Biometrika (1983), **70**, 1, *pp.* 41–55
Printed in Great Britain

The central role of the propensity score in observational studies for causal effects

BY PAUL R. ROSENBAUM

Departments of Statistics and Human Oncology, University of Wisconsin, Madison, Wisconsin, U.S.A.

AND DONALD B. RUBIN

University of Chicago, Chicago, Illinois, U.S.A.

SUMMARY

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications include: (i) matched sampling on the univariate propensity score, which is a generalization of discriminant matching, (ii) multivariate adjustment by subclassification on the propensity score where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations, and (iii) visual representation of multivariate covariance adjustment by a two-dimensional plot.

Some key words: Covariance adjustment; Direct adjustment; Discriminant matching; Matched sampling; Nonrandomized study; Standardization; Stratification; Subclassification.

1. DEFINITIONS

1·1. *The structure of studies for causal effects*

Inferences about the effects of treatments involve speculations about the effect one treatment would have had on a unit which, in fact, received some other treatment. We consider the case of two treatments, numbered 1 and 0. In principle, the ith of the N units under study has both a response r_{1i} that would have resulted if it had received treatment 1, and a response r_{0i} that would have resulted if it had received treatment 0. In this formulation, causal effects are comparisons of r_{1i} and r_{0i}, for example $r_{1i}-r_{0i}$ or r_{1i}/r_{0i}. Since each unit receives only one treatment, either r_{1i} or r_{0i} is observed, but not both, so comparisons of r_{1i} and r_{0i} imply some degree of speculation. In a sense, estimating the causal effects of treatments is a missing data problem, since either r_{1i} or r_{0i} is missing.

This formulation is that used in the literature of experimental design, for example, in the books by Fisher (1951) and Kempthorne (1952), and follows the development by Rubin (1974, 1977, 1978, 1980a); Hamilton (1979) adopts a similar approach. The structure would not be adequate when, for example, the response of unit i to treatment t depends on the treatment given to unit j, as could happen if they compete for resources. The assumption that there is a unique value r_{ti} corresponding to unit i and treatment t has been called the stable unit-treatment value assumption (Rubin, 1980a), and will be

made here. For discussion of some possible violations of this assumption, see Cox (1958, Chapter 2) or Rubin (1978, § 2·3).

In this paper, the N units in the study are viewed as a simple random sample from some population, and the quantity to be estimated is the average treatment effect, defined as

$$E(r_1) - E(r_0), \tag{1·1}$$

where $E(\,.\,)$ denotes expectation in the population.

Let $z_i = 1$ if unit i is assigned to the experimental treatment, and $z_i = 0$ if unit i is assigned to the control treatment. Let x_i be a vector of observed pretreatment measurements or covariates for the ith unit; all of the measurements in x are made prior to treatment assignment, but x may not include all covariates used to make treatment assignments. It is assumed that the numbering of units is done at random, so that the index i contains no information; observed information about unit i is contained in x_i. Throughout, we ignore measure theoretic details.

1·2. *Balancing scores and the propensity score*

In randomized experiments, the results in the two treatment groups may often be directly compared because their units are likely to be similar, whereas in nonrandomized experiments, such direct comparisons may be misleading because the units exposed to one treatment generally differ systematically from the units exposed to the other treatment. Balancing scores, defined here, can be used to group treated and control units so that direct comparisons are more meaningful.

A balancing score, $b(x)$, is a function of the observed covariates x such that the conditional distribution of x given $b(x)$ is the same for treated ($z = 1$) and control ($z = 0$) units; that is, in Dawid's (1979) notation,

$$x \perp\!\!\!\perp z \,|\, b(x).$$

The most trivial balancing score is $b(x) = x$. More interesting balancing scores are many-one functions of x. In § 2 we identify all functions of x that are balancing scores and identify the coarsest function of x that is a balancing score, namely the propensity score. We also show that easily obtained estimates of balancing scores behave like balancing scores. Also, we show that if treatment assignment is strongly ignorable given x, as defined in § 1·3, then the difference between treatment and control means at each value of a balancing score is an unbiased estimate of the treatment effect at that value, and consequently pair matching, subclassification and covariance adjustment on a balancing score can produce unbiased estimates of the average treatment effect (1·1). Moreover in § 3 we see that common methods of multivariate adjustment in observational studies, including covariance adjustment for x and discriminant matching (Cochran & Rubin, 1973), implicitly adjust for an estimated scalar balancing score.

In order to motivate formally adjustment for a balancing score, we must consider the sampling distribution of treatment assignments. Let the conditional probability of assignment to treatment one, given the covariates, be denoted by

$$e(x) = \mathrm{pr}(z = 1 \,|\, x), \tag{1·2}$$

where we assume

$$\mathrm{pr}(z_1, \ldots, z_n \,|\, x_1, \ldots, x_n) = \prod_{i=1}^{N} e(x_i)^{z_i} \{1 - e(x_i)\}^{1 - z_i}.$$

Although this strict independence assumption is not essential, it simplifies notation and discussion. The function $e(x)$ is called the propensity score, that is, the propensity towards exposure to treatment 1 given the observed covariates x. In § 2, $e(x)$ is shown to be the coarsest balancing score.

1·3. *Strongly ignorable treatment assignment*

Randomized and nonrandomized trials differ in two distinct ways because in randomized experiments z_i has a distribution determined by a specified random mechanism. First, in a randomized trial, the propensity score is a known function so that there exists one accepted specification for $e(x)$. In a nonrandomized experiment, the propensity score function is almost always unknown so that there is not one accepted specification for $e(x)$; however, $e(x)$ may be estimated from observed data, perhaps using a model such as a logit model. To a Bayesian, estimates of these probabilities are posterior predictive probabilities of assignment to treatment 1 for a unit with vector x of covariates.

The second way randomized trials differ from nonrandomized trials is that, with properly collected data in a randomized trial, x is known to include all covariates that are both used to assign treatments and possibly related to the response (r_1, r_0). More formally, in a randomized trial, treatment assignment z and response (r_1, r_0) are known to be conditionally independent given x,

$$(r_1, r_0) \perp\!\!\!\perp z \,|\, x.$$

This condition is usually not known to hold in a nonrandomized experiment. Moreover, in a randomized experiment, every unit in the population has a chance of receiving each treatment. Generally, we shall say treatment assignment is strongly ignorable given a vector of covariates v if

$$(r_1, r_0) \perp\!\!\!\perp z \,|\, v, \quad 0 < \mathrm{pr}(z = 1 \,|\, v) < 1 \tag{1·3}$$

for all v. For brevity, when treatment assignment is strongly ignorable given the observed covariates x, that is, when (1·3) holds with $v = x$, we shall say simply that treatment assignment is strongly ignorable. If treatment assignment is strongly ignorable, then it is ignorable in Rubin's (1978) sense, but the converse is not true.

2. Theory

2·1. *Outline*

Section 2 presents five theorems whose conclusions may be summarized as follows.

(i) The propensity score is a balancing score.

(ii) Any score that is 'finer' than the propensity score is a balancing score; moreover, x is the finest balancing score and the propensity score is the coarsest.

(iii) If treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score.

(iv) At any value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a

balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects.

(v) Using sample estimates of balancing scores can produce sample balance on x.

2·2. *Large-sample theory*

The results in this section treat $e(x)$ as known, and are therefore applicable to large samples.

THEOREM 1. *Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is*

$$x \perp\!\!\!\perp z \mid e(x).$$

The above theorem is a special case of Theorem 2. Cochran & Rubin (1973) proved Theorem 1 in the particular case of multivariate normal covariates x; the result holds regardless of the distribution of x.

THEOREM 2. *Let $b(x)$ be a function of x. Then $b(x)$ is a balancing score, that is,*

$$x \perp\!\!\!\perp z \mid b(x), \tag{2·1}$$

if and only if $b(x)$ is finer than $e(x)$ in the sense that $e(x) = f\{b(x)\}$ for some function f.

Proof. First suppose $b(x)$ is finer than $e(x)$. Since $e(x) = \mathrm{pr}(z = 1 \mid x)$, to show $b(x)$ is a balancing score it is sufficient to show

$$\mathrm{pr}\{z = 1 \mid b(x)\} = e(x). \tag{2·2}$$

Now by the definition of $e(x)$,

$$\mathrm{pr}\{z = 1 \mid b(x)\} = E\{e(x) \mid b(x)\}.$$

But since $b(x)$ is finer than $e(x)$,

$$E\{e(x) \mid b(x)\} = e(x),$$

as required, so that $b(x)$ is a balancing score.

Now, for the converse, suppose $b(x)$ is a balancing score, but that $b(x)$ is not finer than $e(x)$, so that there exists x_1 and x_2 such that $e(x_1) \neq e(x_2)$ but $b(x_1) = b(x_2)$. But then, by the definition of $e(\,.\,)$, $\mathrm{pr}(z = 1 \mid x_1) \neq \mathrm{pr}(z = 1 \mid x_2)$, so that z and x are not conditionally independent given $b(x)$, and thus $b(x)$ is not a balancing score. Therefore, to be a balancing score, $b(x)$ must be finer than $e(x)$.

Theorem 1 implies that if a subclass of units or a matched treatment-control pair is homogeneous in $e(x)$, then the treated and control units in that subclass or matched pair will have the same distribution of x. Theorem 2 implies that if subclasses or matched treated-control pairs are homogeneous in both $e(x)$ and certain chosen components of x, it is still reasonable to expect balance on the other components of x within these refined subclasses or matched pairs. The practical importance of Theorem 2 beyond Theorem 1 arises because it is sometimes advantageous to subclassify or match not only for $e(x)$ but for other functions of x as well; in particular, such a refined procedure may be used to obtain estimates of the average treatment effect in subpopulations defined by components of x, for example males, females.

Theorem 3 is the key result for showing that if treatment assignment is strongly

ignorable, then adjustment for a balancing score $b(x)$ is sufficient to produce unbiased estimates of the average treatment effect (1·1).

THEOREM 3. *If treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score $b(x)$; that is,*

$$(r_1, r_0) \perp\!\!\!\perp z \mid x$$

and

$$0 < \mathrm{pr}(z = 1 \mid x) < 1$$

for all x imply

$$(r_1, r_0) \perp\!\!\!\perp z \mid b(x)$$

and

$$0 < \mathrm{pr}\{z = 1 \mid b(x)\} < 1$$

for all $b(x)$.

Proof. The inequality given $b(x)$ follows immediately from the inequality given x. Consequently, it is sufficent to show that

$$\mathrm{pr}\{z = 1 \mid r_1, r_0, b(x)\} = \mathrm{pr}\{z = 1 \mid b(x)\},$$

which by Theorem 2, equation (2·2), is equivalent to showing that

$$\mathrm{pr}\{z = 1 \mid r_1, r_0, b(x)\} = e(x).$$

Now

$$\mathrm{pr}\{z = 1 \mid r_1, r_0, b(x)\} = E\{\mathrm{pr}(z = 1 \mid r_1, r_0, x) \mid r_1, r_0, b(x)\},$$

which by assumption equals $E\{\mathrm{pr}(z = 1 \mid x) \mid r_1, r_0, b(x)\}$, which by definition equals $E\{e(x) \mid r_1, r_0, b(x)\}$, which, since $b(x)$ is finer than $e(x)$, equals $e(x)$ as required.

Theorem 3 also can be proved using Lemmas 4.2(i) and 4.3 of Dawid (1979).

We are now ready to relate balancing scores and ignorable treatment assignment to the estimation of treatment effects.

The response r_t to treatment t is observed only if the unit receives treatment t, that is if $z = t$. Thus, if a randomly selected treated unit, $z = 1$, is compared to a randomly selected control unit, $z = 0$, the expected difference in response is

$$E(r_1 \mid z = 1) - E(r_0 \mid z = 0). \tag{2·3}$$

Expression (2·3) does not equal (1·1) in general because the available samples are not from the marginal distribution of r_t, but rather from the conditional distribution of r_t given $z = t$.

Suppose a specific value of the vector of covariates x is randomly sampled from the entire population of units, that is, both treated and control units together, and then a treated unit and a control unit are found both having this value for the vector of covariates. In this two-step sampling process, the expected difference in response is

$$E_x\{E(r_1 \mid x, z = 1) - E(r_0 \mid x, z = 0)\}, \tag{2·4}$$

where E_x denotes expectation with respect to the distribution of x in the entire population of units. If treatment assignment is strongly ignorable, that is if (1·3) holds with $v = x$, then (2·4) equals

$$E_x\{E(r_1 \mid x) - E(r_0 \mid x)\},$$

which does equal the average treatment effect (1·1).

Now suppose a value of a balancing score $b(x)$ is sampled from the entire population of units and then a treated unit and a control unit are sampled from all units having this value of $b(x)$, but perhaps different values of x. Given strongly ignorable treatment assignment, it follows from Theorem 3 that

$$E\{r_1 \mid b(x), z = 1\} - E\{r_0 \mid b(x), z = 0\} = E\{r_1 \mid b(x)\} - E\{r_0 \mid b(x)\}$$

from which it follows that

$$\begin{aligned} E_{b(x)}[E\{r_1 \mid b(x), z = 1\} - E\{r_0 \mid b(x), z = 0\}] &= E_{b(x)}[E\{r_1 \mid b(x)\} - E\{r_0 \mid b(x)\}] \\ &= E(r_1 - r_0), \end{aligned} \tag{2·5}$$

where $E_{b(x)}$ denotes expectation with respect to the distribution of $b(x)$ in the entire population. In words, under strongly ignorable treatment assignment, units with the same value of the balancing score $b(x)$ but different treatments can act as controls for each other, in the sense that the expected difference in their responses equals the average treatment effect.

The above argument has established the following theorem and corollaries.

THEOREM 4. *Suppose treatment assignment is strongly ignorable and $b(x)$ is a balancing score. Then the expected difference in observed responses to the two treatments at $b(x)$ is equal to the average treatment effect at $b(x)$, that is,*

$$E\{r_1 \mid b(x), z = 1\} - E\{r_0 \mid b(x), z = 0\} = E\{r_1 - r_0 \mid b(x)\}.$$

COROLLARY 4·1. *Pair matching on balancing scores. Suppose treatment assignment is strongly ignorable. Further suppose that a value of a balancing score $b(x)$ is randomly sampled from the population of units, and then one treated, $z = 1$, unit and one control, $z = 0$, unit are sampled with this value of $b(x)$. Then the expected difference in response to the two treatments for the units in the matched pair equals the average treatment effect at $b(x)$. Moreover, the mean of matched pair differences obtained by this two-step sampling process is unbiased for the average treatment effect* (1·1).

COROLLARY 4·2. *Subclassification on balancing scores. Suppose treatment assignment is strongly ignorable. Suppose further that a group of units is sampled using $b(x)$ such that:* (i) *$b(x)$ is constant for all units in the group, and* (ii) *at least one unit in the group received each treatment. Then, for these units, the expected difference in treatment means equals the average treatment effect at that value of $b(x)$. Moreover, the weighted average of such differences, that is, the directly adjusted difference, is unbiased for the treatment effect* (1·1), *when the weights equal the fraction of the population at $b(x)$.*

COROLLARY 4·3. *Covariance adjustment on balancing scores. Suppose treatment assignment is strongly ignorable, so that in particular, $E\{r_t \mid z = t, b(x)\} = E\{r_t \mid b(x)\}$ for balancing score $b(x)$. Further suppose that the conditional expectation of r_t given $b(x)$ is linear:*

$$E\{r_t \mid z = t, b(x)\} = \alpha_t + \beta_t b(x) \quad (t = 0,1).$$

Then the estimator

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\, b(x)$$

is conditionally unbiased given $b(x_i)$ $(i = 1, \ldots, n)$ for the treatment effect at $b(x)$, namely

$E\{r_1 - r_0 \mid b(x)\}$, if $\hat{\alpha}_t$ and $\hat{\beta}_t$ are conditionally unbiased estimators of α_t and β_t, such as least squares estimators. Moreover,

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\bar{b},$$

where $\bar{b} = n^{-1}\Sigma b(x_i)$, is unbiased for the average treatment effect (1·1) *if the units in the study are a simple random sample from the population.*

2·3. *Some small-sample theory*

Usually the propensity scores $e(x_i)$ must be estimated from available data, (z_i, x_i) $(i = 1, \ldots, N)$. Define the sample conditional proportion $\text{prop}(A \mid B)$ as the proportion of those vectors (z_i, x_i) satisfying condition B that also satisfy condition A, leaving $\text{prop}(A \mid B)$ undefined if no vector satisfies condition B. For example, $\text{prop}\{z = 1 \mid x = (1, 0)\}$ is the proportion of the N units with $z = 1$ among all units with $x = (1, 0)$. Estimate $e(x)$ by $\hat{e}(a) = \text{prop}\,(z = 1 \mid x = a)$. If $\hat{e}(a) = 0$ or 1 then all units with $x = a$ received the same treatment. Theorem 5, which parallels Theorem 1, shows that at all intermediate values of $\hat{e}(a)$, that is for $0 < \hat{e}(a) < 1$, there is sample balance. Of course, intermediate values of $\hat{e}(x)$ will exist only when x takes on relatively few values.

THEOREM 5. *Suppose $0 < \hat{e}(a) < 1$. Then*

$$\text{prop}\{z = 0, x = a \mid \hat{e}(x) = \hat{e}(a)\} = \text{prop}\{z = 0 \mid \hat{e}(x) = \hat{e}(x)\}\ \text{prop}\{x = a \mid \hat{e}(z) = \hat{e}(a)\}. \tag{2·6}$$

An analogous theorem about sample balance parallels Theorem 2, and the proofs parallel the corresponding proofs of Theorems 1 and 2 because proportions follow essentially the same axioms as probabilities.

COROLLARY 5·1. *Suppose the N units are a random sample from an infinite population, and suppose x takes on only finitely many values in the population and at each such value $0 < e(x) < 1$. Then with probability 1 as $N \to \infty$, subclassification on $\hat{e}(x)$ produces sample balance, that is,* (2·6) *holds.*

In practice, except when x takes on only a few values, $\hat{e}(a)$ will be either zero or one for most values of a. Consequently, in order to estimate propensity scores, some modelling will be required.

The propensity score can often be modelled using an appropriate logit model (Cox, 1970) or discriminant score.

Clearly,

$$e(x) = \text{pr}\,(z = 1 \mid x) = \frac{\text{pr}\,(z = 1)\,\text{pr}\,(x \mid z = 1)}{\text{pr}\,(z = 1)\,\text{pr}\,(x \mid z = 1) + \text{pr}\,(z = 0)\,\text{pr}\,(x \mid z = 0)}.$$

Elementary manipulations establish the following facts.

(i) If $\text{pr}\,(x \mid z = t) = N_p(\mu_t, \Omega)$ then $e(x)$ is a monotone function of the linear discriminant $x^{\mathrm{T}}\Omega^{-1}(\mu_1 - \mu_2)$. Therefore, matching on $e(x)$ includes discriminant matching (Cochran & Rubin, 1973; Rubin 1976a, b; 1979; 1980b) as a special case. Some related results appear in § 3·2.

(ii) If $\text{pr}\,(x \mid z = t)$ is a polynomial exponential family distribution, i.e. if

$$\text{pr}\,(x \mid z = t) = h(x)\exp\{P_t(x)\},$$

where $P_t(x)$ is a polynomial in x of degree k, say, then $e(x)$ obeys a polynomial logit model

$$\log\frac{e(x)}{1-e(x)} = \log\frac{\text{pr}(z=1)}{1-\text{pr}(z=1)} + P_1(x) - P_0(x)$$

$$= \log\frac{\text{pr}(z=1)}{1-\text{pr}(z=1)} + Q(x),$$

where $Q(x)$ is a degree k polynomial in x. This polynomial exponential family includes the linear exponential family resulting in a linear logit model for $e(x)$, the quadratic exponential family described by Dempster (1971), and the binary data model described by Cox (1972). Related discussion is given by Dawid (1976).

3. THREE APPLICATIONS OF PROPENSITY SCORES TO OBSERVATIONAL STUDIES

3·1. *Techniques for adjustment in observational studies*

The general results we have presented suggest that, in practice, adjustment for the propensity score should be an important component of the analysis of observational studies because evidence of residual bias in the propensity score is evidence of potential bias in estimated treatment effects. We conclude with three examples of how propensity scores can be explicitly used to adjust for confounding variables in observational studies. The examples involve three standard techniques for adjustment in observational studies (Cochran, 1965; Rubin, 1983), namely, matched sampling, subclassification, and covariance adjustment, that is, the three methods addressed by Corollaries 4·1, 4·2 and 4·3.

3·2. *Use of propensity scores to construct matched samples from treatment groups*

Matching is a method of sampling from a large reservoir of potential controls to produce a control group of modest size in which the distribution of covariates is similar to the distribution in the treated group. Some sampling of a large control reservoir is often required to reduce costs associated with measuring the response, for example, costs associated with obtaining extensive follow-up data on patients in clinical studies (Rubin, 1973a; Cohn *et al.*, 1981).

Although there exist model-based alternatives to matched sampling, e.g. covariance adjustment on random samples, there are several reasons why matching is appealing.

(I) Matched treated and control pairs allow relatively unsophisticated researchers to appreciate immediately the equivalence of treatment and control groups, and to perform simple matched pair analyses which adjust for confounding variables. This issue is discussed in greater detail below in § 3·3 on balanced subclassification.

(II) Even if the model underlying a statistical adjustment is correct, the variance of the estimate of the average treatment effect (1·1) will be lower in matched samples than in random samples since the distributions of x in treated and control groups are more similar in matched than in random samples. To verify this reduced variance, inspect the formula for the variance of the covariance adjusted estimate (Snedecor & Cochran, 1980, p. 368, formula 18.2.3), and note that the variance decreases as the difference between treatment and control means on x decreases.

(III) Model-based adjustment on matched samples is usually more robust to

departures from the assumed form of the underlying model than model-based adjustment on random samples (Rubin, 1973b, 1979), primarily because of reduced reliance on the model's extrapolations.

(IV) In studies with limited resources but large control reservoirs and many confounding variables, the confounding variables can often be controlled by multivariate matching, but the small-sample sizes in the final groups do not allow control of all variables by model-based methods.

Ideally, treated and control units would be exactly matched on all covariates x, so that the sample distributions of x in the two groups would be identical. Theorem 2 shows that it is sufficient to match exactly on any balancing score $b(x)$ to obtain the same probability distributions of x for treated and control units in matched samples. Moreover, Corollary 4·1 shows that if treatment assignment is strongly ignorable, exact matching on a balancing score leads to an unbiased estimate of the average treatment effect. Unfortunately, exact matches even on a scalar balancing score are often impossible to obtain, so methods which seek approximate matches must be used. We now study properties of some matching methods based on the propensity score.

A multivariate matching method is said to be equal per cent bias reducing if the bias in each coordinate of x is reduced by the same percentage (Rubin, 1976a, b). Matching methods which are not equal per cent bias reducing have the potentially undesirable property that they increase the bias for some linear functions of x. If matched sampling is performed before the response (r_1, r_0) can be measured, and if all that is suspected about the relation between (r_1, r_0) and x is that it is approximately linear, then matching methods which are equal per cent bias reducing are reasonable in that they lead to differences in mean response in matched samples that should be less biased than in random samples.

The initial bias in x is

$$B = E(x \mid z = 1) - E(x \mid z = 0). \tag{3·1}$$

Let us suppose that we have a random sample of treated units and a large reservoir of randomly sampled control units, and suppose each treated unit is matched with a control unit from the reservoir. Then the expected bias in x in matched samples is

$$B_m = E(x \mid z = 1) - E_m(x \mid z = 0), \tag{3·2}$$

where the subscript m indicates the distribution in matched samples. In general, from Theorem 2, B_m is a null vector if exact matches on a balancing score have been obtained. If $B_m = \gamma B$ for some scalar γ, with $0 < \gamma < 1$, then the matching method is equal per cent bias reducing: the bias in each coordinate of x is reduced by $100(1-\gamma)\%$. If the method is not equal per cent bias reducing, then there exists a vector w such that $wB_m > wB$, so that matching has increased the bias for some linear function of x.

In §2·3 we observed that discriminant matching is equivalent to matching on the propensity score if the covariates x have a multivariate normal distribution. Assuming multivariate normality, Rubin (1976a) showed that matching on the population or sample discriminant is equal per cent bias reducing. We now show that matching on the population propensity score is equal per cent bias reducing under weaker distributional assumptions. It is assumed that the matching algorithm matches each treated, $z = 1$, unit with a control, $z = 0$, unit drawn from a reservoir of control units on the basis of a balancing score, for example, using nearest available matching on a scalar balancing score.

THEOREM 6. *Let $b = b(x)$ be a balancing score. For any matching method that uses b alone to match each treated unit, $z = 1$, with a control unit, $z = 0$, the reduction in bias is*

$$B - B_m = \int E(x \mid b) \{\mathrm{pr}_m(b \mid z = 0) - \mathrm{pr}(b \mid z = 0)\} db, \tag{3·3}$$

where $\mathrm{pr}_m(b \mid z = 0)$ denotes the distribution of b in matched samples from the control group.

Proof. From (3·1) and (3·2) we have

$$B - B_m = \int \{E_m(x \mid z = 0, b)\,\mathrm{pr}_m(b \mid z = 0) - E(x \mid z = 0, b)\,\mathrm{pr}(b \mid z = 0)\}\, db. \tag{3·4}$$

For any matching method satisfying the condition of the theorem,

$$E_m(x \mid z = 0, b) = E(x \mid z = 0, b) \tag{3·5}$$

because any matching method using b alone to match units alters the marginal distribution of b in the control group, $z = 0$, but does not alter the conditional distribution of x given b in the control group. However, by Theorem 2,

$$E(x \mid z = 0, b) = E(x \mid b). \tag{3·6}$$

Substitution of (3·5) and (3·6) into equation (3·4) yields the result (3·3).

COROLLARY 6·1. *If $E(x \mid b) = \alpha + \beta f(b)$ for some vectors α and β and some scalar-valued function $f(\,.\,)$, then matching on b alone is equal per cent bias reducing.*

Proof. The per cent reduction in bias for the ith coordinate of x is, from (3·3)

$$100 \frac{\beta_i [E_m\{f(b) \mid z = 0\} - E\{f(b) \mid z = 0\}]}{\beta_i [E\{f(b) \mid z = 1\} - E\{f(b) \mid z = 0\}]},$$

which is independent of i, as required.

The following corollary shows that if subpopulations are defined using x so that some function $d(x)$ is constant within each subpopulation, then propensity matching within subpopulations is equal per cent bias reducing in each subpopulation.

COROLLARY 6·2. *Let $d = d(x)$ be some function of x. If $E(x \mid b, d) = \alpha_d + \beta_d f_d(b)$ for vectors α_d, β_d, and some scalar-valued functions $f_d(\,.\,)$, then matching on b alone at each value of d is equal per cent bias reducing at each value of d, that is,*

$$E(x \mid d, z = 1) - E_m(x \mid d, z = 0) = \gamma_d \{E(x \mid d, z = 1) - E(x \mid d, z = 0)\}$$

for scalar γ_d.

Proof. Apply Theorem 6 and Corollary 6·1 within subpopulations.

Rubin's (1979) simulation study examines the small-sample properties of discriminant matching in the case of normal covariates with possibly different covariances in the treatment groups. Thus, the study includes situations where the true propensity score is a quadratic function of x but the discriminant score is a linear function of x. Table 1 presents previously unpublished results from this study for situations in which the propensity score is a monotone function of the linear discriminant, so that propensity matching and discriminant matching are effectively the same. The covariates x are

Table 1. *Per cent reduction in bias due to matched sampling based on the sample and population propensity scores*

Ratio of size control reservoir to size of treatment group	Type of score	Initial bias along standardized discriminant 0·25	0·50	0·75	1·00
2	Sample	92	85	77	67
	Population	92	87	78	69
3	Sample	101	96	91	83
	Population	96	95	91	84
4	Sample	97	98	95	90
	Population	98	97	94	89

Assuming bivariate normal covariates with common covariance matrix, parallel linear response surfaces, sample size of 50 in treated and control groups. Estimated per cent reduction in bias from Rubin's (1979) simulation study. The largest estimated standard error for this table is less than 0·03.

bivariate normal with common covariance matrix. In the simulation, 50 treated units are matched using nearest available matching (Cochran & Rubin, 1973) on the sample discriminant with 50 control units drawn from a reservoir of $50R$ potential control units, for $R = 2,3,4$; details are given by Rubin (1979).

Assuming parallel linear response surfaces, Table 1 shows that even in the absence of additional adjustments, propensity, i.e. discriminant, matching alone can remove most of the initial bias if the reservoir is relatively large. Moreover, Table 1 shows that the population and sample propensity scores are approximately equally effective in removing bias, so that no substantial loss is incurred by having to estimate the propensity score. It should be noted that the conditions underlying Table 1 differ from the conditions underlying Theorem 1 because nearest available matching with imperfect matches provides only a partial adjustment for the propensity score.

Propensity matching should prove especially effective relative to Mahalanobis metric matching (Cochran & Rubin, 1973; Rubin, 1976a, b; 1979; 1980b) in situations where markedly nonspherically distributed x make the use of a quadratic metric unnatural as a measure of distance between treated and control units. For example, we have found in practice that if x contains one coordinate representing a rare binary event, then Mahalanobis metric matching may try too hard to match that coordinate exactly, thereby reducing the quality of matches on the other coordinates of x. Propensity matching can effectively balance rare binary variables for which it is not possible to match treated and control units adequately on an individual basis.

3·3. *Subclassification on propensity scores*

A second major method of adjustment for confounding variables is subclassification, whereby experimental and control units are divided on the basis of x into subclasses or strata (Cochran, 1965, 1968; Cochran & Rubin, 1973). Direct adjustment with subclass total weights can be applied to the subclass differences in response to estimate the average treatment effect (1·1) whenever treatment assignment is strongly ignorable, without modelling assumptions such as parallel linear response surfaces; see Corollary 4·2.

As a method of multivariate adjustment, subclassification has the advantage that it involves direct comparisons of ostensibly comparable groups of units within each

subclass and therefore can be both understandable and persuasive to an audience with limited statistical training. The comparability of units within subclasses can be verified by the simplest methods, such as bar charts of means.

A major problem with subclassification (Cochran, 1965) is that as the number of confounding variables increases, the number of subclasses grows dramatically, so that even with only two categories per variable, yielding 2^P subclasses for P variables, most subclasses will not contain both treated and control units. Subclassification on the propensity score is a natural way to avoid this problem.

We now use an estimate of the propensity score to subclassify patients in an actual observational study of therapies for coronary artery disease. The treatments are coronary artery bypass surgery, $z = 1$, and drug therapy, $z = 0$. The covariates x are clinical, haemodynamic, and demographic measurements on each patient made prior to treatment assignment. Even though the covariates have quite different distributions in the two treatment groups, within each of the five subclasses, the surgical and drug patients will be seen to have similar sample distributions of x.

The propensity score was estimated using a logit model for z given x. Covariates and interactions among covariates were selected for the model using a stepwise procedure. Based on Cochran's (1968) observation that subclassification with five subclasses is sufficient to remove at least 90% of the bias for many continuous distributions, five subclasses of equal size were constructed at the quintiles of the sample distribution of the propensity score, each containing 303 patients. Beginning with the subclass with the highest propensity scores, the five subclasses contained 234 surgical patients, 164 surgical patients, 98 surgical patients, 68 surgical patients and 26 surgical patients, repectively.

For each of the 74 covariates, Table 2 summarizes the balance before and after sublassification. The first row describes the 74 F statistics, that is the squares of the usual

Table 2. *Example of increased balance using subclassification on estimated propensity score as summarized by distributions of F statistics for 74 covariates*

	Minimum	Lower quartile	Median	Upper quartile	Maximum
Treatment main effect without subclassification	4·0	6·8	10·9	16·8	51·8
Treatment main effect with subclassification	0·0	0·1	0·2	0·6	3·6
Treatment by subclass interaction	0·0	0·4	0·8	1·2	2·9

two-sample t statistics, for comparing the surgical group and drug group means of each covariate prior to subclassification. The second and third rows describe F statistics for the main effect of treatment and for interaction in a 2×5, treatments by subclasses, analysis of variance, performed for each covariate. Although there is considerable imbalance prior to subclassification, within the constructed subclasses there is greater balance than would have been expected if treatments had been assigned at random within each subclass.

When subclasses are perfectly homogeneous in $b(x)$, Theorem 2 shows that x has the same distribution for treated, $z = 1$, and control, $z = 0$, units in each subclass. Moreover, by Corollary 4·2, if treatment assignment is strongly ignorable, then the directly adjusted estimate with population total weights is unbiased for the average treatment

effect (1·1). However, in this example, and generally in practice, subclasses will not be exactly homogeneous in the balancing score $b(x)$ that was used in subclassification, so the directly adjusted estimate may contain some residual bias due to x.

The corollary to the following theorem shows that direct adjustment based on a balancing score $b = b(x)$ can be expected to reduce bias in each coordinate of x providing the adjustment reduces the bias in b.

Let I_s be the set of values of a balancing score which make up subclass s ($s = 1, \ldots, S$), so that $b(a) \in I_s$ implies that units with $x = a$ fall in subclass s. Suppose the weight applied to subclass s in direct adjustment is w_s.

THEOREM 7. *The bias in x after direct adjustment for the subclasses $(I_s, s = 1, \ldots, S)$ is*

$$B_s = \sum_{s=1}^{S} w_s \int E(x \mid b) \{\mathrm{pr}(b \mid z = 1, b \in I_s) - \mathrm{pr}(b \mid z = 0, b \in I_s)\}\, db,$$

where $b = b(x)$.

COROLLARY 7·1. *If $E(x \mid b) = \alpha + \beta f(b)$ for some vectors α and β and some scalar valued function $f(\,.\,)$ of b, and if the subclasses are formed using b, then the subclassification is equal per cent bias reducing in the sense that the per cent of bias in x remaining after adjustment is the same for each coordinate of x, namely, 100γ, where*

$$\gamma = \frac{\Sigma_s\, w_s \int f(b) \{\mathrm{pr}(b \mid z = 1, b \in I_s) - \mathrm{pr}(b \mid z = 0, b \in I_s)\}\, db}{\int f(b) \{\mathrm{pr}(b \mid z = 1) - \mathrm{pr}(b \mid z = 0)\}\, db},$$

where the sum is over $s = 1, \ldots, S$.

Proof. Apply Theorem 7 and follow the argument of Corollary 6·1.

In parallel with Corollary 6·2 direct adjustment based on a balancing score within subpopulations defined by x can be shown to be equal per cent bias reducing within those subpopulations.

Subclassification on the propensity score is not the same as any of the several methods proposed by Miettinen (1976): the propensity score is not generally a 'confounder' score. For example, one of Miettinen's confounder scores is

$$\mathrm{pr}(z = 1 \mid r_z = 1, x) \neq \mathrm{pr}(z = 1 \mid x) = e(x).$$

Moreover, under strong ignorability,

$$e(x) = \mathrm{pr}(z = 1 \mid x) = \mathrm{pr}(z = 1 \mid r_1, r_0, x) \neq \mathrm{pr}(z = 1 \mid r_z = 1, x),$$

so strong ignorability does not convert a confounder score into the propensity score.

3·4. *Propensity scores and covariance adjustment*

The third standard method of adjustment in observational studies is covariance adjustment. The point estimate of the treatment effect obtained from an analysis of covariance adjustment for multivariate x is, in fact, equal to the estimate obtained from

univariate covariance adjustment for the sample linear discriminant based on x, whenever the same sample covariance matrix is used for both the covariance adjustment and the discriminant analysis. This fact is most easily demonstrated by linearly transforming x to the sample discriminant and components orthogonal to the sample discriminant which by construction have the same sample mean in both groups. Since covariance adjustment is effectively adjustment for the linear discriminant, plots of the responses r_{1i} and r_{0i} or residuals $r_{ki}-\hat{r}_{ki}$, where $\hat{r}_{ki}$ is the value of r_{ki} predicted from the regression model used in the covariance adjustment, versus the linear discriminant are useful in identifying nonlinear or nonparallel response surfaces, as well as extrapolations, which might distort the estimate of the average treament effect. Furthermore, such a plot is a bivariate display of multivariate adjustment, and as such might be useful for general presentation.

Generally, plots of responses and residuals from covariance analysis against the propensity score $e(x)$ are more appropriate than against the discriminant, unless of course the covariates are multivariate normal with common covariance matrix in which case the propensity score is a monotone function of the discriminant. The reason is that, by Corollary 4·3, if treatment assignment is strongly ignorable, then at each $e(x)$ the expected difference in response $E\{r_1 \mid z = 1, e(x)\} - E\{r_0 \mid z = 0, e(x)\}$ equals the average treatment effect at $e(x)$, namely $E\{r_1 \mid e(x)\} - E\{r_0 \mid e(x)\}$. This property holds for the propensity score $e(x)$ and for any balancing score $b(x)$, but does not generally hold for other functions of x; generally, plots against other functions of x are still confounded by x.

Cases where covariance adjustment has been seen to perform quite poorly are precisely those cases in which the linear discriminant is not a monotone function of the propensity score, so that covariance adjustment is implicitly adjusting for a poor approximation to the propensity score. In the case of univariate x, the linear discriminant is a linear function of x, whereas the propensity score may not be a monotone function of x if the variances of x in the treated and control groups are unequal. Intuitively, if the variance of x in the control group is much larger than the variance in the treated group, then individuals with the largest and smallest x values usually come from the control group. Rubin (1973b, Tables 4 and 6, with $r = 1$ and τ_p as the estimator) has shown that with nonlinear response surfaces, univariate covariance adjustment can either increase the bias or overcorrect for bias dramatically if the variances of x in the treated and control groups differ. Unequal variances of covariates are not uncommon in observational studies, since the subset of units which receives a new treatment is often more homogeneous than the general population. For example, in the observational half of the Salk vaccine trial, the parents of second graders who volunteered for vaccination had higher and therefore less variable educational achievement, x, than parents of control children who were parents of all first and third graders (Meier, 1978).

In the case of multivariate normal x, Rubin (1979, Table 2) has shown that covariance adjustment can seriously increase the expected squared bias if the covariance matrices in treated and control groups are unequal, that is, if the discriminant is not a monotone function of the propensity score. In contrast, when the covariance matrices are equal, so that the discriminant is a monotone function of the propensity score, covariance adjustment removes most of the expected squared bias in the cases considered by Rubin (1979, Table 2). In summary, covariance adjustment cannot be relied upon to perform well unless the linear discriminant is highly correlated with the propensity score.

The authors acknowledge valuable discussions with Arthur P. Dempster, A. Philip Dawid and Roderick J. A. Little on the subject of this paper. This work was partially supported by the U.S. Health Resources Administration, the U.S. Environmental Protection Agency, the Educational Testing Service, the U.S. Army Research Office, and the U.S. National Cancer Institute.

References

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. R. Statist. Soc.* A **128**, 234–55.

COCHRAN, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**, 295–313.

COCHRAN, W. G. & RUBIN, D. B. (1973). Controlling bias in observational studies: a review. *Sankhyā* A **35**, 417–46.

COHN, P. F., HARRIS, P., BARRY, W., ROSATI, R. A., ROSENBAUM, P. R. & WATERNAUX, C. (1981). Prognostic importance of anginal symptoms in angiographically defined coronary artery disease. *Am. J. Cardiol.* **47**, 233–7.

COX, D. R. (1958).*The Planning of Experiments*. New York: Wiley.

COX, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

COX, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113–20.

DAWID, A. P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647–58.

DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc.* B **41**, 1–31.

DEMPSTER, A. P. (1971). An overview of multivariate data analysis. *J. Mult. Anal.* **1**, 316–46.

FISHER, R. A. (1951). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

HAMILTON, M. A. (1979). Choosing a parameter for 2 × 2 table or 2 × 2 × 2 table analysis. *Am. J. Epidemiol.* **109**, 362–75.

KEMPHORNE, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.

MEIER, P. (1978). The biggest public health experiment ever: The 1954 trial of the Salk poliomyelitis vaccine. In *Statistics: A Guide to the Unknown*, Ed. J. M. Tanur, *et al.*, pp. 3–15. San Francisco: Holden Day.

MIETTINEN, O. (1976). Stratification by a multivariate confounder score. *Am. J. Epidemiol.* **104**, 609–20.

RUBIN, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics* **29**. 159–83. Correction (1974) **30**. 728.

RUBIN, D. B. (1973b). The use of matching and regression adjustment to remove bias in observational studies. *Biometrics* **29**, 185–203.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psychol.* **66**, 688–701.

RUBIN, D. B. (1976a). Matching methods that are equal percent bias reducing: Some examples. *Biometrics* **32**. 109–20.

RUBIN, D. B. (1976b). Multivariate matching methods that are equal percent bias reducing: Maximums on bias reduction for fixed sample sizes. *Biometrics* **32**, 121–32.

RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Ed. Statist.* **2**, 1–26.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6**, 34–58.

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Am. Statist. Assoc.* **74**, 318–28.

RUBIN, D. B. (1980a). Discussion of paper by D. Basu. *J. Am. Statist. Assoc.* **75**, 591–3.

RUBIN, D. B. (1980b). Bias reduction using Mahalanobis metric matching. *Biometrics* **36**, 293–8.

RUBIN, D. B. (1983). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In *Research Work of William G. Cochran*, Eds. S. Rao and J. Sedransk. New York: Wiley. To appear.

SNEDECOR, G. W. & COCHRAN, W. G. (1980). *Statistical Methods*. Ames, Iowa: Iowa State University Press.

[*Received March* 1981. *Revised October* 1982]