UNIVERSITY OF
Southampton
Southampton Statistical
Sciences Research Institute

Working Paper M11/06
Methodology

Propensity Score Matching With Missing Covariates Via Iterated, Sequential Multiple Imputation

Robin Mitra, Jerome P. Reiter

Abstract

In many observational studies, analysts estimate causal effects using propensity score matching. Estimation of propensity scores is complicated when covariate values intended for collection are in fact missing. To handle the missing data, one approach is to use multiple imputation to create completed datasets, and compute propensity scores from these datasets. However, inaccurate imputation models can result in ineffective matching, thereby limiting reductions in bias. We propose a multiple imputation approach based on chained equations in which the researcher gradually reduces the set of control units used to estimate the imputation models. This approach can reduce the influence of control records far from the treated units' region of the covariate space on the estimation of parameters in the imputation model, which can result in more plausible imputations and better balance in the true covariate distributions. This approach can be conveniently implemented with standard multiple imputation software for missing data. Using simulations, we find that the approach can improve estimation when imputation models are mis-specified; however, it can be ineffective when imputation models are correctly specified. This suggests using the approach as part of sensitivity analysis in causal inference. We apply the approach to an observational study of the effect of breast-feeding onthe child's educational outcomes later in life.

Propensity score matching with missing covariates via iterated, sequential multiple imputation

ROBIN MITRA[*]

School of Mathematics, University of Southampton,

SO17 1BJ, UK, R.Mitra@soton.ac.uk

Tel. (+44) 2380 594550 Fax. (+44) 2380 595147

JEROME P. REITER

Department of Statistical Science, Duke University,

Box 90251, Durham, NC 27708, USA, jerry@stat.duke.edu

[*] *corresponding author*

Abstract

In many observational studies, analysts estimate causal effects using propensity score matching. Estimation of propensity scores is complicated when covariate values intended for collection are in fact missing. To handle the missing data, one approach is to use multiple imputation to create completed datasets, and compute propensity scores from these datasets. However, inaccurate imputation models can result in ineffective matching, thereby limiting reductions in bias. We propose a multiple imputation approach based on chained equations in which the researcher gradually reduces the set of control units used to estimate the imputation models. This approach can reduce the influence of control records far from the treated units' region of the covariate space on the estimation of parameters in the imputation model, which can result in more plausible imputations and better balance in the true covariate distributions. This approach can be conveniently implemented with standard multiple imputation software for missing data. Using simulations, we find that the approach can improve estimation when imputation models are mis-specified; however, it can be ineffective when imputation models are correctly specified. This suggests using the approach as part of sensitivity analysis in causal inference. We apply the approach to an observational study of the effect of breast-feeding on the child's educational outcomes later in life.

Keywords: Missing data, Multiple imputation, Observational studies, Propensity scores

1 INTRODUCTION

When estimating causal effects in observational studies, many data analysts use propensity score matching (Rosenbaum and Rubin, 1983, 1985) to reduce the bias that can result from imbalanced covariate distributions in the treated and full control groups (e.g., Cho *et al.*, 2007; da Veiga and Wilder, 2008; Austin, 2008, 2009a,b). The propensity score for the ith subject, $e(\mathbf{x}_i)$, is the probability that the subject receives the treatment given its vector of covariates $\mathbf{x}_i$; that is, $e(\mathbf{x}_i) = P(T_i = 1|\mathbf{x}_i)$, where $T_i = 1$ if subject i receives treatment and $T_i = 0$ otherwise. Rosenbaum and Rubin (1983) show that when two large groups have the same distributions of propensity scores, the groups should have similar distributions of $\mathbf{x}$. Thus, by selecting control records whose propensity scores are similar to the treated records' propensity scores, analysts can create a matched control group whose covariates are similar to the treated group's covariates. Analysts then base inference on the treated and matched control groups, thereby avoiding bias that results from imbalanced covariate distributions in the two groups, at least for those covariates in $\mathbf{x}$.

In practice, analysts must estimate propensity scores from the data. Typically, this is done by regressing $\mathbf{T}$ on functions of $\mathbf{x}$ and using the estimated probabilities as the propensity scores. Alternatives to logistic regression include generalized additive models (Woo *et al.*, 2008) and machine learning methods (Westreich *et al.*, 2010), among others. Once the analyst estimates all $e(\mathbf{x}_i)$, he or she can create the matched control group using one of many strategies, including, for example, nearest neighbor pair matching with or without replacement, full matching (Rosenbaum, 1991; Stuart and Green, 2008), genetic matching (Diamond and Sekhon, 2005), and matching within pre-specified calipers (Austin, 2009b). Here, we use nearest neighbor pair matching without replacement. Analysts also can use propensity scores for sub-classification (Rosenbaum and Rubin, 1984; Hullsiek and Louis, 2002; Graf, 1997; Drake and McQuarrie, 1993; Zanutto, 2006) and inverse probability weighted estimation (Robins *et al.*, 2000; Lunceford and Davidian, 2004); we do not consider these methods here.

We consider matching scenarios in which some covariate data are missing, so that complete-data propensity score estimation methods cannot be easily fit to the observed data. Several approaches for handling such scenarios have been proposed in the literature. For example, D'Agostino Jr. and Rubin (2000) propose an EM algorithm for estimating the propensity scores. Haviland *et al.* (2007) include missing data indicators as additional covariates in the propensity score model, so that matches are selected within patterns of missing data. Alternatively, analysts can consider doubly robust estimators (Robins *et al.*, 1994; Liang *et al.*, 2004).

We focus on approaches based on multiple imputation (Rubin, 1987), in which the data analyst repeatedly imputes missing covariate values by sampling from predictive distributions conditional on the observed data. The analyst estimates propensity scores in each completed dataset, averages the propensity scores across datasets, and matches on the averaged scores. With multiply-imputed datasets, the analyst can easily pursue further modeling, such as regression adjustment to reduce residual imbalances or sub-domain comparisons (Hill, 2004; Hill *et al.*, 2004). Additionally, the analyst can use propensity score estimation approaches that are not tied to the imputation models, for example algorithmic approaches based on regression trees (Setoguchi *et al.*, 2008). For further discussion of propensity score estimation with multiple imputation, see Hill (2004) and Qu and Lipkovich (2009).

Mitra and Reiter (2010) demonstrated empirically that the bias-reduction from propensity score matching can be limited when the imputed covariate values are implausible. This is because acceptable balance on implausible imputed (and observed) covariates may not equate to acceptable balance on the true values of the covariates. When the study has many covariates spread over a large space, it is all-too-easy for analysts' imputation models to make inappropriate extrapolations that are difficult to check in practice. Indeed, the difficulties of specifying models and the desire to avoid extrapolations in high dimensions motivate propensity score methods in place of regression analysis for causal inference in the first place.

To mitigate these problems, Mitra and Reiter (2010) proposed a latent class, general location mixture model for multiple imputation. This model conceives of the control units as a latent mixture of units whose covariates are drawn from the same distributions as the treated units' covariates and units whose covariates are drawn from different distributions. By using only the cases in the first mixture component to estimate imputation models, this model reduces the influence of control records far from the treated units' region of the covariate space on the estimation of parameters in the imputation model. This can result in more plausible imputations in the region where control and treated units' covariate distributions overlap most, which is where matches are likely to come from. Since matches are based on imputed values, better imputation models can result in better balance in the true covariate distributions.

The latent class, general location mixture model approach requires computationally intensive Markov Chain Monte Carlo sampling and specification of the joint distribution of covariates. Many practitioners are likely to prefer less computationally intense imputation methods and specification of conditional rather than joint models (Van Buuren, 2007).

In this article, we propose to modify sequential regression multivariate imputation (SRMI) (Raghunathan *et al.*, 2001), also known as regression switching (van Buuren *et al.*, 1999), chained equations (Van Buuren and Groothuis-Oudshoorn, 2000), partially incompatible MCMC (Rubin, 2003), and iterative univariate imputation (Gelman, 2004). The basic idea is to repeatedly apply the SRMI on successively smaller sets of control cases, each time tossing out cases that are not plausible matches based on the imputed values. We refer to this as the winnow method; it is described in detail in Section 3. We apply the winnow method to analyze an observational study with missing covariate data on the effects of breast-feeding; the data and application are described in Sections 2 and 5. We evaluate the winnow method and compare it to standard SRMI via simulation studies; these are described in Section 4. Finally, we conclude with general remarks about using the winnow method and multiple imputation for propensity score matching.

2 Motivating application: Breast-feeding and educational outcomes

We motivate the methodology with an observational study concerning the effect of breast-feeding on child's educational outcomes later in life. The data come from the U.S. National Longitudinal Survey of Youth (NLSY). The NLSY has been recruiting youths to the survey from 1979 onwards. It measures a range of social, economic and health related characteristics on these youths. This study was specifically interested in the effect of breast-feeding on cognitive development for children born to these youths. This study also was analyzed by Mitra and Reiter (2010).

The response variable is the Peabody individual assessment test (PIATM) math score administered to children at 5 or 6 years of age. The treatment variable is breast feeding duration in weeks. We dichotomize this variable into a control condition, < 24 weeks, and a treatment condition, ≥ 24 weeks. The 24 week cutoff corresponds to the number that has been given by the American Academy of Pediatrics Chantry *et al.* (2006) and the World Health Organization as a minimum standard for breast feeding duration. There are other ways to define the treatment variable, and the analysis could be repeated with different cut points on the breast feeding duration variable.

We do not pursue these here. Additionally, we cannot determine from these data whether or not the mother used breast feeding exclusively.

Fourteen background covariates are measured. These include five categorical variables: the child's race (Hispanic, black or other), the mother's race (Hispanic, black, Asian, white, Hawaiian/PI/American Indian, or other), child's sex, and two variables indicating whether the spouse or grandparents were present at birth. They also include nine continuous variables, including difference between mother's age at birth and in 1979, mother's intelligence as measured by an armed forces qualification test, mother's highest educational attainment, the weeks worked by the mother in the year prior to giving birth, child's birth weight, the number of weeks the child was born premature, the number of days that the child spent in hospital, the number of days that the mother spent in hospital, and family income.

We include only first born children in the analysis to avoid complications due to birth order and family nesting. In addition, we discard 506 units with missing values in their treatment variable (breast-feeding duration) and 4977 units with a missing outcome variable (PIATM). Excluding these units is reasonable under missing at random assumptions, which may not be true in practice. We do not consider other methods for handling the missing treatment indicators and missing outcome data in the analysis here. The resulting data set comprise $n = 2388$ youths of whom $n_T = 370$ are treated. Of these, 1306 have complete data on all covariates, of whom 216 are treated. Three covariates were completely observed in the study, and nine covariates had missing data rates of less than 10%. The two covariates with the largest rates of missing data were family income (22.4%) and the number of weeks that the mother worked in the year prior to giving birth (23.1%).

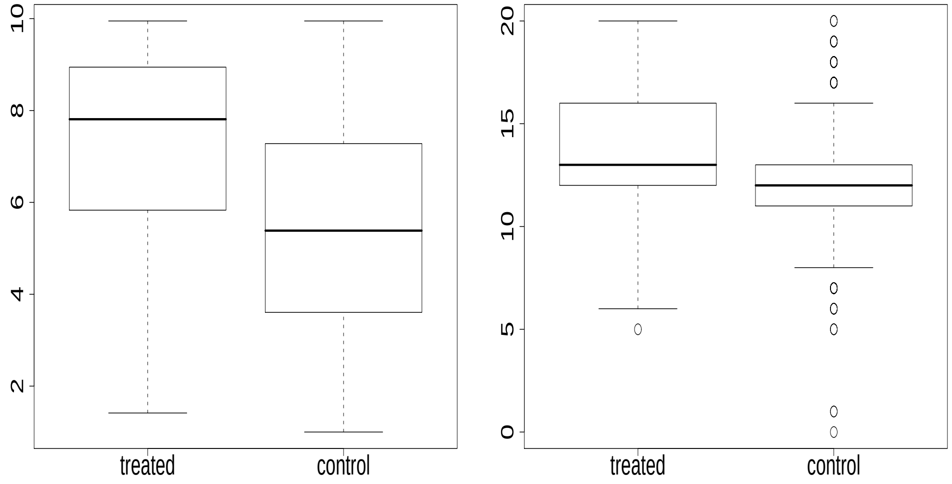

Figure 1: Box plots of mother's intelligence score and mother's years of education respectively for treated and control units

Several of the covariates are imbalanced in the treated and control group. For example, Figure 1 displays box plots of observed mother's intelligence and education for treated and all control units, and Table 1 displays the proportion of treated and control units in each level of child's race. Treated units tend to have higher mother's intelligence scores, more mother's years of education and lower proportions of Hispanics and blacks.

These differences motivate us to use propensity score matching to reduce potential biases from the imbalanced covariates. Since the breast-feeding data are plagued by missing covariate data, we cannot apply complete-data techniques for estimating propensity scores. We therefore use multiple imputation to create completed datasets, with which propensity scores can be estimated.

race	treated	control
Hispanic	0.1378	0.1903
black	0.1108	0.2844
other	0.7514	0.5253

Table 1: Distribution of child's race for treated and control units

3 Sequential imputation and the winnow method

When covariates include categorical and continuous variables, many analysts use sequential regression imputation to create the completed datasets. Routines implementing this approach are available in the software packages R, SAS, and Stata. In this section, we describe SRMI and how it can lead to ineffective matching when the imputation models do not correctly describe the missing data. We then present the winnow method and argue why it might correct these deficiencies.

3.1 Description of SRMI

Suppose that the data comprise n units and p partially observed covariates $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$, where x_{ij} corresponds to the ith unit's value for the jth covariate. Here, the covariates are ordered with increasing percentages of missing data. Let $\mathbf{D}$ represent the set of covariate variables that are fully observed.

In SRMI, we impute missing values using an iterative procedure for $t = 1, \ldots, T$ iterations, where ideally T is large. For $t = 1$, missing values in $\mathbf{x}_1$ are imputed using some predictive distribution $f_1(\mathbf{x}_1|\mathbf{D})$; denote this completed covariate $\mathbf{x}_1^{(1)}$. Missing values in $\mathbf{x}_2$ are imputed using $f_2(\mathbf{x}_2|\mathbf{D}, \mathbf{x}_1^{(1)})$; denote this completed covariate $\mathbf{x}_2^{(1)}$.

This continues until missing $\mathbf{x}_p$ are imputed using $f_p(\mathbf{x}_p|\mathbf{D}, \mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{p-1}^{(1)})$. Here, the form of each f_j depends on the type of $\mathbf{x}_j$. For continuous $\mathbf{x}_j$, f_j is typically a normal linear regression; for categorical $\mathbf{x}_j$, f_j is typically a multinomial regression. Flat prior distributions are used for all imputation model parameters. In subsequent iterations ($t > 1$), the SRMI method cycles through a sequence of conditional regressions, $g_j(\mathbf{x}_j|\mathbf{D}, \mathbf{x}_1^t, \ldots, \mathbf{x}_{j-1}^t, \mathbf{x}_{j+1}^{t-1}, \ldots, \mathbf{x}_p^{t-1})$, to impute missing values in $\mathbf{x}_j$.

SRMI can be viewed as an approximation to a Gibbs sampler. In certain situations, e.g., when the data arise from a multivariate normal distribution, SRMI converges to a stable joint posterior distribution. In other situations, convergence is not guaranteed (Gelman and Speed, 1993). However, the SRMI tends to perform well in practice (Van Buuren *et al.*, 2006). Once the SRMI method has cycled through a sufficient number of iterations, the imputed values generated in the final iteration are used to create an imputed data set. This approach is applied at m different random starting points resulting in m imputed data sets.

3.2 Winnow method

Many researchers apply SRMI with default specifications, e.g., additive effects for each of the predictors in the model without any transformations. However, with high dimensional covariate spaces, default specifications easily could result in inappropriate extrapolations, leading to implausible imputations and ineffective matching. To illustrate this, we summarize some of the results of Mitra and Reiter (2010).

Consider the bivariate distribution of covariates $(\mathbf{x}_1, \mathbf{x}_2)$ in Figure 2, which also was used by Mitra and Reiter (2010). Suppose that the only missing data are in $\mathbf{x}_2$ for some control units. A default application of SRMI is to impute missing $\mathbf{x}_2$ using a linear regression on $\mathbf{x}_1$, which is clearly inaccurate. What can happen when this model is used to impute missing values of $\mathbf{x}_2$? First, consider control units with actual covariate values of both x_1 and x_2 in the treated units' region of the covariate space and above the regression line; these are ideal candidates to be included in the matched control set. When based on default SRMI, imputations of missing x_2 for these control units will tend to be lower than the actual values. As a result, these control units' completed data could be in a different space than the treated units covariates. If propensity score matching is done with the completed data, these control units could be (incorrectly) excluded from the matched control set. Second, consider control units with values of x_1 similar to treated units' values of x_1 but with actual values of x_2 that are smaller than any treated units' values of x_2 (e.g., below the regression line); these are not good candidates for the matched control set. When based on the default SRMI, imputations of missing x_2 for these units will tend to be higher than their true x_2 values, so that their imputed values could be in the same region as the treated units' covariates. Therefore, they could be incorrectly selected as matched controls. We note that control units whose covariates are far away from the treated units' covariate space are not likely to be selected as matches, even with the model mis-specification.

These problems motivate the winnow method. Basically, in the winnow method, we seek to toss out control records that are not plausible matches, and refit imputation models only with the remaining units. In this way, we hope to tailor imputation models to the area of covariate space inhabited by the treated units. With smaller regions, default specifications are less prone to the effects of model mis-specification. For example, in Figure 2, a linear relationship for the covariates is inappropriate over

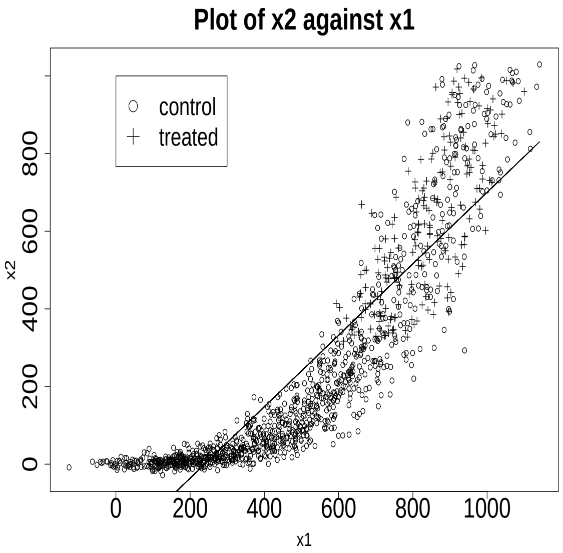

Figure 2: Scatter plot of $\mathbf{x}_2$ against $\mathbf{x}_1$ when a cubic relationship is present, illustrating the effects of using a poor imputation model.

the whole covariate space, but it is not an unreasonable approximation over only the region where the treated units lie.

We toss out unlikely matches by using SRMI iteratively, at each step reducing the set of possible matched controls used to estimate the SRMI. Specifically, the algorithm involves the following steps.

1. Let $\mathbf{Z}$ be the covariate data on all n units in the study. Let $k = 3$.

2. Use default SRMI to create m multiple imputations of the missing covariates in $\mathbf{Z}$. Estimate the propensity scores from the multiply-imputed data.

3. Using these propensity scores, find the k closest controls for each treated unit using nearest neighbor pair matching without replacement.

4. Let $\mathbf{Z}$ now be the n_T treated units and the kn_T matched control units from step 3; that is, delete the non-matched controls from the data. Replace the current

value of k with $k-1$.

5. Repeat step 2 through step 4 until the current value of $k = 0$. The $2n_T$ records at the last step include the treated and matched control units.

We set $k = 3$ for computational convenience, as our experience suggests that value still enables tossing out of outlying control records without excessive refitting of SRMI. The winnow method can be easily adapted to imputation routines other than SRMI, including joint model specification approaches such as those in Schafer (1997).

4 Simulation studies

We now empirically evaluate the performance of the winnow method and compare it with default application of SRMI. To begin, we use the setting of Figure 2, so that the default SRMI is not a plausible imputation model. In this figure, we simulate $n = 1200$ units with two continuous covariates such that

$$x_{i1} = 50 + 0.8i + \epsilon_{i1}, \quad \epsilon_{i1} \sim N(0, 75) \tag{1}$$

$$x_{i2} = 0.000001i^3 + \epsilon_{i2}, \quad \epsilon_{i2} \sim N(0, 10) \tag{2}$$

for $i = 1, \ldots, 1200$. This results in the covariates having a cubic relationship. We assign treatment so that, for $i = 1, \ldots, 1200$, the $P(T_i = 1) = 0.5I(i > 800)$, where $I(\cdot) = 1$ when the condition inside the parentheses holds and $I(\cdot) = 0$ otherwise. In this way, the treated units tend to have larger values of x_1 and x_2, and $n_T = 200$.

Initially, we introduce missing values in $\mathbf{x}_2$ for only the control units using a missing at random mechanism. For each unit i, let $r_i = 1$ indicate that x_{i2} is missing and $r_i = 0$ indicate that x_{i2} is observed. We draw r_i independently across i with

$$P(r_i = 1) = exp\big(-3 + 0.005x_{i1}\big)/\big\{1 + exp\big(-3 + 0.005x_{i1}\big)\big\}. \tag{3}$$

With this mechanism, control units with larger values of x_1 are more likely to be missing x_2. In this way, the missing values for control units are concentrated in the covariate space of the treated units. If missing data were instead concentrated in covariate space that was far from the treated units, the observed cases would be the best matches, and there would be no impact of model mis-specification. Approximately 35% control units are missing x_2 values.

We investigate two scenarios in which treated units also are missing values in x_2. In the first scenario, we randomly delete x_2 for 10% of the treated units; in the second, we randomly delete x_2 for 30% of the treated units. Hence, the treated data are missing completely at random. This mechanism is selected for simplicity, since the treated units' covariates are already in a relatively small region of the covariate space compared to the entire distribution for the control units.

For each scenario, we generate a response variable $\mathbf{y}$ with a linear regression, $y_i \sim N(x_{i1} + x_{i2}, 1)$, for all i. Thus, the true treatment effect $\tau = 0$. We estimate τ with $\hat{\tau} = \bar{y}_T - \bar{y}_{MC}$, where $\bar{y}_T$ is the sample mean of $\mathbf{y}$ in the treatment group and $\bar{y}_{MC}$ is the sample mean of $\mathbf{y}$ in the matched control group.

We replicate this simulation design 100 times per scenario, each time generating new values of $\mathbf{y}$ and $\mathbf{r}$. We generate $m = 5$ datasets using default SRMI (which we label as the once only method) and at each cycle of the winnow method. In each replication, we estimate the treatment effect after propensity score matching without replacement based on the default SRMI and winnow methods. For benchmarking, we also record the treatment effect estimate after propensity score matching from the fully observed data, i.e., prior to introducing missing values. Figure 3 displays box plots of the 100 treatment effect estimates from all three scenarios. The winnow method tends to estimate treatment effects closer to τ compared to the once only method. This trend holds for all three scenarios; in fact, in this simulation, the performance of each method does not change much as we introduce missing data in the treated units.

Of course, when confronted with data like Figure 2, a wise modeler would recognize the inadequacy of the default SRMI from exploratory data analysis and use some other imputation approach, for example transformations. However, in practice, it is not always easy to diagnose imputation model inadequacies with high dimensional data. Furthermore, although unfortunate, many data analysts do not carefully check imputation models, so that they may face the problems from imputation model misspecification.

We next investigate a simulation where the default SRMI is a plausible method to impute the missing values. Following Mitra and Reiter (2010), we simulate x_1 and x_2 so that

$$\begin{aligned} x_{1i} &= 50 + 0.8i + \epsilon_{1i}, \quad \epsilon_{1i} \sim N(0, 75) \qquad (4) \\ x_{2i} &= 50 + 0.8i + \epsilon_{2i}, \quad \epsilon_{2i} \sim N(0, 10) \qquad (5) \end{aligned}$$

for $i = 1, \ldots, 1200$. As shown in Figure 4, this results in a linear relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$; hence, the default SRMI is a reasonable model. The treatment assignment, response, and missing data mechanisms are generated as in the previous set of simulations.

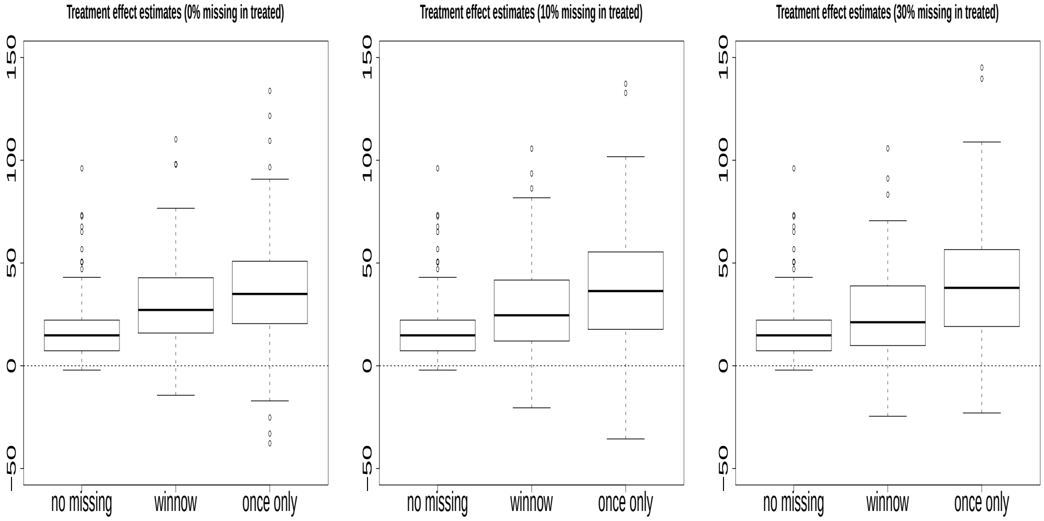

Figure 3: Treatment effect estimates in the model mis-specification simulation when the covariates are fully observed, when using the winnow method and when using the once only method respectively. Missing data are introduced in 0%, 10% and 30% of treated units' x_2 covariate respectively. The dotted line indicates the true treatment effect of 0.

Figure 5 summarizes the treatment effect estimates for all three scenarios. Here, the estimates from the default SRMI method tend to be closer to $\tau = 0$ than the winnow estimates. Hence, applying the winnow method in situations where default SRMI is appropriate reduces the effectiveness of matching.

It is perhaps not surprising that neither method always dominates the other. From these results, we do not believe it is justified to recommend one approach over the other. Instead, we suggest using the winnow method as part of a sensitivity analysis. Analysts can fit SRMI on all n units—ideally after exploratory data analysis to improve imputation model accuracy—and obtain estimates of treatment effects. Then, they run the winnow method to obtain another set of estimates. When the

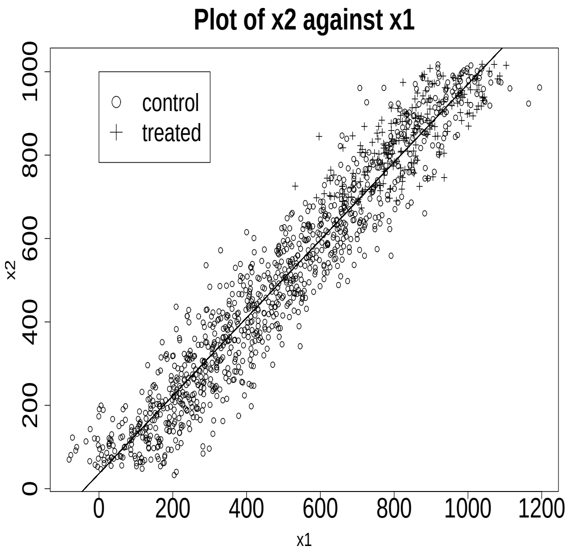

Figure 4: Scatter plot of x_2 against x_1 when a linear relationship is present and the once only imputation method is plausible.

estimates from both approaches are similar, analysts can feel more confident that the inference is not greatly affected by possible imputation model mis-specification. When the estimates differ in ways that have practical significance, analysts may want to redouble their efforts in checking the fit of the imputation models.

5 Application to the breast-feeding study

We now apply default SRMI and the winnow method to impute missing covariates in the study of breast-feeding on child's cognitive development. We generate $m = 5$ imputed datasets for the default SRMI method and at each cycle of the winnow method. For the imputation models, we categorized two of the continuous variables. The variable measuring weeks preterm has a large spike at zero weeks. We therefore categorized the preterm variable into three levels; not preterm (zero weeks), moderately preterm (one to four weeks), and very preterm (five or more weeks), with cut points

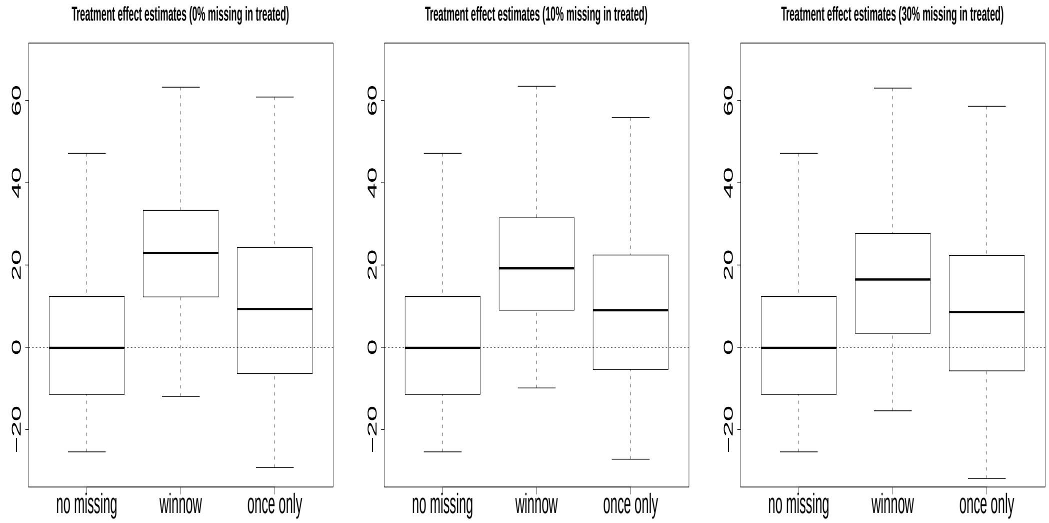

Figure 5: Treatment effect estimates in the 'once only correct' simulation when the covariates are fully observed, when using the winnow method, and when using the once only method respectively. Missing data are introduced in 0%, 10%, and 30% of treated units' x_2 covariate respectively. The dotted line indicates the true treatment effect of 0.

determined from guidelines from the March of Dimes (*www.marchofdimes.com*). The variable measuring number of weeks worked in the year prior to giving birth has a distinct U shape; this would be difficult to model using the default specifications in SRMI. We therefore categorized this variable into four levels (not worked at all, worked between 1 and 47 weeks, worked 48-51 weeks, and worked all 52 weeks). We also applied Box-Cox transformations (Box and Cox, 1964) to several other continuous variables to improve assumptions of normality. Further details on the transformations applied to variables in the study are available in Mitra and Reiter (2010).

We estimate τ with $\hat{\tau} = \bar{Y}_T - \bar{Y}_{MC}$. Using the winnow method, $\hat{\tau} = 1.38$ points of the PIATM score with a matched pairs standard error of 0.95 (two-sample pooled standard error of 0.96). For a discussion of approaches to estimating standard errors from propensity score matching, see Austin (2009c) for matching without replacement contexts. The default SRMI (once only) method estimates the treatment effect to be 1.69 (matched pairs SE = 0.93, two-sample SE = 0.98). Thus, the methods result in less than half a point difference; this may not be practically significant. In this application, the once only SRMI may be reasonable due to the careful modeling of the data prior to imputation.

Because of the modest number of imputations ($m = 5$), the estimates of the treatment effect for the winnow and default SRMI methods can be expected to fluctuate somewhat. However, when we repeated both imputations multiple times, the differences between the winnow and the default SRMI methods remain less than one point of the PIATM. Analysts could reduce the fluctuation by significantly increasing m, although this comes at the price of increased computations.

While the two imputation methods give similar results, the treatment effect estimate from the full sample is 5.23 (two-sample SE = 0.741). This is substantially larger than estimates obtained from either method.

6 Concluding Remarks

The winnow method is a simple approach for multiple imputation of missing covariates to enable propensity score matching. Simulations suggest that it can result in better matches when imputation models are mis-specified; however, it does not perform as well when imputation models are correctly specified. Of course, it is challenging to determine the validity of the imputation models in genuine settings.

Hence, we recommend implementing the winnow method and default SRMI—both with suitable transformations—as a check on the impact of the imputation models on causal estimates.

Acknowledgments

The authors would like to thank Professor Jennifer Hill who provided us with the data on the breast-feeding study. This research was supported by the National Science Foundation [NSF-ITR-0427889].

References

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* **27**, 12, 2037–2049.

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment group in propensity-score matched samples. *Statistics in Medicine* Published online in Wiley InterScience (www.interscience.wiley.com).

Austin, P. C. (2009b). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal* **51**, 1, 171–184.

Austin, P. C. (2009c). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics* **5**, 1.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* **26**, 2, 211–252.

Chantry, C. J., Howard, C. R., and Auinger, P. (2006). Full breastfeeding duration and associated decrease in respiratory tract infection in us children. *Pediatrics* **117**, 2, 425–432.

Cho, Y. B., Lee, K., Suh, K., Kim, Y., Yoon, J., Lee, H., S., H., and Park, B. (2007). Maternal smoking during pregnancy and birthweight: a propensity score matching approach. *Journal of Gastroenterology & Hepatology* **22**, 10, 1643–1649.

da Veiga, P. V. and Wilder, R. P. (2008). Maternal smoking during pregnancy and birthweight: a propensity score matching approach. *Maternal and Child Health Journal* **12**, 2, 194–203.

D'Agostino Jr., R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* **95**, 451, 749–759.

Diamond, A. and Sekhon, J. (2005). Genetic matching for estimating causal effects. presented at the Society for Political Methodology Meeting, FSU, July 21-23 2005.

Drake, C. and McQuarrie, A. (1993). The power of the Mantel-Haenszel test when adjustment is by the true or estimated propensity score. *Biometrical Journal* **35**, 445–449.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association* **99**, 466, 537–545.

Gelman, A. and Speed, T. P. (1993). Characterizing a joint probability distribution by conditionals (Corr: 1999V61 p483). *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **55**, 185–188.

Graf, E. (1997). The propensity score in the analysis of therapeutic studies. *Biometrical Journal* **39**, 297–307.

Haviland, A., Nagin, D., and Rosenbaum, P. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods* **12**, 3, 247–267.

Hill, J. (2004). Reducing bias in treatment effect estimation in observational studies suffering from missing data. *Columbia University Institute for Social and Economic Research and Policy (ISERP)* working paper 04-01.

Hill, J. L., Reiter, J. P., and Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman and X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective*. Wiley.

Hullsiek, K. H. and Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics (Oxford)* **3**, 2, 179–193.

Liang, H., Wang, S., Robins, J. M., and Carroll, R. J. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association* **99**, 466, 357–367.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23**, 19, 2937–2960.

Mitra, R. and Reiter, J. P. (2010). Latent class mixture models to impute missing covariates in observational studies. Statistics in Medicine (forthcoming).

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine* **28**, 9, 1402–1414.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 1, 85–95.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 5, 550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B-Methodological* **53**, 597–610.

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.

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 1, 33–38.

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

Rubin, D. B. (2003). Nested multiple imputation of nmes via partially incompatible mcmc. *Statistica Neerlandica* **57**, 1, 3–18.

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

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety* **17**, 546–555.

Stuart, E. A. and Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology* **44**, 2, 395–406.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16**, 3, 219–242.

van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681–694.

Van Buuren, S., Brand, J., Groothuis-Oudshoorn, C., and Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**, 12, 1049–1064.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2000). Multivariate imputation by chained equations: Mice v1.0 user's manual. TNO Quality of Life.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* **63**, 826–833.

Woo, M.-J., Reiter, J. P., and Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine* **27**, 19, 3805–3816.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science* **4**, 1, 67–91.