UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M11/05
Methodology

A comparison of two methods of estimating propensity scores after multiple imputation

Robin Mitra, Jerome P. Reiter

Abstract

In many observational studies, analysts estimate treatment effects using propensity scores, e.g., by matching or sub-classifying on the scores. When some values of the covariates are missing, analysts can use multiple imputation to fill in the missing data, estimate propensity scores based on the m completed datasets, and use the propensity scores to estimate treatment effects. We compare two approaches to implementing this process. In the first, the analyst estimates the treatment effect using propensity score matching within each completed data set, and averages the m treatment effect estimates. In the second approach, the analyst averages the m propensity scores for each record across the completed datasets, and performs propensity score matching with these averaged scores to estimate the treatment effect. We compare properties of both methods via simulation studies using artifical and real data. The simulations suggest that the second method has greater potential to produce substantial bias reductions than the first

A comparison of two methods of estimating propensity scores after multiple imputation

ROBIN MITRA*

School of Mathematics

University of Southampton, Southampton, SO17 1BJ, UK

Tel. (+44) 2380 594550 Email: R.Mitra@soton.ac.uk

JEROME P. REITER

Department of Statistical Science

Duke University, Box 90251, Durham, NC 27708, USA

Tel. (+1) 919 6685227 Email: jerry@stat.duke.edu

* *corresponding author*

Abstract

In many observational studies, analysts estimate treatment effects using propensity scores, e.g., by matching or sub-classifying on the scores. When some values of the covariates are missing, analysts can use multiple imputation to fill in the missing data, estimate propensity scores based on the m completed datasets, and use the propensity scores to estimate treatment effects. We compare two approaches to implementing this process. In the first, the analyst estimates the treatment effect using propensity score matching within each completed data set, and averages the m treatment effect estimates. In the second approach, the analyst averages the m propensity scores for each record across the completed datasets, and performs propensity score matching with

these averaged scores to estimate the treatment effect. We compare properties of both methods via simulation studies using artifical and real data. The simulations suggest that the second method has greater potential to produce substantial bias reductions than the first.

Keywords: Missing data; Multiple imputation; Observational studies; Propensity score.

1 INTRODUCTION

In many studies of causal effects, analysts can reduce the bias that results from imbalanced covariate distributions, at least for observed covariates, using propensity score matching[1–3]. The propensity score for any subject, $e(\boldsymbol{x}_i)$, is the probability that the subject receives the treatment given its vector of covariates $\boldsymbol{x}_i$; that is, $e(\boldsymbol{x}_i) = P(T_i = 1|\boldsymbol{x}_i)$, where $T_i = 1$ if subject i receives treatment and $T_i = 0$ otherwise.[1] show that, when two large groups have the same distributions of propensity scores, the groups should have similar distributions of $\boldsymbol{x}$. Thus, by selecting control units whose propensity scores are similar to the treated units' 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 $\boldsymbol{x}$. Other approaches to causal inference based on propensity scores include sub-classification[4,5], full matching[6,7] and propensity score weighted-estimation[8]. See[9] for a review of different approaches to causal inference using propensity scores.

Propensity scores are typically estimated via regressions of T on functions of $\boldsymbol{x}$[10–13]. When some covariate data are missing, these complete-data methods cannot be easily applied. Several strategies exist for overcoming this complication[4,14–16]. In this article, we focus on the use of multiple imputation[17] to fill in the missing

covariate data, thus enabling estimation of propensity scores via complete-data methods.

With m completed datasets, the analyst potentially can estimate the propensity scores in each dataset, thus obtaining m values of each unit's propensity score. What should the analyst do with these multiple propensity scores? One approach is to average each unit's m propensity scores, match treated and control units based on their averaged scores, and thereby estimate the treatment effect. We call this the Across approach. Another approach is to match treated and control units within each completed dataset, thereby coming up with m estimates of the treatment effect. These m treatment effect estimates can be averaged to come up with the final estimated treatment effect. We call this the Within approach. Both of these approaches seem intuitively reasonable strategies: which can we expect to be more effective? To our knowledge, this question has been investigated previously only by[18], who expertly pointed out its many complexities.

In this article, we shed further light on this question. To do so, we use two types of simulations: a simple setting with artificial data, and a complicated setting with actual data. In the simulations, the Across method exhibits greater potential for substantial reductions in bias, whereas the Within method results in smaller variance estimates. The remainder of the article is organized as follows. In Section 2, we formally define the Across and Within approaches. In Section 3, we compare the two approaches in simulation studies with artificial data. In Section 4, we use the two approaches to estimate a treatment effect in an observational study of the effect of breast feeding on the child's cognitive development later in life. In Section 5, we conclude with remarks about the Across and Within approaches.

2 Across and Within approaches

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be an $n \times p$ matrix of covariates, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ corresponds to the ith unit's covariates, for $i = 1, \ldots, n$. For each $\boldsymbol{x}_i$, let $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{ip})'$ be a vector of missing data indicators, where $m_{ij} = 1$ indicates x_{ij} is missing, and $m_{ij} = 0$ indicates x_{ij} is observed, for $j = 1, \ldots, p$. Let $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)'$ be the $n \times p$ matrix of missing data indicators for $\boldsymbol{X}$. Let $\boldsymbol{X}_{mis} = \{x_{ij} : (i,j) : m_{ij} = 1\}$ and $\boldsymbol{X}_{obs} = \{x_{ij} : (i,j) : m_{ij} = 0\}$. For each unit i, the binary treatment indicator is $T_i \in \{0, 1\}$, and the outcome is Y_i. Let $\boldsymbol{T} = (T_1, \ldots, T_n)'$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. Here, we assume that $\boldsymbol{T}$ and $\boldsymbol{Y}$ are fully observed.

In multiple imputation, values of $\boldsymbol{X}_{mis}$ are filled in m times with draws from the predictive distribution, $p(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs})$, resulting in m completed datasets $\boldsymbol{X}_{com}^{(1)}, \ldots, \boldsymbol{X}_{com}^{(m)}$. For each $\boldsymbol{X}_{com}^{(k)}$, let $e(\boldsymbol{x}_{i,com}^{(k)})$ be the estimated propensity score for unit i, where $i = 1, \ldots, n$ and $k = 1, \ldots, m$. Here, each $e(\boldsymbol{x}_{i,com}^{(k)})$ is estimated using only the data in $\boldsymbol{X}_{com}^{(k)}$, for example with a logistic regression of $\boldsymbol{T}$ on some function of $\boldsymbol{X}_{com}^{(k)}$.

In the Across approach, we estimate the propensity score for each unit, $e^{A,m}(\boldsymbol{x}_i)$, by averaging $e(\boldsymbol{x}_{i,com}^{(k)})$ over the imputations, so that

$$e^{A,m}(\boldsymbol{x}_i) \quad = \quad \frac{\sum_{k=1}^{m} e(\boldsymbol{x}_{i,com}^{(k)})}{m}. \tag{1}$$

Let $\boldsymbol{e}^{A,m} = (e^{A,m}(\boldsymbol{x}_1), \ldots, e^{A,m}(\boldsymbol{x}_n))'$. Analysts use $\boldsymbol{e}^{A,m}$ to find a matched control set, for example for each treated unit find the control unit with the nearest propensity score. We obtain a matched control set in this way, where we match without replacement. Given the matched set, the analyst estimates the treatment

effect in the Across approach with

$$\hat{\tau}^{A,m} = \bar{Y}_T - \bar{Y}_{mc}^{A,m}, \tag{2}$$

where $\bar{Y}_{mc}^{A,m}$ is the mean of matched control units' outcomes selected in the Across approach.

The Within approach uses the propensity scores estimated from each completed dataset, $\boldsymbol{e}(\boldsymbol{X}_{com}^{(k)}) = (e(\boldsymbol{x}_{1,com}^{(k)}), \ldots e(\boldsymbol{x}_{n,com}^{(k)}))'$, to obtain m matched control sets, one for each $\boldsymbol{X}_{com}^{(k)}$; that is, matching is performed separately in each $\boldsymbol{X}_{com}^{(k)}$. Let $\bar{Y}_{mc}^{(k)}$ be the average of the outcomes for the matched controls in $\boldsymbol{X}_{com}^{(k)}$, where $k = 1, \ldots, m$. Let $\bar{Y}_{mc}^{W,m} = \sum_{k=1}^{m} \bar{Y}_{mc}^{(k)}/m$. The analyst estimates the treatment effect for the Within approach using

$$\hat{\tau}^{W,m} = \bar{Y}_T - \bar{Y}_{mc}^{W,m}. \tag{3}$$

3 Artificial data simulation

We now compare the Across and Within approaches using simulations with artificial data. For each simulation run, we generate two covariates $\boldsymbol{X}$ for $n = 1100$ records such that

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{\mu} = (10, 10)'$, and $\boldsymbol{\Sigma}$ has variances equal to 5 with correlation 0.5. We generate the response Y so that, for all i,

$$Y_i = x_{i1} + x_{i2} + \epsilon_i, \quad \epsilon_i \sim N(0, 1). \tag{5}$$

Hence, the treatment effect $\tau = 0$ for all simulations. We introduce missing data into $\boldsymbol{x}_2$ based on missing at random mechanisms; we leave $\boldsymbol{x}_1$ and $\mathbf{Y}$ fully observed. We consider three mechanisms for assigning treatment, including (i) assignment depends only on $\boldsymbol{x}_1$, (ii) assignment depends only on $\boldsymbol{x}_2$, and (iii) assignment depends equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. As we shall see, the Across and Within methods are differentially effective for these assignment mechanisms.

3.1 Simulation 1: treatment assignment depends only on $\boldsymbol{x}_1$

In this simulation, we assign treatment so that

$$logit\left(P(T_i = 1)\right) = -7.8 + 0.5x_{i1}. \tag{6}$$

Thus, treatment assignment depends only on x_1. In any dataset, this generates approximately 100 treated units and 1000 control units. Figure 1 displays typical covariate patterns that arise from this design.

We consider two mechanisms for introducing missing data in x_2. In the first, we randomly make some control units' x_2 values missing so that

$$logit\left(P(m_{i2} = 1)\right) = -10.1 + 0.9x_{i1}. \tag{7}$$

In this way, units with larger x_1 values, which are the units most likely to be selected as matches, are more likely to be missing their x_2 values. Approximately 30% of control units' values of x_2 are missing. In the second, we use the same missing data patterns for the control units and also introduce missing values into 30% of the treated units' x_2 covariate through a missing completely at random (MCAR) mechanism. We use the MCAR mechanism because the treated units already tend to have large values of x_1.

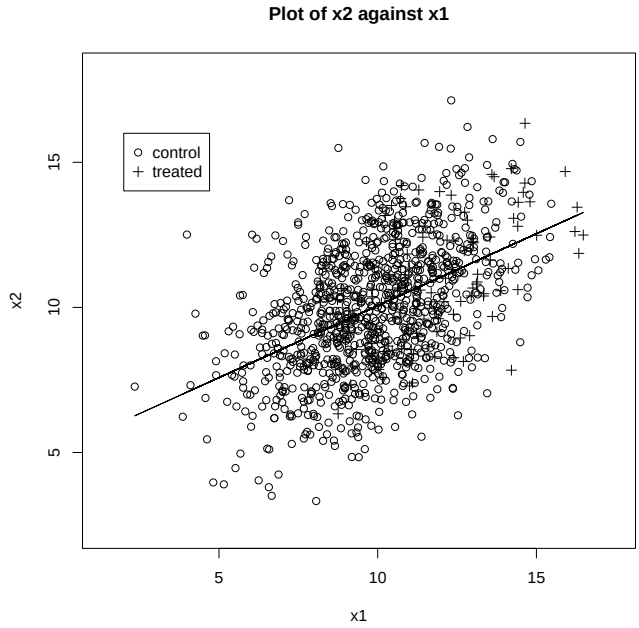

Figure 1: Plot of the covariate distribution in the simulation design where treatment assignment depends on $\boldsymbol{x}_1$ together with the fitted regression line based on a normal linear model for $\boldsymbol{x}_2$.

We impute missing $\mathbf{x}_2$ from a normal linear regression of $\mathbf{x}_2$ on $\mathbf{x}_1$, using the appropriate Bayesian posterior predictive distribution with flat prior distributions. We do not control for Y in the imputations. This is done to remain consistent with the philosophy of propensity score matching: manipulation of covariates and the creation of a matched control set is done without consideration of the outcome values. This is followed, for example, by[14]. In this way, propensity score computations, and any subsequent causal inferences made using the propensity scores, are not affected by assumptions about the outcome variable. That said, it can be advantageous to include the outcome variable in imputation models[19,20]. One could easily modify the imputation models used here to include the outcome variable in the imputation model.

After multiple imputation of $\mathbf{x}_2$, we estimate the propensity scores $e(\boldsymbol{x}_{i,com}^{(k)})$ for each unit i in each of $k = 1, \ldots, m$ completed datasets using a logistic regression of $\mathbf{T}$ on

$(\mathbf{x}_1, \mathbf{x}_2)$. We then compute $\hat{\tau}^{A,m}$ and $\hat{\tau}^{W,m}$ as in Section 2.

We run this simulation design 1000 times to get new values of $(\mathbf{X}, \mathbf{T}, \mathbf{Y}, \mathbf{M})$. Table 1 summarizes the bias and variance of $\hat{\tau}^{A,m}$ and $\hat{\tau}^{W,m}$ across the 1000 simulations for different values of m. Both the Across and Within approaches result in estimates of τ that are close to zero. The bias in $\hat{\tau}^{A,m}$ tends to be slightly smaller than that of $\hat{\tau}^{W,m}$, but its variance is slightly higher. The variance of $\hat{\tau}^{W,m}$ appears to decrease as m increases; the variance trend is non-linear for $\hat{\tau}^{A,m}$.

m	Across estimate	Across variance	Within estimate	Within variance
Only control units missing $\mathbf{x}_2$				
5	0.0710	0.0831	0.0745	0.0540
10	0.0501	0.0811	0.0721	0.0503
15	0.0609	0.0847	0.0738	0.0487
20	0.0651	0.0850	0.0739	0.0486
50	0.0600	0.0866	0.0741	0.0473
Treatment and control units missing $\mathbf{x}_2$				
5	0.0461	0.0838	0.0751	0.0613
10	0.0443	0.0888	0.0749	0.0562
15	0.0493	0.0881	0.0746	0.0548
20	0.0534	0.0857	0.0741	0.0541
50	0.0501	0.0868	0.0737	0.0526

Table 1: Treatment effect estimates from the Across and Within approaches in the simulation design where treatment assignment depends only on $\mathbf{x}_1$. The average treatment effect estimates based on the covariates before introduction of missing data is 0.074.

Since treatment assignment only depends on $\boldsymbol{x}_1$, which is fully observed, propensity scores are determined essentially from $\boldsymbol{x}_1$ only, so that the missing data in $\boldsymbol{x}_2$ play a minor role here. Thus, it is not surprising that both the Across and Within methods result in similar reductions in bias close to the true treatment effect of zero. Nonetheless, this is a situation where the Within approach dominates on mean squared error, at least for these values of m.

3.2 Simulation 2: treatment assignment depends only on $\boldsymbol{x}_2$

In this simulation, we assign treatment so that

$$logit\left(P(T_i = 1)\right) = -7.8 + 0.5x_{i2}. \tag{8}$$

As before, this generates approximately 100 treated units and 1000 control units, but now the treatment assignment depends only on $\boldsymbol{x}_2$. Figure 2 displays a typical covariate distribution for this design. We introduce missing values in x_2 values using the same two scenarios as in Section 3.1, and impute missing values from a normal linear regression as before.

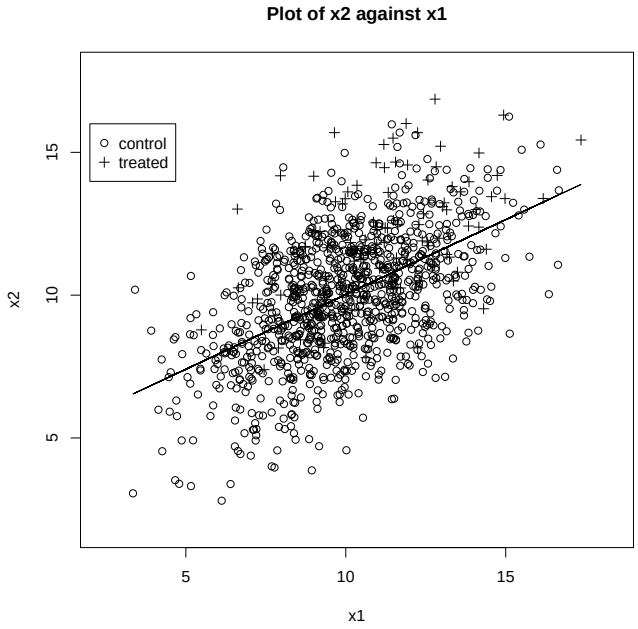

Figure 2: Plot of the covariate distribution in the simulation design where treatment assignment depends on $\boldsymbol{x}_2$ together with the fitted regression line assuming a normal linear model for $\boldsymbol{x}_2$.

Table 2 summarizes the results for 1000 simulation runs for different m. Here, $\hat{\tau}^{A,m}$ consistently has substantially smaller bias than $\hat{\tau}^{W,m}$. The bias in $\hat{\tau}^{A,m}$ tends to decrease as m increases; this is not the case for $\hat{\tau}^{W,m}$. The variance of $\hat{\tau}^{W,m}$

continues to be lower than that of $\hat{\tau}^{A,m}$, and it decreases with m.

m	Across estimate	Across variance	Within estimate	Within variance
Only control units missing $\mathbf{x}_2$				
5	0.6296	0.0740	0.9367	0.0471
10	0.5607	0.0846	0.9372	0.0438
15	0.5478	0.0798	0.9355	0.0412
20	0.5381	0.0831	0.9380	0.0400
50	0.5396	0.0835	0.9392	0.0374
Treatment and control units missing $\mathbf{x}_2$				
5	0.7396	0.0742	1.1526	0.0484
10	0.6461	0.0750	1.1514	0.0443
15	0.6251	0.0743	1.1519	0.0429
20	0.6214	0.0741	1.1529	0.0419
50	0.6064	0.0716	1.1536	0.0408

Table 2: Treatment effect estimates from the Across and Within approaches in the simulation design where treatment assignment depends on $\mathbf{x}_2$

Since treatment assignment depends only on $\boldsymbol{x}_2$, which is partially observed, the imputation of the missing values plays a major role in determining the matched control set. In this case, the Across approach dominates on mean squared error.

3.3 Simulation 3: treatment assignment depends equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

In this simulation, we assign treatment so that

$$logit\left(P(T_i = 1)\right) = -7.8 + 0.255x_{i1} + 0.255x_{i2}. \tag{9}$$

This generates approximately 100 treated units with treatment assignment depending equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Figure 3 displays a typical covariate distribution for this design. We introduce missing values in x_2 values using the same two scenarios as in Section 3.1.1, and impute missing values from a normal linear

regression as before.

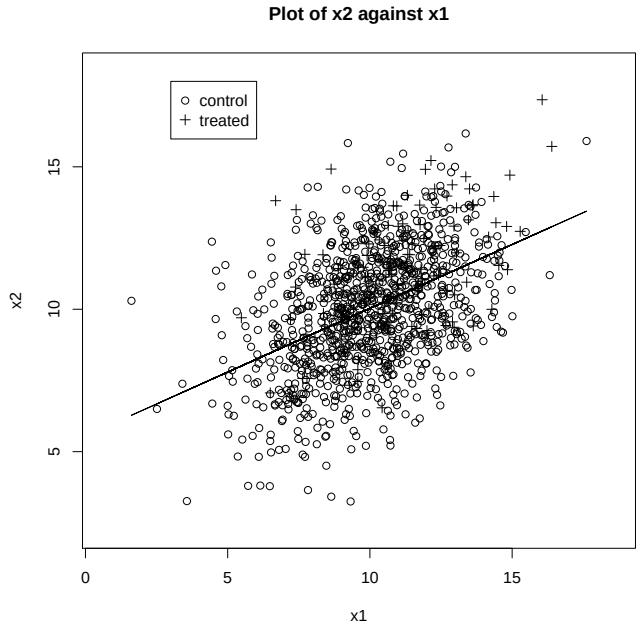

Figure 3: Plot of the covariate distribution in the simulation design where treatment assignment depends equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with the fitted regression line assuming a normal linear model for $\boldsymbol{x}_2$.

Table 3 summarizes the results of 1000 simulation runs for different m. Here, $\hat{\tau}^{A,m}$ again has consistently smaller bias than $\hat{\tau}^{W,m}$. The differences between the two estimators are smaller than observed in Table 2, yet larger than those observed in Table 1. The bias in $\hat{\tau}^{A,m}$ tends to decrease as m increases; this is not the case for $\hat{\tau}^{W,m}$. As before, the variance of $\hat{\tau}^{W,m}$ is smaller than the variance of $\hat{\tau}^{A,m}$, and it appears to decrease with m.

4 Simulation involving actual data

We now apply both the Across and Within approaches using data intended to inform analysis of the effects of breast feeding on child's later cognitive development. The data are a subset of the 1979 U.S. National Longitudinal Survey of Youth, commonly referred to as the NLSY79. This longitudinal survey, begun in

m	Across estimate	Across variance	Within estimate	Within variance
Only control units missing $\mathbf{x}_2$				
5	0.4853	0.0572	0.6133	0.0436
10	0.4619	0.0606	0.6142	0.0405
15	0.4477	0.0580	0.6143	0.0394
20	0.4455	0.0568	0.6137	0.0385
50	0.4360	0.0577	0.6130	0.0376
Treatment and control units missing $\mathbf{x}_2$				
5	0.5546	0.0618	0.7012	0.0495
10	0.5513	0.0668	0.7001	0.0462
15	0.5462	0.0654	0.6996	0.0450
20	0.5423	0.0618	0.7008	0.0440
50	0.5400	0.0626	0.6996	0.0426

Table 3: Treatment effect estimates from the Across and Within approaches in the simulation design where treatment assignment depends equally on $\mathbf{x}_1$ and $\mathbf{x}_2$

1979, interviewed a nationally representative sample of 12686 young adults in the U.S. who were aged between 14 and 22 years at the time. The survey was administered on an annual basis until 1994, after which time the cohort was interviewed biannually. From 1986, detailed information on the children born to women in the NLSY79 were also collected. This study was used by[21] to illustrate a latent class, general location mixture model for multiple imputation of missing covariates. The description of the study below is adapted from that article.

The response variable is the Peabody individual assessment test math score (PIATM) administered to children at 5 or 6 years of age. The treatment variable is breast feeding duration, which is measured 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[22] 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.

We use fourteen potentially relevant background covariates. These include the categorical variables: the child's race (Hispanic, black or other), the mother's race (Hispanic, black, asian, white, Hawaiian/Pacific Islander/American Indian, or other), child's sex, and two variables indicating whether the spouse or grandparents were present at birth. In addition, we categorize the number of weeks the child was born premature 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 of the March of Dimes (*www.marchofdimes.com*). The categorization was used because weeks preterm has a very large spike at zero weeks. We also categorize the number of weeks that the mother worked in the year prior to giving birth into four levels: not worked at all, worked between 1 and 47 weeks, worked 48-51 weeks, and worked all 52 weeks. This variable has a distinct U shaped histogram, which would be difficult to capture with a linear model. See[21] for further details on the transformations. The background covariates also include seven continuous variables, including number of years between 1979 and the mother's age at the child's birth, mother's intelligence as measured by an armed forces qualification test, mother's highest educational attainment, child's birth weight, the number of weeks that the child spent in hospital, the number of weeks that the mother spent in hospital, and family income. We applied Box-Cox transformations[23] to several continuous variables to improve the assumption of normality; details of these are given in[21].

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 breast feeding duration and 4977 units with a missing 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, as the cases with complete treatment and outcome data suffice for our purposes: to examine the implications for treatment effect estimation when using either the across or within method. The resulting data comprise 2388 youths, of whom 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%).

Several covariates in the available data are clearly imbalanced. To illustrate, we focus on three variables. Figure 4 summarizes the distribution of mother's intelligence and education for observed treated and control units, and Table 4 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. Because of these imbalances, we seek to do propensity score estimation and matching in the presence of the missing data.

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

Table 4: Distribution of child's race.

We evaluate the performance of the Across and Within approaches at achieving true covariate balance in a simulation involving the 1306 complete cases. Although this is a smaller sample size, we can introduce missing data, run the methods to estimate treatment effects, and compare how close these estimates are to the estimate obtained from the 1306 complete cases before introduction of missing values. We

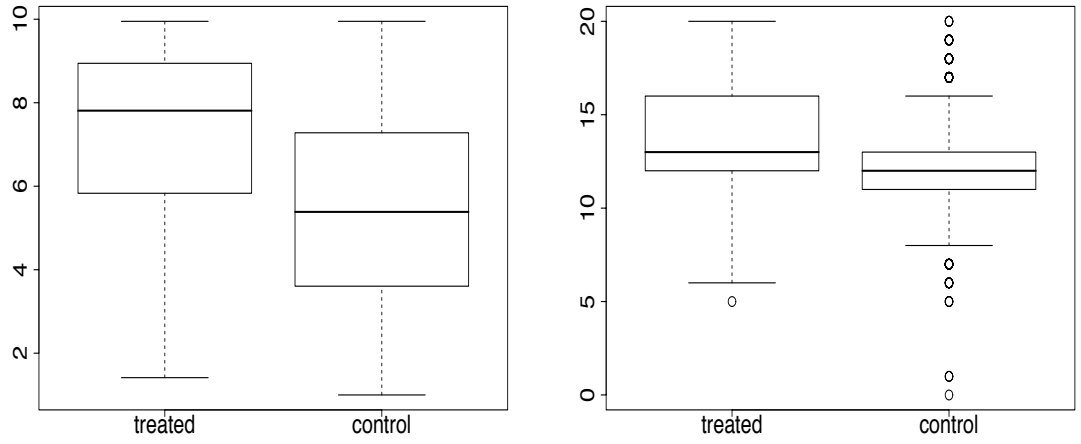

Figure 4: Box plots of mother's intelligence score and mother's years of education respectively for treated and control units before matching.

introduce missing values by randomly sampling with replacement from the missing data patterns present in the original data set. This results in 717 units with fully observed covariates; the remainder have some missing data. For imputation, we use the data augmentation algorithm developed by[24] based on the general location model; this is a convenient modeling strategy to handle missing values in mixed categorical and continuous data. We run the data augmentation algorithm for 200000 iterations after discarding an initial 5000 as burn-in; thus, 200000 imputed data sets were generated here. This is arguably more than most analysts would construct; however, we wanted to minimize simulation noise when comparing the Across and Within approaches.

In addition to $\widehat{\tau}^{A,m}$ and $\widehat{\tau}^{W,m}$, we estimate τ using propensity score matching prior to introducing missing values and all 1306 cases. The resulting estimate is $\widehat{\tau} = 2.32$, whereas $\widehat{\tau}^{A,m} = 1.84$ and $\widehat{\tau}^{W,m} = 1.64$. Thus, $\widehat{\tau}^{A,m}$ is slightly closer to $\widehat{\tau}$, which presumably (although not definitively, since we do not know τ) means that it achieves greater reductions in bias than $\widehat{\tau}^{W,m} = 1.64$. This is consistent with the artificial data simulation results. We repeated the simulation three more times with

new introduction of missing data and running the data augmentation algorithm for 140000 iterations after an initial burn-in of 5000 iterations. For all three replications, the corresponding $|\tau^{A,m} - \hat{\tau}| < |\tau^{W,m} - \hat{\tau}|$.

5 Concluding remarks

In the simulations studied here, the Across approach had the potential for greater bias reduction than the Within approach. This was especially true when treatment assignment depended on the missing covariates. However, the Within approach resulted in smaller variances than the Across approach. Of course, as with any simulation study, these results may have limited generalizability. For some response surfaces, covariate distributions, treatment assignments, or missing data patterns, it may be that one approach always dominates the other. Alternatively, in other settings the two approaches may always give the same answer, for example if data were missing only for control units in a region of covariate space far away from that of the treated units (these units never would be selected as matches). Furthermore, the choice of imputation model also affects treatment effect estimates[21], as might the choice of whether or not to condition on the response in the imputation models[18]. Thus, we recommend that analysts run simulation studies akin to the one done on the complete cases in the breast-feeding simulation study to get a rough guide of the relative potentials of each procedure for bias reduction. When such studies are not possible, we suggest the Across method as a default, since it had the potential for greater bias reductions in the simple simulations.

For the Across method, we noticed that the set of matched controls did not change after m reached some threshhold. This is because the component of variance due to imputation of the propensity scores, and hence the treatment effect estimate, based on the Across method goes to zero as $m \to \infty$ (as for the Within method). For fixed m, in other simulations not reported here we found that one can do better than

either the Across or Within methods by using a hybrid approach. Specifically, we independently generate $r > 1$ sets of $m > 1$ multiply-imputed datasets, use the Across method within each set, and average the Across treatment effect estimates over the r sets. For example, in the simulations in Section 3, we often found that setting $(m = 10, r = 10)$ usually resulted in smaller mean squared error than setting $(m = 100, r = 1)$, which is the Across approach, or $(m = 1, r = 100)$, which is the Within approach. In a sense, this hybrid approach combines the favorable bias properties of the Across approach with the favorable variance properties of the Within approach. We plan to investigate this hybrid approach more thoroughly in future work.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

[2] Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician. 1985;39(1):33–38.

[3] Park GS, Wong WK, Oh M, Khanna D, Gold RH, Sharp JT, et al. Classifying Radiographic Progression Status in Early Rheumatoid Arthritis Patients Using Propensity Scores to Adjust for Baseline Differences. Statistical Methods in Medical Research. 2007;16(1):13–29.

[4] Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association. 1984;79:516–524.

[5] Hullsiek KH, Louis TA. Propensity Score Modeling Strategies for the Causal Analysis of Observational Data. Biostatistics (Oxford). 2002;3(2):179–193.

[6] Rosenbaum PR. A characterization of optimal designs for observational studies. Journal of the Royal Statistical Society, Series B-Methodological. 1991;53:597–610.

[7] Stuart EA, Green KM. Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. Developmental Psychology. 2008;44(2):395–406.

[8] Lunceford JK, Davidian M. Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. Statistics in Medicine. 2004;23(19):2937–2960.

[9] D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine. 1998;17:2265–2281.

[10] Hanley JA, Dendukuri N. Efficient Sampling Approaches to Address Confounding in Database Studies. Statistical Methods in Medical Research. 2009;18(1):81–105.

[11] Woo MJ, Reiter JP, Karr AF. Estimation of Propensity Scores Using Generalized Additive Models. Statistics in Medicine. 2008;27(19):3805–3816.

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

[13] Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and Drug Safety. 2008;17:546–555.

[14] D'Agostino Jr RB, Rubin DB. Estimating and using propensity scores with partially missing data. Journal of the American Statistical Association. 2000;95(451):749–759.

[15] Haviland A, Nagin DS, Rosenbaum PR. Combining Propensity Score Matching and Group-Based Trajectory Analysis in an Observational Study. Psychological Methods. 2007;12(3):247–267.

[16] Qu Y, Lipkovich I. Propensity Score Estimation with Missing Values Using a Multiple Imputation Missingness Pattern (MIMP) Approach. Statistics in Medicine. 2009;28(9):1402–1414.

[17] Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley; 1987.

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

[19] Little RJA. Regression with Missing X's: A Review. Journal of the American Statistical Association. 1992;87:1227–1237.

[20] Moons KGM, Donders RART, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006;59(10):1092–1101.

[21] Mitra R, Reiter JP. Latent class mixture models to impute missing covariates in observational studies; 2010. Statistics in Medicine (forthcoming).

[22] Chantry CJ, Howard CR, Auinger P. Full Breastfeeding Duration and Associated Decrease in Respiratory Tract Infection in US Children. Pediatrics. 2006;117(2):425–432.

[23] Box GEP, Cox DR. An Analysis of Transformations. Journal of the Royal Statistical Society Series B (Methodological). 1964;26(2):211–252.

[24] Schafer JL. Analysis of Incomplete Multivariate Data. London: Chapman & Hall; 1997.