Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Estimating and Using Propensity Scores with Partially Missing Data

Ralph B. D'Agostino Jr. & Donald B. Rubin

To cite this article: Ralph B. D'Agostino Jr. & Donald B. Rubin (2000) Estimating and Using Propensity Scores with Partially Missing Data, Journal of the American Statistical Association, 95:451, 749-759

To link to this article: http://dx.doi.org/10.1080/01621459.2000.10474263

Published online: 17 Feb 2012.

Estimating and Using Propensity Scores With Partially Missing Data

Ralph B. D'AGOSTINO, Jr. and Donald B. RUBIN

Investigators in observational studies have no control over treatment assignment. As a result, large differences can exist between the treatment and control groups on observed covariates, which can lead to badly biased estimates of treatment effects. Propensity score methods are an increasingly popular method for balancing the distribution of the covariates in the two groups to reduce this bias; for example, using matching or subclassification, sometimes in combination with model-based adjustment. To estimate propensity scores, which are the conditional probabilities of being treated given a vector of observed covariates, we must model the distribution of the treatment indicator given these observed covariates. Much work has been done in the case where covariates are fully observed. We address the problem of calculating propensity scores when covariates can have missing values. In such cases, which commonly arise in practice, the pattern of missing covariates can be prognostically important, and then propensity scores should condition both on observed values of covariates and on the observed missing-data indicators. Using the resulting generalized propensity scores to adjust for the observed background differences between treatment and control groups leads, in expectation, to balanced distributions of observed covariates in the treatment and control groups, as well as balanced distributions of patterns of missing data. The methods are illustrated using the generalized propensity scores to create matched samples in a study of the effects of postterm pregnancy.

KEY WORDS: General location model; Ignorability; Iterative proportional fitting; Log-linear model; Matching; Matched sampling; Maximum likelihood estimation; Missing data; Observational study; Pattern-mixture model.

1. INTRODUCTION

1.1 Background on Propensity Scores

Since they were introduced by Rosenbaum and Rubin (1983), propensity scores have been used in observational studies in many fields to adjust for imbalances on pretreatment covariates, $\mathbf{X}$, between a treated group, indicated by $Z = 1$, and a control group, indicated by $Z = 0$ (e.g., D'Agostino 1998; Dehejia and Wahba 1999; Rubin 1997). Propensity scores are a one-dimensional summary of multidimensional covariates, $\mathbf{X}$, such that when the propensity scores are balanced across the treatment and control groups, the distribution of all the covariates, $\mathbf{X}$, are balanced in expectation across the two groups. Typically, matched sampling (e.g., Heckman, Ichimura, Smith, and Todd 1996; Lytle et al. 1999; Rosenbaum and Rubin 1985; Takizawa et al. 1999; Willoughby et al. 1990) or subclassification (e.g., Barker et al. 1998; Conners et al. 1996; Nakamura et al. 1999; Rosenbaum and Rubin 1984; U.S. General Accounting Office 1994) on estimated propensity scores is used, often in combination with model-based adjustments (Curley, McEachern, and Speroff 1998; Lieberman et al. 1996; Rich 1998; Rubin and Thomas 2000; Smith et al. 1998).

The propensity score for an individual is the probability of being treated conditional on the individual's covariate values. To estimate propensity scores for all individuals, one must model the distribution of Z given the observed covariates, $\mathbf{X}$. There is a recent and large technical literature on propensity score methods with complete data (e.g., Gu and Rosenbaum 1993; Rubin and Thomas 1992a,b, 1996). In practice, however, typically some covariate values will be missing, and so it is not clear how the propensity score should be estimated. In addition, the missingness itself may be predictive about which treatment is received in the sense that the treatment assignment mechanism is ignorable (Rubin 1978) given the observed values of $\mathbf{X}$ and the observed pattern of missing covariates but not ignorable given only the former.

Rosenbaum and Rubin (1984) considered using a "pattern mixture" model (Little 1992; Rubin 1986) for propensity score estimation with missing covariate data. Appendix B of Rosenbaum and Rubin (1984) defined a "generalized" propensity score as the probability of treatment assignment given $\mathbf{X}^*$, the vector covariate with an asterisk indicating a missing component of the vector covariate $\mathbf{X}$ (as in Rubin 1976a). This is equivalent to conditioning on the observed values of $\mathbf{X}, \mathbf{X}_{\text{obs}}$, and a missing covariate indicator R ($R = 1$ for observed, $R = 0$ for missing); with discrete covariates, this is equivalent to adding an additional "missing" category to each covariate. Rosenbaum and Rubin (1984) proved that adjustment for the "generalized" propensity score in expectation balances the observed covariate information and the pattern of missing covariates. They suggested that in large enough samples, one can estimate this generalized propensity score by estimating a separate logit model using the subset of covariates fully observed for each pattern of missing data. The practical problem is that typically there are many patterns of missing data with only a few individuals from each of the two treatment groups, thereby making the straightforward pattern-mixture approach infeasible. Thus estimating the generalized propensity score requires smoothing the parameters of Z given $\mathbf{X}_{\text{obs}}$ across the patterns of missing data.

Ralph B. D'Agostino, Jr. is Associate Professor, Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Winston-Salem, NC 27157 (E-mail: *rdagosti@wfubmc.edu*). Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. This work was supported in part by National Cancer Institute grant 1 R01 CA79934. The authors wish to thank their families for their support during the preparation of this manuscript, especially Carey, who spent numerous hours reading drafts.

Our approach is to model the joint distribution of $(Z, \mathbf{X}, R)$. The particular approach that we implement in our application is based on a general location model (Olkin and Tate 1961) accounting for the missing data (Schafer 1997). This modeling implies a conditional distribution for Z given $(\mathbf{X}_{\text{obs}}, R)$; that is, the generalized propensity score: probabilities of $Z = 1$ versus $Z = 0$ for each unit as a function of its observed covariate values $\mathbf{X}_{\text{obs}}$ and its missing-data pattern R. Because $\mathbf{X}$ is missing when $R = 0$, a saturated model for $(\mathbf{X}, R)$ cannot be fit, even with the general location model. We impose log-linear constraints on the categorical variables, which include the missing value indicators for covariates whose missingness is related to treatment assignment. In the special case of no missing data and only continuous covariates, the approach reduces to estimating propensity scores by discriminant analysis, which practically is very close to logistic regression (Rubin and Thomas 1992). Our methods use as basic computational tools the EM (Dempster, Laird, and Rubin 1977) and ECM (Meng and Rubin 1993) algorithms applied to the general location model. We estimate three different propensity score models and use these estimated propensity scores to select matched samples that have similar distributions of observed covariates and missing-value indicators. We also provide suggestions for diagnostic procedures to assess the success of the matching in creating balanced distributions of these observed covariates and missing-value indicators, and use the results from these diagnostics to compare the three propensity score models. We illustrate these procedures in the context of a matched-sampling study of the effects of postterm pregnancy.

It is important to note that our problem is different from most missing-data problems in which the goal is parameter estimation. We are not interested in obtaining one set of estimated parameters for a logistic regression or discriminant analysis, or a posterior distribution for these parameters, or even in drawing inferences about these parameters. Rather, parameters particular to each pattern of missing data serve only in intermediate calculations to obtain estimated propensity scores for each subject. Moreover, the propensity scores themselves serve only as devices to balance the observed distribution of covariates and patterns of missing covariates across the treated and control groups. Consequently, the success of the propensity score estimation is assessed by this resultant balance rather than by the fit of the models used to create the estimated propensity scores. This goal is not special to the case with missing values in covariates, but rather has been the goal with propensity score estimation from the start.

Because the major goal of this article is to present a method for handling missing covariates when estimating propensity scores and to illustrate its use in a real application that involves finding matches for participants in an observational study, we do not consider other matching methods that do not use propensity scores. Such other methods have been addressed by others (e.g., Rosenbaum and Rubin 1985; Rubin 1976b). Here we focus on nearest available matching on the estimated propensity score.

The remainder of Section 1 presents details of the motivating example, which led to the development of our methods. Section 2 provides notation and describes our method. Section 3 applies our methodology to the motivating example and evaluates the resulting balance of three different models.

1.2 The Database and Problem

The original motivation for our work was a particular March of Dimes observational study examining the effects of postterm birth versus term birth on neuropsychiatric, social, and academic achievements of children age 5–10 years. This study is still ongoing, and to date there has not been a published report of any results. Because the focus of this study has been only on the effects of postterm birth versus normal-term birth, children born preterm are not of interest. At the onset of the study, the investigators had available a collection of 9,275 birth records of children born term or postterm at Beth Israel Hospital, Boston, with prenatal and birth history information, including gestational age. About 92% of these children were born term (37–41 weeks), whereas about 8% were born postterm (43+ weeks). Children with gestational ages of 42 weeks are not included, because the medical investigators felt that these children may potentially be a mixture of term and postterm children due to the variability in assessing gestational age, whereas children with gestational ages between 37 and 41 weeks were clearly term and those with gestational ages over 43 weeks were clearly postterm. The investigators were interested in selecting a sample of term and postterm children from this population to be part of their study because it was financially infeasible to follow up all children in the database. Therefore, the initial issue that they faced was how to select the sample to facilitate inference for the effect of being postterm. Because postterm children were relatively rare, essentially all postterm subjects could be followed up, and a matched sample of term children was desired; that is, matched with respect to covariates. A complication was that for some children, some covariates had missing values.

We illustrate our methodology for estimating propensity scores with missing values, and its application to obtain matched samples, using a random sample of 4,500 of the 9,275 subjects in this dataset. The remaining 4,775 subjects were used in validation studies that extend beyond the scope of this article. Of the 4,500 subjects chosen for analysis, 4,155 (92.3%) were term babies and 345 (7.7%) were postterm babies. We estimate several propensity score models using 25 of the covariates that are thought to be scientifically significant for predicting postterm birth and prognostically important for predicting outcomes, and thus if left uncontrolled could confound estimated treatment effects. At this point, it is scientifically important to point out that many of these covariates are not truly proper covariates in the sense of taking their values before "treatment assignment" to term or postterm conditions. That is, in the hypothetical experiment underlying the observational study,

before week 42 a decision could have been made to induce labor for the postterm babies, and the effect of not doing so is the effect we seek. Formally, any covariate measured after 41 weeks thus is an improper covariate, because it could be affected by treatment.

For example, infant's weight had the largest initial imbalance, but can be considered an outcome of being postterm and not a proper covariate. Despite this, as with the other improper covariates, the investigating physicians felt strongly that this variable needed to be controlled. Another example, induction, measures whether there was some form of medical induction performed on the woman during labor (0 = no induction, 1 = elective induction, and 5 = induction due to ruptured membranes, a medical disorder, or a fetal disorder). It too is formally an improper covariate, because it is possibly affected by the hypothesized treatment assignment. Because these improper covariates can be thought of as proxies for unmeasured proper pretreatment covariates that predict fetal disorders, the physicians and investigators felt that they needed to explicitly control these variables as if they were a proper covariates, if useful inferences were to be drawn about policy-relevant advice concerning postterm pregnancies. Regardless, we acknowledge that the inclusion of such improper covariates may actually adjust away part of the true treatment effect. However, this limitation occurs regardless of which method for control is used (i.e., matching or covariate modeling). In any case, the example illustrates the estimation and use of propensity scores with missing covariate data, which could have been applied using only proper covariates.

Tables 1 and 2 present descriptive statistics for the covariates and fitted propensity scores, separately for the term and postterm groups. It is important to emphasize that these statistics are descriptive and not inferential, in the sense that they do not purport to estimate relevant population parameters, but rather simply describe the two samples and their differences. Table 1 presents, for each continuous covariate, the mean and standard deviation using available cases; also presented are standardized percentage differences, defined as the mean difference between postterm and term groups as a percentage of the standard deviation $\{[100(\bar{x}_p - \bar{x}_t)]/\sqrt{[(s_p^2 + s_t^2)/2]}\}$, where $\bar{x}_p$ and $\bar{x}_t$ are the sample means in the postterm and term groups and s_p^2 and s_t^2 are the corresponding sample variances, again based on available cases. Also presented are the variance ratios, s_p^2/s_t^2.

The first two columns of Table 2 present, using available case data for the categorical covariates (first four rows), the proportion of women in each category in the term and postterm groups: Also presented are the corresponding results for the missing-data indicators (last 10 rows) for the 10 covariates with any missing values (either continuous or categorical). The third column displays the absolute differences in percent between the term and postterm groups for each of the categorical covariates and missing-data indicators. Two covariates, delivery mode and labor compli-

Table 1. Means (Standard Deviations), Standardized Differences (Based on Available Cases) in Percent, and Variance Ratios for Continuous Covariates in Both Groups Before Matching

Covariate	Term mean (SD)	Postterm mean (SD)	Initial standardized difference (%)[b]	Variance ratio[c]
Antepartum complications (yes/no)	.72 (.45)	.72 (.45)	1	1
Previous obstetrical history (yes/no)	.47 (.50)	.40 (.49)	−14	.96
Vaginal bleeding (yes/no)	.12 (.33)	.11 (.31)	−4	.88
Second-stage indicator[a]	.81 (.39)	.77 (.42)	−10	1.16
Delivery mode	1.26 (.51)	1.30 (.51)	8	1
Labor complications	.58 (.63)	.66 (.59)	14	.88
Class[a]	2.37 (.77)	2.31 (.77)	−8	1
Diabetes[d]	.15 (1.05)	.11 (.82)	−4	.61
Fetal distress	.04 (.64)	.15 (1.2)	11	3.51
Induction	.17 (.88)	.41 (1.2)	23	1.85
Pelvic adequacy (clinic)[a]	.19 (.68)	.19 (.67)	0	.97
Pelvic adequacy (X-Ray)[a]	1.71 (.94)	1.69 (.79)	−3	.70
Placental problems	.11 (1.04)	.09 (.93)	−2	.80
Previous perinatal mortality[e]	.22 (1.49)	.15 (1.13)	−6	.57
Urinary tract disorders	.11 (.51)	.13 (.53)	4	1.08
Child's age (months from 1980, range 0–48)	23.4 (13.0)	23.9 (11.4)	4	.77
Infant's weight (grams)[a]	3338 (461)	3,626 (533)	58	1.33
Length of first stage (min)[a]	784 (571)	910 (665)	20	1.35
Length of second stage (min)[a]	53.8 (65)	59.5 (66)	9	1.03
Time since membranes ruptured (min)[a]	454 (791)	414 (651)	−6	.68
Mother's age (years)	28.8 (5)	28.2 (5)	−12	1
Parity	.77 (1.0)	.66 (1.1)	−10	1.20
Total length of labor (min)[a]	841 (589)	968 (688)	20	1.37

[a] Covariate suffers from some missing data.

[b] The standardized difference is the mean difference as a percentage of the average standard deviation: $\{[100(\bar{x}_p - \bar{x}_t)]/\sqrt{[(s_p^2 + s_t^2)/2]}\}$, where for each covariate $\bar{x}_p$ and $\bar{x}_t$ are the sample means in the postterm and term groups and s_p^2 and s_t^2 are the corresponding sample variances.

[c] The variance ratio is s_p^2/s_t^2.

[d] Diabetes: 0 = none, 1 = diabetes insipidus or glucosuria, 5 = abnormal glucose tolerance test, and 10 = diabetes mellitus.

[e] Previous perinatal mortality: 0 = no previous child deaths, 5 = previous late death (in the first year of life), 10 = previous stillbirth or neonatal death, and 20 = previous stillborn and previous neonatal death (or any combination of 2 or more perinatal mortalities).

Table 2. Table of Observed Proportions and Percent Differences for Categorical Covariates and Missing-Value Indicators for Initial Data

		Term	Postterm	Difference (in %)
Covariate				
Race	White	.70	.72	2
	Nonwhite	.30	.28	2
Gender	Male	.49	.51	2
	Female	.51	.49	2
	Vertex	.77	.72	5
Delivery mode	Cesarean	.21	.27	6
	Other	.02	.01	1
	No labor (Cesarean)	.08	.06	2
Labor	No complications	.26	.21	5
complications	Some complications	.66	.73	7
Missing-value indicators (proportion observed)				
Pelvic adequacy (x-ray)		.05	.10	5
Length of second stage of labor		.78	.74	4
Race		.95	.95	0
Second stage of labor indicator		.99	1.00	1
Class		.99	.99	0
Pelvic adequacy (clinic)		.85	.90	5
Infant's weight		.99	1.00	1
Length of first stage of labor		.89	.91	2
Time that membranes ruptured		.97	.97	0
Length of labor		.89	.91	2

cations, were considered to be either continuous or categorical, depending on the specific propensity score model and thus appear in both Tables 1 and 2.

Another diagnostic assessment compares the pairwise available-case correlations between the 23 continuous covariates in Table 1. Suppose that we plotted the $253 = 23 \times 22/2$ pairwise correlations in the initial term group against the pairwise correlations for the postterm group. If the two groups had similar distributions of their pairwise correlations, then they would have approximately the same means and variances, and we would see a roughly linear relationship. The mean correlation is .0143 in the term group and .0161 in the postterm group, and the corresponding variances are .016 and .022, indicating slightly larger and more variable correlations in the postterm group; for example, most of the pairwise correlations are between $-.5$ and .5 for the postterm group, but between $-.4$ and .4 for the term group. Moreover, the R^2 value of .61 is not particularly high. Note that we are not recommending the use of available-case correlations to estimate population correlations—rather, we are using them only to summarize aspects of the observed data for comparison across treatment groups.

The initial term versus postterm group differences summarized in Tables 1 and 2 and by the R^2 value indicate the possible extent of biased comparisons of outcomes due to different distributions of observed covariates and patterns of missing data in the initial term and postterm groups. That is, ideally all such descriptive statistics should suggest the same distribution in the term and postterm groups, as they would be in expectation if the treatment indicator (term vs. postterm) had been randomly assigned. As can be seen from these tables, there exists considerable initial bias between the term and postterm groups. For instance, nine of the continuous covariates have initial standardized differences larger than 10%. In addition, there is substantial differences between the groups based on the estimated propensity scores. Among categorical covariates, we see that labor complications and delivery mode are different between the postterm group and the term group. The missingness rates appear similar, except that there seems to be a trend for some indicators of potential complications to be observed more often in the postterm group (e.g., pelvic adequacy, both x-ray and clinical), suggesting a greater need for such medical tests among the postterm subjects. In addition, the missing-data indicator for length of second stage of labor shows that more individuals had this variable observed in the term group than in the postterm group (78% versus 74%).

2. NOTATION

2.1 Estimation of Propensity Scores

With complete data, Rosenbaum and Rubin (1983) introduced the propensity score for subject i $(i = 1, \ldots, N)$ as the conditional probability of receiving a particular treatment $(Z_i = 1)$ versus control $(Z_i = 0)$ given a vector of observed covariates, $\mathbf{x}_i$,

$$e(\mathbf{x}_i) = \text{pr}(Z_i = 1 | \mathbf{X}_i = \mathbf{x}_i), \tag{1}$$

where it assumed that, given the $\mathbf{X}$'s, the Z_i are independent,

$$\text{pr}(Z_1 = z_1, \ldots, Z_N = z_N | \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_N = \mathbf{x}_N) = \prod_{i=1}^{N} e(\mathbf{x}_i)^{z_i} \{1 - e(\mathbf{x}_i)\}^{1-z_i}. \tag{2}$$

Rosenbaum and Rubin (1983) showed that for a specific value of the propensity score, the difference between the treatment and control means for all units with that value of the propensity score is an unbiased estimate of the average treatment effect at that propensity score, if the treatment assignment is strongly ignorable given the covariates. Thus matching, subclassification, or regression (covariance) adjustment on the propensity score tends to produce unbiased estimates of the treatment effects when treatment assignment is strongly ignorable, which occurs when the treatment assignment, Z, and the potential outcomes, Y, are conditionally independent given the covariates $\mathbf{X}$: $\Pr(Z|\mathbf{X}, Y) = \Pr(Z|\mathbf{X})$.

2.2 Propensity Scores With Incomplete Data

Let the response indicator be $R_{ij}, (j = 1, \ldots, T)$, which is 1 when the value of the jth covariate for the ith subject is observed and 0 when it is missing; R_{ij} is fully observed by definition. Also, let $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$, where $\mathbf{X}_{\text{obs}} = \{\mathbf{X}_{ij} | R_{ij} = 1\}$ denotes the observed parts and $\mathbf{X}_{\text{mis}} = \{\mathbf{X}_{ij} | R_{ij} = 0\}$ denotes the missing components of $\mathbf{X}$.

The generalized propensity score for subject i, which conditions on all of the observed covariate information, is

$$e_i^* = e_i^*(\mathbf{X}_{\text{obs},i}, R_i) = \text{pr}(Z_i = 1 | \mathbf{X}_{\text{obs},i}, R_i). \tag{3}$$

Rosenbaum and Rubin (1985) showed that with missing covariate data and strongly ignorable treatment assignment given $\mathbf{X}_{\text{obs}}$ and R, the generalized propensity score e_i^* in (3) plays the same role as the usual propensity score e_i in (1) with no missing covariate data. Treatment assignment is strongly ignorable given $(\mathbf{X}_{\text{obs}}, R)$ if $\Pr(Z|\mathbf{X}, Y, R) = \Pr(Z|\mathbf{X}_{\text{obs}}, R)$. If in addition, the missing-data mechanism is such that $\Pr(R|\mathbf{X}, Z) = \Pr(R|\mathbf{X}_{\text{obs}})$, then $\Pr(Z|\mathbf{X}, Y, R) = \Pr(Z|\mathbf{X}_{\text{obs}})$, and R itself can be ignored in the modeling. It is important to emphasize that, just as with propensity score matching with no missing data, the success of a propensity score estimation method is to be assessed by the quality of the balance in the $(\mathbf{X}_{\text{obs}}, R)$ distributions between term and postterm groups that has been achieved by matching on it. Consequently, the usual concerns with the fit of a particular model (i.e., the general location model) are not relevant if such a balance is achieved.

2.3 General Location Model With Complete Data

The distribution of $(\mathbf{X}, Z)$ is defined by the marginal distribution of the categorical variables, Z, and the categorical covariates, U, and the conditional distribution of continuous covariates, say V, given (U, Z). U_{ij}, Z_i locates the ith subject in one of the m cells of the table formed by (U, Z).

We assume that (U, Z) are iid multinomial random variables, and conditional on U_i, Z_i, we assume that V_i is K-variate normal with mean that depends on the cell but with a common covariance. This is the general location model (Olkin and Tate 1961) with parameters $\Pi =$, cell probabilities from the multinomial distribution, $\mathbf{\Gamma} =$, the matrix of cell means, and $\mathbf{\Omega} =$, the positive-definite covariance matrix common to all cells; $\theta = (\Pi, \mathbf{\Gamma}, \mathbf{\Omega})$.

Krzanowski (1980, 1982), Little and Rubin (1987), and Little and Schlucter (1985), have described restricted general location models with fewer parameters. One way to reduce the number of parameters to be estimated is to constrain Π by a log-linear model (Bishop, Fienberg, and Holland, 1975; Goodman 1968); for example, three-way and higher-order interactions are set to 0. Maximum likelihood (ML) estimates of the parameters for these models have closed-form solutions for many configurations, but if they do not have a closed form, they can be found by using an iterative procedure such as iterative proportional fitting (IPF; Bishop et al. 1975).

A second way to reduce the number of parameters to be estimated in the general location model is to impose analysis of variance (ANOVA)-like restrictions on the means, $\mathbf{\Gamma}$, using a known design matrix A to define Γ in terms of a lower-dimensional matrix of unknown regression coefficients β; if $\mathbf{A}$ is the identity matrix, then there are no restrictions. With standard models and complete data, the parameter estimates for β and $\mathbf{\Omega}$ can be found using standard regression techniques (Anderson 1958, chap. 8).

Although the restrictions described previously reduce the number of parameters to be estimated in the model, we could also generalize the model to increase the number of parameters to be estimated. For instance, the assumption of a common covariance Ω across all cells of the contingency table can be relaxed to allow for possibly different covariance matrices to be estimated in different cells. But this can require substantial sample sizes in each cell. A more useful extension may be to estimate separate covariance matrices only for the treated and the control groups. Other extensions involve proportional covariance matrices and more general ellipsoidal distributions (e.g., t distributions as in Liu and Rubin 1995, 1998).

2.4 Fitting the General Location Model With Missing Data

The basic method for finding estimates for the parameters of the general location model when there are ignorably missing data has been outlined by Little and Rubin (1987, chap. 10) and is based on the EM algorithm (Dempster et al. 1977) and the ECM algorithm (Meng and Rubin 1993), which is used when log-linear restrictions have been placed on the general location model such that IPF is needed with complete data. Of particular importance for our situation, where we want to include explicitly the response indicator R in the modeling, is that R is a fully observed collection of categorical (in fact, binary) covariates. For notational convenience, let $U^* = (U, R)$, with corresponding changes to the other notation. Because $\mathbf{X}_{ij}$ is missing when $R_{ij} = 0$, some restrictions are needed to obtain unique ML estimates of parameters for the joint distribution of $(Z, \mathbf{X}, R)$, which we need to obtain unique ML estimates of the conditional distribution of Z given $\mathbf{X}_{\text{obs}}$ and R.

At each iteration of the EM algorithm, the E step computes the expected values of the complete-data sufficient statistics given the observed data and the current estimates of the parameters, $\theta^{*(t)} = (\mathbf{\Gamma}^{*(t)}, \mathbf{\Omega}^{*(t)}, \Pi^{*(t)})$, where t indexes iterations. The M step computes the ML estimates for the parameters using the estimated values of the sufficient statistics. These become the current estimates of the parameters to be used in the next E-step calculations. The E and M steps are repeated until convergence. When there are log-linear constraints on the categorical covariates, ECM is used instead of EM; the M step of EM is replaced by CM steps, which perform one cycle of IPF. The complete-data sufficient statistics for this model are the raw sums of squares and cross-products of the V's $(\Sigma V_i^T V_i)$, the sums of the V's in each cell (cell totals), and the cell frequencies from the table defined by (Z, U^*).

Once EM or ECM has converged, we have ML estimates, $\hat{\theta}^*$, for the parameters $\theta^* = (\Pi^*, \mathbf{\Gamma}^*, \mathbf{\Omega}^*)$ for the joint distribution of (V, Z, U^*), which we use to calculate an estimated propensity score for each subject, $\hat{e}_i^*$, as in (3) with $\mathbf{X}_{\text{obs},i} = (V_{\text{obs},i}, U_{\text{obs},i})$:

$$\hat{e}_i^* = \text{pr}(Z_i = 1 | V_{\text{obs},i}, U^*_{\text{obs},i}, R_i, \hat{\theta}^*). \tag{4}$$

To find the estimated propensity score (4) from $\hat{\theta}^*$ and the observed data, we simply run one E step using the converged MLE $\hat{\theta}^*$, but now treating Z_i as missing.

2.5 Generalized Propensity Score Estimates: An Illustration

To illustrate the calculation outlined in Section 2.4, consider a March of Dimes example in which there are one categorical covariate, gender (two levels), U_1, with some missing values indicated by R_1; one continuous covariate, mother's age, V_1, with some missing values indicated by R_2; and the two-level treatment indicator Z. Consider the case where the only missing-value indicator that needs to be balanced between the groups is R_1. To estimate the propensity score, we need to estimate the parameters for the joint distribution of (U_1, R_1, V_1, Z). To estimate the cell proportions defined in the $2 \times 2 \times 2$ contingency table defined by (Z, U_1, R_1), we could fit a log-linear model with no three-way interaction. To estimate parameters for V_1 conditional on (Z, U_1, R_1), the mean and variance of V_1 in each of the eight cells, we could place an additive structure on the means so that the linear regression of V_1 on Z, U_1, and R_1 has a constant, a term for Z, a term for U_1, and a term for R_1, and a common variance.

To fit such models with complete data, we would use IPF to estimate the cell probabilities for the three-way table using the sufficient statistics that are the observed cell counts for the $2 \times 2 \times 2$ table of (Z, U_1, R_1), and ordinary least squares for the regression of V_1 on (Z, U_1, R_1), where the sufficient statistics are the within-cell totals and raw sums of squares of V_1. With missing data, we use the ECM algorithm to estimate the parameters for these models.

The E step of ECM gets the expectations of the sufficient statistics given the observed data and the current estimates of the parameters for each pattern of missing data (U_1 and V_1 observed, U_1 observed and V_1 missing, U_1 missing and V_1 observed, and U_1 and V_1 both missing). When both covariates are observed, the estimates of the sufficient statistics are the observed cell counts, cell totals, and sum of squares of V_1. When V_1 is missing, the expected cell totals and raw sum of squares are based on the additive regression model described earlier and involve the expected value both of V_1 and of its square within each cell. When U_1 is missing, the expected cell counts are found by performing a discriminant analysis between male and female to determine the probabilities of being in each cell given the observed value of V_1, the observed value of R_1, and the observed treatment indicator Z. When both covariates are missing, the expected sufficient statistics are found conditionally given $R_1 = 0$ and the observed value of Z. The expected cell counts are found using the conditional probability for being in each cell defined by U_1, based on the current estimates of cell probabilities. The expected within-cell totals and raw sum of squares are based on the regression equations for V_1 conditional on being in one and then the other cell defined by U_1.

Once we have the updated expectations of the sufficient statistics, we compute the complete-data estimates from these sufficient statistics using the CM step of the ECM algorithm. First, we perform one iteration of IPF to get new estimates of the cell proportions based on the log-linear model specified. Then we get estimates for the cell means and within-cell variance parameters from fitting the regression model specified.

We continue iterating through ECM to find the ML estimates for the joint distribution of (Z, U_1, V_1, R_1). Finally, we run the E step of the algorithm one more time with Z considered missing to estimate the propensity scores. This E step, conditional on the final parameters estimates, essentially performs a discriminant analysis to determine the probabilities of being treated or control given $(\mathbf{X}_{\text{obs}}, R_1)$. [Recall that $\mathbf{X}_{\text{obs}}$ corresponds to the observed parts of (U_1, V_1).] The propensity score is then found by summing the probabilities across the four cells of the three-way table (defined by U_1, R_1, and Z) that correspond to $Z = 1$, the treated group.

3. PROPENSITY SCORE MODELS FOR MARCH OF DIMES DATA

3.1 Specific Models

All three propensity score models we fit used all 23 continuous covariates in Table 1 and all 4 categorical covariates in Table 2. However, the models differed in their inclusion of missing-data indicators.

Among the 25 covariates used in our analyses, the missingness on 2 were differentially distributed in the treatment groups and believed to be prognostically important: results of a pelvic x-ray and length of the second stage of labor. These were determined to be prognostically important for the following reasons. During the years 1981–1984, when the birth data for this study were recorded, pelvic x-rays were performed on women who were having difficulties during labor. The x-ray would usually be performed if the obstetrician suspected that the baby was too large to fit through the mother's pelvis. These x-rays were rarely ordered, and the existence of an x-ray order was an indication that the labor would be different than an ordinary one. In our data, out of the 4,500 women, only 234 (or 5.2%) had pelvic x-rays recorded, but their frequency was twice as great among postterm subjects (10.4% vs. 4.8%); the missing-value indicator for a pelvic x-ray may be prognostically important, even as important as the outcome of the x-ray itself. Of course, this is not a proper covariate inasmuch as it may be an outcome of allowing the fetus to continue to grow, but the physicians wanted it to be controlled.

The other missingness indicator considered prognostically important was the length of the second stage of labor; that is, the length of time from full dilation of the cervix to delivery. If a woman does not have a cesarean section, then she must have a second stage of labor. In our data, 1,002 out of 4,500 women did not have a second stage of labor recorded. Of these, 872 had a cesarean section, whereas 130 did not.

We fit three generalized propensity score models with missing-value indicators; the first included the missingness indicator for the pelvic x-ray, the second included the missingness indicator for length of the second stage of labor, and the third included both indicators. The first

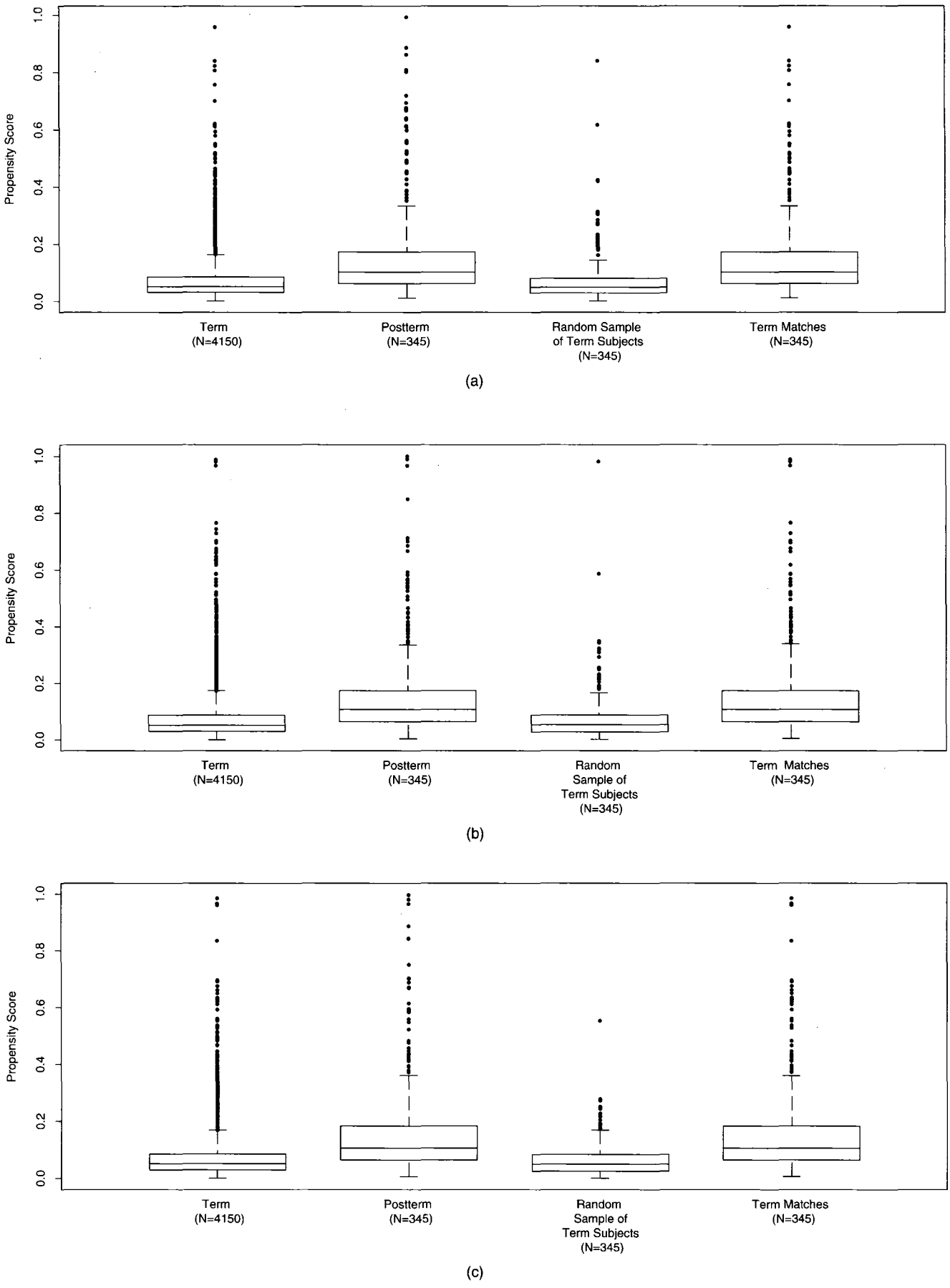

Figure 1. Boxplots Comparing Propensity Score Balance Before and After Matching. (a) Model 1; (b) model 2; (c) model 3.

and second models without restrictions each have 659 parameters. In each model, log-linear constraints were placed on the cell probabilities so that the three-way and higher interactions were set to 0, and thus we estimated four main effects and six two-way interactions. The design matrix relating the means of the continuous variables to the categorical variables includes an intercept, main effects for each of the categorical variables, and terms for each of the two-way interactions of the categorical variables. These constrained models have 539 total parameters: 10 for the contingency table, 253 regression coefficients, and 276 variances and covariances. The third model without any constraints has 1,043 parameters. Here we fit a log-linear model with only main effects and two-way interactions. The design matrix relating the continuous variables to the categorical variables includes an intercept, main effects for each of the categorical variables, and terms for each of the two-way interactions of the categorical variables. The resulting 659 parameters are 15 for the contingency table, 368 regression coefficients, and 276 variances and covariances.

3.2 Matching Using Estimated Propensity Scores

We estimate propensity scores for each of the models using the ECM algorithm as illustrated in Section 2.5. We then use nearest-available matching on these estimated propensity scores to choose matches for the postterm subjects. We randomly order the term and postterm subjects, and then select the term subject with the propensity score closest to the first postterm subject. Both subjects are then removed from the pools of subjects. We repeat this procedure for each postterm subject, which results in selecting a total of 345 term subjects from the 4,150 available ones. We could have used many other approaches to select the matches, but chose this straightforward approach using propensity scores to focus on how well matching based on the different propensity score models succeeds in balancing the distribution of observed covariates and missing-data indicators between the term and postterm groups.

3.3 Resultant Distributions of Propensity Scores and Their Logits

Figure 1 gives, for each model, associated boxplots of estimated propensity scores in the term group ($n = 4,150$), the postterm group ($n = 345$), a group of randomly selected term subjects ($n = 345$), and the matched term group ($n = 345$) for that model. In this figure, for each boxplot the interquartile range is displayed by a rectangular box, with a solid horizontal line within the box representing the median propensity score for each group. Above and below the interquartile ranges are small notches representing the upper and lower values for the 5th and 95th percentiles. Above the upper percentile notch, individual circles represent propensity scores for single subjects beyond the 95th percentile. For example, in model 1 there is a postterm subject with a propensity score nearly equal to 1. With all models, there are two striking features: (1) The median propensity score for the postterm group is larger than the 75th percentile propensity score in the unmatched term group and the randomly selected term group, but nearly equal to the median propensity score in the matched term group, and (2) the spread of propensity scores for the unmatched term group is wide enough to cover most of the propensity scores for the postterm group in each model. This second feature allows us to find matches from the term group with propensity scores close to the propensity scores of the subjects in the postterm group. In Figure 1 the distribution of propensity scores for the random sample of term subjects has few values near those of postterm subjects, whereas the propensity scores for the matched term subjects for each model covers most propensity scores in the postterm group.

Table 3 compares the distributions of the estimated propensity scores and their logits that result from each of the three propensity score models. Columns 1 and 2 present the standardized differences in percent and variance ratios prior to matching. Columns 3–8 present the same statistics after matching for model 1 (columns 3 and 4), model 2 (columns 5 and 6) and model 3 (columns 7 and 8). The initial standardized differences in percentage were quite large, with model 3 having the largest initial difference, 69%, which suggests that this model may be the most reveal-

Table 3. Standardized Differences (in %) and Variance Ratios for Propensity Scores and Their Logits, Before and After Matching Using Each Model

			Results after matching					
			Matching Model 1		Matching Model 2		Matching Model 3	
Propensity score	Initial Standard Difference (%)	Initial variance ratio	Standard difference (%)	variance ratio	Standard difference (%)	variance ratio	Standard difference (%)	variance ratio
Model 1	67	5.24	2	1.13	15	1.58	8	1.22
Model 2	67	3.85	5	1.07	1	1.01	1	.99
Model 3	69	5.00	7	1.28	11	1.46	2	1.19
Logit model 1	85	1.61	2	1.1	13	1.27	8	1.14
Logit model 2	83	1.39	11	1.10	3	1.12	5	1.03
Logit model 3	86	1.59	7	1.26	11	1.28	2	1.18

[a] The standardized difference in % is the mean difference as a percentage of the average standard deviation: $\{[100(\bar{x}_p - \bar{x}_t)]/\sqrt{[(s_p^2 + s_t^2)/2]}\}$ where for each covariate $\bar{x}_p$ and $\bar{x}_t$ are the sample means in the postterm and term groups and s_p^2 and s_t^2 are the corresponding sample variances.

[b] The variance ratio is s_p^2/s_t^2.

ing by maximally separating the term and postterm groups. The initial variance ratios were also quite large with the propensity scores from models 1 and 3 each having ratios of 5.00 or greater. When we compare the standardized differences after matching (columns 3, 5, and 7), we see that matching using any of the three propensity score models performed well in reducing the standardized differences, with model 3 performing best; matching based on propensity scores from this model reduced the standardized differences to below 10% for all models. More explicitly, when we selected matches based on propensity scores estimated from model 3 and then compared the distribution of propensity scores estimated from model 1, we find that the standardized difference was 8% (from column 7). When we compared variance ratios (columns 4, 6, and 8) across the three models, we again found that model 3 performed best; this model produced variance ratios that were on average closer to 1 than the other two models.

3.4 Resultant Covariate Balance After Matching

To further assess the relative success of the propensity score models for creating balanced matched samples, we compare balance on observed covariates and missing-data indicators in the matched samples created by each model. It is important for practice to realize that, as done by Rosenbaum and Rubin (1984, 1985), these assessments can be made before any resources have been committed to collecting outcome data on the matched controls. Also, it is important to realize that because these comparisons involve only observed covariates and their missing-data indicators and not outcome variables, there is no chance of biasing results in favor of one treatment condition versus the other through the selection of matched controls.

Figure 2 compares the standardized differences in percent, after matching, for the continuous covariates. The matching using any of the models performs well in reducing the bias of the background covariates with moderate-to-large initial standardized differences. For instance, the initial standardized difference for the length of first stage of labor variable is 20%, and all models were able to reduce this significantly, ranging from 3% (models 1 and 3) to 5% (model 2). Even the initial standardized difference for infant's weight is substantially reduced by the matching using any of the models, although the difference is 11% using model 3.

From Figure 3, which compares the available-case cell proportions for the categorical covariates and missing-value indicators between the term and postterm groups for the three models, we find that the initial imbalance in delivery mode and labor complications was moderate, with 26% of postterm babies being cesarean births versus 20% of term pregnancies and 73% of postterm pregnancies having some complications versus 66% of term pregnancies. These differences were reduced by the matching for all models. For missing-value indicators, we found that all models improved the balance between the term and postterm groups.

When we compared the pairwise correlations among covariates for the postterm and matched term groups, we found that the means of the correlations for matched samples obtained by models 2 and 3 were closer to the postterm values than those obtained by model 1. The mean in the postterm group was .0161, compared to .0108 with model 1, .0145 with model 2, and .0155 with model 3. The variances in all three models were closer to the variance in the postterm group after matching. In addition, we found that the R^2 values for the samples matched by models 2 and 3 were higher (.85 and .82) than those for the sam-

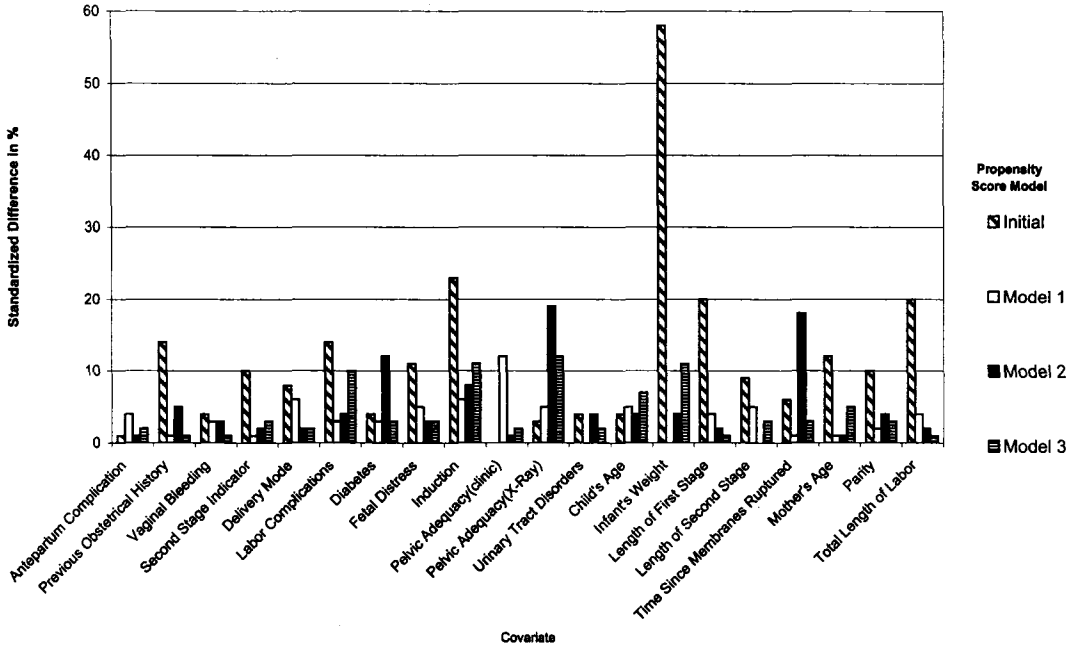

Figure 2. Comparison of Standardized Difference (in %) for Covariates Between Term and Postterm Women, Based on Available Case Means for Each Propensity Score Model.

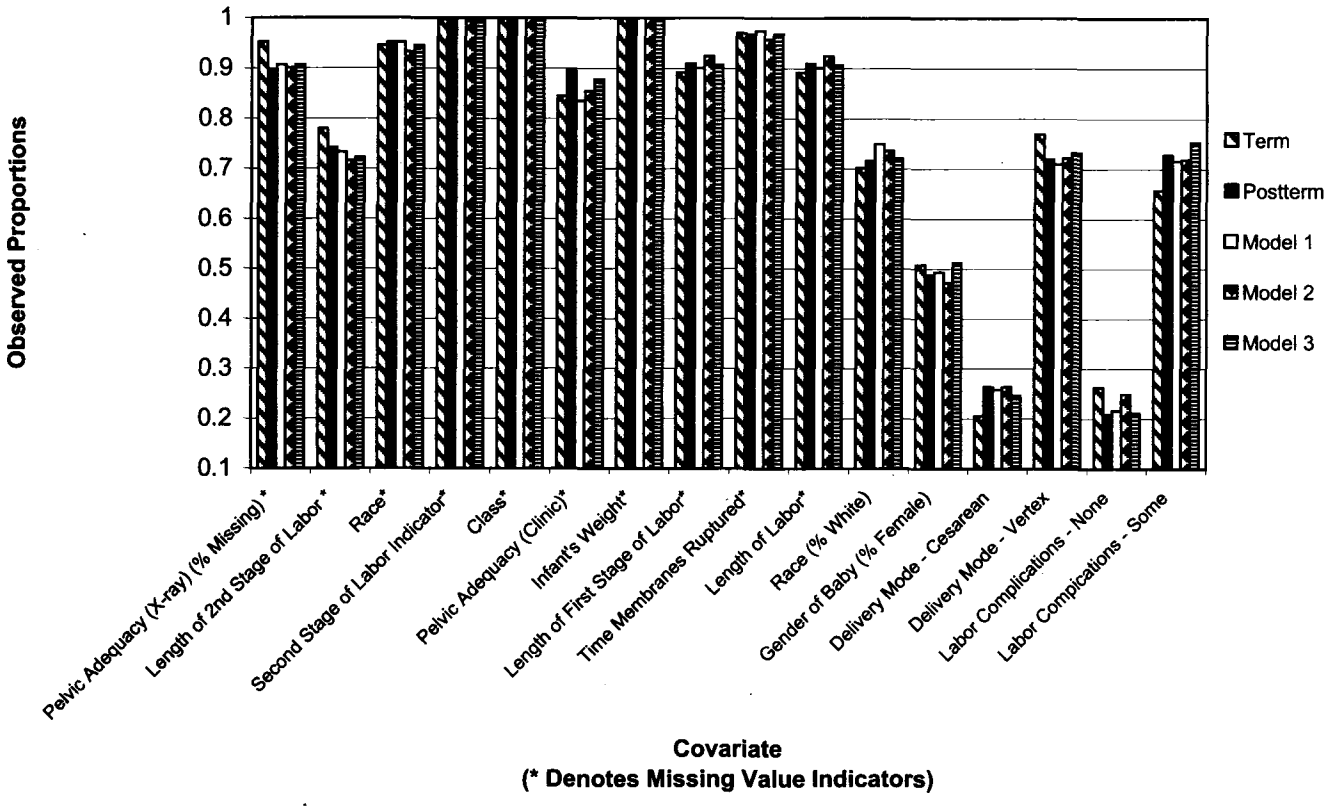

Figure 3. Comparison of Observed Proportions for Categorical Covariates, including Missing-Data Indicators, Between Term and Postterm Women, for Each Propensity Score Model.

ples matched by model 1, which indicates that the pairwise correlations among the matched term subjects selected using these two models more closely resemble the pairwise correlations in the postterm group.

Model 3 (with race, gender, missingness for length of second stage, and missingness for pelvic x-ray as categorical covariates) appears to be the best propensity score model that we fit for producing balanced matched samples on the propensity scores, their logits, and the individual covariates. This model reduces the bias on all covariates with large or moderate initial bias and on the missing-value indicator for the pelvic x-ray exam. Moreover, it revealed the largest initial bias based on the propensity score (69% initial standardized difference).

We acknowledge that many other plausible propensity score models could be constructed using the 25 covariates and their missing-data indicators, and that among these there may exist models that produce better balance than our model 3. Still, this model did give the investigators what they wanted—propensity scores that were used to select matches for the postterm babies from the available pool of term babies, where the bias observed between the term and postterm groups on many covariates and their missingness prior to matching was substantially reduced (and often essentially removed) by the matching.

4. CONCLUSIONS

We have presented an approach for estimating propensity scores in the presence of missing data using the EM and ECM algorithms as computing tools. The framework allows the investigator to impose a structure on the relationships among the covariates in the model, including missing-value indicators for specific effects. We illustrated our approach using March of Dimes data. Simulation studies are underway to examine the effects of specifying different missing-data mechanisms on the data. In addition, we are in the process of developing user-friendly software to perform these analyses.

[Received August 1996. Revised March 2000.]

REFERENCES

Anderson, T. W. (1958), *An Introduction to Multivariate Statistics*, New York: Wiley.

Baker, S. G., and Laird, N. M. (1988), "Regression Analysis for Categorical Variables With Outcome Subject to Nonignorable Nonresponse," *Journal of the American Statistical Association*, 83, 62–69.

Barker, F. G. II, Chang, S. M., Gutin, P. H., Malec, M. K., McDemott, M. W., Prados, M. D., and Wilson, C. B. (1998), "Survival and Functional Status After Resection of Recurrent Glioblastoma Multiforme," *Neurosurgery*, 42, 709–720.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analyses: Theory and Practice*, Cambridge, MA: MIT Press.

Conaway, M. R. (1992), "The Analysis of Repeated Categorical Measurements Subject to Nonignorable Nonresponse," *Journal of the American Statistical Association*, 87, 817–824.

Conners, A. F. Jr., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E. Jr., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J. Jr., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., Knaus, W. A. (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients. SUPPORT Investigators," *Journal of the American Medical Association*, 276, 889–997.

Curley, C., McEachern, J. E., and Speroff, T. (1998), "A Firm Trial of In-

terdisciplinary Rounds on the Inpatient Medical Wards: An Intervention Designed Using Continuous Quality Improvement," *Medical Care*, 36, AS4–12.

D'Agostino, R. B. Jr. (1998), "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Nonrandomized Control Group," *Statistics in Medicine*, 17, 225–2281.

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993), "Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-Ups," *Journal of the American Statistical Association*, 88, 984–993.

Goodman, L. A. (1968), "The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interactions in Contingency Tables With or Without Missing Entries," *Journal of the American Statistical Association*, 63, 1091–1131.

Gu, X. S., and Rosenbaum, P. R. (1993), "Comparison of Multivariable Matching Methods; Structures, Distances, and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420.

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. (1996), "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences of the United States of America*, 93, 13416–13420.

Krzanowski, W. J. (1980), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," *Biometrics*, 36, 493–499.

——— (1982), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis Testing Approach," *Biometrics*, 38, 991–1002.

Lieberman, E., Cohen, A., Lang, J. M., D'Agostino, R. B. Jr., Datta, S., and Frigoletto, F. D. Jr. (1996), "The Association of Epidural Anesthesia With Cesarean Delivery in Nulliparas," *Obstetrics and Gynecology*, 88, 993–1000.

Little, R. J. A. (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York, New York: Wiley.

Little, R. J. A., and Schlucter, M. D. (1985), "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data With Missing Values," *Biometrika*, 72, 497–512.

Liu, C., and Rubin, D. B. (1995), "ML Estimation of the t Distribution Using EM and Its Extensions, ECM and ECME," *Statistic Sinica*, 5, 19–39.

——— (1998), "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data," *Biometrika*, 85, 673–688.

Lytle, B. W., Blackstone, E. H., Loop, F. D., Hotalling, P. L., Arnold, J. H., McCarthy, P. M., and Cosgrove, D. M. (1999), "Two Internal Thoracic Artery Grafts are Better Than One," *Journal of Thoracic and Cardiovascular Surgery*, 117, 855–872.

Meng, X. L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Nakamura, Y., Moss, A. J., Brown, M. W., Kinoshita, M., and Kawai, C. (1999), "Long-Term Nitrate Use may be Deleterious in Ischemic Heart Disease: A Study Using the Databases From Two Large-Scale Postinfarction Studies. Multicenter Myocardial Ischemia Research Group," *American Heart Journal*, 138, 577–585.

Olkin, I., and Tate, R. F. (1961), "Multivariate Correlation Models With Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics*, 32, 448–465.

Park, T., and Brown, M. B. (1994), "Models for Categorical Data With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 89, 44–52.

Rich, S. S. (1998), "Analytic Options for Asthma Genetics," *Clinical and Experimental Allergy*, 28, 108–110.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

——— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

Rubin, D. B. (1976a), "Inference and Missing Data" (with discussion), *Biometrika*, 63, 581–592.

——— (1976b), "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples," *Biometrics*, 32, 109–120.

——— (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of Survey Research Methods Section, American Statistical Association*, pp. 20–28.

——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.

——— (1997), "Estimating Causal Effects From Large Data Sets Using Propensity Scores," *Annals of Internal Medicine*, 127, 757–763.

Rubin, D. B., Schafer, J. L., and Schenker, N. (1988), "Imputation Strategies for Missing Values in Post-Enumeration Surveys," *Survey Methodology*, 14, 209–221.

Rubin, D. B., and Thomas, N. (1992a), "Affinely Invariant Matching Methods With Ellipsoidal Distributions," *The Annals of Statistics*, 20, 1079–1093.

——— (1992b), "Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions," *Biometrika*, 79, 797–809.

——— (1996), "Matching Using Estimated Propensity Scores; Relating Theory to Practice," *Biometrics*, 52, 249–264.

——— (2000), "Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates," *Journal of the American Statistical Association*, 95, 573–585.

Schafer, J. L. (1997), *Analysis of Incomplete Data*, London: Chapman and Hall.

Smith, N. L., Reiber, G. E., Psaty, B. M., Heckbert, S. R., Siscovick, D. S., Ritchie, J. L., Every, N. R., and Koepsell, T. D. (1998), "Health Outcomes Associated with Beta-Blocker and Diltiazem Treatment of Unstable Angina," *Journal of the American College of Cardiology*, 32, 1305–1311.

Takizawa, T., Haga, M., Yagi, N., Terashima, M., Uehara, H., Yokoyama, A., and Kurita, Y. (1999), "Pulmonary Function After Segmentectomy for Small Peripheral Carcinoma of the Lung," *Journal of Thoracic and Cardiovascular Surgery*, 118, 536–541.

U.S. General Accounting Office (1994), "Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and Randomized Studies: Report to the Chairman, Subcommittee on Human Resources and Intergovernmental Relations, Committee on Government Operations, House of Representatives," U.S. General Accounting Office Report GAO-PEMD-95-9.

Willoughby, A., Graubard, B. I., Hocker, A., Storr, C., Vletza, P., Thackaberry, J. M., Gerry, M. A., McCarthy, M., Gist, N. F., Magenheim, M., Berendes, H., Rhoads, G. G. (1990), "Population-Based Study of the Developmental Outcome of Children Exposed to Chloride-Deficient Infant Formula," *Pediatrics*, 85, 485–490.