INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH AND POLICY
COLUMBIA UNIVERSITY
WORKING PAPERS

REDUCING BIAS IN TREATMENT EFFECT ESTIMATION IN OBSERVATIONAL STUDIES SUFFERING FROM MISSING DATA

Jennifer Hill
School of International and Public Affairs
Columbia University

January 2004

ISERP WORKING PAPER 04-01

ISERP
Institute for Social and Economic Research and Policy

* The author gratefully acknowledges funding for this project provided by a Harvard University Graduate Society Fellowship Term Time Award.

Reducing Bias in Treatment Effect Estimation in Observational Studies Suffering from Missing Data

Jennifer Hill *
School of International and Public Affairs
Columbia University
740 IAB
420 W. 118th St.
New York, NY 10027
phone: (212) 854-4474
fax: (212) 854-5765
email: jh1030@columbia.edu

January 13, 2004

Abstract

Matching based on estimated propensity scores (that is, the estimated conditional probability of being treated) has become an increasingly popular technique for causal inference over the past decade. By balancing observed covariates, propensity score methods reduce the risk of confounding causal processes. Estimation of propensity scores in the complete data case is generally straightforward since it uses standard methods (e.g. logistic regression or discriminant analysis) and relies on diagnostics that are relatively easy to calculate and interpret. Most studies, however, have missing data. This paper illustrates a principled approach to handling missing data when estimating propensity scores makes use of multiple imputation (MI). Placing the problem within the framework of the Rubin Causal Model makes the assumptions explicit by illustrating the interaction between the treatment assignment mechanism and the missing data mechanism. Several approaches for estimating propensity scores with incomplete data using MI are presented. Results demonstrating improved efficacy compared with existing methodology are discussed. These advantages include greater bias reduction and increased facility in model choice.

Keywords: propensity scores, matching, missing data, multiple imputation, causal inference

1 Introduction

Suppose you have been given the task of evaluating the effectiveness of a new job training program. Outcome measures, such as employment indicators and earnings, as well as background covariates (labor market history, education level, age, gender) have been

*The author gratefully acknowledges funding for this project provided by a Harvard University Graduate Society Fellowship Term Time Award.

collected for the n individuals who chose to participate in the program. How would you proceed?

Most likely you would want to collect outcome and baseline data on a comparison population. The question at hand is how to best utilize this data on the program participants and comparison population when making causal statements regarding the effect of the job training program. This paper addresses this question first with the additional complications of first missing covariate data and then with the complications of both missing covariate and outcome data.

The paper begins with a brief comparison of observational studies and randomized experiments (Sections 2 and 3). Section 4 introduces the concept of propensity score matching as a potential link between these templates. The complications for implementation of this technique posed by missing data are discussed in Section 5 as the data structure and notation are put in place. Section 6 describes several existing approaches and their assumptions and then contrasts these with a new set of techniques. Simulation design and comparisons of efficacy under one set of assumptions are presented in Sections 7 and 8 followed by a discussion of additional advantages of the new techniques in Section 9. Extensions to the theory and simulations necessary for additionally accommodating missing outcome data are given in Section 10 along with corresponding simulation results. We conclude with a summary and suggestions for direction of future research.

2 Observational Studies versus Randomized Experiments: Intuition

Let's return to our example of evaluating the effectiveness of a new job training program. If we have obtained data on both the treatment group (those who participated in the job training program) and a comparison group (those who didn't participate in the job training program), why not simply compare outcomes across the two groups to estimate program effectiveness? The intuition behind why this simple comparison is generally a bad strategy is that the two groups may differ systematically in ways related to the outcome. In fact it seems reasonable that they are *likely* to differ precisely

because one group chose to participate and the other chose not to participate. Presumably, systematic differences in the study participants caused this difference in choice of "treatment". Due to these differences in pre-treatment characteristics, any difference in outcomes (means, distributions) could be attributed not just to treatment differences (received this job training program or did not) but could also be attributed to any other characteristics that differed across groups. For instance, one group might be older, more educated, or more motivated. The estimation problem arising from this phenomenon of different types of people choosing different types of real-world treatments is referred to in some social science disciplines as *selection bias*.

In order to account for these differences between treatment groups we have to be able to measure the characteristics that differ. In addition, standard approaches to treatment effect estimation such as linear regression require us to be able to model the conditional distribution of the outcomes given these measured covariates. The more similar these covariates are across groups the less important these modeling assumptions are (see, for example, Cochran and Rubin 1973). The more disparate the groups the more we have to worry about extrapolating models across large portions of the covariate space. It would be useful to have an approach to causal inference that could avoid such model dependence.

This is why researchers trust randomized experiments – the randomization creates two groups that, on average, are the same. It balances the characteristics, measured and unmeasured, of the study units across treatment groups. Therefore differences in outcomes across the two groups can be confidently attributed to the treatment without having to rely on models for the outcomes to estimate treatment effects.

3 Observational Studies vs. Randomized Experiments: Probability

We can also answer the question, "Why not simply compare outcomes across the two groups?" using probability statements. First it is helpful to define the concept of "potential outcomes," first introduced by Neyman (1923), and re-introduced into the mainstream by Rubin (1978). Let T denote the treatment indicator: $T = 0$ indicates

assignment to the control group; $T = 1$ indicates assignment to the treatment group. Together, $Y(0)$, $Y(1)$ is the set of potential outcomes such that,

$$\begin{aligned} Y(0) &= Y(T = 0), \text{ and,} \\ Y(1) &= Y(T = 1). \end{aligned}$$

That is, $Y(0)$ represents the outcome resulting from assignment to control ($T = 0$) and $Y(1)$ represents the outcome resulting from assignment to treatment ($T = 1$).

Suppose we want to estimate the average "effect of treatment on the treated" (see Heckman 1997, for a discussion). Using potential outcome notation we can define this effect as

$$E[Y(1) - Y(0) \mid T = 1] = E[Y(1) \mid T = 1] - E[Y(0) \mid T = 1]. \tag{1}$$

Unfortunately, however, *we cannot observe* $Y(0)$ *when* $T = 1$. Therefore we cannot estimate $E[Y(0) \mid T = 1]$ without further assumptions. Similar problems regarding inestimability of Equation 1 occur if we decide try to estimate other possible causal effects such as $E[Y(1) - Y(0) \mid T = 0]$ or $E[Y(1) - Y(0)]$.

3.1 Randomized Experiments

Randomized experiments, however, create a situation where the treatment groups are balanced. That is, for all covariates, measured and unmeasured, X_A,

$$T \perp X_A,$$

and, in addition,

$$T \perp Y(0), Y(1).$$

This allows

$$p(Y(0) \mid T = 1) = p(Y(0) \mid T = 0).$$

So Equation 1,

$$E[Y(1) \mid T = 1] - E[Y(0) \mid T = 1],$$

becomes equivalent to

$$E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0],$$

which is estimable.

Sometimes experimenters have information that leads them to believe that a few key manipulable characteristics are responsible for a large portion of the variability in the outcomes. In this situation it makes sense to randomize within levels (or combinations of levels) of these special covariates (often called "blocks"). This type of design is referred to as a randomized block experiment and it creates a situation where the treatment groups are balanced within subclasses defined by blocking variables B. That is,

$$T \perp\!\!\!\perp X_A \mid B,$$

and, in addition,

$$T \perp\!\!\!\perp Y(0), Y(1) \mid B.$$

This allows

$$p(Y(0) \mid T = 1, B) = p(Y(0) \mid T = 0, B).$$

So Equation 1

$$E[Y(1) \mid T = 1] - E[Y(0) \mid T = 1]$$

becomes equivalent to

$$E_B\Big[E[Y(1) \mid T = 1, B]\Big] - E_B\Big[E[Y(0) \mid T = 0, B]\Big]$$

which is estimable.

In either scenario, randomized experiments allow us to make unbiased estimates of the treatment effect.

4 Propensity Scores: Moving the Observational Study toward the Randomized Experiment Template

How can we attempt to recreate this ideal situation of balanced treatment groups in our observational study? If we assume *strong ignorability* of the assignment mechanism (Rosenbaum and Rubin 1983), then

$$T \perp\!\!\!\perp Y(0), Y(1) \mid X,$$

where X denotes all measured covariates[1]. In this case, estimating

$$E[Y(1) \mid T = 1] - E[Y(0) \mid T = 1],$$

is equivalent to estimating

$$E_X\Big[E[Y(1) \mid T = 1, X]\Big] - E_X\Big[E[Y(0) \mid T = 0, X]\Big]. \tag{2}$$

Strong ignorability of the assignment mechanism asserts basically that the observed covariates are sufficient to explain why people chose one treatment or another. The plausibility of this assumption rests on the amount of information contained in X, so often the higher the dimension of X (that is, roughly, the greater the number of covariates in X), the more plausible we might consider the assumption to be. However, conditioning on a high-dimensional X makes the choices regarding the estimation of Equation 2 less transparent. Generally it will involve modeling assumptions that become less and less plausible the more the treatment groups imbalanced in X.

4.1 Propensity Score Matching

Rosenbaum and Rubin (1983) demonstrated that matching on the "propensity score," defined as

$$e(X) = p(T = 1 \mid X),$$

produces the following results:

1. $T \perp\!\!\!\perp Y(0), Y(1) \mid e(X)$, and,
2. $T \perp\!\!\!\perp X \mid e(X)$.

This solves the multi-dimensionality problem of conditioning on the full X, becaues conditioning on $e(X)$ is sufficient for ignorability. Matching on the univariate propensity score (or conditioning on it in some other way, such as by subclassifying) in effect returns us to the template of the randomized block experiment by virtue of the fact that it works to balance all observed covariates across treatment groups.

[1] Here we temporarily put aside the issue of missing values.

4.2 Matching on Propensity Scores

Generally we can estimate the propensity score,

$$e(X) = p(T = 1 \mid X),$$

fairly trivially using, for example, logistic regression, probit reqression, or discriminant analysis. Then for each treatment group member we choose the control reservoir member with the closest propensity score. If we match without replacement, this control unit is then removed from the reservoir and cannot be used as a match for any other treatment group member. In practice, we tend to match on the logit of the estimated propensity score, $q(x)$, which is linear in the covariates and for the purposes of this paper we will think of $e(x)$ and $q(x)$ interchangeably.

Model choice for the propensity scores is governed by diagnostics such as balance between the treatment group and matched comparison group in the sample analogs of marginal means and, potentially, higher order sample moments of the joint covariate distribution. Therefore model mis-specification should be easily detectable (through resulting covariate imbalance) and is not of great concern.

5 Complications in Estimation with Incomplete Data

What happens, however, when X is not fully observed? Consideration of the hypothetical data matrix displayed in Figure 1 helps to firm up some concepts and introduce notation. As is customary, the columns of this matrix represent variables and the rows represent units.

5.1 Hypothetical Data

In general, X will represent the vector of covariate data intended to be collected for the study. In this hypothetical example $X = (X_1, X_2, X_3)^T$. The check marks in Figure 1 represent values that the researcher observes; the question marks represent missing values. X_{obs} will be used to denote the collection of observed covariate values. The missingness of the two covariates which have missing data is indexed by the indicators R_2 and R_3 (corresponding to X_2 and X_3, respectively); for each, 0 corresponds to an

X1	X2	X3	R2	R3	T	Y(0)	Y(1)
✓	✓	✓	0	0	0	✓	?
✓	✓	✓	0	0	0	✓	?
✓	✓	✓	0	0	1	?	✓
✓	✓	✓	0	0	0	✓	?
✓	?	✓	1	0	1	?	✓
✓	?	✓	1	0	1	?	✓
✓	✓	?	0	1	0	✓	?
✓	✓	?	0	1	1	?	✓
✓	?	?	1	1	0	✓	?
✓	?	?	1	1	1	?	✓

Figure 1: Hypothetical Data Matrix

observed values and 1 to a missing value. The vector of missing data indicators will be denoted R_x. The indicator for treatment receipt is T: $T = 0$ indicates assignment to the control group; $T = 1$ indicates assignment to the treatment group. The potential outcomes (defined in Section 3) are displayed in the last two columns. Notice that the only question marks in these two columns are in cells where it is impossible to observe a particular value, for instance in the column for $Y(0)$ for a unit with $T = 1$. Here outcomes that are observed are labeled "intended" and outcomes that remain unobserved because they correspond to a different treatment assignment than that received are labeled "excluded". Initially, no other forms of "missing" outcome data will be considered in this paper; later the theory will be extended to handle outcomes that are unobserved in the traditional sense (that is, a study participant who failed to respond).

When missing covariate data such as that described in this section exist it is no longer obvious how to estimate propensity scores. Any technique will have to either make a stronger assumption regarding ignorability of the assignment mechanism or will have to make an assumption about the missing data mechanism. These "mechanisms," introduced by Rubin (1978), will be described in more detail in the following section.

5.2 General Model

The following factorization of the joint sampling distribution of all of the variables will help elucidate the assumptions of each of the methods that will be discussed in the

following section:

$$p(Y(0), Y(1), T, R_x, X \mid \theta) =$$
$$p_1(X \mid \theta_1) p_2(R_x \mid X, \theta_2) p_3(T \mid X, R_x, \theta_3) p_4(Y(0), Y(1) \mid X, R_x, T, \theta_4)$$

The covariate generation process can be described by

$$p_1(X \mid \theta_1).$$

The mechanism generating the missing data can be thought of as

$$p_2(R_x \mid X, \theta_2).$$

We can conceive of the process resulting in the choice of treatment or control as

$$p_3(T \mid X, R_x, \theta_3).$$

Conditional on all of the other variables, the response surface can be represented by

$$p_4(Y(0), Y(1) \mid X, R_x, T, \theta_4).$$

As described in Section 3, propensity score matching relies heavily on the assumption of ignorability of the assignment mechanism. Since this ignorability depends on the relationship between T and all of the other study variables, it is crucial to examine our assumptions about all of these mechanisms. For instance, for all of the techniques described below we will need to assume at the minimum that

$$p_4(Y(0), Y(1) \mid X, R_x, T, \theta_4) = p_4(Y(0), Y(1) \mid X, R_x, \theta_4).$$

in order to maintain ignorability of the assignment mechanism.

6 Approaches to the missing data problem

Several possible approaches to the missing data problem exist. Complete case and complete variables strategies (and combinations thereof) are extremely common approaches to missing data. Little and Rubin (1987) outline the potential problems with reliance on these simple but unprincipled approaches. D'Agostino and Rubin (2000, henceforth

DR) developed a strategy precisely for the issue of estimating propensity scores in the presence of covariate missing data. This paper explores the differences in and potential advantages of using one of several combinations of multiple imputation (MI) and propensity score matching.

We now discuss the assumptions implicit in each of the competing methodologies in terms of these generating mechanisms defined in the previous section.

6.1 Complete Cases

Complete-case analyses use only observations where all variables are observed. This means that any unit that has any missing data is removed from the study. In the best of circumstances this will be inefficient. In general, this best case scenario assumes that the units removed, those with missing data, are just a simple random sample of the others. This is a strong assumption, formally referred to as data *missing completely at random* (Little and Rubin 1987). It does not allow the missing data mechanism to depend on any other variables. In this specific context, to make valid causal inferences with this approach we require the following new assumption regarding the above mechanisms:

$$p_2(R_x \mid X, \theta_2) = p_2(R_x \mid \theta_2)$$

This means that the observations removed from the dataset need to be a random sample of the entire dataset with respect to the covariates. Said another way, the joint distribution of the covariates has to be the same across the two groups, those with missing data and those without. The reason we don't have to specify assumptions about the relationships between R_x and T and $Y(0), Y(1)$ is due to the dependence between these variables as well as the ignorability assumption. Intuitively we know that the covariates X are the keys to our causal inference because they are what is needed for ignorability of the assignment mechanism. So if this distribution doesn't change, we can still make causal inferences. Nonetheless, this is a very strong assumption (particularly as the number of covariates needed for ignorability of the assignment mechanism to hold grows).

6.2 Complete Variables

A complete-variables analysis uses only fully the observed variables, denoted X_F. This type of analysis will fail if any of the covariates excluded are not independent of treatment assignment conditional on X_F and R, and are also related to the potential outcomes (again, conditional on X_F and R). Formally, we need

$$\begin{aligned} p_3(T \mid X, R, \theta_3) &= p_3(T \mid X_F, R, \theta_3), \text{ or,} \\ p_4(Y(0), Y(1) \mid X, R, \theta_4) &= p_4(Y(0), Y(1) \mid X_F, R, \theta_4). \end{aligned}$$

In words, we would have to believe either the variables removed were independent of treatment assignment (i.e. already balanced across treatment groups) or that ignorability of the assignment mechanism depends in fact only upon the variables retained.

This approach makes no assumption about the missing data mechanism. However, the omission of any variables with missing data will generally throw away too much information to continue to justify the ignorability of the assignment mechanism.

6.3 ECM–DR

D'Agostino and Rubin (2000) introduced the first principled solution to this problem. Their method (DR) estimates propensity scores using the ECM algorithm (an algorithm that can be used to maximize complicated likelihood functions Meng and Rubin 1993) and takes the following form:

1. Use ECM to fit a model on X_{obs}, R, T corresponding to the complete data following a general location model;
2. After convergence force T to be missing;
3. Do one E-step iteration;
4. The propensity scores are the conditional expectations of T; given the observed data.

The general location model, introduced by Olkin and Tate (1961) and now used commonly for modeling incomplete data (Little and Rubin 1987; Schafer 1997) is a flexible model which can handle both categorical and continuous data. The sampling distribution consists of both a contingency table (multinomial distribution) for the categorical

data and then, conditional on belonging to a cell in this table, a multivariate normal distribution for the continuous variables. Loglinear constraints can be placed on the parameters of the contingency table and ANOVA-like constraints can be placed on the means of the multivariate normal distribution. The variance-covariance matrices of the multivariate normals are constrained to be equal across all cells.

The DR method relies on either of the following assumptions:

$$\begin{aligned} p_3(T \mid X, R_x, \theta_3) &= p_3(T \mid X_{\text{obs}}, R_x, \theta_3), \text{ or,} \\ p_4(Y(0), Y(1) \mid X, R_x, \theta_4) &= p_4(Y(0), Y(1) \mid X_{\text{obs}}, R_x, \theta_4). \end{aligned}$$

One way of thinking of these assumptions is as follows. Within each missing data pattern (defined by R_x), we either need assignment to be independent of the covariates unobserved for that pattern, or we need ignorability to satisfied just on the basis of those covariates observed in that pattern.

The strength of this method is that, in principle, it does not make any assumption about the missing data mechanism, yet still makes weaker assumptions that the assingment mechanism and response surface than the complete variables approach. However, it does assume that either all missing covariate values are already balanced across treatment groups or that they are independent of the potential outcomes (conditional on the observed covariate values and missing data patterns).

Another potential weakness of this method is that since it specifies one model for both handling missing data and estimating propensity scores it will not always possible to incorporate Y into this model, even though it might provide useful information about the missing values. Some standard software packages for MI (for instance those that fit the general location model, which will be discussed later in the paper) cannot force indendence between binary T and continuous Y without forcing independence between T and all continuous covariates as well, clearly an undesirable property.

6.4 MI

The techniques described in the next two sections rely on a methodology known as multiple imputation (MI; Rubin 1987). MI is a methodology developed specifically to handle missing data problems. It generally relies on a Bayesian model for the data, fit

via Data Augmentation (Tanner and Wong 1987). As illustrated in Figure 2, missing values are "replaced" by draws from the predictive distribution. Specifically, for each of M independent draws from the posterior distribution a new "completed" dataset is formed that comprises the observed values and imputed values from the posterior predictive distribution of the missing values. Then any complete-data analysis can be performed on each of the imputed datasets and the results can be combined in a straightforward manner to yield the final estimate. MI for this paper was performed using Joseph Schafer's "mix" software (for the general location model) for Splus.

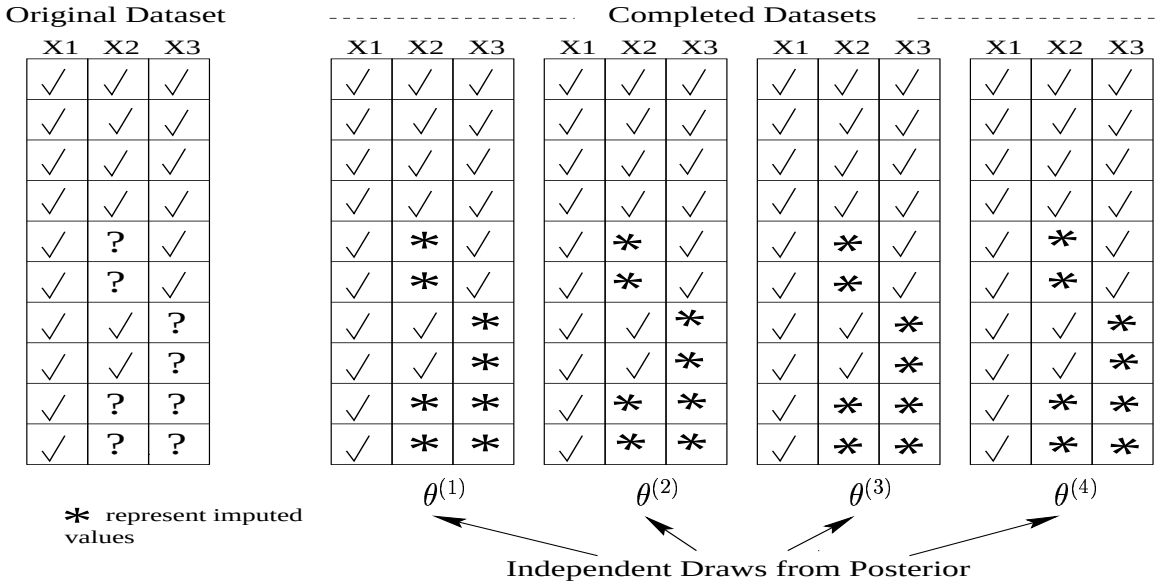

Figure 2: Original Dataset and Four Datasets Completed Using MI

6.4.1 MI Technique I

The first MI technique will be labeled MIPM-I (for MI Propensity Matching I) and consists of the following steps:

1. Use Data Augmentation (Tanner and Wong 1987) to fit a model for X_{obs}, T, Y_{obs} corresponding to the complete data following the general location model;
2. Impute M datasets;
3. For each imputed dataset calculate propensity scores;
4. Combine propensity scores across imputed datasets;

5. Pick a matched control group, calculate causal estimates.

The approach relies on the following assumption:

$$p(R_x \mid X, \theta_2) \quad = \quad p(R_x \mid X_{\text{obs}}, \theta_2)$$

In sum we assume the *latent ignorability* of the assignment mechanism. Latent ignorability was first introduced by Frangakis and Rubin (1999) as an extension of standard ignorability in the context of the missing data mechanism. It describes a situation where the mechanism is ignorable only when conditional on certain latent or missing values (in addition to the observed values). In this case, the assignment mechanism is ignorable only conditional on complete covariate data (which includes, of course, values that in practice are missing). Computationally this is achieved by filling in the missing covariate values using MI.

6.4.2 MI Technique II – MIPM-II

The second MI technique will be labeled MIPM-II (for MI Propensity Matching I) and consists of the following steps:

1. Use Data Augmentation to fit a model for X_{obs}, T, Y_{obs} corresponding to the complete data following the general location model;
2. Impute M datasets
3. For each imputed dataset calculate propensity scores, pick a matched control group, calculate causal estimates
4. Combine causal estimates across imputed datasets

This approach makes the same structural assumptions as MIPM-I.

Each of these MIPM methods will be estimated in two ways: including Y in the model and not including Y in the imputation model. The former will be labeled with (Y) after the name of the method. The latter are included to provide a more direct comparison between the DR and MIPM models.

6.4.3 Comparison of MIPM techniques

It is unclear a priori which of the two MIPM techniques we expect to dominate. One difference is in the form of the estimates (treatment effect estimates versus propensity score estimates) being combined across imputed datasets for each method and how this form fits in with the distributional theory for MI analyses. Simplistically speaking, the closer to normal the quantities being combined, the more we trust our estimates. Theoretically the appropriateness of this assumption could be evaluated for any given dataset.

Averaging the propensity scores in MIPM-I allows us to reduce the variability in the scores prior to matching which could be helpful for finding better matches. MIPM-II, however, has the advantage of averaging over the results from a variety of comparison groups. This would seem to be more robust to possible aberrations resulting from odd matching patterns. Of course use of a more formal matching algorithm (or even matching with replacement) might accomplish this.

Finally, MIPM-I is more flexible in the situation where, for example, we are matching units prospectively. If the study units are expensive and we can only afford to follow a subset of the controls, then MIPM-I could define such a set.

6.4.4 MI With Different Assumptions

Theoretically one could implement either of the above MI techniques in such a way as to satisfy the assumptions of DR. However, there are differences between the MIPM techniques and DR beyond the differences in assumptions that are clearly advantages in favor of MIPM-I and II. These are described in Section 9.

7 Simulations

This section describes the major features of the simulation created to test the efficacy of the MI techniques under the assumptions they require for validity. The number of observations of observations in each simulated dataset was 1000 and the number of imputations performed was 10.

Consider again the data generation process,

$$p(Y(0), Y(1), T, R_x, X \mid \theta) =$$
$$p_1(X \mid \theta_1) p_2(R_x \mid X, \theta_2) p_3(T \mid X, R_x, \theta_3) p_4(Y(0), Y(1) \mid X, R_x, \theta_4).$$

The data was generated from a version of this process that satisfies the assumptions needed for the MIPM methods, therefore it can be rewritten as:

$$p(X, R_x, T, Y(0), Y(1) \mid \theta) =$$
$$p_1(X \mid \theta_1) p_2(R_x \mid X_{\text{obs}}, \theta_2) p_3(T \mid X, R_x, \theta_3) p_4(Y(0), Y(1) \mid X, \theta_4).$$

7.1 Covariates

There are eight covariates in the simulation model: three continuous, $W = (W_1, W_2, W_3)$; and five categorical, $Z = (Z_1, Z_2, Z_3, Z_4, Z_5)$. These covariates jointly , $X = (W, Z)$, follow the general location model, specifically:

1. W_1,W_2, $W_3 \sim$ Multinomial(π), with loglinear constraints on π
2. $Z \sim \text{MVN}(W\delta, \Sigma)$

7.2 Missing Data Mechanism

W_1 and Z_1 are retained as fully observed variables. Response indicators are defined for the remaining six covariates. These indicators are denoted R_{W_2}, and R_{W_3}, respectively, for covariates W_2, W_3 and denoted R_{Z_2}, R_{Z_3}, R_{Z_4} and R_{Z_5}, respectively, for covariates Z_2, Z_3, Z_4, and Z_5. The vector of response indicators is denoted R_x.

I assume an ignorable missing data mechanism. The parameterization of the missing data mechanism can be best described by considering the following factorization

$$\begin{aligned}
p(R_x \mid X) &= p(R_x \mid X_{\text{obs}}) \\
&= p_a(R_{W_2} \mid W_1, Z_1, \lambda_{W_2}) \times \\
&\quad p_a(R_{W_3} \mid W_1, Z_1, W_2 R_{W_2}, \lambda_{W_3}) \times \\
&\quad p_a(R_{Z_2} \mid W_1, Z_1, W_2 R_{W_2}, W_3 R_{W_3}, \lambda_{Z_2}) \times \\
&\quad p_a(R_{Z_3} \mid W_1, Z_1, W_2 R_{W_2}, W_3 R_{W_3}, Z_2 R_{Z_2}, \lambda_{Z_3}) \times \\
&\quad p_a(R_{Z_4} \mid W_1, Z_1, W_2 R_{W_2}, W_3 R_{W_3}, Z_2 R_{Z_2}, Z_3 R_{Z_3}, \lambda_{Z_4}) \times \\
&\quad p_a(R_{Z_5} \mid W_1, Z_1, W_2 R_{W_2}, W_3 R_{W_3}, Z_2 R_{Z_2}, Z_4 R_{Z_4}, Z_4 R_{Z_4}, \lambda_{Z_5})
\end{aligned}$$

where each of the distributions is Bernoulli with expected values equal to the inverse logit of the linear combination of the covariates (or products of covariates and response indicators) multiplied by the elements of the corresponding λ vector. For example

$$p_a(R_{W_3} \mid W_1, Z_1, W_2 R_{W_2}, \lambda_{W_3}) = \text{Binomial}(n, [1 + exp(-(\mathbf{1}, W, Z, W_2 R_{W_2})\lambda_2)]^{-1})$$

where **1** is a column of 1's.

On average across all the covariates with missing data there is a missingness rate which ranges from 13% to 37%; the average of these rates is roughly 23%.

7.3 Treatment Assignment Mechanism

$$T \sim \text{Bin}(p = g(X))$$

where $g(X)$ is a linear function. On average, approximately 23% of the population was assigned to treatment.

7.4 Response Surface

The response surface took the form:

$$\begin{aligned} Y(0) \sim \text{N}(E[Y(0)] &= g(X), \sigma^2) \\ Y(1) \sim \text{Exp}(E[Y(1)] &= 1/g(X)) \end{aligned}$$

where X =(W_1, W_2, Z_1, Z_2) The response surface was chosen so that the surfaces for the potential outcomes are non-parallel overall. If they were parallel and linear throughout the covariate space then regression would handle the selection problem effectively (ignoring the missing data issues) However, parallel response surfaces do not seem to be a convincingly realistic assumption.

8 Simulation Results

8.0.1 Balance Diagnostics – Percent Reduction in Bias

We can judge the adequacy of the matching methods by the balance each produces between the resultant matched groups. If we define bias a quantity Q as

$$\text{Bias}(Q) = |\bar{Q}_a - \bar{Q}_b|$$

where $\bar{Q}_a$ and $\bar{Q}_b$ are the means of the quantity for the groups "a" and "b", respectively. Then reduction in bias between the matched groups (treatment and matched control (mc)) relative to the initial groups (treatment and control reservoir (cr)) can be defined as

$$\text{Reduction in Bias}(Q) = |\bar{Q}_t - \bar{Q}_{cr}| - |\bar{Q}_t - \bar{Q}_{mc}|,$$

and the percent reduction in bias divides this reduction by the initial bias,

$$\text{Percent Reduction in Bias}(Q) = \frac{|\bar{Q}_t - \bar{Q}_{cr}| - |\bar{Q}_t - \bar{Q}_{mc}|}{|\bar{Q}_t - \bar{Q}_{cr}|}.$$

For further discussion see Cochran and Rubin (1973).

Since this is a simulation, we have access to the missing values that the techniques couldn't use, so we use the complete data to calculate our the bias measures thus providing a more accurate picture of the true reductions in bias.

Table 1 presents the percent reduction in bias for each of the techniques discussed. The first row represents the results of propensity score matching had the data been complete and thus acts as a yardstick for the other methods. The first eight columns correspond to the eight covariates; the ninth column represents the results for the "true" sample propensity scores (those calculated with complete data and the correct model). The tenth column actually represents the percent reduction in Mahalanobis distance between the group means (thus it is not a percent reduction in bias measure).

Complete cases performs poorly overall, actually increasing the bias (shown by the negative numbers) in most cases. Complete variables fares even worse; it is severely hampered by the imbalance of the not fully observed variables (all but W_1 and Z_1) that it is forced to ignore in the estimation of the propensity scores. DR performs much better than its predecessors but is compromised by its reliance only on observed values for estimating propensity scores. MIPM-II appears to be marginally superior to DR, and then MIPM-I and MIPM-II(Y) are still better. MIPM-I(Y) performs the best overall though it doesn't dominate for all variables.

8.0.2 Treatment Effect Estimation

Of course increases in balance are just a means to an end, that end being treatment effect estimation. The first column of Table 2 displays the average bias calculated as

Methods	W_1	W_2	W_3	Z_1	Z_2	Z_3	Z_4	Z_5	$q(x)$	MD
Complete Data	85	92	80	80	88	88	88	89	91	97
Complete Cases	0	45	21	-112	53	60	63	69	45	45
Complete Variables	89	17	18	82	17	16	15	12	24	20
DR	84	81	69	79	71	74	77	77	75	90
MIPM-I	86	88	71	81	81	83	85	86	87	94
MIPM-I (Y)	86	85	73	82	88	88	88	88	92	96
MIPM-II	89	82	70	87	73	76	78	80	78	91
MIPM-II(Y)	90	82	75	89	88	87	87	85	86	94

Table 1: Percent Reduction in Bias and Mahalanobis distance across methods and variables. q(x) denotes the (linear portion of) the estimated complete-data propensity scores and MD denotes the Mahalanobis distance between the two comparison group means.

the absolute difference between the true treatment effect[2] and average point estimates of the treatment effect obtained by taking the difference in outcomes means between the treatment group and matched control group. The second column provides the mean squared error of these estimates.

The differences in average bias (absolute difference between the estimated treatment effect estimate and the true treatment effect) in treatment effect estimates appear less severe than the differences in percent reductions in bias, though they follow a similar ordering in terms of performance, with the exception of MIPM-I which performs the worst. The differences in mean squared error (MSE), however, are much more severe. The average bias coupled with the inefficiencies created by using incomplete information produce quite large MSE's for the complete-variables method in particular. For MSE MIPM-I(Y) performs the best, followed by MIPM-II(Y) and MIPM-I, with MIPM-II lagging, just as with the balance diagnostics. All four MIPM methods outperform DR in terms of MSE.

[2]The true treatment effect was 20.18.

Propensity Methods	Bias	MSE
Complete Data	0.41	0.17
Complete Cases	0.87	1.29
Complete Variables	3.37	15.42
DR	0.34	1.24
MIPM-I	0.19	0.32
MIPM-I (with Y)	0.53	0.19
MIPM-II	0.29	0.92
MIPM-II(with Y)	0.26	0.27

Table 2: Bias and Mean Squared Error in Treatment Effect Estimation

9 Additional Advantages of MI method

In addition to the gains in (and sometimes acting as a partial explanation for) bias reduction resulting under the correct assumptions, the MI techniques have other advantages over their competitors:

1. Different models for imputation and propensity scores. This allows the MI techniques to incorporate model features in one model that might be inappropriate for another. For example we can include Y in the imputation model to help predict missing X values, while inclusion of Y in our propensity score model would be a strong violation of our assumptions.

2. Better model diagnostics. Propensity score models are generally chosen based on the balance they produce. This balance cannot be measured correctly using complete variables or complete case methods unless the appropriate assumptions hold. Therefore model diagnostics may be misleading. The MIPM techniques, with no added assumptions, can produce "completed" case diagnostics which can easily be combined across datasets. DR theoretically can be used to calculate expected values for each covariate.

3. Ease of propensity model choice. Missing data models such as the general location model sometimes can be a bit tedious to fit and re-fit especially when there are

many variables involved. Propensity score model fitting often involves iterations through many versions of the model using balance diagnostics to compare models. Using the MIPM techniques the propensity score model fitting takes place on completed datasets. Refitting logistic regressions in this scenario is not nearly so cumbersome. In addition, it allows for use of propensity model-choice strategies such as automated "step-wise" procedures that have been found to produce good balance results at times.

4. Mahalanobis matching within propensity calipers. A combination of methods which has proven to be superior at producing balance than either method alone is Mahalanobis matching within propensity score calipers (Rosenbaum and Rubin 1985; Rubin and Thomas 1996). Propensity scores are used to define neighborhoods for each treatment group member and then Mahalanobis matching on a select group of "most important" variables is used to find the best match with the calipers defining the neighborhood boundaries. The completed datasets provided by the MI techniques make the augmentation of the original propensity score matching trivial. It is unclear how to proceed with this combination using any of the other approaches[3].

5. Allows for final analyses of the outcomes (such as covariance adjustments) which include covariates which are not fully observed. Often the causal estimand of interest is not simply a comparison of outcome means across treatment groups. For instance, treatment effects broken down by subgroups may be of interest. In addition, regression-adjusted results may be desired for increased precision. Such analyses may be difficult or impossible to perform if they involve missing covariate data. The MIPR techniques easily handle any such analyses.

[3]Note that filling in expected values for missing values with the ECM technique and then calculation Mahalanobis distances won't work because the distance is quadratic in the values. Furthermore, treating each distance as a variable and then calculating its expected value would add n new variables to the dataset. Finally, neither of these suggestions would be able to properly account for the uncertainty inherent in matching to a unit with missing data.

10 Extensions to Accommodate Missing Outcome Data

It is unusual in evaluation work to have missing data for baseline characteristics but have fully observed outcomes. This section extends the theory and simulation work from the previous sections to handle incomplete outcome data.

10.1 Theory

When outcome data is incomplete we must also consider the mechanism behind that missingness. Our joint distribution extends to

$$\begin{aligned} p(R_{y0}, R_{y1}, Y(0), Y(1), T, R_x, X \mid \theta) = \\ p_1(X \mid \theta_1) p_2(R_x \mid X, \theta_2) p_3(T \mid X, R_x, \theta_3) p_4(Y(0), Y(1) \mid X, R_x, T, \theta_4) \\ \times p_5(R_{y0}, R_{y1} \mid X, R_x, T, Y(0), Y(1), \theta_5) \end{aligned}$$

and we now need to consider assumptions about

$$p_5(R_{y0}, R_{y1} \mid X, R_x, T, Y(0), Y(1), \theta_5),$$

where $\theta_5 = (\theta_5^0, \theta_5^1)$.

Complete case analyses in the presence of missing data require the additional assumption

$$p_5(R_{y0}, R_{y1} \mid X, R_x, T, Y(0), Y(1), \theta_5) = p_5(R_{y0}, R_{y1} \mid R_x, T, \theta_5).$$

Similar to the discussion about complete cases with covariate missing data, as long as the outcome missing data mechanism is independent of the covariates and potential outcomes then the observations removed from the sample will be a random sample of the entire dataset with respect to those variables. If these distributions remain the same in the observations remaining then analyses can proceed appropriately.

In this scenario, strictly speaking, complete variables analyses cannot be performed (because Y has been removed from the dataset). So, we'll use a combination of complete covariate variables and cases that have the outcome observed. The additional assumption needed here then is the same as for complete cases.

We'll need to combine DR with complete cases for outcomes as well since the DR missing data model cannot incorporate outcomes in general since it is also used to compute propensity scores[4].

MIPM-I and MIPM-II do not model Y (by definition) so these will also be extended by omitting observations with missing outcomes. MIPM-I(Y) includes Y in the model but then retains only estimated propensity scores from each imputed dataset, so this information is lost[5] MIPM-II(Y) extends most naturally to accommodate missing data, relying upon the following new assumption,

$$p_5(R_{y0}, R_{y1} \mid X, R_x, T, Y(0), Y(1), \theta_5) = p_5(R_{y0}, R_{y1} \mid X_{\text{obs}}, R_x, T, \theta_5),$$

which is strictly weaker than the assumption for complete cases.

10.2 Simulations

The only change to the simulations is the addition of a missing data mechanism for $Y(0)$ and $Y(1)$. These mechanisms were specified separately and as follows:

$$\begin{aligned} p(R_{y0} \mid X, R_x, T, Y(0), Y(1), \theta_5^0) &= \text{Bin}(n, (1 + exp(-X^*\zeta_0)^{(-1)}) \\ p(R_{y1} \mid X, R_x, T, Y(0), Y(1), \theta_5^1) &= \text{Bin}(n, (1 + exp(-X^*\zeta_1)^{(-1)}) \end{aligned}$$

where X^* is the X matrix with a column of ones added as the first column. This specification does implicitly make the added assumption that $R_{y0} \perp\!\!\!\perp R_{y1} \mid X$ which has no effect on inference for our super-population parameters of interest (Rubin 1978).

The average missingness rates are approximately 24% for $Y(0)$ and 19% for $Y(1)$. We omit dependence on fully observed R_x for the sake of parsimony.

10.3 Simulation Results

Table 3 presents the percent reduction in bias for each of the techniques now that missing outcomes are also present in the same way that Table 1 did when the outcomes

[4]Alternately, missing outcomes could be imputed through some separate process, but this extension was not performed since it seems to negate the strength of DR – it's simplicity.

[5]Alternately, each imputed dataset could be retained and then analyses performed on each using the combined scores to choose matched datasets. The setting simulated here maps most closely, however, to the real-life setting of using MIPM to choose control group members prospectively – that is, before outcome data has been collected.

Methods	W_1	W_2	W_3	Z_1	Z_2	Z_3	Z_4	Z_5	$q(x)$	MD
Complete Data	85	92	82	79	89	88	88	88	98	97
Complete Cases	-59	8	-9	-178	18	28	24	41	-226	-20
Complete Variables	86	10	14	78	13	11	11	6	90	8
DR	83	75	64	76	66	70	72	73	90	86
MIPM-I	83	84	66	75	76	78	80	83	90	92
MIPM-I (Y)	77	80	65	65	84	84	83	84	90	92
MIPM-II	84	76	65	81	69	72	74	76	91	88
MIPM-II(Y)	89	81	74	87	86	85	85	85	84	94

Table 3: Percent Reduction in Bias and Mahalanobis distance across methods and variables. q(x) denotes the (linear portion of) the estimated complete-data propensity scores and MD denotes the Mahalanobis distance between the two comparison group means.

were fully observed.

Complete case analyses perform extremely poorly overall, actually increasing the bias for several quantities. Complete variables does not fare quite as poorly, however, once again it fails to correct the imbalance of the variables that it is forced to ignore. DR performs much better than its predecessors. The MIPM methods, to varying degrees, perform better than DR again with MIPM-II once again performing most similarly to DR and MIPM-I(Y) performing the best for most measures.

The performance with regard to treatment effect estimation varies much more widely than in the simulation where Y is complete. As seen in Table 4, DR, MIPM-I, and MIPM-II are hurt by the fact that the first can not and the latter two do not use Y in their estimation strategy. MIPM-I (Y) uses Y to impute the covariate data that is subsequently used to choose propensity scores, but its reliance on complete cases based on Y missingness to estimate propensity scores thereafter leads to some bias. In these simulations, MIPM-II (Y) is the clear winner, and it performs quite well given the missing data with which it had to contend.

Propensity Methods	Bias	MSE
Complete Data	0.04	0.20
Complete Cases	2.83	10.15
Complete Variables	4.75	23.52
DR	1.69	3.39
MIPM-I	1.18	1.88
MIPM-I (with Y)	0.75	0.95
MIPM-II	1.64	3.32
MIPM-II(with Y)	0.33	0.21

Table 4: Treatment Effect Estimation

11 Conclusion

Observational data with missing data, both for covariates and outcome variables, are prevalent in the social sciences. Studies using such data to make causal inferences are increasingly making use of propensity score techniques as a means for controlling for observable differences between treatment groups. These methods are complicated, however, by the addition of missing data.

This paper has illustrated two approaches to combining propensity score matching with multiple imputation and discussed the required structural assumptions. Furthermore it has evaluated the potential relative performance of these methods using simulation models with compatible assumptions. For the scenario explored here, the MI methods outperformed not only complete case and complete variables analyses, but also the proposed DR method. Additionally, the MI techniques can accommodate a broader range of missing data models, matching methods, and analysis models.

The relevance of such simulation work is always limited by the degree to which it corresponds to data that would naturally occur. Therefore, further work needs to done. Simulations should be performed under a wider variety of assumptions and different methods for combining propensity scores and MI should be explored. Furthermore, testing the comparative efficacy of these techniques in real-world applications where the true answer is known (see, for instance, Dehejia and Wahba 1999) would yield

potentially even more relevant information for practitioners than these tests using simulated data. Finally, robustness of these approaches to model mis-specification should be investigated. Absent these extensions we have some evidence, however, that MI and propensity score matching can be successfully combined in the context of non-randomized data.

References

Cochran, W. G. and Rubin, D. B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya* 35, 417–446.

D'Agostino, Ralph B., J. and Rubin, D. B. (2000), "Estimating and using propensity scores with partially missing data," *JASA* 95, 749–759.

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

Frangakis, C. E. and Rubin, D. B. (1999), "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes," *Biometrika* 86, 365–380.

Heckman, J. J. (1997), "Instrumental Variables: A Study of the Implicit Assumptions Underlying One Widely Used Estimator for Program Evaluations," *Journal of Human Resources* 32, 441–462.

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

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

Neyman, J. (1923), "On the Application of Probablity Theory to Agricultural Experiments Essay on Principles. Section 9," translated in *Statistical Science* 5, 465–480, 1990.

Olkin, I. and Tate, R. F. (1961), "Multivariate Correlation Models With Mixed Discrete and Continuous Variables (Corr: V36 P343)," *The Annals of Mathematical Statistics* 32, 448–465.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70, 1, 41–55.

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. (1978), "Bayesian Inference for Causal Effects: The role of randomization," *The Annals of Statistics* 6, 34–58.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D. B. and Thomas, N. (1996), "Matching using estimated propensity scores: Relating theory to practice," *Biometrics* 52, 249–264.

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

Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions By Data Augmentation (C/R: P541-550)," *Journal of the American Statistical Association* 82, 528–540.

Recent ISERP Working Papers

03-01: "The Plasticity of Participation: Evidence From a Participatory Governance Experiment," Shubham Chaudhuri, Economics, Columbia University, and Patrick Heller, Sociology, Brown University

03-02: "Factional Politics and Credit Networks in Revolutionary Vermont," Henning Hillmann, Sociology, Columbia University

03-03 " 'Active Patients' in Rural African Health Care: Implications for Welfare, Policy and Privatization," Kenneth L. Leonard, Economics, Columbia University

03-04 "Living at the Edge: America's Low-Income Children and Families," Hsien-Hen Lu, Public Health, Columbia University, Julian Palmer, Younghwan Song, Economics, Union College, Mary Clare Lennon, Public Health, Columbia University, Lawrence Aber, Public Health, Columbia University

02-01 "Alternative Models of Dynamics in Binary Time-Series-Cross-Section Models: The Example of State Failure," Nathaniel Beck, Political Science, UC San Diego, David Epstein, Political Science, Columbia University, Simon Jackman, Political Science, Stanford University and Sharyn O'Halloran, Political Science, Columbia University

02-02 "Substitutability Cross-Stream Between Oriented Markets: Conventions in the Wine Sector of France," Harrison White, Sociology, Columbia University

02-03 "Link, Search, Interact: The Co-Evolution of NGOs and Interactive Technology," Jonathan Bach, Center on Organizational Innovation, Columbia University and David Stark, Center on Organizational Innovation, Columbia University

02-04 "Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks," Peter Bearman, Institute for Social and Economic Research and Policy, Columbia University, James Moody, Sociology, Ohio State, Katherine Stovel, Sociology, University of Washington

02-05 "Permanently Beta: Responsive Organization in the Internet Era," Gina Neff, Center on Organizational Innovation (COI), Columbia University, and David Stark, Center on Organizational Innovation (COI), Columbia University

02-06 "Negotiating the End of Transition: A Network Approach to Political Discourse Dynamics, Hungary 1997," Balázs Vedres, Columbia University, Péter Csigó, Ecole des Hautes Etudes en Sciences Sociales

For copies of ISERP Working Papers
visit http://www.iserp.columbia.edu/initiatives/working_papers/paper_program.html,
write to iserp@columbia.edu or call 212-854-3081

Institute for Social and Economic
Research and Policy
Columbia University
International Affairs Building
420 West 118 Street, 8th Floor
Mail Code 3355
New York, NY 10027
telephone: 212-854-3081
facsimile: 212-854-8925
e-mail: iserp@columbia.edu
URL: http://www.iserp.columbia.edu