Biostatistics (2002), **2**, 4, *pp.* 179–193
Printed in Great Britain

Propensity score modeling strategies for the causal analysis of observational data

KATHERINE HUPPLER HULLSIEK*
Division of Biostatistics, School of Public Health, University of Minnesota, 2221 University Avenue S.E., Minneapolis, MN 55414, USA
kathy-h@ccbr.umn.edu

THOMAS A. LOUIS
The RAND Corporation, Arlington, VA 22202

Summary

Propensity score methods are used to estimate a treatment effect with observational data. This paper considers the formation of propensity score subclasses by investigating different methods for determining subclass boundaries and the number of subclasses used. We compare several methods: balancing a summary of the observed information matrix and equal-frequency subclasses. Subclasses that balance the inverse variance of the treatment effect reduce the mean squared error of the estimates and maximize the number of usable subclasses.

Keywords: Bias reduction; Confounding; Observational data; Propensity score methods.

1. Introduction

In comparative studies, where investigators do not control treatment assignment, the directly estimated treatment effect can be strongly affected by confounding (Rubin, 1991; Sommer and Zeger, 1991). Such confounding can also affect randomized studies analysed by other than intent-to-treat, due to differential enrollment in a suite of studies, differential treatment adherence, and informative censoring (Robins and Finkelstein, 2000; Scharfstein *et al.*, 1999; Ellenberg *et al.*, 1992; Larntz *et al.*, 1996; Sommer and Zeger, 1991). Observational data are often readily available, can be representative of the population of interest, are prevalent even in randomized studies, and may be the only feasible source of information. Therefore, valid and robust methods of analysing observational data are needed to take advantage of this information.

Many methods, such as matching and subclassification (Cochran, 1968; Billewicz, 1965; Rubin, 1979), have been used to control for confounding variables (covariates associated with both treatment assignment/selection and outcome) with observational data. To address confounders, Rosenbaum and Rubin (1983a) developed propensity score methods to make inferences about a binary treatment effect with multiple observed covariates and observational data. The propensity score is the conditional probability of assignment to a treatment group given a vector of observed covariates. Rosenbaum and Rubin show that pair matching, subclassification, and covariance adjustment on propensity scores allow for an unbiased estimate of the treatment effect.

Propensity score methods are frequently used to analyse observational data (Rosenbaum and Rubin, 1984, 1985; Rubin, 1997; Rubin and Thomas, 1992, 1996), and the efficacy of subclassification on the

*To whom correspondence should be addressed propensity score has been well documented (Roseman, L. D., 1998, Ph.D. Dissertation, Unpublished; Drake, 1993; Rubin and Thomas, 1992, 1996; Rosenbaum and Rubin, 1983b). However, little attention has been given to the formation of optimal subclasses for propensity score methods; equal-frequency subclasses are generally used. Our goals are to select propensity score subclass boundaries to balance a within-subclass feature (frequency or variance of the estimated treatment effect), and to select the number of subclasses. Section 2 gives an overview of propensity score methods and Section 3 discusses methods and rationales for forming subclass boundaries. Section 4 presents a simulation study to compare the methods. The results are summarized in Section 5.

2. PROPENSITY SCORE METHODS

We use the following notation: Y is a response vector of interest, x is a vector of covariates, and z is a vector of binary treatment assignment indicators, with $z_i = 1$ for treatment and $z_i = 0$ for control. There are three links of interest: between covariates and response ($x \to Y$), between covariates and treatment ($x \to z$), and between treatment and response ($z \to Y$). Robins *et al.* (1992) show that a correctly specified model for either of the $x \to z$ or $x \to Y$ links will provide asymptotically normal and unbiased estimates of the treatment effect. In randomized studies the $x \to z$ link is known: x is statistically independent of z, and thus unadjusted analyses of the data are valid.

In observational studies x and z cannot be assumed independent, and even small to moderate differences in the distribution of the covariates between the treatment groups can have substantial biasing effects. If we have measured all confounders and correctly model the association between covariates and response ($x \to Y$) we can obtain unbiased estimates of the treatment effect without considering the association between covariates and treatment ($x \to z$). However, if we cannot correctly model this association, or if there are unmeasured confounders, then ignoring the association between covariates and treatment assignment can bias the estimate for the treatment effect.

Cochran (1968) showed that adjustment by subclassification is an effective method for removing bias due to a single confounding covariate. Subclassification divides the population into strata so that observations within a stratum have a similar distribution for the confounding variable. When there are many confounding covariates, however, subclassification may be impossible. Propensity score methods allow subclassification to be used with multiple categorical or continuous covariates.

With x denoting the vector of covariates, the propensity score $e(x)$ is the conditional probability of treatment group assignment given a vector of observed covariates: $e(x) = \Pr(z = 1|x)$. Rosenbaum and Rubin (1983a, 1984) and Rubin (1997) show that if treatment assignment is strongly ignorable (i.e. all variables related to both outcome and treatment assignment are included in the vector of measured covariates) and the propensity score is correctly formulated, then treatment assignment and the vector of covariates are conditionally independent given the propensity score. In that case, for units with the same propensity score the two treatment groups have the same distribution of covariates, and a treatment effect can be estimated. Subclassification on the propensity score includes placing each unit in a subclass according to its propensity score and then estimating a treatment effect separately for each subclass. If all units in a subclass have the same propensity score and the propensity score is correct, then the weighted mean of propensity score subclass-specific estimates is an unbiased estimate of the overall treatment effect. Note that in a completely randomized study, the propensity score is identically 0.5, all units are in a single subclass, and the unadjusted analysis is valid.

Of course, generally the propensity score must be estimated and one cannot produce completely homogeneous subclasses. Thus, model-based covariate adjustment along with subclassification on the propensity score is frequently used to account for small within-subclass differences in the covariate distributions for the two treatment groups (Roseman, L. D., 1998, Ph.D. Dissertation, Unpublished;

D'Agostino, 1998; Rubin and Thomas, 2000). To implement the approach, one needs to estimate the propensity score, determine the number of propensity score subclasses, and form the subclass boundaries. Though Drake (1993) finds that using subclassification with five equal-frequency subclasses based on an estimated propensity score is effective, the number of subclasses should depend on sample size and empirical assessment of bias reduction. With a large dataset it may be desirable to form more than five propensity score subclasses. Alternatively, with a small dataset it may not be possible to form more than two or three propensity score subclasses and still obtain approximate balance of the covariate distribution within subclasses (Rubin, 1997).

3. Determining propensity score subclasses

To obtain an overall treatment effect estimate Rosenbaum and Rubin (1983a) propose a weighted average of propensity score subclass-specific estimates, with weights equal to the fraction of the population within a subclass. However, if equal-frequency subclasses are formed we may find that most observations in the lowest subclass have $z_i = 0$, while most observations in the upper subclass have $z_i = 1$. That imbalance of treatment assignment indicates that there may be insufficient overlap to estimate an effect. Even if the treatment effect is estimable, the results may be highly variable. With weights equal to the fraction of the population, those subclass-specific estimates with high variance would be given equal weight in the calculation of the overall treatment effect estimate.

If our primary interest is in the overall treatment effect, another weighting scheme has weights equal to the inverse variance of the treatment effect, so that a subclass estimate with high variability will be given less weight in the overall treatment effect calculation. But care must be taken with weights equal to the inverse variance of the treatment effect, as any highly variable estimates would be effectively eliminated from the calculation of the overall estimate. Though five subclasses may have been used, the effective bias control may only be equivalent to three optimally constructed subclasses.

We investigate propensity score subclasses formed by balancing on the inverse variance of the subclass-specific treatment effects, rather than by the number of observations within a subclass. If we form propensity score subclasses this way and then estimate an overall treatment effect using the weighted mean of subclass-specific estimates (weighted by the inverse variance of the treatment effect) each subclass will be given similar weight. Our method should use more of the information in a dataset by forcing balance between the treatment groups within each subclass, and will preserve the number of effective propensity score subclasses. We contrast this method with using equal-frequency subclasses.

In forming subclass boundaries we order units by their propensity score and then select boundaries. We have two goals. To reduce bias due to confounding variables, we want subclasses in which the propensity score is similar for the treatment groups, arguing for a large number of subclasses. And, to maximize the effective bias control, we want subclasses to have similar variances of the estimated treatment effect.

3.1 *Balancing subclasses on estimated variance*

Let X be a design matrix with the ith row equal to $[1\ z_i\ x_{i1}\ x_{i2}]$, $i = 1, 2, \ldots, N$, where z_i is the binary treatment assignment indicator and x_{i1} and x_{i2} are continuous covariates. Assume that the mean of the response Y is related to $X\beta$, where $\beta = (\beta_0,\ \beta_z,\ \beta_x\ \beta_{x2})^T$. To balance propensity score subclasses using the inverse variance of the treatment effect across G subclasses assuming N observations, order the propensity scores from smallest to largest, labeling the corresponding observations in the design matrix $n_1, n_2, \ldots, n_N$. Next find an initial target value for balancing the subclasses according to the inverse variance by first calculating $\widehat{\text{Var}}^{-1}(\hat{\beta}_z)$ for each subclass, $g = 1, 2, \ldots, G$ using equal-frequency subclasses. Use the average of the equal-frequency subclass-specific inverse variances,

$\frac{1}{G}\sum_{g=1}^{G}\widehat{\mathrm{Var}}_g^{-1}(\hat{\beta}_z)$, as the initial target value. Find boundaries for the G subclasses to form the subclasses in the order $1,\ G,\ 2,\ (G-2),\ 3,\ (G-3),\ldots,(G/2)$. Forming subclasses by alternating between the left and right extremes forces the imbalance towards the middle subclasses. To find a boundary point for a subclass formed from the left, place it at $\max(n_i)$ meeting the condition that the information contained in the newly formed subclass is less than or equal to the target value. For a subclass formed by coming in from the right, place it at $\min(n_i)$ meeting the same condition. Finally, check the summary for the middle subclass. If it is not approximately equal to the summaries for the more extreme subclasses (according to a predetermined tolerance) then adjust the target value and iterate back to the previous step.

The method for calculating $\widehat{\mathrm{Var}}_g^{-1}(\hat{\beta}_z)$ will depend on the data. For example, if $Y \sim MVN(X\beta, \sigma^2 I)$, the observed information matrix $I(\beta)$ is proportional to $(X^T X)$, where X is the known design matrix. To balance subclasses according to the inverse variance of the treatment effect, $\widehat{\mathrm{Var}}^{-1}(\hat{\beta}_z)$ is computed as the appropriate element of $I^{-1}(\beta)$. For nonlinear response data, the observed information matrix cannot generally be given in closed form. In that case, to balance the inverse variance of the treatment effect across subclasses, maximum-likelihood estimates for β and the associated estimated covariance matrix are found using iterative procedures. The appropriate element of the estimated covariance matrix is then used to form the propensity score subclasses.

4. SIMULATION STUDY

We use synthetic datasets with propensity score subclass boundaries formed by two different methods: balancing subclasses using the inverse variance $[\widehat{\mathrm{Var}}^{-1}(\hat{\beta}_z)]$ of the treatment effect as a summary (method V) and forming equal-frequency propensity score subclasses (method F). For each dataset we estimate propensity scores, form subclass boundaries for the true and estimated propensity scores according to both methods, and finally compare the bias, variance, and mean squared error (MSE) of the overall treatment effects.

Linear and nonlinear response datasets Y_{ijk} are generated according to the following factorial: $i = 1$ if the true propensity score vector depends on x_1 and 2 otherwise; $j = 1$ if the response Y depends on x_1 and 2 otherwise; $k = 1$ if there is a linear relationship between Y and $X\beta$ and $k = 2$ if the relationship is nonlinear. All response datasets depend on the generated treatment assignment vector z; none depend on x_2.

In order to build the synthetic datasets we first generate the design matrix X. For each observation, we generate two uncorrelated bivariate normal covariates with mean zero and variance equal to one, and assemble them into the $N \times 2$ matrix $\tilde{X} = [x_1\ x_2]$, where $x_k = (x_{1k},\ x_{2k},\ldots,x_{Nk})^T$ for $k = 1,\ 2$. A simulation study (Gu and Rosenbaum, 1993) found that propensity scores behave very differently with 20 covariates than with two covariates. However, the two-covariate case plays quite a general role. In theory, a dataset can be partitioned into two components: those covariates truly associated with treatment selection, and those that are not so associated. The covariates x_1 and x_2 here can each be thought of as linear combinations of several regressors.

To generate the treatment assignment vector we consider two different types of propensity score vectors, and denote them as 'true' propensity score vectors. The first true propensity score vector, generated with the linear logistic model $e_{t1}(\tilde{X}) = \frac{\exp(\gamma_1 x_1)}{1+\exp(\gamma_1 x_1)}$, is used to form the treatment assignment vector for datasets Y_{11k} and Y_{12k}. The second true propensity score vector is represented by the $e_{t2}(\tilde{X})$, which varies across observations but does not depend on x_1 or x_2; its components are independent draws from the Uniform(0, 1) distribution. That propensity score vector is used to generate the treatment assignments for datasets Y_{21k} and Y_{22k}. For each of the true propensity score vectors, the associated treatment assignment vector is formed by comparing the propensity score vector with a vector of

Table 1. *Summary of simulation parameters*

Dataset	Relationship with $X\beta$*	Propensity score	γ	β^T
Y_{111}	†	$\odot$	2.0	$(1, -1, 1, 0)$
Y_{121}	†	$\odot$	2.0	$(1, -1, 0, 0)$
Y_{211}	†	$\oplus$	NA	$(1, -1, 1, 0)$
Y_{221}	†	$\oplus$	NA	$(1, -1, 0, 0)$
Y_{112}	§	$\odot$	0.5	$(1.5, -1, 1, 0)$
Y_{122}	§	$\odot$	0.5	$(1.5, -1, 0, 0)$
Y_{212}	§	$\oplus$	NA	$(1.5, -1, 1, 0)$
Y_{222}	§	$\oplus$	NA	$(1.5, -1, 0, 0)$

* Design matrix $X_{N\times 1}$, with ith row = $[1\ z_i\ x_{i1}\ x_{i2}]$; $\beta^T = (\beta_0\ \beta_z\ \beta_{x1}\ \beta_{x2})$.

† $MVN(X\beta, \sigma^2 I)$, with $\sigma^2 = 0$ and 0.224.

§ $h(y|X) = \frac{\delta}{\alpha(X)}\left(\frac{t}{\alpha(X)}\right)^{\delta-1}$; $\alpha(X) = \exp(X\beta)$.

$\odot$ $e_{t1}(\tilde{X}) = \frac{\exp(\gamma_1 x_1)}{1+\exp(\gamma_1 x_1)}$.

$\oplus$ $e_{t2}(\tilde{X})$: no dependence on X.

$\hat{e}_{10}(\tilde{X})$: estimated propensity score based on x_1 only.
$\hat{e}_{11}(\tilde{X})$: estimated propensity score based on x_1 and x_2.

Uniform(0, 1) draws. For each observation i, if the ith element of the true propensity score vector is less than or equal to the corresponding number in the comparison vector then that observation has $z_i = 1$, otherwise it has $z_i = 0$.

For the linear datasets the true underlying relationship is $Y \sim MVN(X\beta, \sigma^2 I)$, where the ith row of $X = [1\ z_i\ x_1\ x_2]$ and $\beta = (\beta_0,\ \beta_z,\ \beta_x,\ \beta_{x2})^T$. To generate the treatment assignment vector for datasets Y_{111} and Y_{121} we use $e_{t1}(\tilde{X})$ with $\gamma_1 = 2$. For datasets Y_{111} and Y_{211} we use $\beta^T = (1,\ -1,\ 1,\ 0)$; for datasets Y_{121} and Y_{221} we use $\beta^T = (1,\ -1,\ 0,\ 0)$. For all cells in the factorial design we generate datasets with $\sigma^2 = 0$ and with $\sigma^2 = 0.224$.

For the nonlinear response datasets we draw outcomes from a Weibull distribution with hazard function $h(y|X) = \frac{\delta}{\alpha(X)}\left(\frac{t}{\alpha(X)}\right)^{\delta-1}$, where $\delta > 0$ is the shape parameter and $\alpha(X) > 0$ is the scale parameter. We use an increasing hazard function with $\alpha(X) = \exp(X\beta)$. We assume we know the exact lifetimes of all observations, and generate only uncensored data. The propensity score vector $e_{t1}(\tilde{X})$ with $\gamma_1 = 0.5$ is used to generate the treatment assignment vector for nonlinear response datasets Y_{112} and Y_{122}. For datasets Y_{112} and Y_{212} we use $\beta^T = (1.5,\ -1,\ 1,\ 0)$; for datasets Y_{122} and Y_{222} we use $\beta^T = (1.5,\ -1,\ 0,\ 0)$. Table 1 summarizes the simulation parameters.

We note that the propensity score vectors used to generate the treatment assignment vectors may define two extreme cases: with $e_{t2}(\tilde{X})$ the covariates x_1 and x_2 play no role in treatment assignment. With $e_{t1}(\tilde{x})$, $\gamma = 2$, and $x_1 \sim N(0, 1)$ for the linear datasets, the odds of receiving treatment at $x_1 = 2$ are much greater than at $x_1 = -2$. With the nonlinear datasets and $\gamma = 0.5$, the odds of receiving treatment at $x_1 = 2$ are 7.4 times greater than at $x_1 = -2$, a more reasonable association.

To investigate precisely the magnitude and direction of any treatment effect bias, the first simulation for each type of response data (linear and nonlinear) consists of one sample of size 2000 for each response dataset, with propensity score subclass boundaries formed for up to 10 subclasses. To consider the trade-off between bias and variance inflation with smaller samples, another simulation consists of 1500 samples of size 100 for each response dataset, with up to five propensity score subclasses.

For each dataset logistic regression is used to estimate two propensity scores: $\hat{e}_{10}(\tilde{X})$, based on x_1 only, and $\hat{e}_{11}(\tilde{X})$, based on x_1 and x_2. Propensity score subclass boundaries are determined separately using the vectors of true and estimated propensity scores for the two methods V and F, where the true propensity score vector is defined as the one used to generate the treatment assignment. For method V, propensity score subclasses are balanced according to the inverse variance of the subclass-specific treatment effects, from the model assuming that the response Y depends on treatment only.

Once the propensity score subclasses are formed, estimates of the treatment effect are obtained for each dataset by fitting the model depending on treatment assigment only separately to each subclass. For the nonlinear response models, the population estimands are assumed constant over subclasses. Weighted averages of the subclass-specific estimates are calculated to estimate the overall treatment effect, with weights equal to the inverse variance of the subclass-specific treatment effect estimates. If an overall treatment effect is not estimable using all j subclasses, then the estimate with $j - 1$ subclasses is used instead. The overall treatment effect estimate ignoring the propensity score is compared to the weighted mean estimates from the two methods.

Models assuming that the response depends on treatment assignment only ($Y \sim \beta_0 + \beta_z z$ for the linear case, and the exponential model with $\log \alpha(X) = \beta_0 + \beta_z z$ for the nonlinear datasets) are both poor fits for datasets Y_{11k} and Y_{21k}, which do depend on x_1. With those datasets we investigate whether subclassification on the propensity score helps reduce the bias and variance of the estimated treatment effect. The models are good fits for datasets Y_{12k} and Y_{22k}, which do not depend on x_1. There we determine if there is a ‘penalty’ for subclassifying on the propensity score when it is not necessary.

4.1 *Results*

4.1.1 *Linear response with $N = 2000$.* We first consider the linear response datasets Y_{111}, where both the propensity score and the response depend on x_1. We have two cases: response data generated with $\sigma^2 = 0$ and with $\sigma^2 \neq 0$. With these datasets we expect the unadjusted treatment effect estimate to be biased because treatment assignment is correlated with x_1, and x_1 is not in the response model. Since the propensity score does depend on x_1 we expect subclassification on the propensity score to reduce the bias. Figure 1 shows the bias of the overall treatment effect estimates by the number of subclasses used to calculate the estimate for the dataset generated with $\sigma^2 = 0$ and subclass boundaries formed using method V with the true propensity score. The dotted lines represent 95% confidence intervals based on the standard errors of the weighted means. The bias of the estimates decreases as the number of subclasses is increased; the variability of the estimates also generally decreases. Using only two propensity score subclasses reduces bias by over 50%; using at least five propensity score subclasses produces generally unbiased estimates with low variability. The bias of the overall estimates are similar when method F is used (not shown) to form the subclass boundaries. The bias patterns are the same when the datasets are generated with $\sigma^2 \neq 0$, but the variability is increased.

Table 2 gives the bias, variance, and MSE of the overall treatment effect estimate for a subset of the number of propensity score subclasses used to calculate the estimate from the datasets generated with $\sigma^2 = 0$. For dataset Y_{111} and subclass boundaries formed according to the true propensity score, the bias decreases for both methods as the number of propensity score subclasses is increased. With either method V or F, when more than six propensity score subclasses are used to calculate the treatment effect the variance of the estimate is less than the variance with no subclassification, and there is a hundredfold reduction in the bias. When three or more subclasses are used, the estimates using method F have more bias but less variability than those using method V. The MSE decreases as the number of subclasses is increased with both methods. Subclassification on the propensity score gives at least a sixfold reduction of the MSE compared to no subclassification. The MSEs are similar for the two methods.

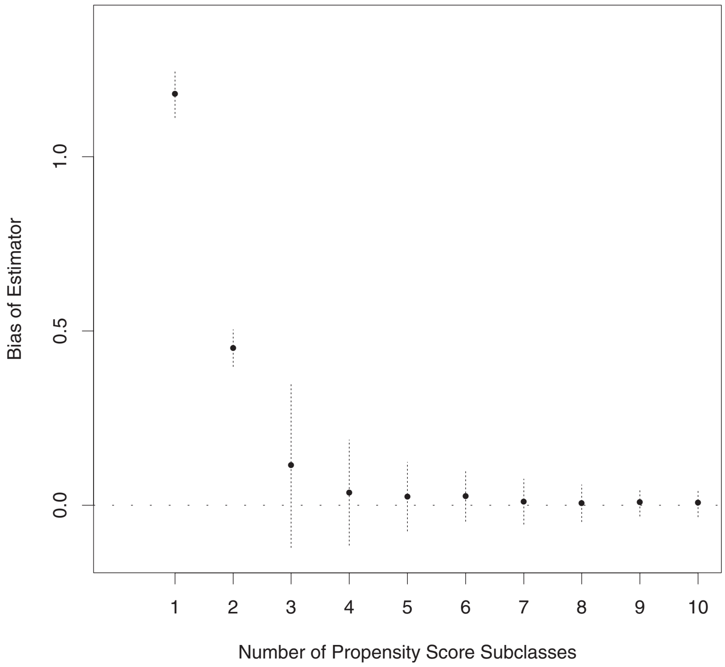

Fig. 1. Bias (with 95% confidence intervals) of the overall treatment effect estimates by the number of propensity score subclasses used to calculate the estimate for the linear model with dataset Y_{111}, $\sigma^2 = 0$, $N = 2000$, and up to ten subclasses. Method V and the true propensity score are used to form the subclasses boundaries.

For dataset Y_{111}, using an estimated propensity score to form the subclass boundaries does not change the patterns of bias and MSE (data not shown). Subclass boundaries formed according to $\hat{e}_{10}(\tilde{X})$ are identical in this large dataset to those using the true propensity score, so the bias and MSE results are the same. The MSEs of the weighted means with subclasses formed according to $\hat{e}_{11}(\tilde{X})$ are only slightly larger than when using the true propensity score.

Although our primary interest is in the overall treatment effect estimates, we are interested in seeing how the subclass-specific estimates contribute to the overall estimate. Figure 2 shows the subclass-specific estimates from dataset Y_{111} generated with $\sigma^2 = 0$ when seven and ten propensity score subclasses are used with the true propensity score and both methods. The bias of the overall treatment effects is shown to the left of the origin for methods F and V, respectively. As with the overall estimates, both the bias and the variability of the subclass-specific estimates decrease as the number of subclasses increases with method V. Using five (not shown) or more propensity score subclasses produces generally unbiased subclass-specific estimates except for the extreme propensity score subclasses. Those subclasses are inherently the most unstable because most of the observations usually have the same treatment assignment. The bias is similar when method F is used to form propensity score subclasses, but when using seven or more subclasses the standard errors of the extreme subclass estimates with method F are at least twice as large as when method V is used. The large variance for the extreme subclasses with method F effectively eliminates them from the overall treatment effect estimate, since the inverse variance is used as the weight. The results from the dataset generated with $\sigma^2 \neq 0$ have a similar bias pattern but increased variability of the estimates.

With dataset Y_{211} the propensity score does not depend on x_1, so subclassification on the propensity

Table 2. *Bias* (*absolute value* $\times$ 10^{-3})*, variance* ($\times 10^{-4}$)*, and MSE* ($\times 10^{-4}$) *of the overall treatment effect estimates by the number of propensity score subclasses formed to calculate the estimate. Results are from the linear response model with datasets* Y_{111} *and* Y_{211}*,* $\sigma^2 = 0$*, and* $N = 2000$

				Number of subclasses					
Data	PS	Method		1	2	3	5	7	10
Y_{111}	$e_{t1}(\tilde{X})$	V	Bias	1 181	451	115	24	10	7
			Var	12	7	144	25	10	4
			MSE	13 950	2040	276	31	12	5
		F	Bias	1 181	451	147	73	22	20
			Var	12	6	81	20	6	2
			MSE	13 950	2042	299	73	11	6
Y_{211}	$e_{t2}(\tilde{X})$	V	Bias	26	35	30	32	40	38
			Var	19	28	93	71	170	159
			MSE	26	40	102	81	186	173
		F	Bias	26	30	27	39	48	43
			Var	19	42	30	38	108	65
			MSE	26	51	38	52	130	83
	$\hat{e}_{10}(\tilde{X})$	V	Bias	26	9	13	5	2	2
			Var	19	3	15	7	4	2
			MSE	26	4	17	7	5	2
		F	Bias	26	12	12	2	4	2
			Var	19	3	13	7	4	2
			MSE	26	4	14	7	4	1

score cannot be expected to reduce bias of the treatment effect estimates. On the other hand, subclassification should not significantly increase the bias since x_1 and treatment assignment were generated without correlation. When subclasses are formed according to the true propensity score $e_{t2}(\tilde{X})$ the results are in contrast to what was seen with dataset Y_{111}; the MSE generally increases as the number of subclasses is increased, and is usually greater than the MSE with no subclassification (Table 2). The treatment effect estimates are biased because of the randomly non-zero correlation between the realized treatment assignment and x_1. If we simply look at the MSE of the estimates, method F may appear to perform better. However, the subclass-specific estimates (not shown) are highly variable with method F for the extreme subclasses, effectively removing them from the calculation of the overall treatment effect. Results are similar when subclasses are formed according to $\hat{e}_{11}(\tilde{X})$ (not shown). Dramatic reductions in the MSE are seen when the subclass boundaries are formed using propensity score $\hat{e}_{10}(\tilde{X})$, estimated from the propensity score model adjusting only for x_1, because that model adjusts for the empirical correlation between treatment assignment and x_1. With subclasses formed according to a well estimated propensity score, the results are similar for both methods. The MSEs for dataset Y_{211} generated with $\sigma^2 \neq 0$ follow a similar pattern.

Datasets Y_{121} and Y_{221} are generated with no dependence on x_1. Thus the linear model $Y \sim \beta_0 + \beta_z z$ has a good fit, and subclassification on the propensity score is unnecessary. For those datasets all estimates are close to unbiased with small variance estimates. The maximum MSE for datasets Y_{121} and Y_{221} with up to 10 propensity score subclasses is 26×10^{-4}.

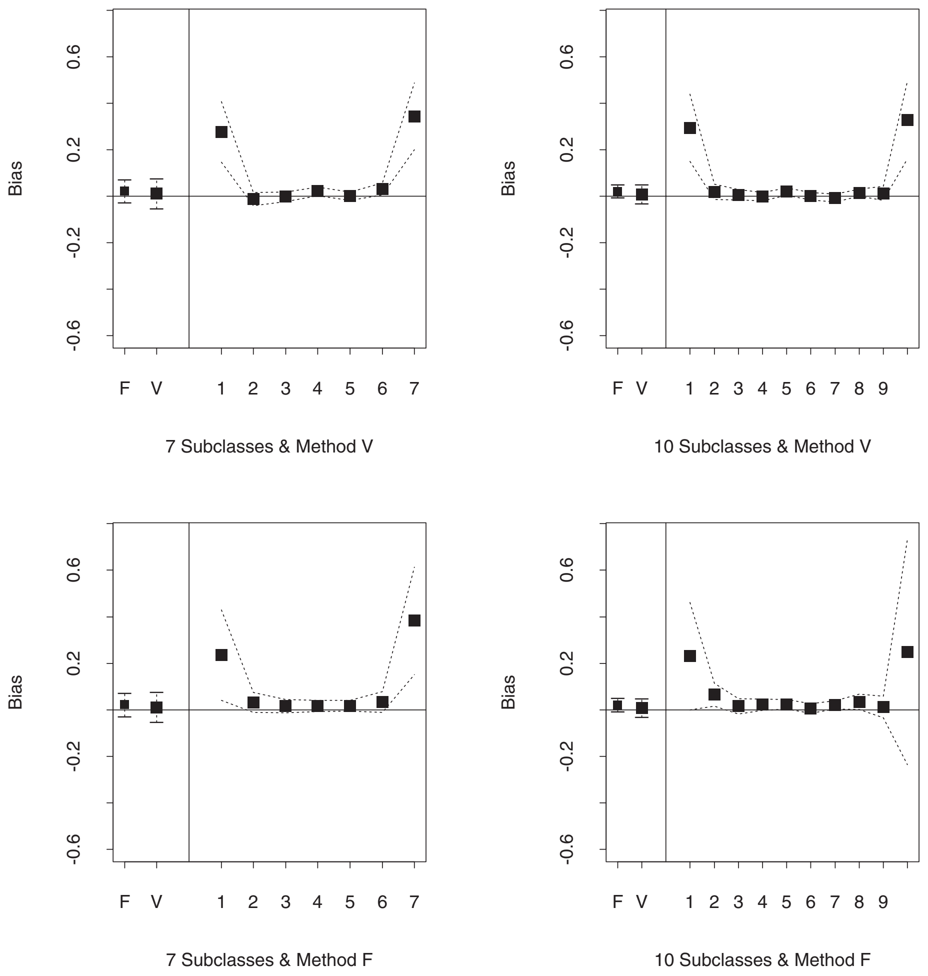

Fig. 2. Subclass-specific bias for seven and ten subclasses with the linear dataset Y_{111}, $\sigma^2 = 0$, and $N = 2000$. The true propensity score is used with methods V and F to form subclasses. The bias of the overall treatment effects are shown to the left of the origin for methods F and V, respectively.

4.1.2 *Linear response with* $N = 100$. Simulation results with $N = 100$ have the same patterns of bias as the $N = 2000$ simulation, although the variability and MSE of the estimates are larger. Table 3 gives the simulation average bias, variability, and MSE for the weighted means of subclass-specific treatment effect estimates for datasets Y_{111} and Y_{211}.

For dataset Y_{111}, the MSE decreases as the number of subclasses is increased when both the true and estimated propensity scores are used to form the boundaries. The reduction in the MSE compared to no propensity score subclassification is similar to the result with $N = 2000$; using only two subclasses gives a sixfold reduction, while using five subclasses reduces the MSE dramatically. Using the estimated propensity score only slightly increases the MSE compared to using the true propensity score (data not shown). With more than two propensity score subclasses, the MSE for method V is lower than the MSE for method F because the algorithm for determining propensity score subclasses with method V ensures design matrices that are of full-column rank. For dataset Y_{111} and $N = 100$, with method F it was

Table 3. *Bias* (*absolute value* × 10^{-3}), *variance* (× 10^{-3}), *and MSE* (× 10^{-4}) *of the average overall treatment effect estimates by the number of propensity score subclasses formed to calculate the estimate. Results are from the linear response model with datasets* Y_{111} *and* Y_{211}, $\sigma^2 = 0$, *and* $N = 100$

				Number of subclasses				
Data	PS	Method		1	2	3	4	5
Y_{111}	$e_{t1}(\tilde{X})$	V	Bias	2414	949	163	70	26
			Var	103	57	89	40	25
			MSE	60 250	10 050	1390	540	340
		F	Bias	2 414	941	335	221	180
			Var	103	51	56	42	38
			MSE	60 250	9 820	2080	1280	1 060
Y_{211}	$e_{t2}(\tilde{X})$	V	Bias	20	21	18	17	22
			Var	162	215	487	274	1 034
			MSE	3 210	4 260	7370	9860	13 140
		F	Bias	20	13	19	13	7
			Var	162	219	470	703	907
			MSE	3 210	4 300	7120	9560	11 720
	$\hat{e}_{10}(\tilde{X})$	V	Bias	8	9	1	1	1
			Var	162	54	43	33	29
			MSE	3 210	1 190	680	460	400
		F	Bias	8	12	2	0	1
			Var	162	56	41	32	27
			MSE	3 210	1 180	640	430	350

impossible to form four subclasses in over 15% of the simulations; in over 30% of the simulations we were unable to form five subclasses. In those cases the weighted mean with fewer propensity score subclasses, which generally has higher bias and variability, was included to calculate the simulation average. With method V we were able to form five subclasses with all simulations.

For dataset Y_{211} with $\sigma^2 = 0$, Table 3 shows again the divergent pattern of MSE using the true versus estimated propensity scores. When the true propensity score is used to form subclass boundaries, the MSE increases as the number of subclasses is increased, and is greater than the MSE with no subclassification. Results are similar when subclasses are formed according to $\hat{e}_{11}(\tilde{X})$ (not shown). With the estimated propensity score $\hat{e}_{10}(\tilde{X})$ the MSE is much smaller overall, and decreases as the number of subclasses is increased. Method F failed with dataset Y_{211} for 6% of the simulations with four subclasses, and 20% of the simulations with five subclasses. No simulations failed with method V.

The estimates from datasets Y_{121} and Y_{221} generated with $\sigma^2 = 0$, where the response model has a good fit, are close to unbiased and have small variances. The average of the MSEs for the weighted means of subclass-specific treatment effect estimates for datasets Y_{121} and Y_{221} generated with $\sigma^2 \neq 0$ range from 40×10^{-4} when no subclassification is used up to 180×10^{-4} when five subclasses are formed. In contrast, with $N = 2000$ the maximum MSE for datasets Y_{121} and Y_{221} and up to five subclasses is 11×10^{-4}.

4.1.3 *Nonlinear response models.* Figure 3 shows the bias of the overall treatment effect estimates by the number of propensity score subclasses formed to calculate the estimate for datasets Y_{112} and Y_{212} with

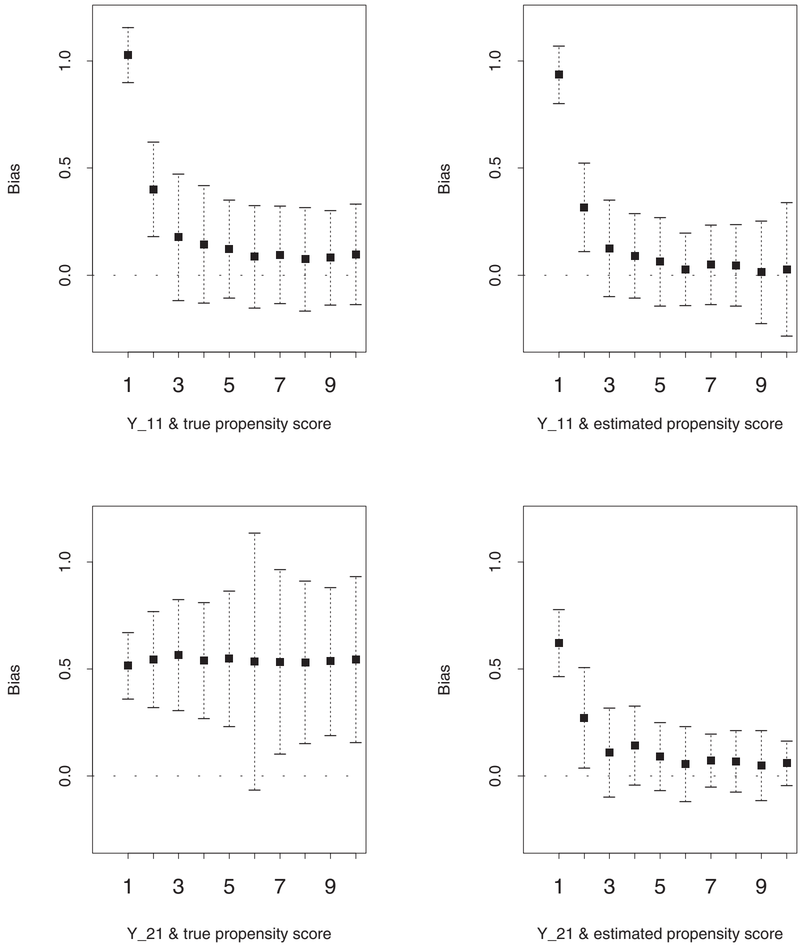

Fig. 3. Bias (with 95% confidence intervals) of the overall treatment effect estimates by the number of propensity score subclasses used to calculate the estimate for the nonlinear datasets Y_{112} and Y_{212} with method V, $N = 2000$, and subclasses formed with the true and estimated propensity scores.

method V and $N = 2000$. For dataset Y_{112}, the bias decreases as the number of subclasses is increased, while the variance of the estimates remains fairly constant. The bias and the variability are similar when the true and estimated propensity scores are used to form subclass boundaries. By contrast, for dataset Y_{212}, the bias is not affected by the number of propensity score subclasses formed to calculate the estimate when subclasses are formed according to the true propensity score, but decreases almost to zero when the estimated propensity score is used.

Figure 4 shows the subclass-specific estimates from dataset Y_{112} when subclass boundaries are formed with method V and the true propensity score for two, four, seven, and ten propensity score subclasses. As with the linear case, the bias tends to decrease as the number of subclasses is increased, although the subclasses with the highest propensity scores remain biased. Unlike the linear case, here the variance increases as the number of subclasses is increased. The subclass-specific results are similar when method F and the estimated propensity scores are used to form the subclass boundaries (not shown).

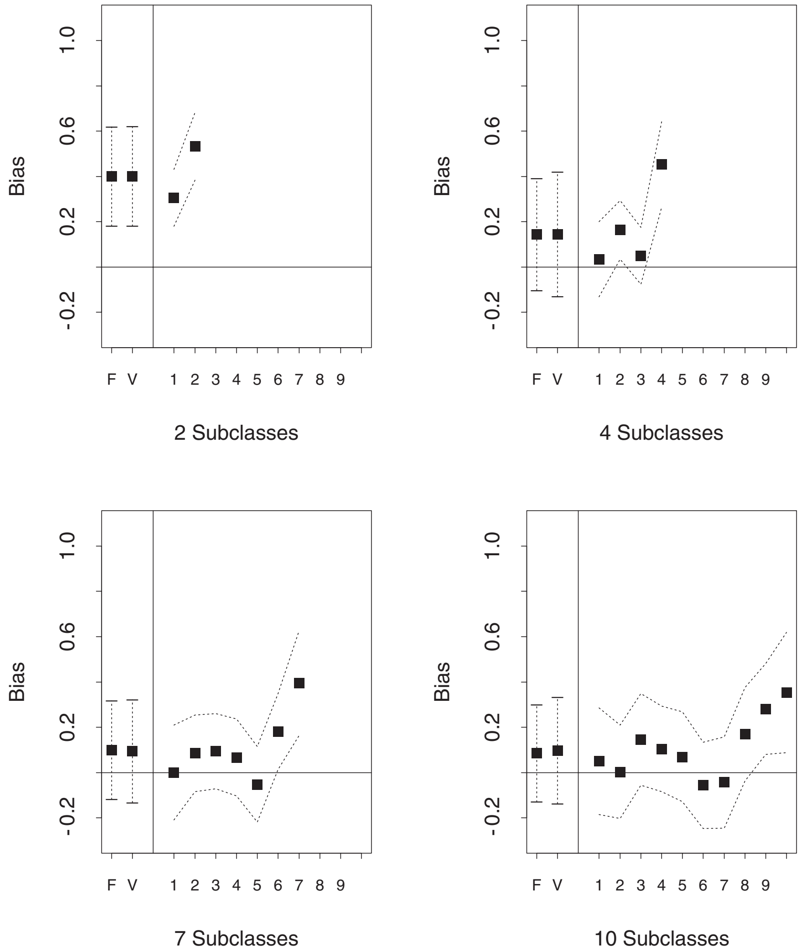

Fig. 4. Subclass-specific bias by the number of subclasses used from the nonlinear response model with dataset Y_{112} and $N = 2000$. Method V and the true propensity score are used to form subclasses. The bias of the overall treatment effect is to the left of the origin for methods F and V, respectively.

With dataset Y_{212}, where the true propensity score $e_{t2}(\tilde{X})$ was generated independently of the covariate x_1, the subclass-specific results are similar to the linear case (not shown). The bias remains constant and the variability of the estimates increases as the number of subclasses is increased with subclassification according to $e_{t2}(\tilde{X})$. The estimates from the extreme propensity score subclasses are very variable with method F, effectively removing those subclasses from the weighted mean calculation. When the subclassification is on the estimated propensity score $\hat{e}_{10}(\tilde{X})$, which adjusts for the empirical correlation between treatment assignment and x_1, the bias and variability of the subclass-specific treatment effect estimates are both small.

As with the linear response models, results from the simulations with $N = 100$ have the same patterns of bias as the $N = 2000$ simulation, with increased variability and MSE of the estimates. Table 4 gives the bias, variance, and MSE of the average overall treatment effect estimates by the number of propensity score subclasses formed to calculate the estimate for dataset Y_{112}. Using propensity score subclasses to

Table 4. *Bias (absolute value* $\times\ 10^{-3}$*), variance* $(\times\ 10^{-3})$*, and MSE* $(\times\ 10^{-3})$ *of the average overall treatment effect estimates by the number of propensity score subclasses formed to calculate the estimate. Results are from the nonlinear response model, dataset* Y_{112}*, and* $N = 100$*, using the true and estimated propensity scores to form subclasses. Without subclassification on the propensity score the bias treatment effect is* 1896×10^{-3}*, the variance is* 65×10^{-3}*, and the MSE is* 3751×10^{-3}

			Number of subclasses			
PS	Method		2	3	4	5
$e_{t1}(\tilde{X})$	V	Bias	659	229	45	35
		Var	153	214	292	292
		MSE	653	346	433	422
	F	Bias	647	263	204	192
		Var	143	120	135	145
		MSE	663	284	284	293
$\hat{e}_{11}(\tilde{X})$	V	Bias	665	225	48	41
		Var	152	236	310	308
		MSE	661	376	462	448
	F	Bias	654	269	217	206
		Var	146	126	144	152
		MSE	644	298	301	308

calculate the overall treatment effect estimate, the bias is reduced fivefold with method V rather than method F when four or more subclases are formed, primarily because method V preserves the number of allowable subclasses. With datasets Y_{112} and Y_{212}, less than 4% of the simulations failed with method V to produce five subclasses. With method F, we were unable to form three, four, and five subclasses in 13, 51, and 77% of the simulations, respectively.

With datasets Y_{122} and Y_{222} the response models were a good fit, making subclassification on the propensity score unneccessary. The bias and variability of the estimates are low for propensity score subclasses formed by both methods and when either the true and estimated propensity score is used to determine subclass membership (data not shown). For these datasets, method V failed to form five subclasses in 1% of the simulations, while method F failed in over 20% of the simulations.

5. Discussion

Simulations indicate that when propensity scores can be expected to help, subclassification on a well estimated propensity score performs best when boundaries are formed by balancing the inverse variance of the treatment effect, because that method preserves the number of effective subclasses. Forming propensity score subclasses via equal frequencies within subclasses often leads to highly variable estimates for the extreme subclasses; those extreme subclasses therefore contribute little to the overall weighted mean. Using the inverse variance of the treatment effect as a balancing summary for propensity score subclasses is computationally more intensive, but the method ensures that all subclasses are included in the overall estimate of the treatment effect, and with similar weights.

To achieve maximal bias reduction in the context of a given model specfication for the propensity score and the covariance adjustment model, we suggest using as many propensity score subclasses as possible. That is, push onward by increasing the number of propensity score subclasses until the treatment effect cannot be estimated in at least one subclass, and then reduce the number of subclasses by one. Though the plot of subclass-specific treatment effects will be noisy, the inverse-variance-weighted overall treatment effect should be maximally bias-reduced. Using our method of forming propensity score classes will, in general, allow for considerably more subclasses than by using equal-frequency subclasses. The foregoing approach will not be ideal for diagnosing variation in subclass-specific true treatment effects, because at the maximal subclassification the standard errors of the subclass-specific estimated treatment effects will be large.

Another method for determining the number of propensity score subclasses to use is a visual inspection of a plot of estimated treatment effects by the number of propensity score subclasses formed to calculate the estimate, similar to figures 1 and 3. When subclassification on the propensity score can be expected to help, the estimated treatment effects will move towards the truth as the number of subclasses is increased.

The simulations included here have a number of limitations that warrant future work. The propensity score vectors used to generate treatment assignment for the linear response datasets force the covariate x_1 to have either an extremely strong association with treatment, or no association at all. Using a more moderate propensity score vector deserves further study. The simulations also did not address how robust the method of using the inverse variance of the treatment effect as a balancing summary for propensity score subclasses is to model misspecification.

ACKNOWLEDGEMENTS

The authors wish to thank Dr James Hodges for his helpful suggestions, and two editorial reviewers for their comments. This research was supported in part by NIH grants 5U01AI42170-07 and 5T32AI07432-05.

REFERENCES

BILLEWICZ, W. Z. (1965). The efficiency of matched samples: an empirical investigation. *Biometrics* **21**, 623–643.

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

D'AGOSTINO, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* **17**, 2265–2281.

DRAKE, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* **49**, 1231–1236.

ELLENBERG, S. S., FINKELSTEIN, D. M. AND SCHOENFELD, D. A. (1992). Statistical issues arising in AIDS clinical trials. *Journal of the American Statistical Association* **87**, 562–569.

GU, X. S. AND ROSENBAUM, P. (1993). Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* **2**, 405–420.

LARNTZ, K., NEATON, J. D., WENTWORTH, D. N. AND YURIK, T. (1996). Data analysis issues for protocols with overlapping enrollment. *Statistics in Medicine* **15**, 2445–2453.

ROBINS, J. M. AND FINKELSTEIN, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trail with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56**, 779–788.

ROBINS, J. M., MARK, S. AND NEWEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–495.

ROSENBAUM, P. R. AND RUBIN, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

ROSENBAUM, P. R. AND RUBIN, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society* Series B **45**, 212–218.

ROSENBAUM, P. R. AND RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516–524.

ROSENBAUM, P. R. AND RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33–38.

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318–328.

RUBIN, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **47**, 1213–1234.

RUBIN, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internet Medicine* **127**, 757–763.

RUBIN, D. B. AND THOMAS, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* **79**, 797–809.

RUBIN, D. B. AND THOMAS, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics* **52**, 249–264.

RUBIN, D. B. AND THOMAS, N. (2000). Combining propensity score matching with additional adjustment for prognostic covariates. *Journal of the American Statistical Association* **95**, 573–585.

SCHARFSTEIN, D. O., ROTNITZKY, A. AND ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1146.

SOMMER, A. AND ZEGER, S. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10**, 45–52.

[*Received 22 September, 2000*; *first revision 22 March, 2001*; *second revision 30 April, 2001; accepted for publication 8 May, 2001*]