential responses to treatment,  $y_t$ , and control,  $y_c$ . The basic challenge for measurement of effects is that at most one of these two outcomes is observed, according as the subject did ( $z = 1$ ) or did not ( $z = 0$ ) in actuality receive the treatment. This difficulty is most conclusively surmounted when the potential outcomes arise through stable, repeatable processes with known chance properties, as in some physical science laboratories with controlled experimental conditions, and when treatment conditions are decided by a stable, repeatable process with known chance properties, as in social and medical studies with carefully controlled, random assignment to treatment (Holland 1986).

Rosenbaum and Rubin (1983; 1984; 1985) showed that modeling the observed pattern of treatment ( $z$ ) as a function of covariates ( $\mathbf{x} = (x'_1, \dots, x'_k)'$ ) collapses those covariates into a scalar, the propensity score, upon which it is beneficial to condition. The demonstration requires no ostensible assignment mechanism to exist outside the statistical model, and suggests that even when treatment assignment models are misspecified, propensity-score stratification is likely to reduce bias, if not eliminate it. Now if casual models of treatment assignment favorably reduce the dimension of  $\mathbf{x}$ , then dimension reductions of  $\mathbf{x}$  that arise from modeling  $p(y_c|\mathbf{x})$ , even noncommittally, should be favorable as well, either as analogues or as complements to propensity scores.

## 1.1 Assumptions; outline.

To postpone commitment to specifications for  $p(y_t, y_c|\mathbf{x})$ , one fits a model of  $p(y_c|\mathbf{x})$  to the control group, extrapolating this fit to the remainder of the sample. The paper offers a diagnostic for this fit, along with arguments, intuitive and theoretical, that with a favorable diagnosis conditioning on the reduced covariate should reduce or eliminate bias (§ 2). These arguments suppose that the joint density of potential responses  $(y_t, y_c)$ , treatment assignment  $z$ , and covariates  $\mathbf{x}$  factorizes as:

$$p(y_t|\mu_c, m(\mathbf{x}))p(y_c|\mu_c)p(\mu_c|\mathbf{x})p(z|\mathbf{x})p(\mathbf{x}), \tag{1}$$

where  $m$  is some measurable vector- or scalar-valued function, possibly degenerate; or more simply as

$$p(y_t|y_c, m(\mathbf{x}))p(y_c|\mathbf{x})p(z|\mathbf{x})p(\mathbf{x}). \quad (2)$$

Section 3 presents an example and applications. Observe that implicitly, (1) and (2) posit the absence of hidden bias,

$$(Y_t, Y_c) \perp Z | \mathbf{X}. \quad (3)$$

Section 4, which analyzes propensity and prognosis scoring in combination, assumes (3) but not (1) or (2). Section 5 addresses potential reservations about the paper’s proposals, while Section 6 mounts a simulation study — the results of which are quite promising.

## 1.2 The prognosis score

If  $\Psi(\mathbf{X})$  is sufficient for  $Y_c$ , then  $\Psi(\mathbf{X})$  is a *prognosis score*. Should  $Y_c$  follow a shift model,  $p(y_c|\mathbf{x}) \equiv f_c(y_c - h(\mathbf{x}))$  for some fixed  $f_c$ , then  $h(\mathbf{X})$  is a prognosis score. Should  $Y_c|\mathbf{X}$  follow a generalized linear model (McCullagh and Nelder 1989), then the linear predictor of  $Y_c$  given  $\mathbf{X}$  is a prognosis score, as is the scalar  $\mathbf{E}(Y_c|\mathbf{X})$ . Conditioning on this score may suffice for estimating the distribution of  $Y_t - Y_c$  if  $Y_t|\mathbf{X}$  follows a GLM belonging to the same family. Should  $Y_c$  have a linear regression on  $\mathbf{X}$ , but with non-constant variance that is linear in  $\mathbf{X}$ , then the regression and variance functions taken together constitute a prognosis score.

These regression-type models conform to (1). Typical models invoked in permutation inference follow the simpler form (2). For instance, models with additive or multiplicative effects,  $y_t \equiv y_c + \tau$  or  $y_t \equiv \gamma y_c$ , conform to (2). Interpreting (2) so as to permit the likelihood contributions  $p(y_t|y_c)$  to vary between (independent) observations accommodates randomization inference for attributable effects (Rosenbaum 2001). In all cases, one  $p(y_c|\mathbf{x})$  is assumed to govern all observations, and prognosis scores are its sufficient statistics, or estimates of them.

Now (1) and (2) assign asymmetric roles to treatment and control-condition potential responses, giving the distribution of  $y_t$  in terms of  $y_c$  (or  $\mu_c$ ). One could assume the opposite. In this case, extract prognosis scores from the fitting of submodels for  $p(y_t|\mathbf{x})$ , rather than  $p(y_c|\mathbf{x})$ , and interpret this paper reversing the roles of  $t$  and  $c$ .

### 1.3 Precedents

Peters (1941), Belson (1956), Cochran (1969) and Rubin (1984) suggest extracting  $\hat{\mathbf{E}}(Y_{ci}|\mathbf{X} = \mathbf{x}_i)$  from a parametric model fit to the control group, then estimating the treatment effect as the mean of  $(y_i - \hat{\mathbf{E}}(Y_{ci}|\mathbf{X} = \mathbf{x}_i)|z_i = 1)$ . Miettinen (1976), and more recently Zhao (2004), propose post-stratification on estimates of  $\mathbf{E}(Y_c|\mathbf{X} = \mathbf{x})$ . The Peters-Belson method differs from prognosis scoring in using its  $\hat{\mathbf{E}}(Y_c|\mathbf{X})$  to estimate the response surface directly, rather than for subclassification; Miettinen’s and Zhao’s methods use treatment and control subjects together to estimate  $\mathbf{E}(Y_c|\mathbf{X} = \mathbf{x})$ , whereas prognosis scoring uses only one of the groups.

These differences matter. Sections 4 and 6 advantageously combine prognosis and propensity score adjustment, in ways that are possible only if prognosis scores are used for post-stratification, rather than direct estimation of a response surface. Using one but not both groups for fitting the score minimizes the analyst’s commitment to the model used for fitting it; Miettinen’s score skips this precaution, and has been found to risk introducing bias rather than eliminating it (Pike et al. 1979; Cook and Goldman 1989; reviewed in § 5.1, below).

## 2 Conditioning on Prognosis Scores: Theory

### 2.1 Prognostic balance

Conditioning on the propensity score,  $\phi(\mathbf{x}) \equiv \mathbf{E}(Z|\mathbf{X} = \mathbf{x})$ , secures a form of covariate balance:

$$\mathbf{X} \perp Z | \phi(\mathbf{X})$$

(Rosenbaum and Rubin 1983, Thm. 1). Within level sets of the propensity score no covariate associates with membership in the treatment or the control group. A quintessential benefit of experimental randomization is its tendency to impose this absence of association, here called “propensity balance.” An observational study exhibiting propensity balance on scientifically important covariates is in an important sense experiment-like.

In a second experimental model to which an observational study could be compared, it is the process by which outcomes are generated that is repeatable, understood, and carefully controlled, not the process of assigning units to treatment. This model can be understood as using experimental control in the interest of removing associations between covariates and potential outcomes, not treatment assignment. If in advance of studying a new experimental manipulation, an investigator conducts

tests absent the new manipulation in order to better understand accompanying conditions and their influence on the outcome, then it is this second model that informs her procedure. Such preparations may fall short of controlling all factors other than the new manipulation that might influence the outcome of the trial, but they are a success nonetheless if they enable the investigator to bring about conditions within which *uncontrolled variation does not associate systematically with trial outcomes*. Should she subsequently succeed in documenting a systematic association between experimental manipulations and trial outcomes, her preparations will have reduced the possibility that it was the uncontrolled enviroing conditions, not the treatment, that produced her result.

This form of balance — similarity among the covariate distributions of trials or subjects with contrasting potential outcomes — is quite distinct from propensity balance. Call it *prognostic balance*.

**Proposition 1** *Let  $Y_c$  be potential response to control and let  $\Psi(\mathbf{X})$  be an associated prognosis score. Then conditioning on  $\Psi(\mathbf{X})$  induces prognostic balance —*

$$\mathbf{X} \perp Y_c | \Psi_c(\mathbf{X}), g(\mathbf{X}) \tag{4}$$

— where  $g(\cdot)$  may be any measurable function, possibly degenerate.

Under factorization (2) with  $m(\mathbf{X})$  as the possible effect modifier, in addition to (4) one has  $\mathbf{X} \perp Y_t | \Psi_c(\mathbf{X}), m(\mathbf{X}), g(\mathbf{X})$  for all measurable  $g(\cdot)$ .

**Proof of Proposition 1.** Write [*condition*] for the function that is 1 if *condition* is true and 0 otherwise, and let  $A_x, A_y$  represent measurable sets. Then

$$\begin{aligned} \mathbf{E}([\mathbf{X} \in A_x][Y_c \in A_y] | \Psi(\mathbf{X}), g(\mathbf{X})) &= \mathbf{E}(\mathbf{E}([\mathbf{X} \in A_x][Y_c \in A_y] | \mathbf{X}) | \Psi(\mathbf{X}), g(\mathbf{X})) \\ &= \mathbf{E}([\mathbf{X} \in A_x] \mathbf{E}([Y_c \in A_y] | \mathbf{X}) | \Psi(\mathbf{X}), g(\mathbf{X})) \\ &= \mathbf{E}([\mathbf{X} \in A_x] f(\Psi(\mathbf{X})) | \Psi(\mathbf{X}), g(\mathbf{X})) \\ &= \mathbf{pr}(\mathbf{X} \in A_x | \Psi(\mathbf{X}), g(\mathbf{X})) f(\Psi(\mathbf{X})), \end{aligned}$$

where  $f$  is a measurable function such that  $\mathbf{pr}(Y_c \in A_y | \mathbf{X}) = f(\Psi(\mathbf{X}))$ . This establishes the conditional independence, given  $\Psi(\mathbf{X}), g(\mathbf{X})$ , of the events  $\{\mathbf{X} \in A_x\}$  and

$\{Y_c \in A_y\}$  and, because  $A_x$  and  $A_y$  were arbitrary, of the random variables  $\mathbf{X}$  and  $Y_c$ .  
 $\square$

Just as the propensity is the “coarsest” function of covariates that balances covariates between treatment and control groups, prognosis scores can be recovered by coarsening from any function of covariates that has the prognostic balancing property.

**Proposition 2** *Suppose that  $b(\mathbf{X})$  induces prognostic balance, in the sense that*

$$\mathbf{X} \perp Y_c | b(\mathbf{X}), \quad (5)$$

and let  $\Psi(\mathbf{X})$  be a prognosis score. Then with probability one,  $\mathbf{E}(\Psi(\mathbf{X})|b(\mathbf{X})) = \Psi(\mathbf{X})$ .

**Proof of Proposition 2.** For bounded, measurable, scalar-valued  $g$ ,

$$\begin{aligned} \text{cov}(\mathbf{E}\{g(Y_c)|\mathbf{X}\}, g(Y_c)|b(\mathbf{X})) = \\ \text{cov}(\mathbf{E}\{g(Y_c)|\mathbf{X}\}, g(Y_c) - \mathbf{E}\{g(Y_c)|\mathbf{X}\}|b(\mathbf{X})) + \text{var}(\mathbf{E}\{g(Y_c)|\mathbf{X}\}|b(\mathbf{X})). \end{aligned} \quad (6)$$

The first term at right in (6) is, for  $b(\mathbf{X}) = b(\mathbf{x})$ , an inner product of a perpendicular projection with the residual from the same projection. Such a product must vanish. By the assumption that  $b(\mathbf{X})$  has the prognostic balancing property (5), the covariance on the left side of (6) must be zero also. Consequently,  $\text{var}(\mathbf{E}\{g(Y_c)|\mathbf{X}\}|b(\mathbf{X})) = 0$ .

Now any scalar component  $f(\mathbf{X})$  of  $\Psi(\mathbf{X})$  can be written as  $\mathbf{E}(g(Y_c)|\mathbf{X})$ , for some  $g$ . Since the  $g$  of (6) was arbitrary, the theorem follows.  $\square$

**Corollary 1** *Suppose that  $b(\mathbf{X})$  induces prognostic balance (5). Then in the absence of hidden bias, (3), for arbitrary measurable  $g, g'$  one has*

$$Y_c \perp \mathbf{X} | b(\mathbf{X}), g(\mathbf{X}), g'(Z).$$

**Proof of Corollary 1.** By (3), for bounded measurable  $f$  and  $h$ ,  $\mathbf{E}(f(Y_c)h(\mathbf{X})|\mathbf{X}, Z) = \mathbf{E}(f(Y_c)h(\mathbf{X})|\mathbf{X})$ . Certainly  $\mathbf{E}(f(Y_c)h(\mathbf{X})|\mathbf{X}) = h(\mathbf{X})\mathbf{E}(f(Y_c)|\mathbf{X}) = h(\mathbf{X})\tilde{f}(\Psi(\mathbf{X}))$ , for some appropriate  $\tilde{f}$ . Thus  $\mathbf{E}(f(Y_c)h(\mathbf{X})|b(\mathbf{X}), g(\mathbf{X}), g'(Z)) = \mathbf{E}(\tilde{f}(\Psi(\mathbf{X}))h(\mathbf{X})|b(\mathbf{X}), g(\mathbf{X}), g'(Z))$ . Since  $\tilde{f}(\Psi(\mathbf{X}))$  is a function of  $b(\mathbf{X})$ , the result follows.  $\square$

## 2.2 Absence of confounding within prognostically balanced strata

Stratifications of the population in terms of covariate characteristics,  $\mathbf{x}$ , can be identified with many-one functions  $b$  on  $\mathbf{x}$ -space.

**Proposition 3** *Absent hidden bias (3), stratifications  $b$  that exhibit prognostic balance,*

$$Y_c \perp \mathbf{X} | b(\mathbf{X}), \quad (7)$$

*are also unconfounded, in the sense that*

$$Y_c \perp Z | b(\mathbf{X}).$$

In light of Proposition 1, it is a corollary of Proposition 3 that stratifications on prognosis scores  $\Psi(\mathbf{X})$  (and perhaps other axes of variability in  $\mathbf{X}$ ) render  $Y_c$  and  $Z$  unconfounded if conditional on the unreduced set of covariates  $\mathbf{X}$ ,  $Y_c$  and  $Z$  are not confounded.

**Proof of Proposition 3.** Assume (3) and (7). For arbitrary  $A_y$  and  $A_z$ ,

\*\*\*

$$\begin{aligned} \mathbf{pr}(Y_c \in A_y, Z \in A_z | b(\mathbf{X})) &= \mathbf{E}(\mathbf{E}([Y_c \in A_y] \cdot [Z \in A_z] | \mathbf{X}) | b(\mathbf{X})) \\ &= \mathbf{E}(\mathbf{pr}(Y_c \in A_y | \mathbf{X}, b(\mathbf{X})) \mathbf{pr}(Z \in A_z | \mathbf{X}) | b(\mathbf{X})) \\ &= \mathbf{E}(\mathbf{pr}(Y_c \in A_y | b(\mathbf{X})) \mathbf{pr}(Z \in A_z | \mathbf{X}) | b(\mathbf{X})) \\ &= \mathbf{pr}(Y_c \in A_y | b(\mathbf{X})) \mathbf{E}(\mathbf{pr}(Z \in A_z | \mathbf{X}) | b(\mathbf{X})) \\ &= \mathbf{pr}(Y_c \in A_y | b(\mathbf{X})) \mathbf{pr}(Z \in A_z | b(\mathbf{X})). \square \end{aligned}$$

## 2.3 Direct adjustment with prognosis scores

Under certain assumptions about the structure of the underlying model, in the absence of hidden bias treatment effects can be estimated without bias after conditioning on a prognosis score.

**Proposition 4** *Suppose that  $0 < \mathbf{pr}(Z = 1 | \Psi(\mathbf{X})) < 1$  with probability one and that conditional on  $\mathbf{X}$ ,  $(Y_t, Y_c)$  and  $Z$  are not confounded (3). If  $p(y_t | y_c, \mathbf{x}) = p(y_t | y_c)$ , so that the likelihood factorizes as (2) with degenerate effect modifier  $m(\cdot)$ , then*

$$\mathbf{E}(Y_t - Y_c) = \mathbf{E}\{\mathbf{E}(Y | Z = 1, \Psi(\mathbf{X})) - \mathbf{E}(Y | Z = 0, \Psi(\mathbf{X}))\}. \quad (8)$$

If  $p(y_t|y_c, \mathbf{x}) = p(y_t|y_c, m(\mathbf{x}))$ , so that the likelihood factorizes as (2), then

$$\mathbf{E}(Y_t - Y_c) = \mathbf{E}\{\mathbf{E}(Y|Z = 1, \Psi(\mathbf{X}), m(\mathbf{X})) - \mathbf{E}(Y|Z = 0, \Psi(\mathbf{X}), m(\mathbf{X}))\}. \quad (9)$$

A similar principle governs, for example, the difference of conditional logits when that difference is presumed constant,  $\text{logit}(\mathbf{E}(Y_t|\Psi(\mathbf{X}))) - \text{logit}(\mathbf{E}(Y_c|\Psi(\mathbf{X}))) \equiv c$ , and (1) is assumed.

To prove Proposition 4, note first that  $Y_c \perp Z|\Psi(\mathbf{X})$  follows from (3) and Proposition 3. Since  $p(y_t|y_c, \mathbf{x}) = p(y_t|y_c)$  entails  $Y_t \perp Z|(Y_c, \Psi(\mathbf{X}))$ , also  $Y_t \perp Z|\Psi(\mathbf{X})$ . Since  $Y = Y_t$  when  $Z = 1$  and  $Y = Y_c$  when  $Z = 0$ , (8) follows. Similarly  $p(y_t|y_c, \mathbf{x}) = p(y_t|y_c, m(\mathbf{x}))$  entails  $Y_t \perp Z|(Y_c, \Psi(\mathbf{X}), m(\mathbf{x}))$ , and invoking (3) and Proposition 3 leads to (9).

## 2.4 A principle to support conditioning on prognosis scores

The conditionality principle endorses inference conditional on a statistic if, roughly, it is uninformative about the parameter being estimated. Population prognosis scores  $\Phi(\mathbf{X})$  are informative about the distribution of  $Y_c$ , but not about the distribution of the contrast  $Y_t - Y_c$ ; their use is in accord with the conditionality principle. Under what conditions are estimated prognosis scores uninformative in the requisite sense?

Pace and Salvani (1997, §4.2) articulate a conditionality principle for models with both an interest parameter ( $\tau$ ) and a nuisance parameter  $\zeta$ . Then  $u$  is a *partial distribution constant* if  $p(u|\tau, \zeta) = p(u|\zeta)$ ; and the principle recommends conditioning on such statistics. To relate the definition to prognosis scores, let  $\zeta$  parameterize  $p(y_c, x_1, \dots, x_k)$ , and let  $\tau$  parameterize the conditional distribution of  $Y_t|Y_c, X_1, \dots, X_k$ . If  $\Psi(\mathbf{X})$  is a prognosis score, then  $p(\Psi(\mathbf{X})|\tau, \zeta) = p(\Psi(\mathbf{X})|\zeta)$ , as required for  $\Psi(\mathbf{X})$  to be a distribution constant. The same holds for sample prognosis scorings,  $\hat{\Psi}(\mathbf{X})$ , that have been fitted to the control group. However, taken together  $\mathbf{E}(Y_c|X_1, \dots, X_k)$  and  $\mathbf{E}(Y_t|X_1, \dots, X_k)$  are *not* generally distribution constant in the sense required for the conditionality principle. Intuitively speaking, conditioning on both at once pre-judges the outcomes of treatment-control comparisons, something that conditioning on the one or the other need not do; § 5.1 discusses how this can bias estimated treatment effects.

### 3 Applications

Since prognosis scores are to be estimated using the control group only, the approach will best reduce bias in studies with far larger control than treatment groups. Another promising application is to studies in which treatment and control groups have little overlap on the propensity score.

Alone or in combination with other adjustments, the prognosis score can be expected to confer benefits other than bias reduction. Since prognosis score adjustment reduces within-stratum heterogeneity in the ways that matter most for responses, it can be expected to improve the precision of estimates. By the same token, it should reduce the susceptibility of the study to bias caused by omitted variables; see also Rosenbaum (2005) in this regard.

Only controls can be used to test for prognostic balance, but otherwise checking for prognostic balance with an estimated prognosis score is similar to checking for propensity balance after conditioning on a propensity score. If the prognosis score is a scalar, then partitioning the control group into subclasses along an estimated prognosis score should induce prognostic balance, more or less.

#### 3.1 An example to illustrate prognostic balance

Powers and Rock (1999) selected a representative sample of takers of the SAT, the U.S. college entrance exam, to assess benefits of commercial test-coaching programs. The College Board, which produces the test, had long maintained that benefits of coaching are modest, averaging perhaps 20 to 40 of 1600 possible points, while prominent coaching companies claimed average benefits exceeding 100 points in combined score. (Powers and Rock's study was supported by The College Board.) Hansen (2004) analyzed this sample using propensity scores, arranging some 4000 of 4200 respondents to a coaching questionnaire into 500 propensity-matched sets. Since controls are relatively plentiful (3500), the data should be amenable to prognosis scoring. I build directly on the analysis of Hansen (2004), using the same data cleaning and preparations described there in order to prepare two prognosis scores, one each for the math and verbal SAT outcome variables: linear regressions of math and verbal posttest scores on covariates, fitted to the control group and then applied to the combined sample of control and treatment group subjects. To ascertain the prognostic balancing capabilities of this technique, test versions of the prognosis scores are fitted to a subgroup of the controls, the "pseudo-controls," so that balance can

be checked using these and the remaining controls that were not used in the fitting of the prognosis model. The division of the control group into pseudo-control and pseudo-treatment groups attempted to recreate patterns of difference between actual control and treatment groups, and is detailed in Appendix A.1.

The scores' balancing properties are most readily discerned by treating them one at a time. As Table 1 shows, subclassifying controls into bins of equal widths greatly improves prognostic balance on most covariates. For each category of these nominal variables, the table gives estimated increases in the proportion of subjects in that category per 100 point increase in SAT-M score, with and without conditioning on prognosis subclass; it is evident both that  $y_c$  associates strongly with each of these covariates, and that the prognosis score adjustment largely eliminates the association. The table also reports outcomes of permutation tests of no association, as detailed in Appendix A.2. The bin width of .2 s.d.'s was the largest multiple of  $sd/10$  at which the hypothesis of prognostic balance conditional on subclass membership was sustained. The SAT coaching example is discussed further in § 6.

### 3.2 A use for concordant strata in matched case-control studies

Retrospective studies sampling controls,  $y = 0$ , that have been matched to cases,  $y = 1$ , on the basis of covariates often match on fewer covariates than are available at the time of analysis. Suppose that  $\mathbf{x}$  represents the full set of covariates, but matching attended only to making  $\tilde{\mathbf{x}} \equiv b(\mathbf{x})$  homogeneous within matched sets. Considered as *a priori* assumptions, (3) is at least as plausible as

\*\*\*

$$(Y_t, Y_c) \perp Z | \tilde{\mathbf{X}} \tag{10}$$

and may be a good deal more so. However, when strata are uniform in  $\tilde{\mathbf{x}}$  but not  $\mathbf{x}$ , (3) alone does not suffice for the hypotheses tested by such common procedures as McNemar's test, or the Cochran-Mantel-Haenszel test (McNemar 1947; Cochran 1954; Mantel and Haenszel 1959), to merit interpretation as tests of the causal effect of treatment. The causal interpretation of these tests requires (10).

According to Propositions 3 and 2, (10) follows from (3) if conditioning on  $b(\mathbf{x})$  induces prognostic balance. So one might justify the analysis from assumptions that (3) holds and that  $b(\mathbf{X})$  imposes prognostic balance, rather than with (10) as a base assumption.

The advantage of this is that the premise that  $b(\mathbf{X})$  is a balancing score is subject

Covariate (x)	range of values	percent of sample	Slope of $x/s_x$ 's regression... on $y_c/100$ , net of prognosis subclass			
			on $y_c/100$			
	20–43	18	−.40	***	.02	
Math	44–51	17	−.04	*	−.04	
Section of	52–57	17	.22	***	−.04	.
PSAT	58–80	15	.51	***	.05	***
	not taken	33	−.20	***	.00	
Mean SAT at respondent's first choice college	787–987	18	−.24	***	−.00	
	988–1060	17	−.05	**	.00	
	1061–1123	16	.08	***	−.02	
	1124–1336	15	.28	***	.01	
	no response	36	−.04	**	.00	
	high school	43	−.21	***	−.00	
Father's education	A.A. or B.A.	26	.07	***	.01	
	graduate	23	.21	***	.02	
	no response	8	−.06	***	−.03	

Table 1: Prognostic balance on selected covariates within the control group, before and after subclassifying on a prognosis score for SAT-M. The score was fit to a “pseudo-control” group within the control group and extrapolated to remaining controls. Significance stars give results of permutation tests of no association between  $y_c$  and the covariate; these and tests for the remaining 24 covariates combine to  $\chi^2 = 2300$  on 69 d.f. ( $p = 0$ ) prior to subclassification, or to  $\chi^2 = 73$  on 69 d.f. after the subclassification ( $p = .35$ ). Bin widths for the subclassification is .2 sd's of the prognosis score.

to empirical test. If stratification on  $b(\mathbf{X})$  balances covariates  $\mathbf{X}$ , and if  $Y_c \perp Z|\mathbf{X}$ , then Corollary 1 says that for  $z = 0$  or  $1$ ,  $Y_c \perp \mathbf{X}|b(\mathbf{X}), Z = z$ . In this case, then within pooled, independent samples from  $\{u : Y_c(u) = 1, b(\mathbf{X}(u)) = b, Z = 0\}$  and  $\{u : Y_c(u) = 0, b(\mathbf{X}(u)) = b, Z = 0\}$ , there should be no association between  $Y_c$  and covariates  $\mathbf{X}$ . That is, among unexposed units that have been matched on  $b(\mathbf{X})$ , no covariate or function of covariates should be predictive of outcomes. If a covariate or function of covariates is found to associate, conditionally on  $b(\mathbf{X})$ , with  $Y_c$ , then conditioning on  $b(\mathbf{X})$  is probably not sufficient for causal inference. Conversely, the absence of such confounding suggests that in conditioning on  $b(\mathbf{X})$  one has implicitly conditioned on a prognosis score as well.

### 3.3 Regression discontinuity designs

Like randomized experiments and nonexperimental cohort designs, the regression discontinuity design (RDD) (Campbell and Stanley 1966) compares treatment and control groups in order to estimate treatment effects. In a classic RDD, a pretest is administered before treatment, perhaps a remedial education program, is given to all and only the subjects whose pretests scores,  $X_1$ , fall below a specified cutoff point,  $c$ ; eventually treatment and control groups are compared in terms of their posttest scores,  $Y$ . The design yields information about the limits of  $\mathbf{E}(Y_c|X_1 = x)$  as  $x \downarrow c$  and of  $\mathbf{E}(Y_t|X_1 = x)$  as  $x \uparrow c$ . Most analyses estimate and compare these limits (Visser and De Leeuw 1984; Berk and de Leeuw 1999), resulting in inferences that are sensitive to the specifications of these surfaces (Rubin 1977, pp.1,11; Little and Rubin 2000, p.132).

Prognosis score adjustment suggests an alternative estimation strategy that may remove this sensitivity. Consider an analysis that subclassifies on  $\hat{\Psi}(\mathbf{X})$ , a prognosis-balancing reduction of  $\mathbf{X}$ . The subclassification may be so constructed that at least one subclass contains subjects with  $x_1$ -values both above and below  $c$ , *i.e.* treatment and control subjects; then prognosis-adjusted comparisons are possible within this subclass. Such a comparison has the simpler form  $\hat{\mathbf{E}}(Y_t|A(\mathbf{X})) - \hat{\mathbf{E}}(Y_c|A(\mathbf{X}))$  — it requires no comparison of limits.

If in addition the analyst possess good prognostic variables in addition to  $X_1$ , then the range of values of  $\hat{\Psi}(\mathbf{X})$  at which both treatment and control samples are well represented will tend to be broader. That is, adjustment by prognosis scores confers a reward for richer covariate information: the smaller  $X_1$ 's role in determining  $\mathbf{E}(Y_c|\mathbf{X})$ , the more likely the spread of  $\hat{\Psi}(\mathbf{X})$  within the treatment group will overlap with

that of the controls. With propensity adjustment, in contrast, including additional covariates invites worsening of the separation between groups.

## 4 Bias as a function of propensity and prognosis scores

This section develops a perspective on the bias of effect estimates and test statistics that are commonly combined with matched or subclassified analysis. It assumes that the stratification is made on the basis of covariates, but not that it explicitly separate subjects with differing prognosis scores. The sample may have been stratified along some other function of covariates, including the propensity, a combination of propensity and prognosis scores, or neither of those.

Write  $Y_{tsi}, Y_{csi}, Z_{si}$  and  $\mathbf{x}_{si}$  for the potential responses, treatment assignment and covariates of the  $i$ th subject in stratum  $s$ , and  $Y_{ts} = (Y_{ts1}, \dots, Y_{tsn_s})', Y_{cs} = (Y_{cs1}, \dots, Y_{csn_s})', etc..$  Here and in subsequent sections, let  $Y_t = (Y'_{t1}, \dots, Y'_{tS})', y_t = (y'_{t1}, \dots, y'_{tS})', etc..$

### 4.1 Direct adjustment

The direct adjustment estimate,  $\hat{\tau}_d(y, z)$ , is

$$\sum_s \frac{n_s}{n} \left\{ \frac{y'_s z_s}{1'_s z_s} - \frac{y'_s (1_s - z_s)}{1'_s (1_s - z_s)} \right\} = \sum_s \frac{n_s}{n} \left\{ \frac{y'_{ts} z_s}{1'_s z_s} - \frac{y'_{cs} (1_s - z_s)}{1'_s (1_s - z_s)} \right\}, \quad (11)$$

where  $1_s$  is an  $n_s$ -vector of 1's.

For each  $s$  and  $i$ , write  $\mu_{tsi}, \mu_{csi}$  for  $\mathbf{E}(Y_{tsi} | \mathbf{X}_{tsi} = \mathbf{x}_{tsi})$  and  $\mathbf{E}(Y_{csi} | \mathbf{X}_{tsi} = \mathbf{x}_{tsi})$ , respectively, and  $\pi_{si} = \mathbf{E}(Z_{si} | \mathbf{X}_{si} = \mathbf{x}_{si}, Z'_s 1_s = z'_s 1_s)$ ;  $\mu_{[t/c]s} = (\mu_{[t/c]s1}, \dots, \mu_{[t/c]sn_s})'$ ,  $\pi_s = (\pi_{s1}, \dots, \pi_{sn_s})'$ . Then for any stratum  $s$ ,

$$\mathbf{E} \left\{ \frac{y'_{ts} z_s}{1'_s z_s} - \frac{y'_{cs} (1_s - z_s)}{1'_s (1_s - z_s)} \middle| \mathbf{X}_s; 1'_s Z_s = 1'_s z_s \right\} = \bar{\mu}_{ts} - \bar{\mu}_{cs} + \frac{\mu'_{ts} (\pi_s - \hat{p}_s 1_s)}{1'_s z_s} + \frac{\mu'_{cs} (\pi_s - \hat{p}_s 1_s)}{1'_s (1_s - z_s)},$$

$\hat{p}_s := 1'_s z_s / n_s$ . Conditionally, the bias of the direct adjustment estimate (11) is

$$\sum_s \frac{\text{COV}(\mu_{ts}, \pi_s)}{\hat{p}_s} + \frac{\text{COV}(\mu_{cs}, \pi_s)}{1 - \hat{p}_s}. \quad (12)$$

Let  $\theta_{si} = \text{logit}(\pi_{si})$ , and suppose the sample has been stratified on an estimate of  $\theta$  or  $\pi$ . Strata that are relatively homogeneous in estimated propensities are likely to be roughly homogenous in  $\pi$ . To first order,

$$\pi_{si} \approx \hat{p}_s + \frac{n_s}{n_s - 1} \hat{p}_s (1 - \hat{p}_s) (\theta_{si} - \bar{\theta}_s),$$

so that in terms of  $\theta_{si}$  (the logits of the true conditional probabilities of assignment to treatment) the bias of direct adjustment is roughly

$$\sum_s \frac{n_s}{n_s - 1} \{(1 - \hat{p}_s) \text{cov}(\mu_{ts\cdot}, \theta_{s\cdot}) + \hat{p}_s \text{cov}(\mu_{cs\cdot}, \theta_{s\cdot})\}. \quad (13)$$

This says, unsurprisingly, that the bias is small if there is little difference among propensity scores within strata. But it also shows that the bias will be small if (i) within strata there is little variation in prognosis scores, or if (ii) prognosis and propensity scores are unassociated within strata. Indeed, according to (13), whenever a coarse propensity stratification suffices to disassociate propensity and prognosis scores, direct adjustment after propensity score stratification will have little bias. This is so even if the propensity has been estimated with some error, and even if stratification leaves substantial variation among propensity scores within strata.

#### 4.2 Inference from rank tests

The basic idea of (12) extends to inferences for the effect of treatment other than estimation of it via direct adjustment, although the expressions required can be more complicated. For instance, consider the nonparametric inferences recommended by Rosenbaum (2002). Rather than pairing a point estimate with an estimate of its variance, these delimit the range in which a treatment effect might fall by posing and testing sharp null hypotheses specifying which subjects experienced a treatment effect, and specifying the magnitude of the effect they experienced. The test statistics are often simple linear rank statistics, of the form  $z'f(y)$ , and their reference distributions condition on stratum-wise proportionate membership in treatment group, *i.e.*  $(\hat{p}_s : s)$ , and stratum-wise order statistics of responses, doing so under the assumption that propensities to fall in the treatment group are uniform within strata. (Inferences using Fisher's test, the Mantel-Haenszel test and its extension, Wilcoxon's rank-sum test and the Hodges-Lehmann aligned rank test all fit this mold; see Rosenbaum (2002), chapters 2 and 5.)

Write  $\tilde{\mu}_{csi}$  for  $\mathbf{E}(f(Y_{csi}) | \mathbf{X}_{csi}; Y_{cs[1]}, \dots, Y_{cs[n_s]})$ , where  $y_{cs[1]}, \dots, y_{cs[n_s]}$  are the order statistics of  $y_{cs1}, \dots, y_{csn_s}$ . If this specification is in error, so that propensity scores vary within strata, then the (conditional) expectation of  $Z'Y$  under the null differs

from its assumed expectation by

$$\sum_s \text{cov}(\tilde{\mu}_{cs}, \pi_{s.}) \approx \sum_s \frac{n_s}{n_s - 1} \hat{p}_s (1 - \hat{p}_s) \text{cov}(\tilde{\mu}_{cs}, \theta_s).$$

Since  $\mu_c \mapsto \tilde{\mu}_c$  is typically a smooth, monotone transformation, the basic lessons of (12) and (13) apply, despite the change of inferential machinery. The less the within-stratum covariance between propensity and prognosis scores, the better the assumed model tracks the true null distributions of the test statistics.

### 4.3 Implications for multivariate distance matching

To reveal implications for matching, suppose that  $\text{cov}(\mu_{ts}, \theta_s)$  and  $\text{cov}(\mu_{cs}, \theta_s)$  are equal. (If for some  $\tau$ ,  $\mu_{tsi} = \mu_{csi} + \tau$ , this is automatically so; otherwise, it is a simplifying pretense.) Then the bias of direct adjustment is

$$\sum_s \hat{p}_s^{-1} (1 - \hat{p}_s)^{-1} \text{cov}(\mu_{cs}, \pi_{s.}). \quad (14)$$

Putting this in terms of logit-scale propensities  $\theta_{si}$ , a Taylor expansion about a vector  $\theta^*$  that is constant within matched sets gives a bias of, to first order,

$$\sum_s \frac{\sum_i \mu_{csi} (\theta_{si} - \theta_s^*)}{n_s - 1} = \sum_s \frac{n_s}{n_s - 1} \text{cov}(\mu_{cs}, \theta_s). \quad (15)$$

(The error of this approximation to (14) can be shown to be no more than

\*\*\*

$$4 \frac{n_s \min(\sum_i z_{si}, n_s - \sum_i z_{si})^3}{(\sum_i z_{si})(n_s - \sum_i z_{si})} \max_i (\theta_{si} - \theta_s^*)^2 \max_i (\mu_{csi} - \bar{\mu}_{cs}) \quad (16)$$

per stratum  $s$ . In matched analyses, including pair matching, matching with multiple controls, and full matching,  $\min(\sum_i z_{si}, n_s - \sum_i z_{si}) \equiv 1$ ; so in these cases the sum over  $s$  of (16) simplifies to  $4 \sum_s \frac{n_s}{n_s - 1} \max_i (\theta_{si} - \theta_s^*)^2 \max_i (\mu_{csi} - \bar{\mu}_{cs})$  .)

Comparing (14) and (15), it appears that bias is more nearly determined by the variation (within matched sets) in  $(\theta_{si})$  than in  $(\pi_{si})$ , since the matched-set scaling factors of (15),  $n_s/(n_s - 1)$ , will be more nearly uniform across strata than those of (14). This suggests penalizing discrepancies  $\hat{\theta}$ 's rather than  $\hat{\pi}$ 's. (15) also suggests that within-matched-set standard deviations of  $\theta_s$  or  $\mu_{cs}$  matter more for bias in treatment effects than do the radii of the sets  $\{\theta_{si} : i\}$  or  $\{\mu_{csi} : i\}$ . This suggests

matching on a Mahalanobis distance combining  $\hat{\theta}$  and  $\hat{\mu}_c$ , potentially in combination with logit-scale propensity calipers, restrictions that  $i$  and  $j$  be matched only if  $-c < \hat{\theta}_i - \hat{\theta}_j < c$ , some  $c > 0$ , or with prognosis score calipers (on the response scale).

## 5 Hazards of prognosis adjustment

It has been shown that approximate matching on *population* prognosis scores reduces bias; what about matching on sample scores? Insofar as sample prognosis scores approximate their population analogues, one expects similar benefits to accrue to matching on them — unless errors of estimation should introduce new biases that counter the benefits of near-matching on the population score. Such bias can't be ruled out categorically, but the methods of this paper embody effective precautions.

### 5.1 The problem with using both groups to fit prognosis scores

In particular problems, and relative to particular assumptions, there may be distribution-constant statistics similar to prognosis scores that are formed using both control and treatment group responses: for instance, consider the parametric model

$$\begin{aligned} Y_t &= \alpha_t + Y_c \\ Y_c &= \alpha_0 + \alpha_1 X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I); \\ \text{pr}(Z|X) &= \text{inverse.logit}(\beta_0 + \beta_1 X). \end{aligned}$$

Consider estimates  $\hat{\alpha}_1$  derived from the ordinary least squares regression  $y \sim 1 + z + x$ . These estimates are unchanged by adding a real constant  $c$  to each response  $y_i$  from a member of the treatment group,  $z_i = 1$ . It follows that the distribution of  $\hat{\alpha}_1$  is free of  $p(y_t|y_c)$  and its parameter  $\alpha_t$ :  $\hat{\alpha}_1$ , and thus  $\hat{\alpha}_1 X$ , are partial distribution constants in the relevant sense, despite both treatment and control groups' having contributed covariate and response data to their estimation. Zhao (2004) discusses matching on scores of this type.

However, if this model is only slightly misspecified, say if the regressions of  $Y_t$  and  $Y_c$  on  $X$  are not precisely parallel or if the error structure varies with  $X$ , then  $\alpha_t$  may well influence the sampling distribution of an  $\hat{\alpha}_1$  fit by ordinary least squares to both treatment and control groups. In such a case no conditionality principle warrants conditioning on  $\hat{\alpha}_1 X$ . The hazards are not merely theoretical; the critique of Miettinen's confounder score proposal by Pike et al. (1979) generated binary  $Y$ 's

so as to make the dependence of  $\mathbf{pr}(Y|Z, X)$  linear on the logit scale, then used ordinary linear regression to estimate  $X$ 's contribution to  $\mathbf{pr}(Y|Z, X)$ . In effect, this is model misspecification; the population linear regression of  $Y_t$  on  $X$ , given that  $Z = 1$ , need not be parallel to that of  $Y_c$  on  $X$  as restricted to  $Z = 0$ , and the error variance will depend on  $\mathbf{pr}(Y|Z, X)$ . Pike et al. (1979) find that conditioning on  $X$ 's estimated contribution to the regression of  $Y$  on  $Z$  and  $X$  then leads to inflated type I errors. They find the error is greatest when  $\mathbf{pr}(X|Z = 1)$  differs markedly from  $\mathbf{pr}(X|Z = 0)$ , a condition coinciding with sharpest departure from parallelism in  $Y_t$ 's and  $Y_c$ 's population linear regressions on  $X$ . Indeed, in a replication and extension of Pike et al. (1979)'s evaluation of confounder scores, Cook and Goldman (1989) report substantial inflation of Type I error only for unusually discrepant  $\mathbf{pr}(X|Z = 1)$  and  $\mathbf{pr}(X|Z = 0)$ .

To summarize, when the likelihood factorizes as (2), population prognosis scores are partial distribution constants (so far as estimation of  $p(y_t|y_c, \mathbf{x})$  is concerned). For this reason, the conditionality principle recommends conditioning on them. The same is true of estimates of prognosis scores that were fitted using the control group only, but not (except in special cases) of prognosis scores fitted using both treatment and control groups.

## 5.2 The threat due to overfitting

In personal communication with the author, P. Rosenbaum raises the following concern. With many covariates, estimates of  $\mathbf{E}(Y_c|\mathbf{X})$  made from control but not treatment subjects are likely to embody some overfitting, assigning explanatory roles to features of the covariate that associate with  $Y_c$  in the control group but not in the sample as a whole. Suppose for simplicity that  $Y_c$  is a scalar, and that  $\hat{\psi}(\mathbf{X}) := g(\hat{\mathbf{E}}_{\{Z=0\}}(Y_c|\mathbf{X}))$ , for some increasing function  $g : \mathfrak{R} \mapsto \mathfrak{R}$ . If  $l$  is a relatively low value of  $\hat{\psi}(\mathbf{X})$ , then control group subjects with  $\hat{\psi}(\mathbf{X}) = l$  will have a spuriously low average value of  $Y_c$ , lower than their treatment group counterparts with  $\hat{\psi}(\mathbf{X}) = l$ ; symmetrically, if  $u$  is relatively high, then control group subjects with  $\hat{\psi}(\mathbf{X}) = u$  will on average have higher  $Y_c$  values than treatment group subjects with  $\hat{\psi}(\mathbf{X}) = u$ . Even absent any treatment effect whatsoever, there would appear to the analyst to be pronounced effects, varying in sign and magnitude with  $\hat{\psi}(\mathbf{X})$ . Figure 1 depicts this possibility.

Now prognostic models are widely used in medicine, and modern methods of constructing them scrupulously minimize overfitting (Moons et al. 2004); the use

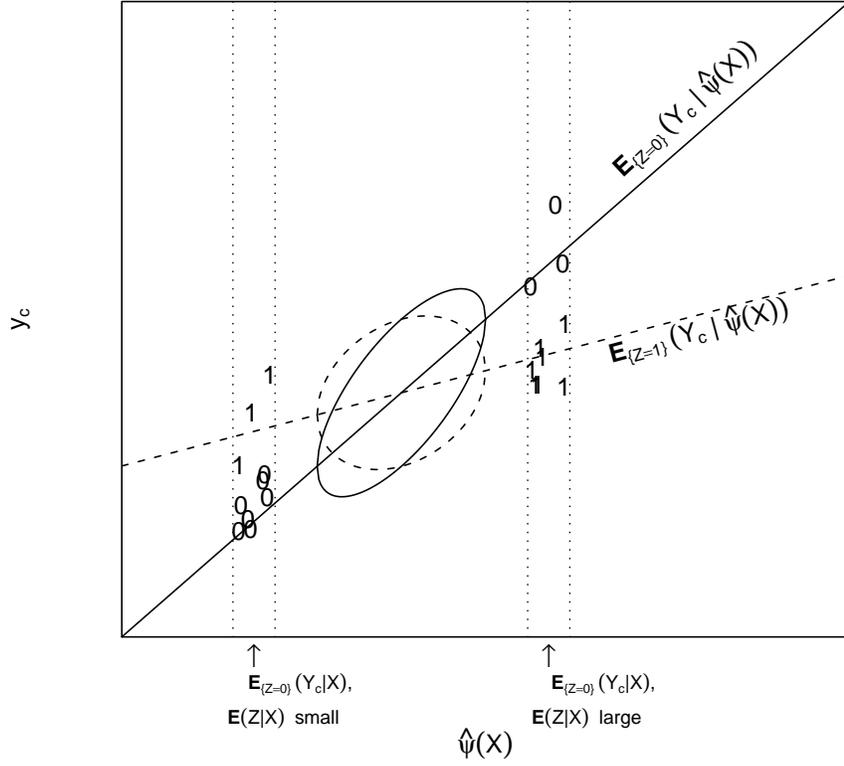


Figure 1: A potential hazard of prognosis score adjustment: errors of estimation of  $\psi(\mathbf{X}) = \mathbf{E}(Y_c|\mathbf{X})$  act differentially on treatment ( $Z = 1$ ) and control ( $Z = 0$ ) groups, because estimation uses only the control group, and naturally induces some overfitting; the result is that  $y_c$  values in the treatment and control groups do not overlap, particularly for relatively large or small values of  $\hat{\psi}(\mathbf{X})$ . Here, this overfitting spuriously suggests that the prognosis score is an effect modifier. In addition, estimates of  $\mathbf{E}(Y_t - Y_c|Z = 1)$ , the average effect of treatment on those treated, would be downwardly biased.

of such methods to fit prognosis scores might conceivably avoid the threat due to overfitting. With conventional techniques, however, such overfitting is sufficiently probable as to make it unwise to interpret variation in  $\hat{\mathbf{E}}(Y_t - Y_c | \hat{\psi}(\mathbf{X}))$  as evidence that treatment interacts with subjects' prior prognosis. This possibility is a limitation of the technique.

A distinct threat is that overfitting should bias estimates of treatment effect overall. The possibility of estimates of effects on the treatment group being downwardly biased is manifest in Figure 1. There treatment group subjects with high values of  $\hat{\mathbf{E}}_{\{Z=0\}}(Y_c | \mathbf{X})$  are inappropriately compared to controls whose  $Y_c$  values exceed theirs. At the other end of the figure,  $\hat{\mathbf{E}}_{\{Z=0\}}(Y_c | \mathbf{X})$  is relatively low and the difference of treatment and control outcomes is spuriously high; but there are fewer treatment group subjects at left than at right, so the errors don't cancel. Can prognosis scoring diagnostics identify and ward off scenarios like these?

Figure 1 exhibits deficits in both prognostic and propensity balance, deficits which, though subtle, are needed for the possibility of bias. In the figure,  $Z \not\perp Y_c | \hat{\psi}(\mathbf{X})$ . Assuming (3) that  $Z \perp Y_c | \mathbf{X}$ , there must be transformations  $f(\mathbf{X}), g(\mathbf{X})$  such that  $f(\mathbf{X}) \perp Y_c | \hat{\psi}(\mathbf{X})$  and  $Z \perp g(\mathbf{X}) | \hat{\psi}(\mathbf{X})$  — conditioning on  $\hat{\psi}(\mathbf{X})$  fails to prognostically balance  $f(\mathbf{X})$ , nor does it propensity balance  $g(\mathbf{X})$ . One ought to avoid the pitfall of Figure 1 if one could ensure either that most every transformation of the covariate is prognostically balanced, *or*, that most every transformation of it is propensity balanced.

## 6 Case study

The bias Figure 1 depicts requires that  $\hat{\mathbf{E}}_{\{Z=0\}}(Y_c | \mathbf{X})$  associate with  $\hat{\mathbf{E}}(Z | \mathbf{X})$ , or that prognosis in the absence of treatment correlates with propensity to receive the treatment. This is certainly true of the SAT coaching study; Powers and Rock (1999) found that students paying for test preparation were on the whole more ambitious, had better educated parents, and were stronger students than those who didn't. Their study shall provide the basis for a test of whether prognosis scoring meaningfully reduces bias — or whether errors of estimation generate biases overwhelming whatever biases are removed.

Propensity and prognosis scores were fitted using linear logistic and, as in § 3.1, ordinary linear regression. (I deliberately selected the simplest method of fitting prognosis scores, despite the fact that other methods might better control overfitting, in

order to more exactly test the method.) The propensity and the math and verbal prognosis scores were then combined into a Mahalanobis distance. The simulation study, to be described presently, also considers the combination of this Mahalanobis distance with a caliper on the propensity score, math prognosis score, or verbal prognosis score. Each of these distances is combined with full matching (Rosenbaum 1991; Gu and Rosenbaum 1993) to give a matched design.

### 6.1 Matching the SAT coaching example on combinations of scores

Does matching on the prognosis score reliably improve prognostic balance, even when the matching criterion involves measures other than the prognosis score? To address this, I: (i) divide the control group, in which  $y_c$ -responses are available, into “pseudo-treatment” and “pseudo-control” groups; (ii) fit a pseudo-propensity and two pseudo-prognosis scores, extrapolating the prognosis scores from pseudo-controls to the pseudo-treatment group, and in general reproducing as closely as possible the scoring methods to be used for the whole sample; (iii) combine these scores into distance measures analogous to those to be used on the entire sample, in this case Mahalanobis distances with and without calipers; and (iv) use this distance to match pseudo-treatments to pseudo-controls by the same method to be used to match actual treatments to controls, in this case full matching. ( Appendix A.1 describes step (i) in detail.)

After this “dry run,” each matching variation implemented can be checked for prognostic balance, since  $y_c$ -values are in actuality available for pseudo-treatments as well as pseudo-controls. The check used here, described in Appendix A.2, culminates in a  $p$ -value for prognostic balance on several  $x$ -variables taken together, rejecting if any linear combination of these is sufficiently imbalanced; I use it to assess the same  $x$ -variables that contribute to the prognosis score.

Prior to matching, prognostic balance for the math test outcome was poor:  $\chi^2 = 2300$ , on 69 d.f.. Full matching on a Mahalanobis distance combining the propensity and the two prognosis scores reduces this to  $\chi^2 = 470$ , on 64 d.f. — much less imbalance, if still enough that the null is soundly rejected. (Degrees of freedom are reduced because matching introduces collinearity among some covariates.) In contrast, the same matching reduced propensity imbalance from  $\chi^2 = 308$  on 69 d.f. ( $p = 0$ ) to  $\chi^2 = 72.5$  on 64 d.f. ( $p = .22$ ). The technique sharply improves balance in both senses, although the far greater prognostic imbalance is not reduced to insignificance. (The prognosis score suffers two handicaps relative to the propensity

\*\*\*

score in this contest: first, as previously noted, it requires extrapolation of a fit from pseudo-controls to pseudo-treatments, which the propensity does not; second, it relies on a specification for an unknown response surface, whereas in the simulation study the propensity specification is known. To somewhat mitigate the latter disparity, I selected pseudo-treatments and controls using a probability that involved interaction terms, whereas the propensity score fitted during the dry run included only first-order terms; this may not entirely compensate.)

Adding a caliper of .1 sd's in math prognosis score gives sufficient improvement that the hypothesis of balance is no longer rejected ( $\chi^2 = 64.3$  on 64 d.f.;  $p = .45$ ), although this comes at the cost of excluding 10% of the subjects who could not be matched while meeting the caliper restriction. In sum, including a simple prognosis score in the matching criterion has sharply improved prognostic balance, despite the pseudo-treatments' scores being extrapolated from pseudo-controls'; and insisting on close matching on it reduces prognostic imbalance to statistical insignificance.

## 6.2 A simulation study to appraise bias

After the matchings just discussed, the hypothesis that  $Z \perp \mathbf{X}|M$ , with  $M$  a factor recording matched set, would be sustained, as would the hypothesis that  $Y_c \perp \mathbf{X}|M_c$ , where  $M_c$  records matched set after Mahalanobis matching with the prognosis score caliper. Both can be seen as the terminus of a procedure that begins with a Mahalanobis match, then introduces calipers on either propensity or prognosis scores, narrowing these calipers until hypotheses of propensity or prognosis balance, respectively, would be sustained. (It so happened that in the dry run discussed, the first procedure terminated at the first step, with no caliper needed for propensity balance.) Related matching proscriptions are simply to Mahalanobis match on the propensity and the two prognosis scores, without additional measures to secure balance, or to improve the balance of Mahalanobis matching by matching only in or near the region of common support on the sample propensity score. (I operationalize the latter with a relaxation of the caliper condition: starting with large  $w$ , exclude from the match those subjects separated from every potential match by more than  $w$  on the propensity score, then match and check for propensity balance, reducing  $w$  and repeating until balance is achieved. The difference between this and caliper matching is that the latter also requires of included subjects that they be matched only within the caliper.)

To assess the benefit of these procedures for estimating treatment effects, I im-

plement them within repeated dry runs, each with a different, randomly selected pseudo-treatment and control group, estimating treatment effects in each run. (The selection of these groups is described in Appendix A.1.) Because neither group received treatment, the treatment effect is known, and is zero. I track estimates over 650 dry runs, after each run doing permutation tests of hypotheses of constant treatment effects  $\tau_m, \tau_v$  on math and verbal responses, varying  $\tau_m$  and  $\tau_v$  between  $\pm 15$ . Each iteration records the best sustained  $\tau_m$  and  $\tau_v$  (Hodges-Lehmann estimates) and  $p$ -values attaching to the hypotheses that  $\tau_m = 0$  and  $\tau_v = 0$ . Results appear in Table 2. For comparison purposes, the table also shows results from full matching on the propensity score alone and from propensity score full matching with structural restrictions (Hansen 2004).

*Results of the simulation experiment.* The risk that overfitting of prognosis scores would introduce bias (§ 5.2) did not materialize. Far from introducing downward bias, as the reasoning of § 5.2 would entail, the inclusion of estimated prognosis scores in the matching criterion appears to have mitigated the downward bias of matching on propensity scores alone. In accord with with § 4, bias was best reduced with a combination of propensity and prognosis scoring — despite the matching criterion’s use of sample rather than population scores.

Error estimation is good for all methods considered, if somewhat conservative for propensity alone. Hypothesis tests after propensity matching are somewhat conservative; after matching on propensity and prognosis scores together their sizes are closer to nominal levels, with Mahalanobis matching alone and Mahalanobis matching modified by a propensity support condition being the closest. Including the prognosis score in these ways dramatically improves propensity score matching, reducing propensity score matching’s mean squared confidence interval width by more than half, 56% (respectively, 56%) for the SAT-M effect interval and 51% (resp., 50%) for the SAT-V interval, without degrading significance levels.

### **6.3 Estimates of coaching effects.**

Applied to the comparison of coached to uncoached students in the full sample, Mahalanobis matching on prognosis and propensity scores gives coaching effect estimates of 23.0 and  $-0.3$  for SAT-M and SAT-V outcomes, respectively, with 95% confidence intervals  $[16, 30]$  and  $[-8, 7]$ . These results are qualitatively similar to, but sharper than, Hansen’s (2004) propensity matching estimates from the same sample.

Matching:	—		Propensity score				Propensity + prognosis scores							
Modified with:	—		—				restric- tions		—		prop. caliper		support condition	
Response:														
SAT-	M	V	M	V	M	V	M	V	M	V	M	V	M	V
Point estimation														
Bias	8.7	5.6	-1.1	-0.9	-1.3	-1.1	-0.8	-0.5	-1.1	-0.7	-0.9	-0.4		
s.d.	5.3	6.0	4.2	4.3	4.1	4.2	3.6	3.9	3.6	3.9	3.6	3.9		
Rejection rate, $H_0 : \mu = 0$														
$\alpha = .05$	.42	.23	.01	.02	.02	.02	.04	.05	.04	.06	.04	.05		
$\alpha = .10$	.54	.33	.03	.04	.04	.04	.09	.10	.10	.11	.09	.10		

Matching:	—		Prop.				Propensity + prognosis scores							
Caliper:	—		—				prop. score		prognosis for ...		SAT-M		SAT-V	
Response:														
SAT-	M	V	M	V	M	V	M	V	M	V	M	V	M	V
Point estimation														
Bias	8.7	5.6	-1.1	-0.9	-0.8	-0.5	-1.1	-0.7	-0.9	-2.0	0.6	0.6		
s.d.	5.3	6.0	4.2	4.3	3.6	3.9	3.6	3.9	3.7	4.3	3.9	3.8		
Rejection rate, $H_0 : \mu = 0$														
$\alpha = .05$	.42	.23	.01	.02	.04	.05	.04	.06	.05	.08	.05	.05		
$\alpha = .10$	.54	.33	.03	.04	.09	.10	.10	.11	.11	.15	.10	.10		

Table 2: Comparative performance of propensity matching, with and without the prognosis score. All matches are optimal full matches (Rosenbaum 1991), with structural restrictions (Hansen 2004) where indicated, either on the propensity score or on a Mahalanobis distance combining the propensity score and scores for the prognosis of Math and Verbal SAT outcomes. Some Mahalanobis matches are modified by calipers on the propensity or on one of the prognosis scores, or by a common propensity support condition. Best performers appear to be Mahalanobis matching on propensity and prognosis scores, with or without the support condition. Based on 650 dry runs.

## 7 Discussion

The theory of propensity scores (Rosenbaum and Rubin 1983; Rubin and Thomas 1992) suggests, and a growing applied literature confirms, that the technique does not require that there be a concrete, ostensible treatment assignment mechanism in order for it to be beneficial. With or without such a mechanism, exact conditioning on a propensity score supports unbiased inference of causal effects if the covariate information is sufficiently rich. With or without such a mechanism, approximate conditioning on an estimated propensity score induces propensity balance on observed covariates (Rosenbaum and Rubin 1984), stably accommodating many of them (Rubin and Thomas 1996); and is more robust to model misspecification than covariance adjustment (Rubin and Thomas 2000). Related benefits flow from the tentative endorsement of a model for the dependence of response on covariates. In particular, there is a form of balance, prognostic balance, that is promoted by such a strategy, and sufficient balance of this type eliminates bias in the estimation of treatment effects.

Unlike propensity balance, which can be checked directly in a given sample, prognostic balance within a sample has to be checked indirectly. (This is done by checking in dry runs the balancing capabilities of the specific scoring technique in use, as in § 3.1 or 6.1.) Another limitation of the technique is that errors of estimation in the prognosis score may spuriously suggest that the prognosis score modifies treatment effects. On the other hand, careful balance checking helps to ward off such false indications, and the potential for spurious interactions can be assessed using dry run validation (§ 6).

The method holds especial promise for samples with controls well outnumbering treatment subjects, and for samples in which treatment and control groups are well separated on propensity scores. Theory (§ 4) and a case study involving real data and simulations (§ 6) show that combining it with the propensity scores is potentially quite advantageous, reducing both bias and variance as opposed to propensity scoring alone. Indeed, in the case study, propensity and prognosis scores combined to *halve* mean square confidence interval length as compared to propensity score adjustment alone, holding bias to a fraction of estimators' standard deviations and leading to accurate representations of standard error.

**Acknowledgements.** The author is grateful to Jake Bowers, Jennifer Hill, Gary King, Paul Rosenbaum and Donald Rubin for helpful discussions and comments.

## A Details of validation study

### A.1 Creation of pseudo-treatment and pseudo-control groups

In dividing the control group into pseudo-treatments and pseudo-controls, I attempt to recreate differences between actual treatments and controls. This involves using the whole sample to estimate propensities to fall in the treatment group, then using this propensity to select pseudo-treatments from among the controls. For the illustration and dry runs described in § 3.1, § 6, I first chose a regression specification using the forward-backward stepwise interpolation between logistic regression models with no and with all second-order interactions, guided by the AIC. In § 3.1 and § 6, estimated propensities are those fitted by bias-reduced logistic regression (Firth 1993). § 6.2 mounts repeated dry runs, with varying probabilities used to select pseudo-treatment and control groups. In order to obtain these probabilities, a logistic regression model was first fit, with the maximum likelihood estimates of center and scale used for a multivariate Normal approximation to the posterior distribution entailed by uniform priors on the coefficients. Each dry run began with sampling a coefficient vector from this distribution, which was then used to generate a score vector and transformed to the probability scale. In both cases, the resulting probability vectors were then used to make  $\lfloor \{n_t/(n_t + n_c)\}n_c \rfloor$  selections, sequentially and without replacement, into the pseudo-treatment group (where  $n_t$  and  $n_c$  are the sizes of treatment and control groups overall).

### A.2 Tests for propensity and prognostic balance

A nonparametric test is used for both propensity and prognosis balance. It takes numeric  $x$ -variables  $\mathbf{x}$ , a stratification of the sample, and a comparison variable  $v$  — which is the treatment assignment variable,  $z$ , in checks of propensity balance, and is  $y_c$  in the checks of prognostic balance — and generates a test statistic that is asymptotically  $\chi^2$ -distributed under the null of within-stratum independence of  $\mathbf{x}$  and  $v$ . To meet the requirement that the  $x$ 's be numeric, nominal  $x$ 's are decomposed into separate indicator variables. When the comparison variable is  $z$ , it coincides with a test shown by Hansen (2006) to dominate regression-based checks of propensity balance.

The test begins by measuring balance on each  $x$ -variable  $x_k$  by its partial covariance with the comparison variable:  $(v^{\perp s})'x_k^{\perp s}$ , where  $v^{\perp s}, x_k^{\perp s}$  are given by  $(v_{si} - \bar{v}_s : s, i), (x_{ksi} - \bar{x}_{ks} : s, i)$ , respectively. With perfect balance, each of these covariances

would be zero, although in practice this goal can be only roughly attained. Permuting  $v$  within matched sets generates a joint distribution for these partial covariances; for instance it can be shown, following Hájek et al. (1999, 3.3.4), that

$$\text{cov}_{\text{perm}}((v^{\perp s})'x_k^{\perp s}, (v^{\perp s})'x_l^{\perp s}) = \sum_s s^2(\{v_{si} : i \leq n_s\}) \sum_{j=l}^{n_s} (x_{ksj} - \bar{x}_{ks.})(x_{lsj} - \bar{x}_{ls.}),$$

where  $s^2(x_{s.}) := \{\sum_1^{n_s} (x_{si} - \bar{x}_s)^2\} / (n_s - 1)$ . From this formula a null covariance matrix  $C$  for the balance measurements  $b = (v^{\perp s})'x_1^{\perp s}, \dots, (v^{\perp s})'x_K^{\perp s}$  is calculable, and  $(v^{\perp s})'x_1^{\perp s}, \dots, (v^{\perp s})'x_K^{\perp s}$  combine into a  $\chi^2$  statistic  $b^t C^{-1} b$  on  $\text{rank}(C)$  d.f.. The null hypothesis of balance is then rejected if one or more of the balance measurements contributing to  $b$  is particularly large. Note that the same statistic would be obtained if balance were assessed and then combined on a set of variables  $x'_1, \dots, x'_{K'}$  with the same linear span as  $x_1, \dots, x_K$ . It follows that the test detects imbalance in any linear combination of  $x_1, \dots, x_K$ , rejecting if it exceeds levels that would have applied were treatment randomized within matched sets.

## References

- Belson, W. A. (1956), “A technique for studying the effects of a television broadcast,” *Applied Statistics*, 5, 195–202.
- Berk, R. A. and de Leeuw, J. (1999), “An Evaluation of California’s Inmate Classification System Using a Generalized Regression Discontinuity Design,” *Journal of the American Statistical Association*, 94, 1045–1052.
- Campbell, D. and Stanley, J. (1966), *Experimental and Quasi-Experimental Designs for Research*, Houghton Mifflin.
- Cochran, W. G. (1954), “Some methods of strengthening the common  $\chi^2$  tests,” *Biometrics*, 10, 417–451.
- (1969), “The Use of Covariance in Observational Studies,” *Applied Statistics*, 18, 270–275.
- Cook, E. F. and Goldman, L. (1989), “Performance of Tests of Significance Based on Stratification by a Multivariate Confounder Score or by a Propensity Score,” *Journal of Clinical Epidemiology*, 42, 317–324.

- Cox, D. (1958), *The Planning of Experiments*, John Wiley.
- Firth, D. (1993), “Bias reduction of maximum likelihood estimates,” *Biometrika*, 80, 27–38.
- Gu, X. and Rosenbaum, P. R. (1993), “Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms,” *Journal of Computational and Graphical Statistics*, 2, 405–420.
- Hájek, J., Šidák, Z., and Sen, P. K. (1999), *Theory of rank tests*, New York: Academic Press, 2nd ed.
- Hansen, B. B. (2004), “Full matching in an observational study of coaching for the SAT,” *Journal of the American Statistical Association*, 99, 609–618.
- (2006), “Appraising Covariate Balance after Assignment to Treatment by Groups,” Tech. Rep. 436, University of Michigan, Statistics Department.
- Holland, P. W. (1986), “Statistics and Causal Inference (with discussion),” *Journal of the American Statistical Association*, 81, 945–970.
- Little, R. and Rubin, D. (2000), “Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches,” *Annual Review of Public Health*, 21, 121–145.
- Mantel, N. and Haenszel, W. (1959), “Statistical aspects of the analysis of data from retrospective studies of disease,” *Journal of the National Cancer Institute*, 22, 719–748.
- McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models (Second edition)*, Chapman & Hall Ltd.
- McNemar, Q. (1947), “Note on the sampling error of the differences between correlated proportions or percentage,” *Psychometrika*, 12, 153–157.
- Miettinen, O. S. (1976), “Stratification by a Multivariate Confounder Score,” *American Journal of Epidemiology*, 104, 609–620.

- Moons, K., Donders, A. R. T., Steyerberg, E., and Harrell, F. (2004), “Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example,” *Journal of Clinical Epidemiology*, 57, 1262–1270.
- Neyman, J. (1990), “On the application of probability theory to agricultural experiments. Essay on principles. Section 9,” *Statistical Science*, 5, 463–480, reprint. Transl. by Dabrowska and Speed.
- Pace, L. and Salvan, A. (1997), *Principles of statistical inference: from a neo-Fisherian perspective*, vol. 4 of *Advanced series on statistical science & applied probability*, Singapore: World Scientific.
- Peters, C. C. (1941), “A method of matching groups for experiment with no loss of population,” *Journal of Educational Research*, 34, 606–612.
- Pike, M., Anderson, J., and Day, N. (1979), “Some insights into Miettinen’s multivariate confounder score approach to case-control study analysis,” *Epidemiology and Community Health*, 33, 104–106.
- Powers, D. and Rock, D. (1999), “Effects of Coaching on SAT I: Reasoning Test Scores,” *Journal of Educational Measurement*.
- Rosenbaum, P. R. (1991), “A Characterization of Optimal Designs for Observational Studies,” *Journal of the Royal Statistical Society*, 53, 597–610.
- (2001), “Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot,” *Biometrika*, 88, 219–231.
- (2002), *Observational Studies*, Springer-Verlag, 2nd ed.
- (2005), “Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies,” *The American Statistician*, 59, 147–152.
- 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.
- (1984), “Reducing Bias in Observational Studies using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79, 516–524.

- (1985), “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score,” *The American Statistician*, 39, 33–38.
- Rubin, D. B. (1977), “Assignment to Treatment Group on the Basis of a Covariate (Corr: V3 P384),” *Journal of Educational Statistics*, 2, 1–26.
- (1984), “William G. Cochran’s Contributions to the Design, Analysis, and Evaluation of Observational Studies,” in *W. G. Cochran’s Impact on Statistics*, Wiley (New York), pp. 37–69.
- Rubin, D. B. and Thomas, N. (1992), “Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions,” *Biometrika*, 79, 797–809.
- (1996), “Matching Using Estimated Propensity Scores: Relating Theory to Practice,” *Biometrics*, 52, 249–64.
- (2000), “Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates,” *Journal of the American Statistical Association*, 95, 573–585.
- Visser, R. A. and De Leeuw, J. (1984), “Maximum Likelihood Analysis for a Generalized Regression-discontinuity Design,” *Journal of Educational Statistics*, 9, 45–60.
- Zhao, Z. (2004), “Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence,” *The Review of Economics and Statistics*, 86, 91–107.