

An approach to impute missing covariates in observational studies when estimating treatment effects with propensity score matching

Robin Mitra and Jerome P. Reiter

Abstract In many observational studies, researchers estimate treatment effects using propensity score matching techniques. Estimation of propensity scores are complicated when some values of the covariates are missing. We can use multiple imputation to create completed datasets, from which propensity scores can be computed; however, we may be sensitive to the accuracy of the imputation models. We propose an iterative approach to multiply impute the missing covariates that gradually winnows the set of control units, until only the matched control units remain. This approach 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, which can result in more plausible imputations and better balance in the true covariate distributions. In addition, this approach can be conveniently implemented together with standard multiple imputation software for missing data. We illustrate the benefits of this approach with simulations and with an observational study of the effect of breastfeeding on the child's educational outcomes later in life.

Key words: Missing data, Multiple imputation, Observational studies, Propensity scores

1 INTRODUCTION

In many studies of causal effects, researchers use observational data in which the treatment and control conditions are not randomly assigned to subjects. Typically in such studies, the subjects in the treated group look different than those in the control group on several covariates. When these covariates are related to the outcome of interest, any observed differences in the two groups' outcome distributions may reflect the differences in the groups' covariates rather than only effects of the treatment Cochran and Chambers (1965); Rubin (1974).

Researchers can reduce the bias that results from imbalanced covariate distributions, at least for observed covariates, using propensity score matching Rosenbaum and Rubin (1983, 1985). The propensity score for any 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 any bias that results from imbalanced covariate distributions in the two groups, at least for those covariates in $\mathbf{x}$. In addition to one-to-one matching, analysts can make causal inferences by using full matching Rosenbaum (1991); Stuart and Green (2008), by sub-classifying on the propensity scores, Rosenbaum and Rubin (1984); Hullsieck and Louis (2002), and by using the the propensity scores for

Robin Mitra
Southampton Statistical Sciences Research Institute, University of Southampton, UK, SO17 1BJ, e-mail: R.Mitra@soton.ac.uk

Jerome P. Reiter
Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708, USA, e-mail: jerry@stat.duke.edu

inverse probability weighted estimation Robins *et al.* (2000); Lunceford and Davidian (2004). For a review of approaches to causal inference using propensity scores, see D’Agostino Jr. (1998).

Propensity scores are rarely known exactly and must be estimated from the data. Typically, this involves fitting regressions with T as the dependent variable and functions of $\mathbf{x}$ as the independent variables, and using the estimated probabilities as the propensity scores. See, for example, Woo *et al.* (2008). In this paper, we consider scenarios in which some covariate data are missing, so that standard regression models cannot be easily fit on the observed data. There are several strategies in the literature for overcoming this difficulty. The analyst could base propensity score estimation only on the complete cases; however, this could result in biased estimates when the data are not missing completely at random. It also shrinks the pool of potential matches. The analyst could match within patterns of missing data; however, with many patterns there may not be adequate matches. The analyst could apply the model-based approach of D’Agostino Jr. and Rubin (2000). They use an EM algorithm to find the maximum likelihood estimates of the parameters in a general location model fit to the data (X_{obs}, T) . After the algorithm converges, they estimate propensity scores as the predicted probabilities in the regression of T on X_{obs} .

We propose to estimate propensity scores using multiple imputation of missing data (Rubin, 1987). In this approach, the data analyst repeatedly imputes missing values by sampling from their posterior predictive distributions, conditioning on the observed covariate data. The analyst estimates propensity scores in each completed dataset, averages the propensity scores across datasets, and matches on the averaged scores. The averaged scores also could be used for sub-classification or weighting. The multiple imputation approach has some advantages over the maximum likelihood approach. With multiple imputation, the analyst’s model for the propensity scores is not tied to the model for imputations. That is, the analyst can try different propensity score models, for example including interaction and higher order effects that may not be in the complete data model or using models other than logistic regression. With completed 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).

A commonly used approach to impute missing covariates is the sequential regression multivariate imputation (SRMI) method proposed by Raghunathan *et al.* (2001). Such an approach is easy to implement with routines available in many statistical software packages. However, in observational studies there are typically many variables spread over a large multivariate space and thus checking whether this is a plausible model for the data is non-trivial. It is thus possible that analysts obtain implausible imputations with a mis-specified imputation model. As matched controls are selected based on these imputed values, this in turn could adversely affect the set of matched controls selected and hence result in an increased bias in estimating the treatment effect.

When imputing missing covariates, we propose an iterative method that results in a gradual winnowing of the set of control units available to be matched to treated units. This strategy 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, which can result in more plausible imputations in the region 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. In addition, this method is compatible with standard imputation software that implements procedures such as SRMI; thus, there is little additional burden for analysts to implement this method. We refer to this method in the paper as the winnow method.

The remainder of the article is organized as follows. In Section 2, we illustrate some difficulties that can arise when using a standard model for imputation. In Section 3, we describe the winnow method and illustrate its improved performance in the setting described in Section 2. In Section 4, we illustrate the use of this method to handle missing covariate data in an observational study of the effect of breastfeeding on the child’s educational outcomes later in life. Finally, in Section 5, we provide some general remarks about the multiple imputation modeling approach for estimation of propensity scores.

2 Potential difficulties with standard approaches

In this section we illustrate problems in finding plausible matches when we use a standard imputation model. The standard imputation model we consider is the SRMI method of Raghunathan *et al.* (2001). We first briefly describe this model and then illustrate potential inadequacies with this model in imputing covariates for a propensity score analysis in observational studies.

2.1 Description of the SRMI approach

The SRMI approach is one of the most commonly used approaches to handle missing data in surveys and observational studies. The routine is easy to implement and available in many standard software packages; for example, the MICE package in R is based on the SRMI idea to impute missing values. For this reason this is a very popular approach frequently employed by analysts to deal with missing data.

Consider the situation where we have n units and p partially observed covariate variables x_j , $j = 1, \dots, p$ where $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})'$ with x_{ij} corresponding to the i th unit's value for the j th covariate. Let D represent the set of covariate variables that are fully observed. The SRMI approach models the joint distribution of the incomplete data through a sequence of conditionals,

$$f(x_1, \dots, x_p | D, \theta_1, \dots, \theta_p) = f_1(x_1 | D, \theta_1) f_2(x_2 | D, x_1, \theta_2) \dots f_p(x_p | D, x_1, \dots, x_{p-1}, \theta_p) \quad (1)$$

Where, f_j , $j = 1, \dots, k$ is the conditional distribution for x_j and θ_j is a vector of parameters in the distribution (e.g. regression coefficients and dispersion parameter). The conditional distribution f_j takes the form of an appropriate regression model dependent on the type of variable x_j . For example, if x_j is continuous f_j will be a normal linear regression model, while if x_j is categorical, f_j will be a multinomial regression model. A standard modeling strategy is to only include additive effects for each of covariates $x_1, \dots, x_{j-1}$ (as well those in D) in model f_j . The covariates in (1) are typically ordered with increasing fractions of missing data.

The SRMI model imputes missing values using an iterative procedure for $t = 1, \dots, T$ iterations. In the first iteration, $t = 1$, missing values are imputed for each covariate variable from posterior predictive distributions using the conditional distributions in (1) with flat priors for all the parameters. So missing values in x_1 are imputed using the distribution $f_1(x_1 | D, \theta_1)$ with flat priors for θ_1 , denote this completed covariate $x_1^{(1)}$. Next missing values are imputed for x_2 using $f_2(x_2 | D, x_1^{(1)}, \theta_2)$ with flat priors for θ_2 and we also condition on the imputed values generated for missing x_1 values. This continues until imputed values for missing x_p are generated from $f_p(x_p | D, x_1^{(1)}, \dots, x_{p-1}^{(1)}, \theta_p)$ again conditioning on the imputed values in $x_1, x_2, \dots, x_{p-1}$ and with flat priors for θ_p . In subsequent iterations ($t > 1$), the SRMI method cycles through a sequence of conditional regressions to impute missing covariates in x_j from the predictive density corresponding to the conditional distribution

$$g_j(x_j | D, x_1^t, \dots, x_{j-1}^t, x_{j+1}^{t-1}, \dots, x_p^{t-1}, \phi_j)$$

where ϕ_j are the unknown regression parameters with diffuse priors. So at the t th iteration, we impute missing values in x_j from a regression model conditional on observed and imputed covariate values for $x_1, \dots, x_{j-1}$ from the t th iteration, and conditional on observed and imputed covariate values for $x_{j+1}, \dots, x_p$ from the $t - 1$ th iteration. van Buuren *et al.* (1999) propose a very similar approach which they call the switching regression method to multiply impute missing values, we do not discuss this method here.

This approach can be viewed as an approximation to a Gibbs sampler where the parameters $\theta_1, \dots, \theta_p$ are not updated from their conditional posterior distributions, but rather the parameters ϕ_j in each conditional distribution are updated with diffuse priors. In certain situations, e.g. if the data arise from a multivariate normal distribution then this approach will result in imputed values converging to a stable posterior distribution. In other situations, such convergence is not guaranteed (Gelman and Speed, 1993). However, this problem is not typically encountered in most practical situations. Once the SRMI method has cycled through a sufficient number of iterations for convergence, the algorithm stops and 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 the creation of m imputed data sets, and allows the multiple imputation framework to be applied.

We now consider a situation where a standard application of the SRMI model is not a plausible imputation model for the missing values. When inferring about the effect of a treatment variable using propensity score matching, we illustrate the problems this imputation model causes on true covariate balance between the covariates of the treated units and the matched controls selected.

2.2 Difficulty of getting the modeling right

We simulate 1200 units with two continuous covariates x_{1i} and x_{2i} in the following way:

$$x_{i1} = 50 + 0.8i + \varepsilon_{i1}, \quad \varepsilon_{i1} \sim N(0, 75) \quad (2)$$

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

for $i = 1, \dots, 1200$. This results in the covariates having a cubic relationship. Each unit is assigned a binary treatment indicator T_i with $T_i = 1$ indicating the unit to be treated otherwise it is a control. We assign treatment in the following way

$$p(T_i = 1) = 0.5I(i > 800), \quad i = 1, \dots, n. \quad (4)$$

So that treated units tend to have larger values of x_1 and x_2 and we set $n_T = \sum_{i=1}^n T_i = 200$. A plot of the covariates is presented in Figure 1.

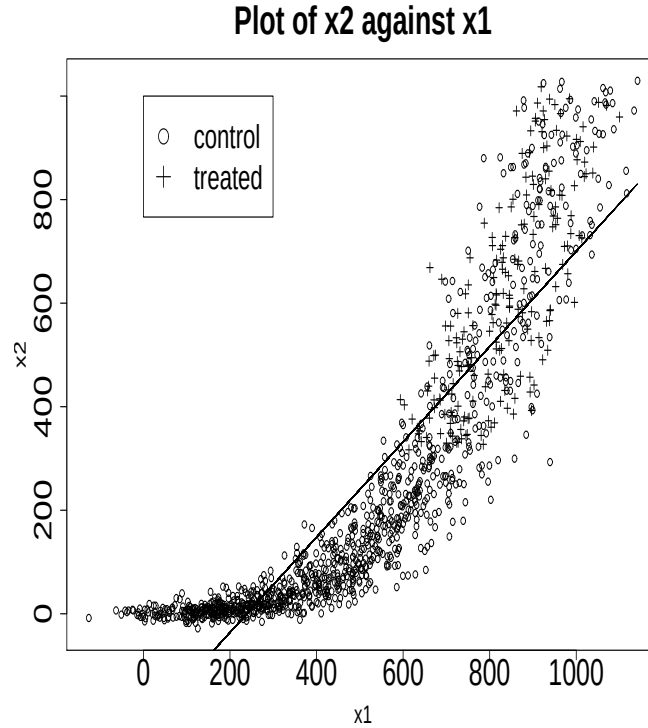


Fig. 1 Scatter plot of x_2 against x_1 when a cubic relationship is present, illustrating the effects of using a poor imputation model.

We introduce missing values in control units' x_2 covariate using a MAR mechanism. To do this, we assign missing data indicators m_i for each unit i , where $m_i = 1$ indicates that x_{i2} is missing and $m_i = 0$ indicates that x_{i2} is observed. This is done using the inverse logit link function,

$$P(m_i = 1) = \frac{\exp(-3 + 0.005x_{i1})}{1 + \exp(-3 + 0.005x_{i1})}. \quad (5)$$

We see with this mechanism that control units with larger values of x_1 are more likely to have missing x_2 values. In this way, there are more missing values for control units in the covariate space of the treated units. In total there are approximately 35% control units with missing x_2 values.

Using a standard application of the SRMI model, we multiply impute missing x_2 values to generate $m = 5$ imputed datasets. We estimate each unit's propensity score in each dataset using a logistic regression with main effects for x_1 and x_2 . We then average each unit's m propensity scores, and perform matching without replacement based on the averaged propensity scores.

Note that a standard application of the SRMI model would assume a linear relationship between x_1 and x_2 , and impute missing x_2 values using a linear regression model. We can clearly see from the fitted regression line indicated in Figure 1 that a linear relationship between the variables is implausible.

Now consider the issues that can arise when using this line to impute missing values of x_2 . If there are missing x_2 values in the region of the space where the treated units lie, they will tend to be imputed with lower values than their true value. In fact, they may be imputed into a region away from the region of the treated units' covariates. Another pitfall can occur when there are missing x_2 values close to the region of the treated units with slightly lower true values of x_2 than those units lying in the treated units' covariate space. Using the fitted line to impute these missing values will result in imputed values tending to be higher than their true x_2 values. Thus those units may be imputed to lie in the matched region.

Figure 2 presents box plots of the distribution of true x_1 and x_2 values for the treated and matched control units selected after multiply imputing missing values using the SRMI model. We see that there is a longer lower tail in the box plots of the matched controls for both x_1 and x_2 . This is due to the model tending to impute missing x_2 values higher than their true values for controls not in the matched region. The issues encountered here motivates our development of the winnow method discussed in the next section.

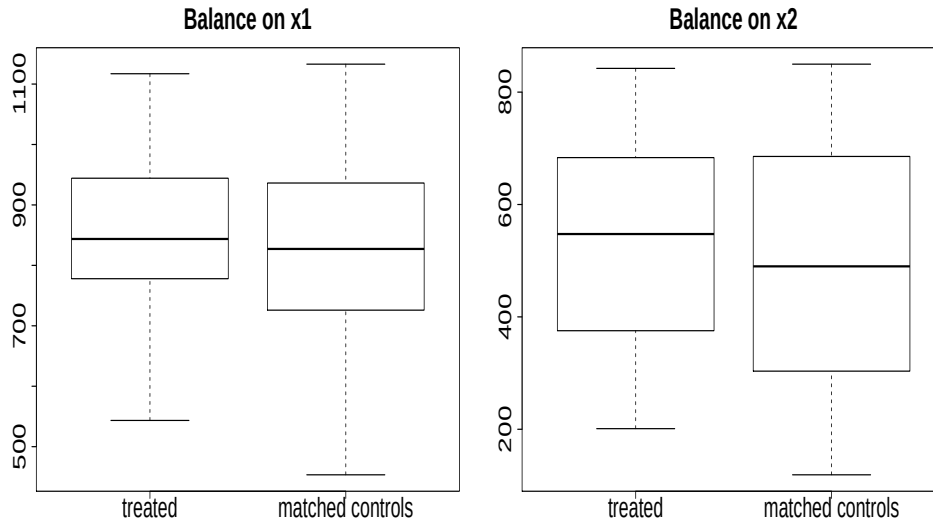


Fig. 2 Box plots of treated and matched control units for x_1 and x_2 respectively. Missing values are imputed using the SRMI model.

3 Winnowing method

In this section we first describe the winnow method, we then illustrate its performance in the simulation design of the previous section.

3.1 Description of method

The winnow method finds a matched control set by iteratively reducing the set of possible matched controls. The algorithm is described below.

1. First, multiply impute the missing values m times using standard multiple imputation software and estimate the propensity scores.
2. Next, with these propensity scores, find the k -closest controls for each treated unit using a without replacement pair matching scheme. This results in a reduced data set comprising $(k + 1)n_T$ units.

3. Re-impute missing values in this reduced data set again using standard MI software and now find the $k - 1$ -closest controls.
4. Continue this process until $k = 1$ and a matched control set is obtained.

This method thus first discards a set of control units unlikely to be selected as matches. It then gradually reduces the set of remaining controls until a matched control set is obtained. We can view the method used to select matched controls in Section 2 as an approach that skips steps 2-4 and finds the matched control set after only the first round of multiple imputation; henceforth, we call this the once only method.

The first iteration in the algorithm discards control units that we believe are unlikely to be selected as matches and that are likely to be far from the covariate space of the treated units. In doing so we hope to reduce the influence of outlying control units on the imputation of missing covariates in the region of the treated units' covariate space. We believe that such an approach will result in more plausible imputations for these units' missing values. In Figure 1 we see that a linear relationship for the covariates, while inappropriate over the whole covariate space is not so unreasonable over a restricted region where the treated units lie.

The winnow method is also desirable as it is easily compatible with analysts' preferred choice of imputation software. Analysts can choose any imputation method they prefer in Steps 1 and 3 of the algorithm. We have used the MICE package in R which is based on the SRMI method as this is a commonly used package to deal with missing data.

3.2 Performance in the simulation

We illustrate the winnow method for estimating propensity scores and selecting matched controls in the simulation design in Section 2. We apply the winnow method to this design with $k = 3$ initially, generating $m = 5$ imputed data sets at each iteration. Figure 3 summarizes the true covariate balance on x_1 and x_2 , for both the winnow and once only methods. The once only method tends to have longer lower tails in the distribution of its matched controls, particularly for x_1 . The winnow method thus results in a better balanced matched control set.

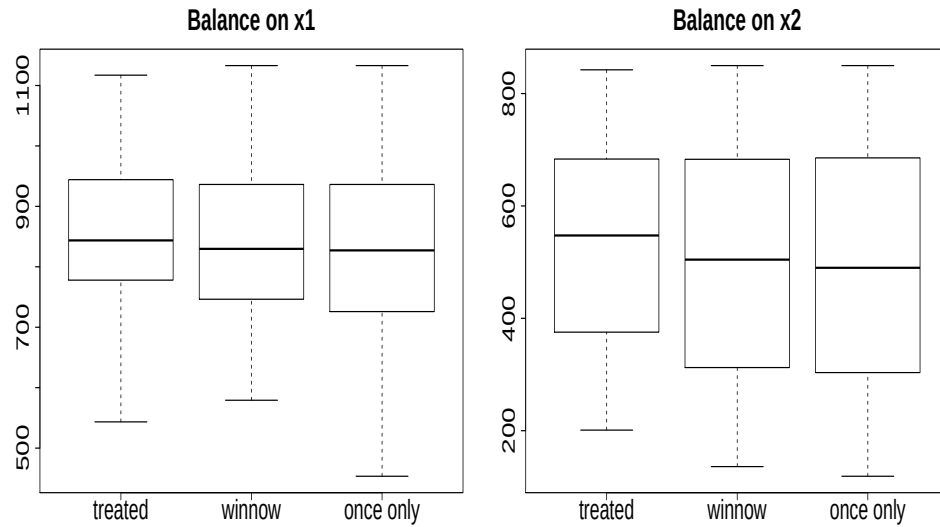


Fig. 3 Boxplots of the treated units and matched controls obtained from the winnow and once only methods for x_1 and x_2 respectively

Of course covariate balance is a preliminary step in estimating treatment effects. In order to assess the performance of the winnow method in estimating treatment effects we therefore simulate a response variable y with a simple response surface, given by

$$y_i = x_{i1} + x_{i2} + \varepsilon_i \quad \varepsilon_i \sim N(0, 1) \quad (6)$$

where $i = 1, \dots, 1200$. Note that there is no treatment effect simulated here, so the true treatment effect, τ , is 0. We estimate τ with $\hat{\tau}$, where

$$\hat{\tau} = \bar{y}_T - \bar{y}_{MC} \quad (7)$$

where $\bar{y}_T$ is the sample mean of y in the treatment group and $\bar{y}_{MC}$ is the sample mean of y in the matched control group. These matched controls, and hence the treatment effect estimate, could be obtained either by using the winnow method or by using the once only method.

We replicate this simulation design 500 times, each time estimating the treatment effect from the winnow and once only methods. We also consider the treatment effect estimated with fully observed covariates in each replication, i.e. prior to introducing missing values. The results are summarized in Figure 4. We see that the treatment effect estimates based on fully observed covariates tends to be closest to the truth as expected. The winnow method tends to get closer to the truth compared to the once only method and thus performs better in this situation.

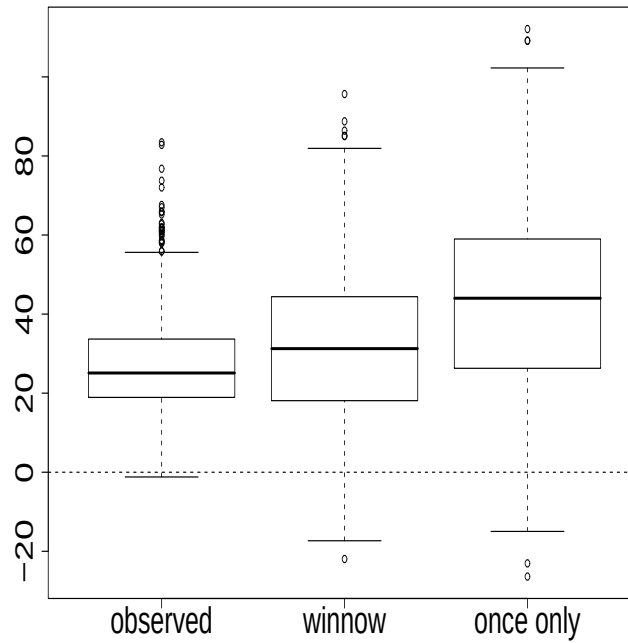


Fig. 4 Treatment effect estimates over repeated simulations when the covariates are fully observed, when using the winnow method and when using the once only method respectively. The dotted line indicates the true treatment effect of 0.

4 Application to study of breastfeeding

We now apply the winnow method to impute missing covariates in a genuine data set. The study assesses the effect of breastfeeding on child's cognitive development using variables from the National Longitudinal Survey of Youth (NLSY).

4.1 Description of study

The NLSY has been recruiting youths to the survey from 1979 onwards. The NLSY measures a range of social, economic and health related characteristics on these youths. This study was specifically interested in the effect of breastfeeding on cognitive development for children born to these youths. The response variable was taken to be the Peabody individual assessment test math score (PIATM) administered to children at 5 or 6 years of age.

breastfeeding duration in weeks is measured for each child, which is used to form the treatment variable. We define a child to be treated if he or she was breastfed for 24 weeks or more. This corresponds to advice that has been given by the American Academy of Pediatrics (AAP, 2005) and the WHO standard. There are other ways to define the treatment variable, and the analysis could be repeated with different cut points on the breastfeeding duration variable. For example a unit could be defined to be treated if breastfeed for 3 months or more, or treated units could be those that are breastfed at all.

Fourteen background covariates are measured. Five categorical variables are included in the analysis: 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. In addition, two continuous variables were categorized. A variable measuring weeks preterm has a large spike at zero weeks as seen in its histogram displayed in Figure 5. This preterm variable was categorized into three levels; not preterm (zero weeks), moderately preterm (one to four weeks), and very preterm (five or more weeks), with cut points determined from guidelines in the March of Dimes study (www.marchofdimes.com). Also, a variable which measures number of weeks worked in the year prior to giving birth has a distinct U shape as its histogram displays in Figure 6. This variable was categorized into four levels (not worked at all, worked between 1 and 47 weeks, worked 48-51 weeks, and worked all 52 weeks). We also measure seven continuous variables, including the difference between mother's age at birth and in 1979, mother's intelligence as measured by an armed forces qualification test, mother's highest education attainment, child's birth weight, weeks that the child spent in hospital, weeks that the mother spent in hospital, and family income. Transformations were applied to certain continuous variables as suggested by the Box-Cox procedure (Box and Cox, 1964), square root transformations were applied to the variables measuring the difference between mother's age at birth and in 1979 and mother's intelligence. Log transformations were applied to the variables measuring the weeks the child spent in hospital, the weeks the mother spent in hospital, and family income.

In this study, with the treatment variable as defined above, there are several clearly imbalanced covariates. To illustrate, we focus on mother's intelligence, mother's years of education, and child's race. Box plots of observed mother's intelligence and education for treated and control units are presented in Figure 7. Table 1 displays the proportion of treated and control units in each level of child's race. We can clearly see the imbalance in these variables. Treated units tend to have higher mother's intelligence scores, more mother's years of education and lower proportions of Hispanics and blacks.

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

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

There were numerous post treatment variables which are not used in estimating the treatment effect. For example, we do not include child care arrangements or weeks worked by the mother in the years after pregnancy. These variables may be affected by the treatment, hence we do not wish to treat them as covariates. We also 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 (breastfeeding duration) and 4977 units with a missing outcome variable (PIATM). With the resulting data set, we have 2388 units with 370 treated units.

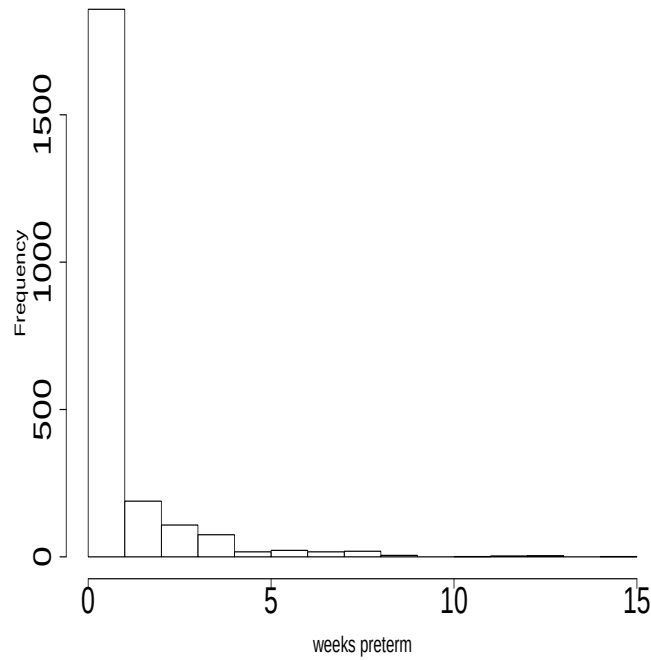


Fig. 5 Histogram of weeks preterm for subjects in the breast feeding study

4.2 Application to the breastfeeding data

For this data set, we apply both the winnow and once only approaches to estimate the treatment effect τ . As in Section 3, we estimate τ with $\hat{\tau} = \bar{Y}_T - \bar{Y}_{MC}$. Using the winnow method the treatment effect is estimated to be 1.28 points of the PIATM score with a two-sample pooled SE of 0.97. For alternative approaches to estimating standard errors from propensity score matching, see Hill and Reiter (2006). For the full sample $\bar{Y}_T - \bar{Y}_C = 5.23$ (SE 0.741), where $\bar{Y}_C$ is the sample mean of the PIATM score in the control group. We are thus significantly closer to zero when applying the winnow method to estimate the treatment effect. Similar data were analyzed by Der *et al.* (2006), who took a model based approach to infer about the treatment effect. They also found that the effect of breastfeeding was minimal.

The once only method estimates the treatment effect to be 2.29 (SE 1). There is thus a one point difference between using both methods; this may not be practically significant, but is perhaps suggestive that using the winnow method can lead to different treatment effect estimates than the once only method. In this application the once only method may not be an unreasonable strategy to estimate treatment effects, due to the careful modeling of the data prior to imputation. Nevertheless, we can view this as a situation where using the winnow method does not adversely affect estimates of the treatment effect.

5 Concluding Remarks

In this paper we have described a method to impute missing covariates in observational studies that is more robust to the effect of outlying control units. When analysts estimate treatment effects using propensity score matching, this method can lead to better covariate balance over standard strategies to impute covariates and estimate propensity scores, as illustrated in the simulations. The simulations further showed that the winnow method can obtain treatment effect estimates closer to the true treatment effect than estimates obtained from the afore mentioned standard strategy. The analysis of the breastfeeding study, while not conclusive in its favour of the proposed method did not appear to suffer from adopting this method either. This was thus an

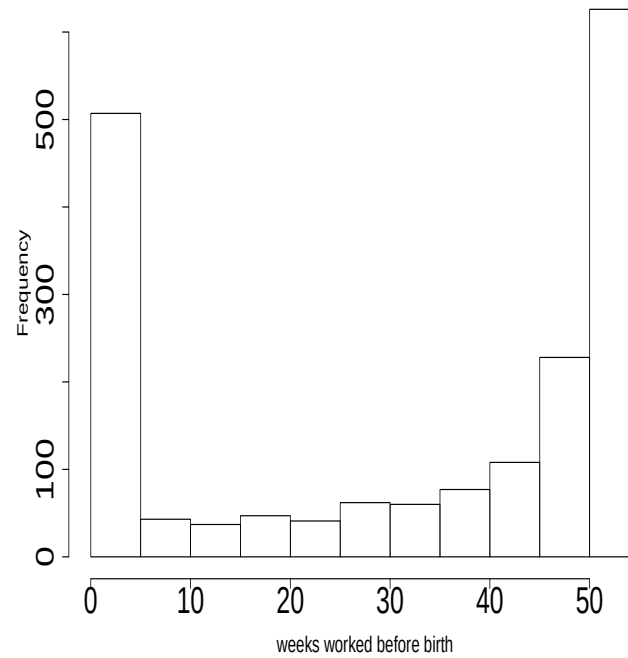


Fig. 6 Histogram of weeks mother worked in the year before giving birth for subjects in the breastfeeding study

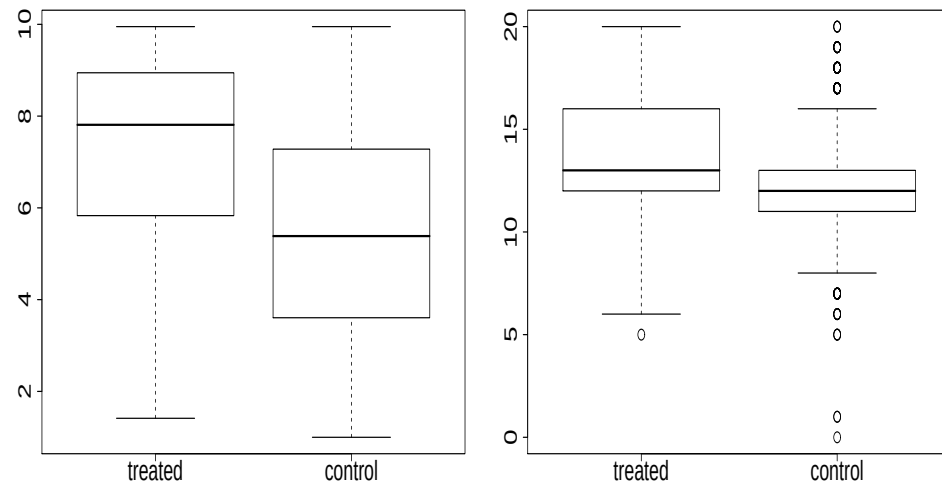


Fig. 7 Box plots of mother's intelligence score and mother's years of education respectively for treated and control units

example of a situation where the winnow method was not strictly necessary but there was little adverse effect. We conjecture that the winnow method can limit the impact of bad imputation models while not losing too much when the method is not required.

A more formal approach to this problem has been taken by Mitra and Reiter (2009) who develop a latent class mixture model for imputations. A latent class indicator variable for each unit is proposed, which corresponds to whether or not the unit's covariates come from a distribution similar to the distribution of the treated units' covariates. Mitra and Reiter (2009) show that using this model for imputing missing values can

lead to improved true covariate balance and closer estimates of treatment effects to the true treatment effect. In order to implement this method, analysts would be required to use computational methods such as Gibbs Sampling to impute missing values; thus, this may not be ideal for analysts who wish to avoid excessive computation. In contrast, the winnow method proposed in this article, while an informal method to deal with the same problem, is relatively easy to implement and is compatible with standard multiple imputation software.

It would also be interesting to evaluate how to utilize the multiply-imputed datasets most effectively. For example, one could compare treatment effect estimates when matching within each completed dataset versus matching on a summary of the propensity scores across all completed datasets. This issue has not been addressed in the literature on multiple imputation for missing data and propensity score matching. Another area of investigation would be to incorporate missing outcome data in this methodology. This is a complicated issue, as propensity score matching is done without consideration of the values of outcome variables. All of these are challenging projects with the potential to have a substantial impact on how practitioners in the epidemiological, medical, and social sciences make causal inferences from observational data.

Acknowledgements

The authors would like to thank Professor Jennifer Hill who provided us with the data on the breastfeeding study that was analysed in Section 4. This research was supported by the National Science Foundation [NSF-ITR-0427889].

References

- AAP (2005). Breastfeeding and the use of human milk. *Pediatrics* **115**, 496–506.
- 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.
- Cochran, W. G. and Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)* **128**, 2, 234–266.
- D’Agostino Jr., R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* **17**, 2265–2281.
- 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.
- Der, G., Batty, G. D., and Deary, I. J. (2006). Effect of breast feeding on intelligence in children: prospective study, sibling pairs analysis, and meta-analysis. *BMJ* **333**.
- 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.
- 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. and Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine* **25**, 2230–2256.
- 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.
- Hullisiek, K. H. and Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics (Oxford)* **3**, 2, 179–193.
- 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. (2009). Latent class mixture models to impute missing covariates in observational studies. Southampton Statistical Sciences Research Institute Working Paper.
- 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.
- 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. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 5, 688–701.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- 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., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681–694.
- 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.