UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M09/13

Methodology

[Estimating Propensity Scores With Missing Covariate Data Using General Location Mixture Models]

[Robin Mitra, Jerome P. Reiter]

Abstract

[In many observational studies, researchers estimate causal effects using propensity scores, e.g., by matching or sub-classifying on the scores. Estimation of propensity scores is complicated when some values of the covariates are missing. We propose to use multiple imputation to create completed datasets, from which propensity scores can be estimated, with a general location mixture model. The model assumes that the control units are a latent mixture of (i) units whose covariates are drawn from the same distributions as the treated units' covariates and (ii) units whose covariates are drawn from different distributions. This formulation reduces the influence of control units outside 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. We illustrate the benefits of

the latent class modeling approach with simulations and with an observational

study of the effect of breast feeding on children's cognitive abilities.

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

Estimating propensity scores with missing covariate data using general location mixture models

ROBIN MITRA*

School of Mathematics

University of Southampton, Southampton, SO17 1BJ, UK

JEROME P. REITER

Department of Statistical Science

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

*E-mail: R.Mitra@soton.ac.uk.

Abstract

In many observational studies, researchers estimate causal effects using propensity scores, e.g., by matching or sub-classifying on the scores. Estimation of propensity scores is complicated when some values of the covariates are missing. We propose to use multiple imputation to create completed datasets, from which propensity scores can be estimated, with a general location mixture model. The model assumes that the control units are a latent mixture of (i) units whose covariates are drawn from the same distributions as the treated units' covariates and (ii) units whose covariates are drawn from different distributions. This formulation reduces the influence of control units outside 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. We illustrate the benefits of

the latent class modeling approach with simulations and with an observational study of the effect of breast feeding on children's cognitive abilities.

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

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, $e(x_i)$, is the probability that the subject receives the treatment given its vector of covariates x_i. That is, $e(x_i) = P(T_i = 1|x_i)$, 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 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 x. Other approaches to causal inference based on propensity scores include sub-classification (Rosenbaum and Rubin, 1984; Hullsiek and Louis, 2002), full matching (Rosenbaum, 1991; Stuart and Green, 2008) and propensity score weighted-

estimation (Lunceford and Davidian, 2004). For a review of different approaches to causal inference using propensity scores, see D'Agostino (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 x as the independent variables, and using the estimated probabilities as the propensity scores. See, for example, Woo *et al.* (2008).

In this article, we consider scenarios in which some covariate data are missing, which complicates estimation of propensity scores. There are several strategies in the literature for overcoming this complication. 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 (Rosenbaum and Rubin, 1984); however, with many patterns there may not be adequate matches. The analyst could apply the model-based approach of D'Agostino 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 conditional 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. Multiple imputation approaches have some advantages over maximum likelihood approaches. 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 using nonparametric models or including interaction and higher order effects that may not be in the complete data model. With com-

pleted datasets, the analyst can easily pursue further modeling, such as sub-domain comparisons or regression adjustment to reduce residual imbalances (Hill, 2004; Hill *et al.*, 2004).

For imputation, we propose a general location model, i.e. the categorical variables follow a log-linear model and the continuous variables follow a multivariate normal distribution within each category, with a novel twist. We introduce a latent indicator variable that captures the notion, "if we had complete data, these units would be good candidates for the matched control group." More precisely, we assume that the control units are a 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. This formulation reduces the influence of control units outside 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. The latent variable is never observed for control units. However, because all treated units are by definition in the treated units' covariate space, there is sufficient information to estimate the posterior distributions of the latent indicators using Markov Chain Monte Carlo techniques.

The remainder of the article is organized as follows. In Section 2, we illustrate the impact on covariate balance of using standard imputation models that generate implausible imputations. Most standard imputation models use the same parameter estimates for all units; we call these one class models. In Section 3, we describe a simple latent class mixture model and illustrate its improved performance over one class models in the settings of Section 2. In Section 4, we present the general location latent class mixture model, which we utilize in Section 5 to handle missing covariate data in an observational study of the effect of breastfeeding on children's cognitive outcomes later in life. In Section 6, we conclude with general remarks about these approaches to propensity score matching.

2 Potential inadequacies of one class models

To illustrate some potential problems with imputations from one class models, we suppose that the covariates, $x_i = (x_{i1}, \ldots, x_{ip})'$, for each unit $i = 1, \ldots, n$ are all continuous. In this context, a standard one class multiple imputation approach presumes that $x_i \sim N(\mu, \Sigma)$, with non-informative prior distributions on (μ, Σ), such as $p(\mu, \Sigma) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)}$. The full conditional distributions of the parameters and missing values are available in closed form, so that draws from the joint posterior distribution of all unknowns can be obtained with Gibbs samplers. The draws of the missing values serve as multiple imputations. This process is identical to data augmentation (Tanner and Wong, 1987).

2.1 Impact of model mis-specification

With high dimensional covariate spaces, it is unfortunately all-to-easy to misspecify regression models. Indeed, this issue motivates propensity score matching in place of regression analysis for causal inference in the first place. Misspecified models can generate implausible imputations, which in turn can negatively impact covariate balance in propensity score matching, as we now demonstrate.

We simulate $p = 2$ continuous covariates for $n = 1200$ units as shown in Figure 1. The $n_T = 200$ treated units tend to have larger values of x_1 and x_2 than the $n_C = 1000$ controls. We introduce missing values in control units' x_2 data with a missing at random mechanism so that units with large values of x_1 are more likely to be missing x_2. Thus, there are many missing values among control units living in the same covariate space as the treated units. Approximately 40% of control units are missing x_2. The models used to create these simulations are in Appendix C. We impute missing x_2 using data augmentation via the one class multivariate normal model. After $m = 100000$ imputations, 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

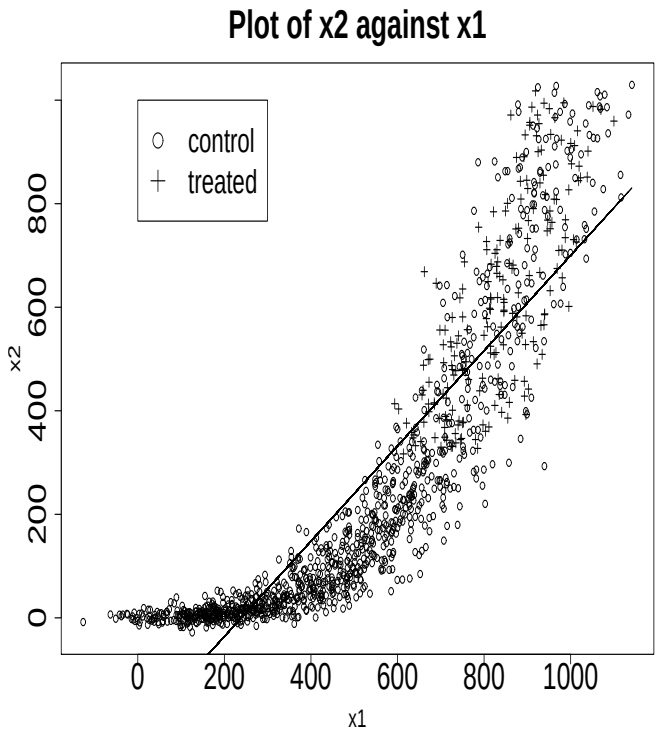

Figure 1: Scatter plot of x_2 against x_1 when a cubic relationship is present, illustrating the effects of using a poor imputation model.

based on the averaged propensity scores.

The one class multivariate normal imputation model implies a linear relationship between x_1 and x_2, which is clearly inappropriate as indicated by the estimated regression line in Figure 1. What can happen when using this regression model to impute the missing x_2? First, consider control units with actual covariate values in the treated units' region of the covariate space; these are ideal candidates for the matched control set. When based on the one class model, imputations of 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 will 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 smaller actual values of x_2. When using the one class model, imputations of x_2 for these units will tend to be higher than their true x_2 values. The imputations may put these control

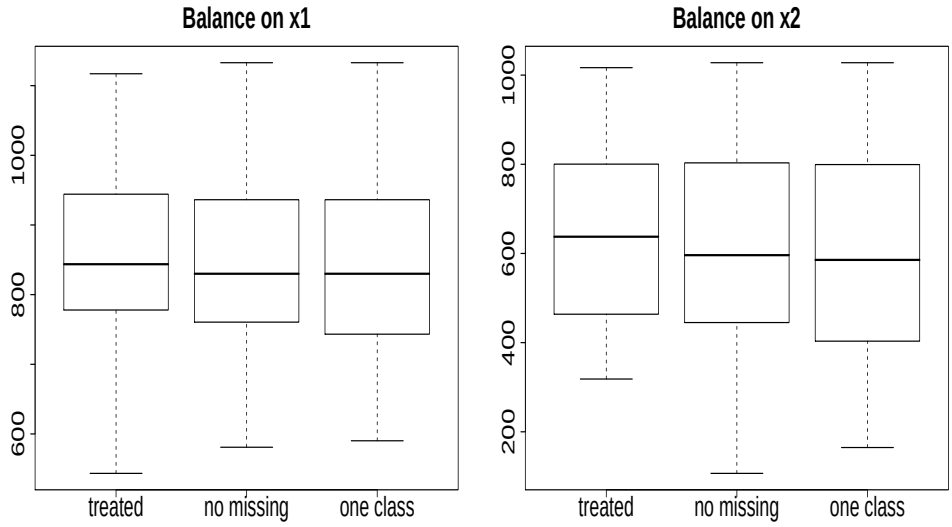

Figure 2: Box plots of x_1 and x_2 for the treated units, matched control units selected with no missing covariate data, and matched control units from the one class model in the model mis-specification simulation design.

units in the same region as the treated units' covariates and, therefore, incorrectly make them 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.

Figure 2 displays the distributions of true x_1 and x_2 values for the treated and matched control units with multiple imputation using the one class model. These matched controls are also compared with the matched controls selected if there are no missing x_2 values. For both x_1 and x_2, the lower tails for the matched controls from the one class model are longer than the lower tails for the treated units and matched controls selected from the fully complete data. This is because the model tends to impute missing x_2 values higher than their true values for control units just outside the treated units' covariate space.

Of course, a wise modeler would recognize the inadequacy of the multivariate normal model and use some other imputation approach. We use an obvious mis-specification in this example to illustrate the impacts of implausible imputations on covariate balance. In problems with many covariates, it is not always easy to diagnose

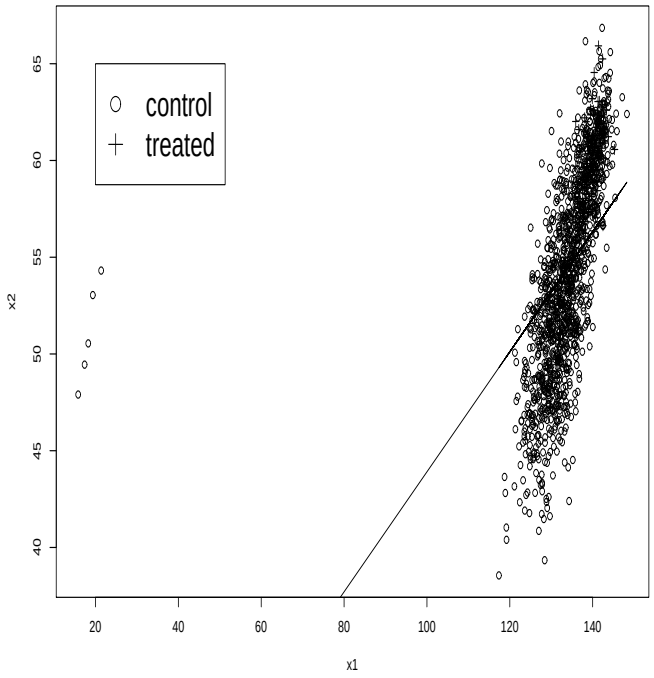

Figure 3: Plot of the covariates in the impact of outliers simulation with the fitted regression line.

model inadequacies. Furthermore, although unfortunate, many data analysts default to multivariate normal imputation procedures, so that they may face the problems from imputation model mis-specification.

2.2 Impact of outliers

Frequently, some units have unusual values of covariates. These values could have undue influence on the parameter estimates of the imputation model, which in turn could result in implausible imputations. In this section, we illustrate this phenomenon.

We simulate $p = 2$ continuous covariates for $n = 1400$ units as shown in Figure 3. There are $n_T = 100$ treated units with ample numbers of overlapping control units. Five control units have severely outlying values of (x_1, x_2). We again introduce missing data in control units' x_2 under a MAR scheme so that units with large x_1 are more likely to be missing x_2. Approximately 40% of control units are missing x_2.

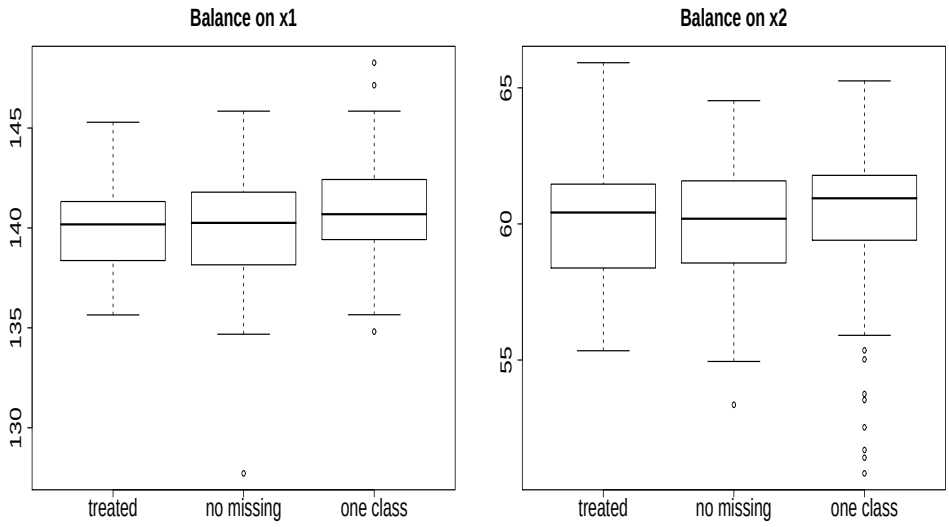

Figure 4: Box plots of x_1 and x_2 for treated units, matched control units selected with no missing covariate data, and matched control units from the one class model in the impact of outliers simulation.

We again use $m = 100000$ imputations from the one class model. Propensity scores are estimated by a logistic regression with main effects and an interaction of the two covariates. We find the matched controls as in Section 2.1. The models used to create these simulations are in Appendix C.

The estimated parameters of the one class model are strongly affected by the outliers, as indicated by the estimated regression line in Figure 3. As a result, imputations based on the estimated regression do not accurately reflect the relationship between x_1 and x_2 in the region of potential control matches. Thus, the types of problems discussed in Section 2.1 can arise when finding matches based on the implausible imputations.

Figure 4 displays the distributions of true x_1 and x_2 values for the treated units, matched control units selected when there is no missing data, and matched control units selected when using the one class model for imputations. While the covariate distributions of the the treated units and matched control units from the fully observed covariate data are similar, the matched control units from the the one class model tend to have larger values for both x_1 and x_2 than the treated units. This in-

dicates that using a one class model for imputations here can have an adverse impact on true covariate balance.

Once again, the wise imputer might account for these outliers when estimating imputation models, for example by tossing out these five clearly visible non-matches. In general, however, multivariate outliers can be challenging to detect in high dimensions.

3 Latent class mixture model

The simulations in Section 2 demonstrate that implausible imputations can negatively impact covariate balance. They further illustrate that inappropriate use of a one class model can lead to these problems. In this section, we propose an approach that attempts, in some sense, to mitigate these problems automatically through latent class mixture modeling. We note that Beunckens *et al.* (2008) also use latent class models for multiple imputation, but not in the context of propensity score matching.

The motivation underlying the use of latent class models in this context is as follows. Ideally, we want to select matched controls that look like the treated units on relevant covariates. When covariate data are missing, we are unsure which control units are in this region of potential matches. However, if we did know which control units were in the potential match region, we could toss out the control units outside the potential match region and, therefore, fit imputation models using only the relevant covariate space. In this way, imputation of missing covariates in the treated units' covariate space would not be affected by outlying controls, as happened in Section 2. Since we do not know which control units are in the potential matched region, we introduce a latent class indicator such that one class corresponds to units lying in the potential matched region and the other class corresponds to all other units. By definition we know the latent class indicators for all treated units, so that there is information to estimate the distribution of the latent class indicators for the control units.

There is great flexibility in this mixture modeling approach. For example, the model might include two or more classes for the treated and matched control units, and two or more classes for the other control units. Such models may result in more plausible imputations than using just two latent classes. Here, we focus on the two class model for its simplicity.

Regardless of the number of latent classes, a key feature of our approach is that any treated and potential matched control units' missing data are imputed based on the same parameter values. This is preferable to imputing treated and control units separately. For example, when imputing separately, in any one imputation run the drawn values of parameters of the imputation model for the control units could differ greatly from the drawn values of the parameters for the treated units, which might lead to comparatively poor matches.

We now demonstrate that the latent class model can improve the problems seen in Section 2. We begin by describing a latent class mixture model for continuous variables only. We present a general location latent class mixture model for continuous and categorical variables in Section 4.

3.1 Latent class model for continuous data only

For each unit i, let $z_i \in \{0, 1\}$ be the latent class indicator, where $z_i = 1$ corresponds to unit i lying in the treated units' covariate space and $z_i = 0$ otherwise. We model each unit's covariate data conditional on z_i with class specific parameters, so that,

$$x_i|z_i \sim N(\mu^{z_i}, \Sigma^{z_i}). \tag{1}$$

The distribution of the latent class indicators conditional on treatment is

$$p(z_i = 1|T_i = 0) = \pi^* \tag{2}$$

$$p(z_i = 1|T_i = 1) = 1. \tag{3}$$

As in the one class model we place non-informative priors on $(\mu^{z_i}, \Sigma^{z_i})$, so that

$$p(\mu^{z_i}, \Sigma^{z_i}) \quad \propto \quad |\Sigma^{z_i}|^{-\left(\frac{p+1}{2}\right)}. \tag{4}$$

This can lead to an improper posterior when $z_i = z_j$ for all (i, j). However, this possibility is rare in practice. If this does occur, a one class model may be adequate. Analysts can adopt the approach of Diebolt and Robert (1994), also recommended by Wasserman (2000), and use a data dependent prior distribution that restricts imputation of z_i so that sufficient numbers of units are in both classes. We place a Beta prior distribution on π^*, $p(\pi^*) = Be(a, b)$, where (a, b) are specified hyper-parameters. Common choices for (a, b) include $a = b = 1$, implying a uniform prior for π^*, and $a = b = 0.5$ for the Jeffrey's prior.

With this model specification, the full conditional distributions are available in closed form. It is straightforward to sample from the joint posterior distribution of all unknowns using a Gibbs sampler, thus creating multiple imputations of the missing covariate values.

3.2 Performance in simulations

We now apply the latent class model in the settings of Section 2. We also add a scenario in which the one class model is appropriate in order to illustrate the effect on covariate balance when the latent class model is inefficient compared to the one class model. In all scenarios, we run the Gibbs sampler for 100000 iterations after a burn-in period of 1000 runs. Thus, we create 100000 multiply-imputed datasets in each scenario. In each dataset, we estimate the propensity scores using the logistic regressions described in Section 2. We then compute the average propensity scores for each unit across the 100000 datasets, and match treated to control units without replacement using nearest neighbor matching.

Figure 5 and Figure 6 summarize the covariate balance on x_1 and x_2 for the simulation with model mis-specification and the simulation with outliers, respectively.

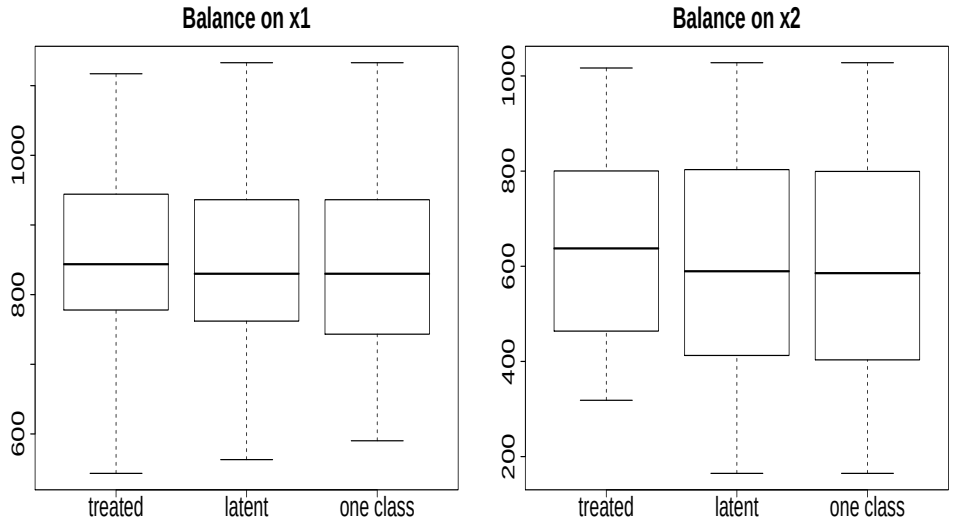

Figure 5: Box plots of treated and matched control units for x_1 and x_2 respectively in the model mis-specification simulation design. Matched controls from both the latent and one class approach are presented.

In both scenarios, true covariate balance is generally improved when using the latent class model as compared to using the one class model. Notably, obtaining more plausible imputations of x_2 not only helps balance x_2 more effectively, it results in better balance on x_1.

In the model misspecification scenario, the latent class model still is not the correct model for $f(x_2|x_1)$. However, as evident from Figure 1, using a linear model for imputations is not unreasonable for units lying in the treated units' region of the covariate space. This points to a general advantage of adding the latent indicators: assumptions of linearity or other simplifications, while possibly inappropriate over the whole covariate space, may be reasonable on a smaller region where the treated units lie.

In the outlier scenario, the latent class model puts the outliers in the class corresponding to the non-matched region. Hence, these points will not unduly influence the estimated regression coefficients for the units in the region of potential matches. This is not the case with the one class model, which does not separate those units out as part of the estimation process. This points to a second general advantage of

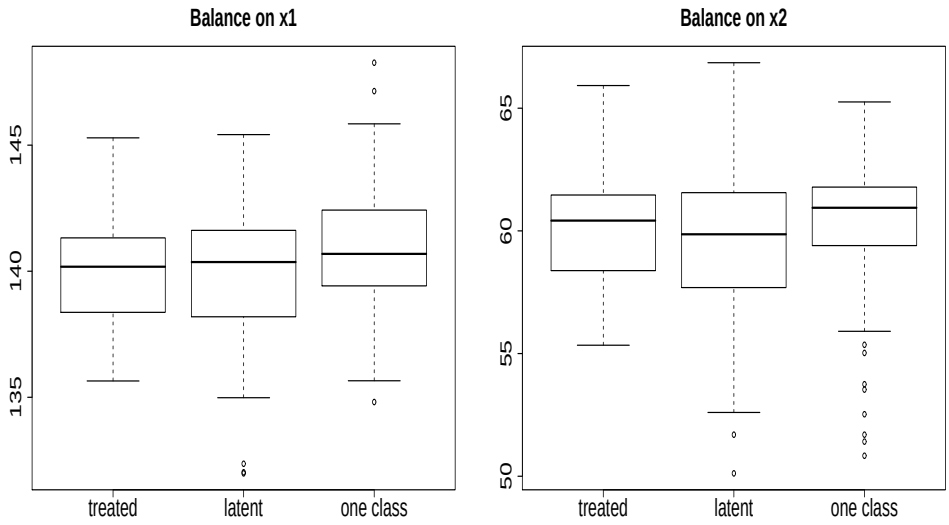

Figure 6: Box plots of treated and matched control units for x_1 and x_2 respectively in the impact of outliers simulation. Matched controls from the latent and one class approach are presented.

adding the latent indicators: the impact of outliers that are not likely to be matches can be mitigated.

Of course, covariate balance is an intermediate step in causal inference. The ultimate goal is to estimate treatment effects. We therefore simulate a response variable, y, for each scenario with a simple response surface, namely

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

Here, the 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 y in the treated group and $\bar{y}_{MC}$ is the sample mean of y in the matched control group.

Table 1 summarizes the estimates for three independent simulations of Y, including $\hat{\tau}$ for the latent and one class models and $\hat{\tau}$ estimated before introduction of missing data. In both scenarios, imputations with the latent class model result in values of $\hat{\tau}$ closer to zero than imputations with the one class model. Thus, the gains in covariate balance from the one class model translate to better estimates of

Estimate	Rep 1	Rep 2	Rep 3
Model misspecification scenario			
No missing	30.6	42.7	22.8
Latent class	43.0	52.6	34.8
One class	52.4	62.7	54.6
Outliers scenario			
No missing	0.25	0.46	-0.09
Latent lass	0.60	0.09	0.06
One class	-1.13	-1.49	-1.64
Truly one class scenario			
No missing	14.4	18.8	39.9
Latent class	18.5	25.5	45.9
One class	16.5	22.8	44.9

Table 1: Three replicates of treatment effect estimates after matching based on no missing data, on multiple imputation with the latent class model, and on multiple imputation with the one class models. The true treatment effect equals zero in all three designs. For the misspecification scenario, the $SE(\bar{Y}_T) \approx 22$. For the outliers scenario, the $SE(\bar{Y}_T) \approx 0.38$. For the truly one class scenario, the $SE(\bar{Y}_T) \approx 14$.

treatment effects for this response surface.

One might ask what happens when the one class model is correct for the covariates, but imputations are done with the latent class model. To explore this scenario, we add a simulation in which (x_1, x_2) have a linear relationship. The $n_T = 200$ treated units tend to have larger values of x_1 and x_2 than the $n_C = 1000$ controls. This simulation design is summarized in Figure 7. As in the previous simulations we introduce missing values in control units' x_2 data with a missing at random mechanism so that units with large values of x_1 are more likely to be missing x_2. The models used to create these simulations are in Appendix C. Here, a one class model is appropriate for imputing the missing x_2. We impute $m = 100000$ datasets using both the one class and the latent class models.

Figure 8 summarizes the covariate balance on x_1 and x_2 for both the one class and the latent class models. Propensity scores are estimated from $m = 100000$ datasets logistic regressions with main effects of x_1 and x_2. Both imputation approaches result in similar balance on x_1, whereas for x_2 the one class model is slightly better balanced.

We also simulate a response surface as in (5) and estimate treatment effects. The

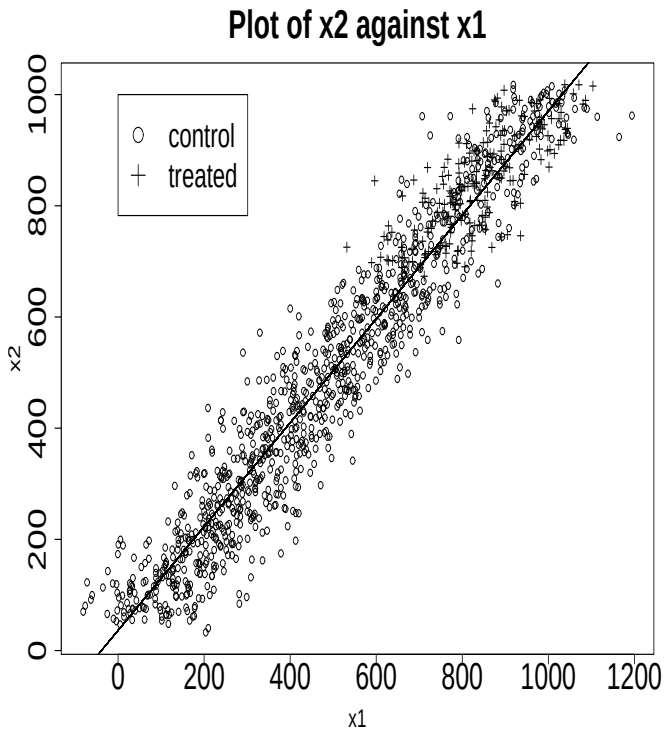

Figure 7: Scatter plot of x_2 against x_1 when a linear relationship is present and the one class model holds.

results are in Table 1. The one class estimates are slightly closer to $\tau = 0$ than the latent class estimates. This is due to the loss of efficiency in estimating parameters in the imputation model with an unnecessary latent class. Essentially, the latent class model estimates the parameters for the treated/matched class using only a fraction of the control units, whereas the one class model appropriately uses all control units. Despite this inefficiency, the treatment effect estimates do not differ substantially.

The three simulation results suggest that the latent class model can help analysts to avoid bias from poor matches caused by implausible imputations, without substantial penalties when they are inefficient compared to one class models.

4 General location latent class mixture model

When covariates include both categorical and continuous variables, the general location model is often used for imputation of missing data. In this section, we extend

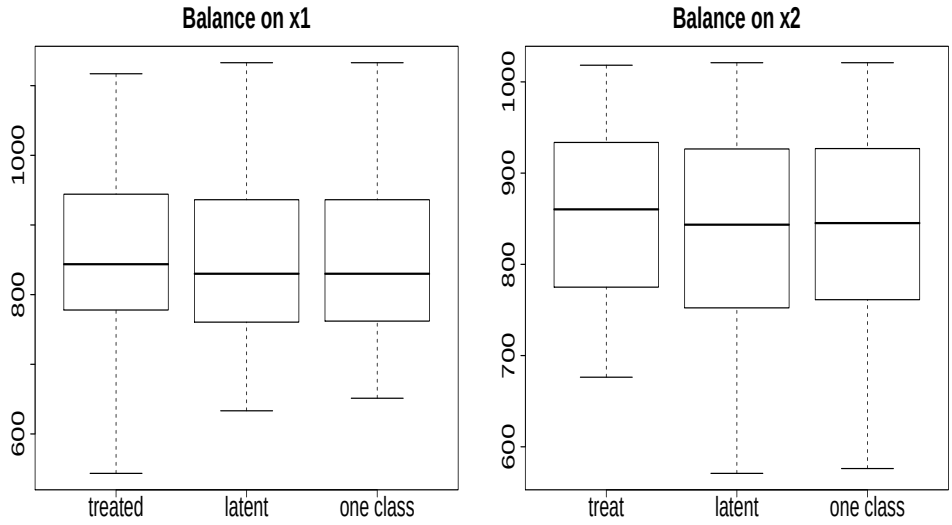

Figure 8: Box plots of treated and matched control units for x_1 and x_2 respectively in the model mis-specification simulation design. Matched controls from the latent and one class approach are presented.

the general location model to include latent indicators for propensity score matching.

4.1 One class general location model

Let X be an $n \times p$ matrix of covariate data for the individuals in the study comprising q continuous variables, $W = (W_1, \ldots, W_q)$, and r categorical variables, $V = (V_1, \ldots, V_r)$, where $q + r = p$. Each V_j takes on d_j distinct values. Thus, each unit can be classified into one of $D = \prod_{j=1}^{r} d_j$ cells of an r-dimensional contingency table.

Let $f = \{f_d : d = 1, \ldots, D\}$ be the resulting set of cell counts, assuming an appropriate (e.g. anti-lexicographical) ordering of cells. We assume that f has a multinomial distribution with probability vector $\pi = \{\pi_d : d = 1, 2, \ldots D\}$. Within each cell d, we assume that W follows a multivariate normal distribution, $p(W|\mu_d, \Sigma) = N(\mu_d, \Sigma)$. Here, μ_d is the q-vector of means for cell d, and Σ is the $q \times q$ covariance matrix assumed equal for all d. We use a Dirichlet prior distribution for π with pre-specified hyper-parameters $\alpha = (\alpha_1, \ldots, \alpha_D)$. We use a non-informative prior distribution on (μ, Σ), i.e., $p(\mu, \Sigma) \propto |\Sigma|^{-\left(\frac{q+1}{2}\right)}$.

In many applications, D is quite large, possibly exceeding n. With large D many

cells are empty or sparsely populated. To allow estimation of the parameters, analysts can restrict π and $\mu = (\mu_1, \ldots, \mu_D)$. For π, a typical approach is to use log linear constraints. Specifically, let C be a $D \times s$ matrix such that $s \leq D$. The log linear model requires π to satisfy $log(\pi) = C\lambda$. Typically, C contains main effects for each V_j and possibly interactions among selected (V_i, V_j). For μ, analysts can specify a linear model on the categorical variables. This model frequently mimics the structure of C, including main effects and interactions among $V_1, \ldots, V_q$.

Analysts can use a Gibbs sampler to sample from the joint posterior distribution of unknowns. A convenient approach for obtaining posterior draws of π is Bayesian iterative proportional fitting; see, Schafer (1997, Ch. 4) and Gelman *et al.* (1995). Conditional on parameter draws, missing categorical data are imputed from multinomial distributions and missing continuous data are imputed from multivariate normal distributions.

4.2 Adding the latent class indicators

As in Section 3.1, let z_i be the latent class indicator for each unit i. We model the distribution of latent class indicators as in (2) and (3), with the same considerations for the prior distribution on π^*. Given the latent class indicators $z = \{z_1, \ldots, z_N\}$, we can partition the data into two groups. Let $X = (X^0, X^1)$, where $X^0 = \{x_i : z_i = 0, i = 1, \ldots, n\}$ and $X^1 = \{x_i : z_i = 1, i = 1, \ldots, n\}$ correspond to covariates for units belonging to latent classes 0 and 1 respectively. As in Section 4.1 we can further partition the data into its continuous and categorical components, $X^0 = (V^0, W^0)$ and $X^1 = (V^1, W^1)$.

Essentially, the mixture model specifies separate general location models for X^0 and X^1. Let $\theta^0 = (\pi^0, \mu^0, \Sigma^0)$ and $\theta^1 = (\pi^1, \mu^1, \Sigma^1)$ be the parameters of the general location model for X^0 and for X^1 respectively. Then,

$$p(X|\theta^*, z) = p(X^0|\theta^0)p(X^1|\theta^1) \tag{6}$$

where $p(X^0|\theta^0)$ and $p(X^1|\theta^1)$ are modeled as described in Section 4.1. Cell counts are still modeled with multinomial distributions, but now cell probabilities depend on latent class membership. Similarly, the continuous data are still modeled as multivariate normal, but the mean and covariance matrix depend on the latent class.

As in Section 4.1, we use Dirichlet prior distributions for π^1 and non-informative prior distributions for $(\mu^1, \Sigma)^1)$, and similarly for (π^0, μ^0, Σ^0). As in Section 3.1, with non-informative prior distributions, sufficient numbers of units are required in both classes to estimate the parameters.

The full conditional distributions for all unknowns are available in closed form. We describe in detail the data augmentation steps needed to impute missing values in Appendix A.

5 Application to study of breast feeding

We now apply the latent class model to impute missing covariates and perform propensity score matching in a study of the effect of breast feeding on child's cognitive development. The data are a subset of the National Longitudinal Survey of Youth.

5.1 Description of study

The response variable, y, 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 (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.

We use fourteen potentially relevant background covariates. These include five 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. They also include seven 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, 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 (Box and Cox, 1964) to several continuous variables to improve the assumption of normality; see Appendix B. We also 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 as seen in its histogram displayed in Figure 16 in Appendix B. Finally, we 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 normal model; see Figure 17 in Appendix B.

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 (MAR) assumptions, which may not be true in practice. We do not consider extensions to handling the missing treatment indicators and missing outcome data in the analysis here. 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

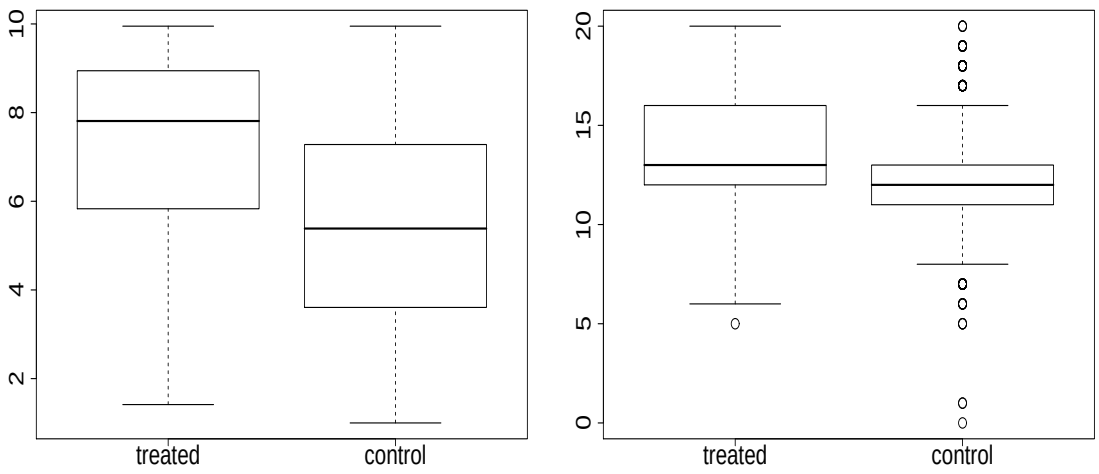

Figure 9: Box plots of mother's intelligence score and mother's years of education respectively for treated and control units before matching.

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

Table 2: Distribution of child's race.

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 9 summarizes the distribution of mother's intelligence and education for observed treated and control units, and Table 2 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 matching in the presence of the missing data.

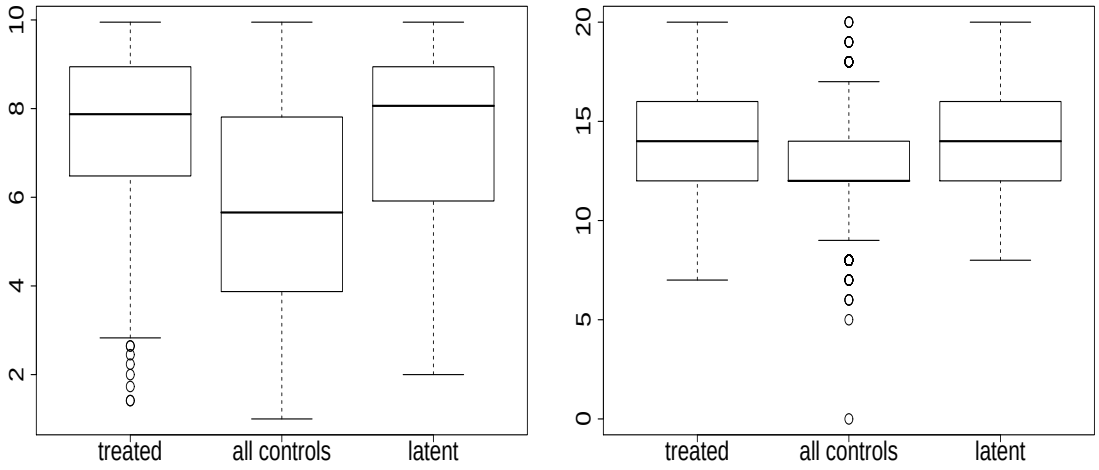

Figure 10: True covariate balance on mother's intelligence score and mother's years of education respectively in the simulation involving the complete cases.

5.2 Complete Case simulation

We first evaluate the performance of the latent class model at achieving true covariate balance in a simulation involving the 1306 complete cases. Although this is a much smaller sample size, we can introduce missing data, run the model, and examine covariate balance using the true data. We 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 the latent class imputation model, we use a main effects only log linear model for the categorical variables. We use the Box-Cox transformations to normality for the continuous data and relate the within-category means using a linear model with main effects of the categorical variables. We run the Gibbs sampler for 200000 iterations after discarding an initial 5000 as burn-in.

Figure 10 displays the distributions of mother's intelligence and years of education for the treated units, full control reservoir, and matched control units. For both variables, the imbalance has been greatly reduced after imputation and matching. More detailed examination of balance on these covariates is presented in Table 4 and Table 5 in Appendix B.

We also compare proportions of child's race for treated and control units before and after matching in Table 3. Once again, covariate imbalance is greatly reduced here. Similar examinations with other variables indicate that the latent class model results in a well balanced matched control set with respect to these covariates.

race	treated	all controls	latent
Hispanic	0.1528	0.1844	0.1296
black	0.0926	0.2697	0.1111
other	0.7546	0.5459	0.7593

Table 3: True covariate balance on child's race in the simulation involving the complete cases.

5.3 Application to the full data

We now apply the latent class model on the original data set of 2388 units. Similar restrictions are imposed on the cell probabilities and within cell means as in the simulation involving the complete cases. We again run the Gibbs sampler for 200000 iterations with an additional burn-in of 5000 iterations.

We estimate the treatment effect with $\bar{Y}_T - \bar{Y}_{MC} = -0.059$, with a conservative two-sample, pooled standard error of 0.94. For alternative approaches to estimating standard errors from propensity score matching, see Hill and Reiter (2006). This is noticeably different than the treatment effect estimate based on all controls, which is 5.23 (SE = 0.74). The treatment effect after matching is thus significantly closer to zero. Similar results were obtained by Der *et al.* (2006), who used a regression approach to infer that the effect of breast feeding is minimal.

For comparisons, we also used the one class model to impute missing values. The estimated treatment effect is 0.96 (SE = 0.96). Thus, there is approximately a one point (and one standard error) difference in the treatment effect estimates from the two imputation approaches. The difference is modest primarily because, on average across imputations of z_i, approximately 85% of control units are imputed to lie in the latent class for the treated/matched control region, so that there is not much

difference between the two models. We note, however, that certain outlying controls, e.g., one mother spent many more weeks in the hospital than others, are never in the latent class for treated units. This is in line with the outlier simulations of Section 3: the latent class model can moderate the impact of outlying units on imputations for units in the region of plausible matches.

We also repeated the complete case simulation with the one class model. It was difficult to distinguish a clear winner on covariate balance between the one class and latent class models. Both greatly improved covariate balance over use of the full control reservoir.

6 Concluding Remarks

When analysts estimate treatment effects using propensity score matching, using a latent class mixture model can result in better covariate balance than using a one class model. Essentially, the latent class model allows the data analyst to estimate imputation models specifically for the region of interest. In this way, control units that have minimal relevance for treatment effect estimation also have minimal relevance for imputation of missing data. Even when one class models are correct, we anticipate that the loss in efficiency from using the latent class model will be minor, particularly when compared to the reductions in bias achievable from avoiding implausible imputations. This was borne out in the simulations in Section 3, and to some extent in the breast feeding study in which the one class and latent class approaches gave results with only modest differences.

We did not control for outcome data in the imputation steps. We avoided controlling for outcomes to be consistent with the philosophy of propensity score matching: the control group should be constructed without consideration of response variables. Of course, one could easily modify our procedure to include outcome variables in the imputation models.

In future research, we plan to investigate methodology that utilizes latent class in-

dicators with multiple levels. Presumably, this could improve imputation models even more than using binary latent variables, albeit at increased computational complexity. This avenue of research leads to questions of how to specify the latent classes, which could be explored with semi-parametric approaches such as Dirichlet process mixture models. We also plan 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 the average of the propensity scores across all completed datasets (as we do here). We tried both approaches in our analyses and found that matching on the average propensity score resulted in more accurate treatment effect estimates than matching within each dataset.

Acknowledgments

The authors would like to thank Professor Jennifer Hill who provided us with the data from the breast feeding study analyzed in Section 5. This research was supported by the National Science Foundation [NSF-ITR-0427889].

Appendicies

In Appendix A we present the data augmentation steps to impute missing values when using the general location latent class mixture model described in Section 4.2 of the main text. In Appendix B, we provide further details about the application of these models to the analysis of breast feeding data. In Appendix C, we present details of the simulation designs described in Sections 2 and 3 of the main text.

A Data augmentation steps to impute missing values with the general location mixture model

We describe here the steps required to impute missing covariates with the general location mixture model. We first explicitly describe the model and then present the I and P steps in a data augmentation algorithm.

A.1 Model specification

Let $X = (x_1, \ldots, x_n)'$, where $x_i = (x_{i1}, \ldots, x_{ip})'$ are the ith unit's covariates. For each unit, let $z_i \in \{0, 1\}$ be a latent class indicator, where $z_i = 1$ when unit i lies in the covariate space occupied by the treated units, and $z_i = 0$ otherwise. Let $z = (z_1, \ldots, z_n)'$. We partition the covariate data into two groups by the latent class. Let $X^1 = \{x_i, i : z_i = 1\}$ and $X^0 = \{x_i, i : z_i = 0\}$ correspond to the covariates for units belonging to latent classes 1 and 0, respectively.

We assume there are q continuous variables and r categorical variables with $q + r = p$. We partition X^1 into its categorical variables, $V^1 = (V_1^1, \ldots, V_r^1)$, and its continuous variables, $W^1 = (W_1^1, \ldots, W_q^1)$. Similarly, we partition X^0 into V^0 and W^0. The distribution of X is then

$$p(X|z) = p(X^1)p(X^0), \tag{7}$$

where we model both X^1 and X^0 using general location models.

Specifically, for X^1 we have

$$p(X^1) = p(V^1, W^1) = p(V^1)p(W^1|V^1). \tag{8}$$

The categorical data V^1 can be summarized using a contingency table. If each variable V_j^1 takes on d_j distinct values, $j = 1, \ldots, r$, then each unit can be classified into one of $D = \prod_{j=1}^{r} d_j$ cells of the r-dimensional contingency table. Denote the resulting set of cell counts by $f^1 = \{f_d^1 : d = 1, \ldots, D\}$ where an appropriate (e.g. anti-lexicographical) ordering of cells is assumed. The distribution of V^1 is a multinomial distribution on the cell counts f^1,

$$p(f^1|\pi^1) \sim M(n^1, \pi^1), \tag{9}$$

where $n^1 = \sum_{i=1}^{n} z_i$ and $\pi^1 = \{\pi_d^1 : d = 1, 2, \ldots D\}$ is an array of cell probabilities. For $p(W^1|V^1)$, we use

$$W^1 = U^1\mu^1 + \epsilon^1, \tag{10}$$

where $U^1 = (u_1^1, \ldots, u_{n^1}^1)'$ is a $n^1 \times D$ matrix, with row u_i^1 containing a one in position d if unit i falls into cell d and zeros elsewhere, and $\epsilon^1 = (\epsilon_1^1, \ldots, \epsilon_{n^1}^1)'$ is a $n^1 \times q$ matrix of error terms such that, $\epsilon_i^1 \sim N(0, \Sigma^1)$.

We similarly model X^0 using a general location model,

$$p(X^0) = p(V^0, W^0) = p(V^0)p(W^0|V^0). \tag{11}$$

Let $f^0 = \{f_d^0 : d = 1, \ldots, D\}$ be the cell counts from the contingency table formed by the cross-classification of the variables in V^0. Its distribution is multinomial with

$$p(f^0|\pi^0) \sim M(n^0, \pi^0) \tag{12}$$

where $n_0 = n - n_1$ and $\pi^0 = \{\pi_d^0 : d = 1, 2, \ldots D\}$ is an array of cell probabilities. For $p(W^0|V^0)$, we use

$$W^0 = U^0\mu^0 + \epsilon^0, \tag{13}$$

where $U^0 = (u_1^0, \ldots, u_{n^0}^0)'$ is a $n^0 \times D$ matrix, with row u_i^0 containing a one in position d if unit i falls into cell d and zeros elsewhere, and $\epsilon^0 = (\epsilon_1^0, \ldots, \epsilon_{n^0}^0)'$ is a $n_0 \times q$ matrix of error terms such that, $\epsilon_i^0 \sim N(0, \Sigma^0)$.

We model the distribution of the latent class indicator conditional on treatment so that,

$$p(z_i = 1|T_i = 0) \quad = \quad \pi^* \tag{14}$$

and,

$$p(z_i = 1|T_i = 1) \quad = \quad 1. \tag{15}$$

To complete the Bayesian specification, we place prior distributions on the parameters. For π^1 and π^0, we use

$$\pi^1 \sim Dir(\alpha^1), \;\; \pi^0 \sim Dir(\alpha^0).$$

For (μ^1, Σ^1), we use

$$p(\mu^1, \Sigma^1) \propto |\Sigma^1|^{-\left(\frac{q+1}{2}\right)}. \tag{16}$$

We use similar non-informative prior distributions for (μ^0, Σ^0). Finally the prior distribution for π^* is

$$p(\pi^*) = Be(a, b). \tag{17}$$

With this model specification, the full conditional distributions of all unknowns are available in closed form. This allows imputations to be drawn using data augmentation. We describe the I and P steps of the data augmentation in the next section.

A.2 I and P steps

First we define notation to characterize the missing and observed portions of the dataset. For unit i with $z_i = 1$, define its covariates by $x_i^1 = (x_{i1}^1, \ldots, x_{ip}^1)'$. We separate these into categorical and continuous covariates using $v_i^1 = (v_{i1}^1, \ldots, v_{ir}^1)'$ and $w_i^1 = (w_{i1}^1, \ldots, w_{iq}^1)'$, respectively. Let $m_i^{1^v}$ and $m_i^{1^w}$ be missing data indicators with ones for variables with missing values and zeros otherwise. Let $m_i^1 = (m_i^{1^v}, m_i^{1^w})$. Let the observed and missing data parts of the categorical variables be $v_{obs,i}^1 = \{v_{ij}^1, j : m_{ij}^{1^v} = 0\}$ and $v_{mis,i}^1 = \{v_{ij}, j : m_{ij}^{1^v} = 1\}$ respectively. Similarly, let $w_{obs,i}^1 = \{w_{ij}, j : m_{ij}^{1^w} = 0\}$ and $w_{mis,i}^1 = \{w_{ij}^1, j : m_{ij}^{1^w} = 1\}$ be the observed and missing values in the continuous data for individual i. Similarly for unit i with $z_i = 0$ define its covariates by x_i^0, with categorical covariates v_i^0 and continuous covariates w_i^0. As before denote the observed and missing data parts of the categorical variables as $v_{obs,i}^0$ and $v_{mis,i}^0$ respectively. Similarly, define $w_{obs,i}^0$ and $w_{mis,i}^0$ as the observed and missing values in the continuous data for individual i.

In addition, for each individual i where $z_i = 1$, denote the set of cells that agree with $v_{obs,i}^1$ as $O_i^1(d)$. For each unit i, partition μ_d^1 and Σ^1 by the observed and missing portions of w_i^1. Define $\mu_{d,i}^{1^o}$ and $\Sigma_i^{1^o}$ as the sub-vector and square sub-matrix of μ_d^1 and Σ^1, respectively, corresponding to $w_{obs,i}^1$. Similarly, define $\mu_{d,i}^{1^m}$ and $\Sigma_i^{1^m}$ as the sub-vector and square sub-matrix of μ_d^1 and Σ^1, respectively, corresponding to $w_{mis,i}^1$. Define $\Sigma_i^{1^{om}}$ as the $k_i^1 \times (q - k_i^1)$ sub-matrix with rows of Σ_i^1 corresponding to $w_{obs,i}^1$ and columns corresponding to $w_{mis,i}^1$ where, $k_i^1 = \sum_{j=1}^q (1 - m_{ij}^{1^w})$, and define $\Sigma_i^{1^{mo}} = \Sigma_i^{1^{om'}}$.

Similarly for unit i with $z_i = 0$ define $O_i^0(d)$, $\mu_{d,i}^{0^o}$, $\Sigma_i^{0^o}$, $\mu_{d,i}^{0^m}$, $\Sigma_i^{0^m}$, $\Sigma_i^{0^{om}}$, and $\Sigma_i^{0^{mo}}$ in the same way.

The I and P steps in the data augmentation algorithm used to impute the missing values are then as follows. First, impute missing covariates for unit i with $z_i = 1$. Impute $v^1_{mis,i}$ from a single multinomial trial with probability that unit i falls into cell d given by

$$p(i = d \mid v^1_{obs,i}, w^1_{obs,i}, \pi^1) = \frac{\exp(\delta^{1^o}_{d,i})}{\sum_{O^1_i(d)} \exp(\delta^{1^o}_{d,i})} \tag{18}$$

$$\tag{19}$$

where

$$\delta^{1^o}_{d,i} = \mu^{1^{o'}}_{d,i} \left(\Sigma^{1^o}_i\right)^{-1} w^1_{obs,i} - \frac{1}{2}\mu^{1^{o'}}_{d,i} \left(\Sigma^{1^o}_i\right)^{-1} \mu^{1^o}_{d,i} + \log(\pi^1_d) \tag{20}$$

for cells d that agree with $O^1_i(d)$ and zero otherwise. Let the imputed cell for unit i be $d^1_{com,i}$ and the corresponding vector of categorical variables be $v^1_{com,i}$. We then define a corresponding $n^1 \times D$ matrix $U^1_{com} = (u^1_{com,1}, \ldots, u^1_{com,n^1})'$, where $u^1_{com,i}$ contains a one in position $d^1_{com,i}$ and zeros elsewhere.

Next impute $w^1_{mis,i}$ from a multivariate normal distribution conditional on $w^1_{obs,i}$, $d^1_{com,i}$, and μ^1, Σ^1. We have

$$\begin{aligned} p(w^1_{mis,i} | w^1_{obs,i}, d^1_{com,i}, \mu^1, \Sigma^1) &= N(\tilde{\mu}_{d_{com,i}}, \tilde{\Sigma}_i) && (21) \\ \tilde{\mu}_{d_{com,i}} &= \mu^{1^m}_{d_{com,i}} - \Sigma^{1^{mo}}_i \left(\Sigma^{1^o}_i\right)^{-1} (w^1_{obs,i} - \mu^{1^o}_{d_{com,i}}) && (22) \\ \tilde{\Sigma}_i &= \Sigma^{1^m}_i - \Sigma^{1^{mo}}_i \left(\Sigma^{1^o}_i\right)^{-1} \Sigma^{1^{om}}_i. && (23) \end{aligned}$$

Let the imputed continuous variables for unit i be $w^1_{com,i}$. The completed covariate data set for units with $z_i = 1$ is then $X^1_{com} = (V^1_{com}, W^1_{com})$, where $V^1_{com} = (v^1_{com,1}, \ldots, v^1_{com,n^1})'$ and $W^1_{com} = (w^1_{com,1}, \ldots, w^1_{com,n^1})'$. Let f^1_{com} be the cell counts from the table formed by V^1_{com}.

We impute missing covariates for unit i with $z_i = 0$ similarly. First, impute $v^0_{mis,i}$

from a single multinomial trial with probability that unit i falls into cell d as

$$p(i = d \mid v^0_{obs,i}, w^0_{obs,i}, \pi^0) = \frac{\exp(\delta^{0^o}_{d,i})}{\sum_{O^0_i(d)} \exp(\delta^{0^o}_{d,i})} \tag{24}$$

where

$$\delta^{0^o}_{d,i} = \mu^{0^{o'}}_{d,i} \left(\Sigma^{0^o}_i\right)^{-1} w^1_{obs,i} - \frac{1}{2}\mu^{0^{o'}}_{d,i} \left(\Sigma^{0^o}_i\right)^{-1} \mu^{0^o}_{d,i} + \log(\pi^0_d) \tag{25}$$

for cells d that agree with $O^0_i(d)$ and zero otherwise. Denote the imputed cell for unit i to be $d^0_{com,i}$ and corresponding vector of categorical variables $v^0_{com,i}$. We then define a corresponding $n^0 \times D$ matrix $U^0_{com} = (u^0_{com,1}, \ldots, u^0_{com,n^0})'$, where $u^0_{com,i}$ contains a one in position $d^0_{com,i}$ and zeros elsewhere.

We next impute $w^0_{mis,i}$ from a multivariate normal distribution conditional on $w^0_{obs,i}$, $d^0_{com,i}$, and μ^0, Σ^0. We have

$$\begin{aligned} p(w^0_{mis,i} | w^0_{obs,i}, d^0_{com,i}, \mu^0, \Sigma^0) &= N(\tilde{\mu}_{d_{com,i}}, \tilde{\Sigma}_i) && (26) \\ \tilde{\mu}_{d_{com,i}} &= \mu^{0^m}_{d_{com,i}} - \Sigma^{0^{mo}}_i \left(\Sigma^{0^o}_i\right)^{-1} (w^0_{obs,i} - \mu^{0^o}_{d_{com,i}}) && (27) \\ \tilde{\Sigma}_i &= \Sigma^{0^m}_i - \Sigma^{0^{mo}}_i \left(\Sigma^{0^o}_i\right)^{-1} \Sigma^{0^{om}}_i. && (28) \end{aligned}$$

Let the imputed continuous variables for unit i be $w^0_{com,i}$. The completed covariate data set for units with $z_i = 0$ is then $X^0_{com} = (V^0_{com}, W^0_{com})$, where $V^0_{com} = (v^0_{com,1}, \ldots, v^0_{com,n^0})'$ and $W^0_{com} = (w^0_{com,1}, \ldots, w^0_{com,n^0})'$. Let f^0_{com} denote the cell counts from the table formed by V^0_{com}.

Finally we impute the latent class indicators z_i using

$$p(z_i | T_i = 0, \pi^*, \pi^1, \mu^1, \Sigma^1, \pi^0, \mu^0, \Sigma^0, X^0_{com}, X^1_{com}) = Ber(\hat{\pi}^*_i), \tag{29}$$

where

$$\begin{aligned}
\hat{\pi}_i^* &= \frac{\exp(\delta^1)\pi^*}{\exp(\delta^1)\pi^* + \exp(\delta^0)(1-\pi^*)}, \\
\delta^1 &= \mu^{1'}_{d_{com,i}}(\Sigma^1)^{-1}w_{com,i} - \frac{1}{2}\mu^{1'}_{d_{com,i}}(\Sigma^1)^{-1}\mu^1_{d_{com,i}} - \log(|\Sigma^1|) + \log(\pi^1_{d_{com,i}}), \\
\delta^0 &= \mu^{0'}_{d_{com,i}}(\Sigma^0)^{-1}w_{com,i} - \frac{1}{2}\mu^{0'}_{d_{com,i}}(\Sigma^0)^{-1}\mu^0_{d_{com,i}} - \log(|\Sigma^0|) + \log(\pi^0_{d_{com,i}}).
\end{aligned}$$

When any treated units have missing data, they are always imputed to be in class $z = 1$.

Conditional on X^1_{com}, X^0_{com} we update the parameters in the following P steps. First, update π^1 conditional on X^1_{com} using

$$p(\pi^1|X^1_{com}) \sim Dir(\alpha^1 + f^1_{com}). \tag{30}$$

When the total number of categories is large, for example exceeding n^1, analysts may need to impose log linear constraints on the cell probabilities. Specifically, define a $D \times s$ matrix C^1 where $s \leq D$. The log linear model requires π^1 to satisfy,

$$\log(\pi^1) = C^1\lambda^1 \tag{31}$$

Typically C^1 contains main effects of each V^1_j and possibly interactions among selected (V^1_i, V^1_j). Analysts can obtain posterior draws of π^1 using Bayesian iterative proportional fitting; see Schafer (1997, Ch. 4) and Gelman *et al.* (1995) for details. We used the R statistical software package "cat" (http://www.stat.psu.edu/ jls/misoftwa.html), developed by Joseph L. Schafer to obtain posterior draws for π^1. Conditional on π^1 and X^1_{com}, we then update (μ^1, Σ^1) in a block using

$$\Sigma^1|\pi^1, X^1_{com} \sim W^{-1}(N_1 - D, (\hat{\epsilon}^{1'}\hat{\epsilon}^1)^{-1}), \tag{32}$$

$$\mu^1|\pi^1, \Sigma^1, X^1_{com} \sim N(\hat{\mu}^1, \Sigma^1 \otimes (U^{1'}_{com}U^1_{com})^{-1}), \tag{33}$$

where $\hat{\epsilon}^1 = W^1_{com} - U^1_{com}\hat{\mu}^1$ is the matrix of estimated residuals and $\hat{\mu}^1 = (U^{1'}_{com}U_{com})^{1^{-1}}U^{1'}_{com}W^1_{com}$ is the least squares estimate of μ^1. Again when D is large a linear model for μ^1 on V^1 can be specified. Define a $D \times t$ design matrix A^1, where $t \leq D$. We re-express equation (10) as

$$W^1 = U^1 A^1 \beta^1 + \epsilon^1, \tag{34}$$

where β is a (reduced) $t \times q$ matrix of regression coefficients. As with C^1, columns of A^1 are typically chosen to reflect the structure of V^1, with main effects of each V^1_i and possibly interactions among selected (V^1_i, V^1_j). Posterior draws of β^1 and Σ^1 can be sampled as in the P-steps in (32) and (33), replacing U^1_{com} with $U^1_{com}A^1$.

We update parameters (π^0, μ^0, Σ^0) conditional on X^0_{com} in a similar manner, using

$$\begin{aligned} p(\pi^0|X^0_{com}) &\sim Dir(\alpha^0 + f^0_{com}) && (35)\\ \Sigma^0|\pi^0, X^0_{com} &\sim W^{-1}(N_0 - D, (\hat{\epsilon}^{0'}\hat{\epsilon}^0)^{-1}), && (36)\\ \mu^0|\pi^0, \Sigma^0, X^0_{com} &\sim N(\hat{\mu}^0, \Sigma^0 \otimes (U^{0'}_{com}U^0_{com})^{-1}), && (37) \end{aligned}$$

Again with large D, analysts can place a log linear model for π^0 that requires π^0 to satisfy

$$\log(\pi^0) = C^0\lambda^0 \tag{38}$$

for a $D \times s$ matrix C^0 with $s \leq D$. Analysts can draw values of π^0 by Bayesian iterative proportional fitting using the "cat" routine. Also, as with μ^1, analysts can specify a linear model for μ^0 on V^0. Let A^0 be a $D \times t$ design matrix, where $t \leq D$. We re-express equation (13) as

$$W^0 = U^0 A^0 \beta^0 + \epsilon^0, \tag{39}$$

where β^0 is a (reduced) $t \times q$ matrix of regression coefficients. Posterior draws of

β^0 and Σ^0 can be sampled as in the P-steps in (36) and (37), replacing U^0_{com} with $U^0_{com}A^0$.

Finally we update parameter π^* by,

$$p(\pi^*|T, a, b, z) = Be\left(a + \sum_{i:T_i=0} z_i,\ b + \sum_{i:T_i=0}(1 - z_i)\right). \tag{40}$$

B Further details of the breast feeding data analysis

We present here further details of the analysis of the breast feeding data considered in Section 5. In that analysis, we transformed several continuous variables prior to estimating the imputation models. Graphical displays revealing why these transformation were necessary are presented here. We also provide more detailed summaries of balance on the variables measuring mother's intelligence and mother's years of education, two variables that were badly imbalanced in the treated group and the observed full control reservoir.

B.1 Transformations of variables

Several continuous covariates had clearly non-normal distributions. To make imputations based on the general location models more plausible, we transformed variables using simple transformations suggested by the Box-Cox procedure (Box and Cox, 1964). Figures 11 – 15 summarize these transformations. The transformed variables are more reasonably described by normal distributions than the raw variables. As noted in the main text, we also categorized two continuous variables because of their highly non-normal distributional shapes. 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 as seen in its histogram displayed in Figure 16. Finally, we categorized 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 normal model; see Figure 17.

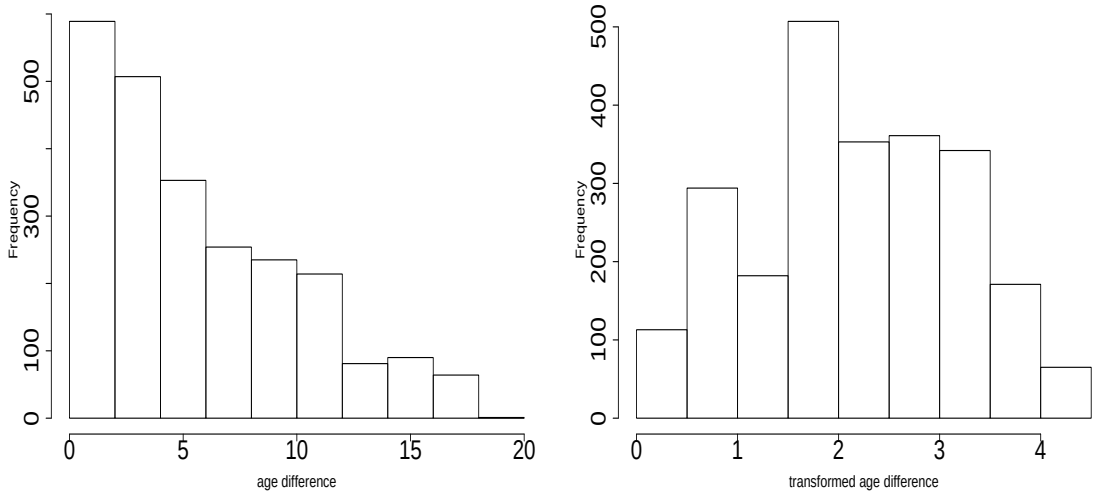

Figure 11: Histograms of difference between mother's age at birth and in 1979 before and after square root transformation.

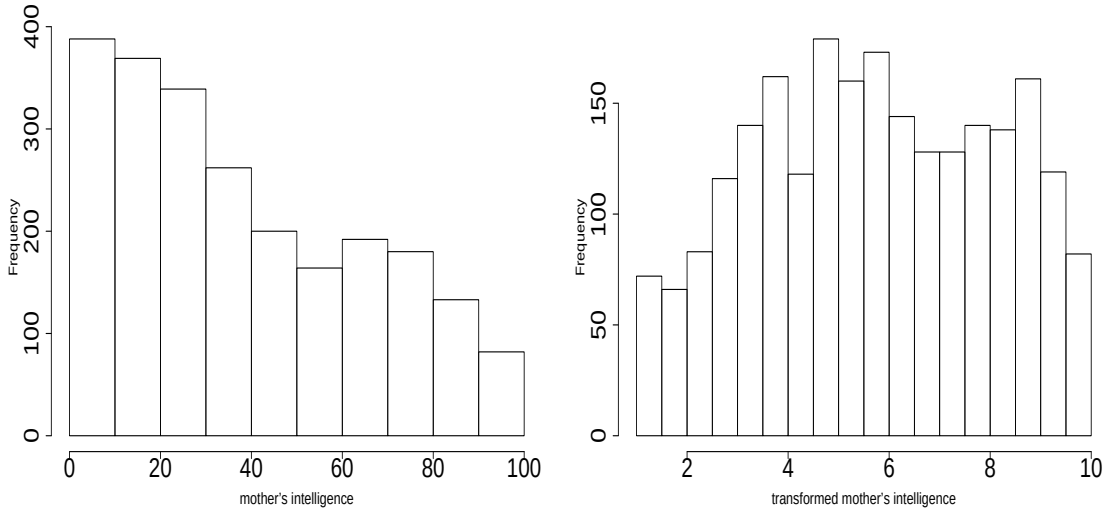

Figure 12: Histograms of mother's intelligence before and after square root transformation.

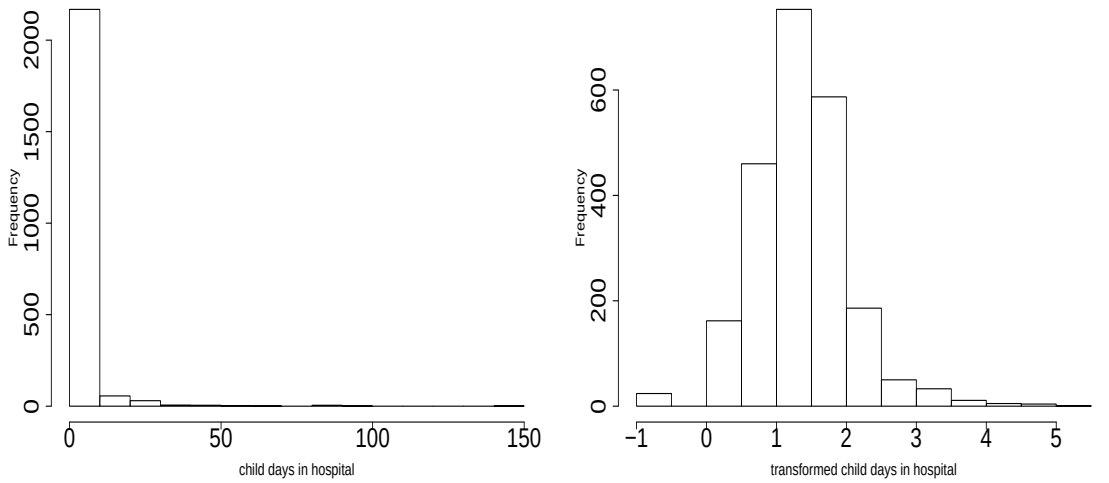

Figure 13: Histograms of child days in hospital before and after log transformation.

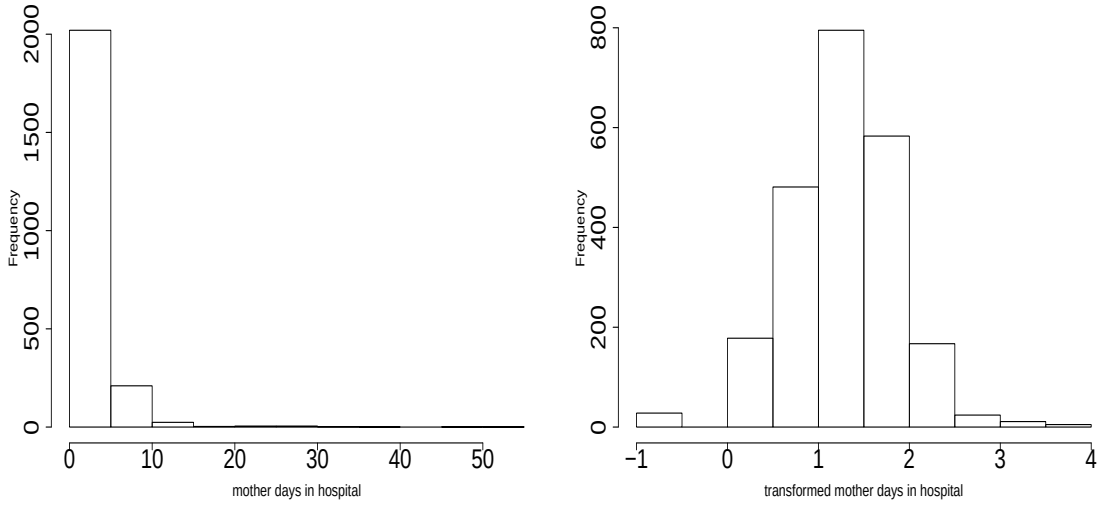

Figure 14: Histograms of mother days in hospital before and after log transformation.

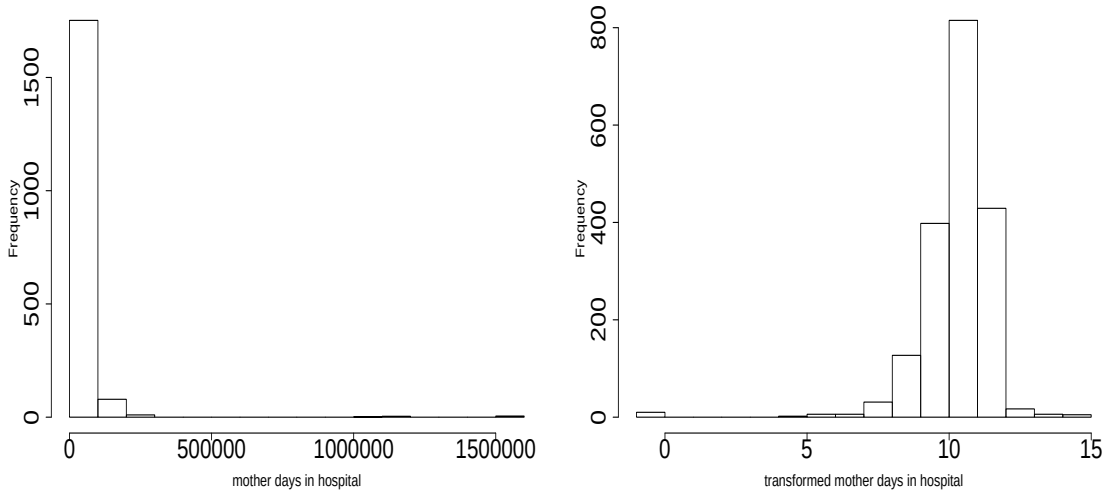

Figure 15: Histograms of family income before and after log transformation.

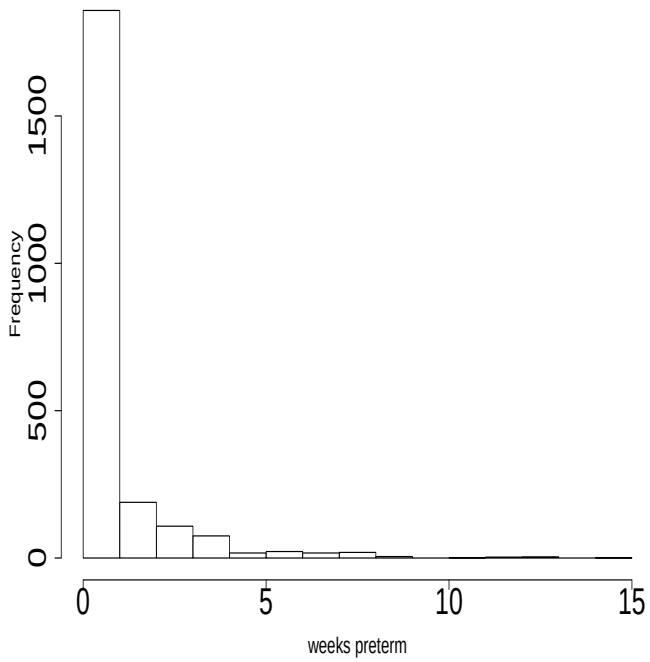

Figure 16: Histogram of weeks preterm for subjects in the breast feeding study.

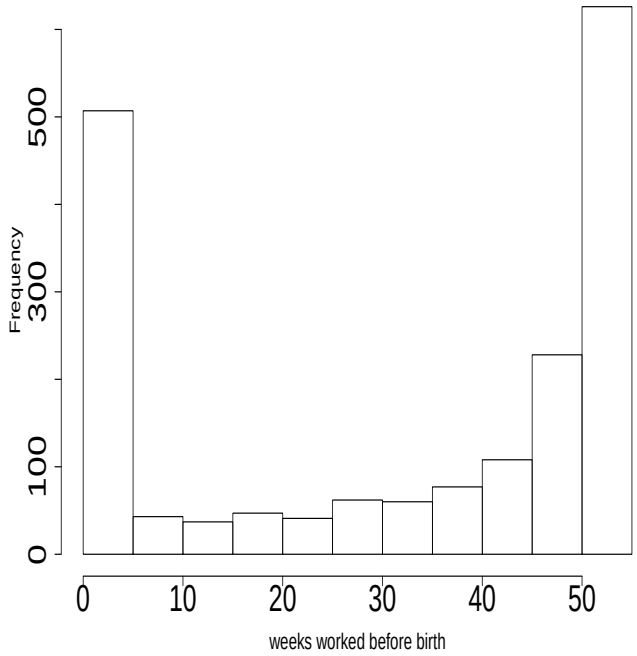

Figure 17: Histogram of weeks mother worked in the year before giving birth for subjects in the breast feeding study.

B.2 Balance on key covariates

In Section 5, we present box plots summarizing the distributions of mother's intelligence and mother's years of education for treated and matched control units based on the latent class general location model. Tables 4 and 5 display more detailed percentiles for the distributions of these variables. The treated and matched controls are well balanced with respect to these covariates.

Percentile	Treated	Matched Control
5	2.83	3.16
10	4.00	4.18
15	4.82	4.92
20	5.66	5.48
25	6.48	5.92
30	6.78	6.44
35	7.21	6.84
40	7.42	7.55
45	7.62	7.92
50	7.87	8.06
55	8.19	8.19
60	8.43	8.37
65	8.60	8.59
70	8.83	8.77
75	8.94	8.94
80	9.22	9.17
85	9.43	9.26
90	9.62	9.46
95	9.80	9.75

Table 4: Percentiles of mother's intelligence for treated and matched controls from the latent class general location model.

Percentile	Treated	Matched Control
5	10	11
10	12	12
15	12	12
20	12	12
25	12	12
30	12	12
35	12	12.25
40	13	13
45	14	14
50	14	14
55	15	14
60	15	15
65	16	15.75
70	16	16
75	16	16
80	16	16
85	16	16
90	17	17
95	18	18

Table 5: Percentiles of mother's education in years for treated and matched controls from the latent class general location model

C Design of the simulation studies

We simulated the covariates described in Figure 1 as follows.

$$x_{i1} = 50 + 0.8i + \epsilon_{i1}, \quad \epsilon_{i1} \sim N(0, 75) \tag{41}$$

$$x_{i2} = 0.000001i^3 + \epsilon_{i2}, \quad \epsilon_{i2} \sim N(0, 10) \tag{42}$$

for $i = 1, \ldots, 1200$. This results in the covariates having a cubic relationship. Each unit is assigned a binary treatment indicator T_i using

$$p(T_i = 1) = 0.5I(i > 800), \quad i = 1, \ldots, 1200. \tag{43}$$

To introduce missing values in x_2, 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 logistic regression,

$$logit\big(P(m_i = 1)\big) = -3 + 0.005x_{i1}. \tag{44}$$

We simulated the covariates described in Figure 3 as follows. We simulate covariates for 400 units using

$$(x_{i1}, x_{i2})' \sim N(\mu, \Sigma), \tag{45}$$

where $\mu = (140, 60)'$ and Σ is such that both x_{i1} and x_{i2} have a variance of 5 with correlation 0.6. Of these units, 100 are randomly allocated to treatment. We simulate covariates for 995 units from a multivariate normal distribution with $\mu = (132, 52)'$ and Σ such that both x_{i1} and x_{i2} have a variance of 20 with correlation 0.6. Finally we simulate covariates for five outlying units from a multivariate normal distribution with $\mu = (20, 52)'$ and Σ is such that both x_{i1} and x_{i2} have a variance of 5 and are uncorrelated. To introduce missing values in x_2, we assign missing data indicators m_i

using the logistic regression,

$$logit\big(P(m_i = 1)\big) = -34 + 0.25x_{i1}. \tag{46}$$

We simulated the covariates described in Figure 7 as follows. For $i = 1, \ldots, 1200$, we have

$$\begin{aligned} x_{1i} &= 50 + 0.8i + \epsilon_{1i}, \quad \epsilon_{1i} \sim N(0, 75) && (47) \\ x_{2i} &= 50 + 0.8i + \epsilon_{2i}, \quad \epsilon_{2i} \sim N(0, 10). && (48) \end{aligned}$$

This results in the covariates having a linear relationship. Each unit is assigned a binary treatment indicator T_i using

$$p(T_i = 1) = 0.5I(i > 800), \quad i = 1, \ldots, 1200. \tag{49}$$

To introduce missing values in x_2 we assign missing data indicators m_i for each unit i using the logistic regression,

$$logit\big(P(m_i = 1)\big) = -3 + 0.005x_{i1}. \tag{50}$$

References

Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal gaussian data. *Biometrics* **64**, 96–105.

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.

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, 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, R. B. J. 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**.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 2, 363–375.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

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.

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.

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.

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.

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

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.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Wasserman, L. (2000). Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 1, 159–180.

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.