ABSTRACT

Title of Document: DIAGNOSTICS FOR MULTIPLE IMPUTATION BASED ON THE PROPENSITY SCORE.

Jia Wang, MPH, 2010

Directed By: Assistant Professor Guangyu Zhang, Department of Epidemiology and Biostatistics

Abstract: Multiple imputation (MI) is a popular approach to handling missing data, however, there has been limited work on diagnostics of imputation results. We propose two diagnostic techniques for imputations based on the propensity score (1) compare the conditional distributions of observed and imputed values given the propensity score; (2) fit regression models of the imputed data as a function of the propensity score and the missing indicator. Simulation results show these diagnostic methods can identify the problems relating to the imputations given the missing at random assumption. We use 2002 US Natality public-use data to illustrate our method, where missing values in gestational age and in covariates are imputed using Sequential Regression Multiple Imputation method.

DIAGNOSTICS FOR MULTIPLE IMPUTATION BASED ON THE PROPENSITY SCORE

By

Jia Wang

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Public Health
2010

Advisory Committee:
Assistant Professor Guangyu Zhang, Chair
Assistant Professor Tongtong Wu
Assistant Professor Xin He

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Missing data is a ubiquitous problem in the analysis of survey data. Missing data for individual variables can occur due to nonresponse for sensitive or difficult items (e.g. income measures), mistakes in responding to survey questions (e.g. incorrect skips) or nonresponse to complete phases of a multi-phase survey (e.g. refusal of medical examination in NHANES). Two potential problems with the analysis of incomplete data are: (1) loss of information or power due to missing data; and (2) potential bias due to systematic differences between observed data and the unobserved data (Barnard and Meng, 1999).

1.1 Missing Data Mechanisms

There are three types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR occurs if the probability of missingness is the same for all units and missing occurs completely at random. In other words, a missing response is independent of both observed and missing values (Rubin 1976; Little and Rubin, 2002). A second condition is MAR, where missingness depends on only the observed characteristics of a participant, but not on the missing values themselves (Rubin 1976; Little and Rubin, 2002). Lastly, MNAR mechanism implies that missingness is related to the unobserved values of the variables with missing data. In such situation, the probability of missingness varies and cannot be characterized by available predictors (Rubin 1976; Little and Rubin, 2002).

1.2 Existing Approaches to Missing Data

There are several approaches to handling missing data. The most simple and convenient method is complete case analysis, by which only individuals with complete information on all variables are included in the statistical analysis. Available case analysis (pairwise deletion) includes all available data under analysis, instead of removing the entire cases that have missing values on any of the variables. However, the inference may base on different subjects for different estimators. The main drawback of these two methods is that MCAR assumption must be held. Otherwise, they may lead to biased results.

Maximum likelihood estimation method obtains the maximum likelihood parameter estimates of interest by maximizing the observed data likelihood given that the missing values are MAR. The disadvantages of this method are that it requires fairly sophisticated computations and they are specific to the model being applied.

Imputation procedures are techniques for assigning plausible values to missing data. Imputation techniques range from the simplest mean imputation to multiple imputation (MI), all of which produce a completed data set that can be analyzed using standard complete data software procedures. Furthermore, unlike maximum likelihood estimation that is problem-specific and may require totally different and complicated computational procedures to for different models, using the same imputation approach to handling missing data on public-use datasets provides consistency across different scientific questions (Parker and Schenker,2007).

Single imputation methods, including mean imputation, regression imputation, hot-deck imputation, stochastic imputation, can be viewed as precursors of multiple imputation. For single imputation, only one plausible value is imputed for each missing observation.

The main disadvantage of this method is that the imputed values are treated as if they were true values, so that it fails to account for the added uncertainty due to the assignment of a plausible, yet not actual, value for each missing value. Therefore, the parameter estimate and variance may be underestimated.

1.3. Multiple Imputation

The ideas of multiple imputation for missing data were first proposed by Donald Rubin in 1977 and now a variety of statistical software packages have capabilities to conduct MI. Multiple imputation method meets two requirements to develop accurate parameter estimates and variances: (1) the imputation should be model-based in a way that the distributions of the variables and the relationships among the variables can be captured; (2) the imputation method should account for the uncertainty in the imputed values. The multiple repetitions of imputation procedures enable the estimation of variance that is added due to imputing missing values in the data set.

Multiple imputation, an extension of the single imputation method, comprises three steps: (1) each missing value is replaced by a set of K>1 plausible values to generate K complete data sets (Sinharay and Russell, 2001). The critical component of this step is the imputation model selection, which is defined by a set of variables available to the imputation process and the distributional assumptions; (2) each of K complete data sets are then analyzed using standard statistical analyses. The results are K point estimates and their corresponding estimated variances; (3) the results from the K completed data sets are combined to create parameter estimates and standard errors. Estimates of population parameter are computed using an average of the parameter estimates of l=1, …, K completed data set from step 2

$$\bar{\theta} = \frac{1}{K} \sum_{l=1}^{K} \hat{\theta}_l ,$$

where $\hat{\theta}_l$=estimate of θ from the completed data set l=1, … , K.

The corresponding variance for $\bar{\theta}$ is estimated by a simple combination of the average of the K variance estimates and the variance of the K point estimates.

$$\text{var}(\bar{\theta}) = \bar{U} + \left(\frac{K+1}{K}\right) \times B,$$

where $\bar{U}$ = within-imputation variance= $\frac{1}{K}\sum_{l=1}^{K} \text{var}(\hat{\theta}_l)$,

B=between-imputation variance=$\frac{1}{K-1}\sum_{l=1}^{K}(\hat{\theta}_l - \bar{\theta})$.

Rubin (1987) showed that the efficiency of an estimate in MI analysis is approximately

$$(1 + \frac{\gamma}{K})^{-1} ,$$

where γ is the fraction of missing values and K is the number of multiple repetitions of the imputation process. For example, consider a dataset with 25% missing values, K=5 imputations gives 95.2% efficiency. Virtually all of the desirable efficiency can be achieved by using K=5 to K=10 independent repetitions of the imputation process.

The success of MI depends on two required assumptions. First, an important step in generating multiple imputations is to assume an imputation model, which is defined by a set of variables and the distributional assumption. The selection of variables to include in the imputation model directly affects the quality of imputations. A general rule of thumb is to incorporate as many as possible the available data and the possible variables correlated with the analysis variables, and at the same time keep the model building and fitting feasible (Barnard and Meng, 1999). Usually, the set of variables included in the imputation model for an MI analysis is much larger and broader in scope than the variables required for the analytic model. Failure to include one or more variables in the

imputation model can yield less accurate imputed values. Except for the selection of variables, in order to generate imputations, one must assume a probability model on the complete data. This multivariate model must preserve the associations among the many variables included in the imputation model. A variety of algorithms like Markov Chain Monte Carlo (MCMC) method, or Sequence Regression Model, can be used to generate the imputations.

Second, MI assumes that the missing data are missing at random (MAR), that is, the probability that an observation is missing only depends on the observed values, but not on the missing value (Rubin, 1976). Let Y be a data matrix, Y_{mis} be the missing part of Y and Y_{obs} be the observed part of Y. Suppose M is a missing data indicator matrix of the same dimension of Y, where the elements are zero or one depending on whether the corresponding elements of Y are observed or missing. MAR implies that $P(M|Y) = P(M|Y_{obs})$. In principle, it is impossible to test the assumption of MAR without additional data collection, since information that would be used to make such a test is unavailable (Abayomi, Gelman and Levy 2008). Therefore, due to the belief that imputed values are merely the guesses of the unobserved data, which are unknown, few attempts have been made to check the quality of imputed data.

We propose a diagnostic method based on the propensity score to check the quality of multiple imputations described in Section 2, and conduct a simulation study to show how this diagnostic method can serve as a reliable method for assessing the problems relating to the imputation model given the assumption of MAR in Section 3. Then we apply this diagnostics method to imputations of gestational age in 2002 US Natality public-use data in Section 4. We conclude the thesis in Section 5.

Chapter 2: Diagnostic Method for Multiple Imputation

2.1 Existing Diagnostics for Multiple Imputation

Abayomi, Gelman and Levy (2008) developed diagnostics for random imputations, based on two arguments: (a) imputations can be checked by using a standard of reasonability: the differences between observed and missing values and the distribution of the completed data as a whole can be checked to see whether they make sense in the context of the problem being studied; (b) imputations are typically generated by using models that are fitted to observed data, and the fit of these models can be checked.

They first checked if there were unusual patterns that might suggest problems with imputation (e.g. the histogram of the completed data of a variable was bimodal because the imputed data markedly differed from the observed data). Next, they compared and flagged the difference between the distributions of observed and imputed data values. Finally, they checked the fit of the observed data to the imputation model that was used to create the imputations (check whether the pattern of residuals versus expected values was random). The non-random pattern of residual plots may flag the problems in terms of the violation of the missingness assumptions, and thus the imputation model.

One limitation of their methods is that it only works well if there are dramatic differences between the imputed and observed data. However, differences in distributions do not necessarily suggest a problem with the imputations or a violation of MAR, because such differences might be explained by other variables in the data set. We will illustrate this issue in detail in the simulation study.

2.2 Diagnostics Based on the Propensity Score

Propensity score is a conditional probability of assignment to a particular group given a vector of observed covariates (Rubin, 1983). The key attribute of the propensity score is that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates (Rubin, 1983). Therefore, the basic idea of our diagnostic method is that for multiple imputation assuming MAR, the observed data and the imputed data have the same conditional distribution given on the propensity score. If the two distributions differ, it suggests that the imputation results are questionable.

Let $(Y, X_1, X_2, \ldots, X_p)$ be a vector of variables with Y having missing values and $X_1, X_2, \ldots, X_p$ fully observed variables. Let m denote a missing indicator with m=1 when Y is missing and m=0 when Y is observed. The propensity score, or the probability of missingness, is denoted as P.

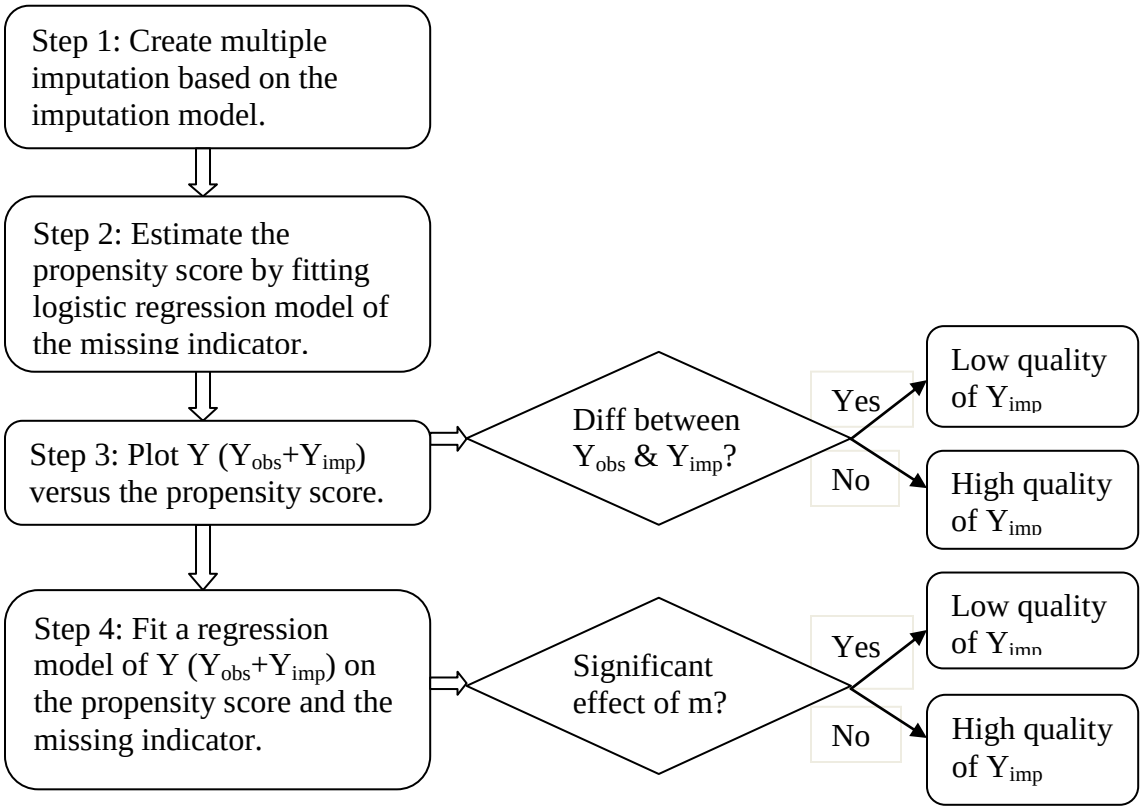
$$\text{logit}(P) = \text{logit}(P(m=1) \mid X_1, X_2, \ldots, X_p))$$

The missingness of Y only depends on $X_1, X_2, \ldots, X_p$. Thus, the missing data mechanism is MAR. We construct a set of imputations by using multiple imputation procedure (PROC MI in SAS 9.1), and then apply the diagnostic method to the imputed data sets. The estimates of the propensity score can be obtained by fitting a logistic regression model of m on $X_1, X_2, \ldots, X_p$, yielding the predicted values of m.

The first diagnostics is to compare the distributions of the imputed and observed values against the propensity score. We look for differences in the conditional distributions, which suggest the inaccuracy of the imputations because the missingness of Y is not a random sample of the original data given on the same propensity score.

We then fit regression models of Y (Y_{obs}+Y_{imp}) as a function of the propensity score (P) and the missing indicator (m). An insignificant association between m and Y implies that the missingness of Y is independent of the values of this variable after adjusting for the propensity score. If the missingness can completely be explained by the propensity score, it indicates the MAR assumption holds true and the imputation model used to generate imputations enables to preserve the associations among all available variables in the data set.

The following is the flowchart of our diagnostic method.



Chapter 3: Simulation Study

We illustrate our method by a simulation study. We generate a dataset with 500 subjects. Table 1 and 2 show the models we use to generate the data, to create the missing values, to estimate propensity score, and to impute missing data. The following procedures describe how we conduct the simulation with data set 1 as an example.

(1) Generate a dataset with a sample size of 500 and three variables (Y, X_1, X_2) based on the following model:

$Y = 1 + X_1 + 2 * X_2 + \varepsilon$, where X_1, X_2, and ε all follow standard normal distribution with mean 0 and variance 1. In addition, create another ten variables X_3, X_4,…, X_{12} in the data set, all of which follow either a normal or a uniform distribution.

(2) Generate missing values of Y from the response propensity model:

logit (P (m=1|X_1, X_2)) = $\gamma_0 + \gamma_1 * X_1 + \gamma_2 * X_2$. The different percentage of missing values depends on the arbitrary assignments of parameters γ_0, γ_1, and γ_2. The missingness of Y only depends on the values of X_1 and X_2, thus, the missing mechanism is MAR.

(3) Impute missing values and develop 5 completed data sets by fitting a correct model of Y given X_1 and X_2 using PROC MI procedure in SAS9.1, denoted as [X_1, X_2].

(4) Impute missing values and develop 5 new completed data sets by using an overfitted model of Y given X_1, X_2 ,…, X_{12}, denoted as [X_1, X_2, …, X_{12}]

(5) Impute missing values and develop another 5 completed data sets by fitting an incorrect model of Y only given X_1, denoted as $[X_1]$.

(6) Estimate two sets of propensity scores by fitting a correct logistic regression model of m given X_1 and X_2 and by fitting a overfitted regression model of m given X_1, X_2 ,…, X_{12}.

(7) Plot the true propensity scores versus two sets of estimated propensity scores from Step 6 and plot the estimated propensity scores from the correct model against the ones from the overfitted model.

(8) Plot three sets of Ys from Step 3-5 against the estimated propensity scores from the correct model (red: Y_{obs}; blue: Y_{imp}). Compare the distribution of the imputed Ys and observed Ys. In addition, true values of Ys (from step 1) versus the propensity score are plotted as well.

(9) Plot the observed and imputed Ys from step 3 and 5 against one of covariates (X_1). Then plot the histograms of Y to compare the distributions of Y at two levels of missingness as Abayomi et al. (2008) did.

(10) Fit a linear regression model of Y from step 3, 4 and 5 on the estimated propensity score and m. The results from the 5 completed data sets are combined to create parameter estimates and standard errors using PROC MIANALYZE procedure in SAS 9.1.

Diagnostic methods are applied not only across percentage of missingness (from low to high), but also across three different mean functions of Y. Repeat step 2 to step 10, except changing the propensity model, correct imputation model, overfitted and incorrect imputation model correspondingly as shown in Table 1 and 2.

The simulation study compares the conditional distributions of the observed and imputed Ys given the propensity score when the imputations are created from correct, overfitted or incorrect imputation model. Theoretically the similarities of the patterns between the observed and imputed Ys suggest the high quality of imputations. Furthermore, fitting regression models of Y (created by correct, overfitted or incorrect model) on the propensity score and m will quantitatively check the effectiveness of our diagnostic method.

The purpose of using overfitted imputation model to generate imputations is to test the robustness of our diagnostic method. Practically, there is no way to know the correct imputation model, a determinant factor affecting the quality of imputations. Overfitted imputation model can be a common situation in real imputation process, because by following the general rule we usually incorporate as many as possible the variables that might be associated with Y. Estimating the propensity score by fitting the overfitted model is also due to the fact that the correct propensity model is always unknown in practice. Therefore, the scatter plots of propensity scores and a regression model of Y on the propensity score from the overfitted model and m are used to test the reliability of our diagnostic method.

The scatter plots of the true propensity score versus the estimated propensity score from the correct propensity model and the overfitted propensity model are shown in Figure 1 and Figure 2 respectively, blue points when Ys are missing and red when Ys are observed. Figure 3 shows the scatter plots of the propensity score from the correct model versus the ones from the overfitted model. In each scatter plot, although there is a larger variation among the points in Figure 2 and 3 because of the noise in the overfitted model,

all of the points almost lie on a 45-degree straight line. It suggests that the estimated propensity scores either from the correct or the overfitted model are reliable to reflect the true probability of missingness of Y.

The results from this simulation study support our statement that the graphical diagnostics proposed by Abayomi et al. (2008) has its own limitation. Figure 5 shows the histograms of the completed data of Y (from Step 3 and 5) at two levels of missingness. These histograms illustrate that the distributions of Y can be different between observed and imputed values. The distributions of observed and imputed Ys against one of the covariates in the imputation model are plotted in Figure 4. In these scatter plots, some deviations between observed and imputed Ys do exist under MAR. Such differences can come from the effects of other variables on the missingness of Y in the data set. For example, in Figure 4 (when $\mu_y = 1+X_1+2*X_2$), due to the effect of X_1, the larger values of Y are more likely to be missing than the smaller values of Y. Therefore, such differences between observed and imputed data can not necessarily flag the potential problems relating to the imputation model.

Our graphical diagnostics can avoid this problem by adjusting for the propensity score. The conditional distributions of Y given the propensity score can actually remove the overall effects of the covariates in the data set. Figure 6 plots three sets of Ys versus the estimated propensity score. These bivariate scatter plots, including the smooth curves, present the comparisons of conditional distributions of the imputed and the observed Ys given the propensity score. Observed data are shown in red and imputed data in blue. Our diagnostics can be applied to each of 5 completed data sets. In this paper, we only show the plots for a single randomly chosen completed data set. There are obvious differences

between the distributions of the observed and imputed Ys conditional on the propensity score, when the missing values are imputed using incorrect imputation model (Figure 6(d)). In contrast, when the correct (Figure 6(a)) or overfitted imputation models (Figure 6(c)) are used to generate MIs, there are only slight deviations between two curves. Additionally, these patterns of distribution are similar to the true distributions when we plot the true Ys against the propensity score (Figure 6(e), blue: m=1; red: m=0). When the estimated propensity scores from the overfitted model are applied in the scatter plot (Figure 6(b)), they show the similar results.

Results of fitting linear regression model of completed Y on propensity score and m are shown in Table 3. When the correct or overfitted imputation model is used to impute missing values, from low (16%) to high (64%) percentage of missing values across three mean functions of Y and no matter if the propensity scores are estimated from correct or overfitted models, the effect of the missing indicator is insignificant, while the propensity score has a significant association with Y. However, when the incorrect imputation model is implemented to create imputations, in most of cases in this simulation study, both the propensity score and the missing indicator have significant effects on the values of Y. This simulation study is empirical evidence that our graphical diagnostic approach to checking the imputation model is robust. It can be functioned as indirect method to identify potential problems relating to the imputation model. Obvious deviations between distributions of observed and imputed values conditional on the propensity score do occur when the imputation model that is used to generate imputations fails to preserve the associations of all important variables. Then this model would be flagged because of the marked differences.

In addition to the graphic presentation of observed and imputed data, we fit a regression model of the completed Y as a function of the propensity score and the missing indicator. The results suggest that we can assess the goodness of imputations by examining the relationship of m with Y. Given the assumption of MAR, the statistically significant effect of m on Y, after adjusting for the propensity score, indicates a deficiency in the imputation model, which fails to preserve all the associations among the variables with the dependent variable. Thus the model underlying the inaccurate imputations should be suspected.

The results from Figure 6 (c) and Table 3 confirm the notion that the inclusion of as many as possible the variables in the imputation model can improve the imputations, even though it might be overfitted. When missing values are imputed by using the overfitted model, there is no significant difference in the conditional distribution given propensity score between observed and imputed Y. Moreover, the insignificant effect of m indicates the sufficiency of imputation model to capture the associations among all variables.

This simulation study illustrates where and how our diagnostics can serve as effective method for assessing the imputation model that is used to generate the imputed data given the assumption of MAR. In both steps of diagnostic procedures, the graphical display and statistical analysis based on the propensity score can flag the inaccurate imputation model.

Chapter 4: Application to 2002 US Natality Public-Use Data

4.1 Data Source: the 2002 Natality Public-Use Data

We apply our method to 2002 US Natality public-use dataset produced by the National Center for Health Statistics (NCHS). The NCHS collects Natality data from Standard Certificate of Live Birth for all living births in the United States every year and releases them to the public. The 1989 version of US Standard Certificate of Live Birth provides a wide variety of information on maternal and infant health characteristics, including information on general items, occurrence, residence, prenatal, child, mother, pregnancy history, father, medical and health data (NCHS, 2002). The 2002 public-use Natality data consists of 4,027,376 live births within the United States to residents and non-residents. Our study sample includes a subset of 2002 US Natality data. We randomly select 40,274 newborns, 19,730 females (48.99%) and 20,544 males (51.01%).

4.2 Variable of Interests: Gestational Age (DGESTAT)

The high incidence rate and consequences of preterm births make it necessary to correctly determine the important factors that affect preterm delivery in order to establish guidelines for monitoring and treatment plans for expectant mothers who are most susceptible to preterm labor (Hammad, 2009). However, missing data and inaccurate information on gestational age have affected the utility of the US Natality public-used datasets (Parker and Schenker, 2007).

The period of gestation is defined as beginning with the first day of the last normal menstrual period (LMP) and ending with the day of the birth (NCHS, 2002). In 2002 Natality file, gestational age information contains four parts: (a) computed using date of

birth of child and last normal menses; (b) imputed from LMP date; (c) the clinical estimate; or (d) unknown when there is insufficient data to impute or no valid clinical estimate (NCHS, 2002). The primary measure (Part (a)) used to determine the gestational age of the newborn is the interval between the first day of the mother's LMP and the date of birth. It is subject to error due to reasons including imperfect maternal recall or misidentification of the LMP because of post conception bleeding, delayed ovulation, or intervening early miscarriage (Martin, et al., 2003). The clinical estimate is used in three situations: (1) if the LMP date is not reported; (2) when the computed gestation is outside the 17-47 code range; (3) normal weight births come with apparently short gestations and very-low-birth weight births reported to be full term. There are 4.6 percent of the births in 2002 Natality data based on the clinical estimate of gestation. The NCHS also publishes the imputed weeks of gestation for records with missing day of LMP when there is a valid month and year. Although LMP-based gestational ages are edited for obvious inconsistence with the infant's plurality and birth weight, reporting problems for this item persist and may occur more frequently among some subpopulations and among births with shorter gestations (Alxandra & Allen, 1996). Some research is ongoing to address these data deficiencies.

In order to avoid dealing with the intricacies of misspecified gestational ages, we set gestational age (DGESTAT in the data set) to missing if computed gestation is different from its clinical estimate by more than 2 weeks or they are replaced with the clinical estimations or the imputed gestational age created by the NCHS. After these alterations, there are 18.69% missing values for DGESTAT among all subjects in the final dataset used in the analysis.

4.3 Multiple Imputation using Sequential Regressions

2002 US Natality public-use data set consists of 213 variables, including the recoded ones. They have many types of variables, such as continuous (birth weight, age of mother, etc.), categorical (race of mother, marital status, etc.), count (number of prenatal visits), or mixed variables (number of cigars per day, number of drinks per week). Some of these variables have small percentages of missing values which need to be imputed as well. In addition, there are certain reasonable bounds for specific variables with missing values, which must be incorporated in the imputation process. For example, the imputed values for "Age of Mother" must be greater than 10 and less than 54 and imputations for "Age of Father" must be greater than 10.

Because of the complex data structure of US Natality public-use file, we choose sequential regression multiple imputation (SRMI) method by using publicly available software (IVEware, available at http://www.isr.umich.edu/src/smp/ive) to handle both missing and implausible gestational ages, as well as missing values in the covariates. The basic strategy of SRMI is to create imputations through a sequence of multiple regressions on a variable by variable basis, varying the type of regression model by the type of variable being imputed. Covariates include all other variables observed or imputed for that individual (Raghunathan, et al., 2001). SRMI imputes the least missing variables before the most missing at each round of the procedure and then continued in a cyclical manner, each time overwriting previously drawn values, building interdependence among imputed values and exploiting the correlational structure among covariates (Raghunathan, et al., 2001).

The rule for the selection of variables in the imputation model is to include as many as possible the variables that are possibly correlated with the period of gestation. Therefore, the imputation model in our study includes variables from all 10 categories regarding the newborn and maternal characteristics mentioned above. Some of the variables are summed into one category to be used in the imputation model. These summary variables include: the total number of medical risk factors (MEDRK), the total number of obstetric procedures (OBSTET), the total number of complications of labor or delivery (LABOR), the total number of abnormal conditions of the newborn (NEWBN), and the total number of congenital anomalies (CONGN). Descriptive statistics for all variables included the imputation model are listed in Table 6 (categorical variables) and Table 7 (continuous, count and mixed variables). Three variables (Number of Live Birth, Now Living; Number of Live Birth, Now Dead; Number of Other Termination) and five summary variables are classified as categorical variables, because high percentages of value 0 of these variables can lead to unstable results in the SRMI procedures if they are treated as continuous or count variables.

The Pearson correlation for continuous variables and Spearman correlation for five discrete variables (CORR procedure in SAS 9.1) are used to check the possible colinearity. Correlation matrixes are shown in Table 4 and Table 5 respectively. As can be seen is Table 5, two pairs of variables (Number of Live Births, Now Living (NLBNL) vs. Detailed Total Birth Order (DTOTORD), Number of Live Births, Now Living (NLBNL) vs. Detailed Live Birth Order (DLIVORD) are highly correlated ($|r|=0.876$, 0.994), which imply that there is possible colinearity between two variables. Thus, both DTOTRORD and DLIVORD are excluded from the imputation model.

4.4 Diagnostic Procedures for Imputations of Gestational Age

We created M=5 SRMIs, repeating the process with 10 iterations (seed=2010). We assume MAR, and estimate the propensity scores by fitting a logistic regression of the missing indicator on all variables in the imputation model without any missing values. The variables used in the propensity model are presented in Table 8.

We apply two steps of our diagnostic method to 5 sets of SRMIs as follows:

(1) Plot the gestational age (DGESTAT) vs. propensity scores (red: observed values, blue: imputed values).

(2) Fit a linear regression model of gestational age (DGESTAT) on the propensity score (P) and the missing indicator (m) for 5 imputed data sets. Then the results from 5 data sets are combined to create one parameter estimates and corresponding standard errors by using PROC MIANALYZE in SAS 9.1 as the methods described earlier.

Figure 7 provides a snapshot of the distributions of the observed and imputed gestational age against the propensity score. There is no significant difference between red and blue curves and the patterns of two sets of points are quite similar, although the variation of imputed values is slightly smaller than that of observed values.

Results from the regression model are summarized in Table 9. All p-values of m are greater than 0.05, in other words, m has insignificant effect on the values of gestational age after adjusting for the propensity score. It implies that the imputation model we created enables to preserve the associations among all variables with gestational age given MAR and, thus, the missingness of DGESTAT can totally explained by the propensity score. By applying two steps of checking procedures, we can conclude that the

imputations under our imputation model can sufficiently reflect the true distribution of gestational age for those newborns with missing values.

Chapter 5: Discussion

Very little attention has been given to the development of diagnostic techniques for multiple imputation (Abayomi, Gelman & Levy, 2008). The aim of this research is to develop diagnostic method based on the propensity score to identify potential problems with the imputations. We propose two steps of diagnostic method for imputations: (1) comparisons of the distributions of observed and imputed data against propensity scores, which are used to reveal differences between the observed and imputed data; and (2) fitting regression model of completed data on the propensity score and the missing indicator. In addition, we apply our method to the 2002 US public-use Natality data published by the NCHS.

In simulation study, when the missing values are imputed by using incorrect imputation model, there are apparent differences between the conditional distributions of the observed and imputed Ys given the propensity score, and a significant association is found between the values of Y and the missing indicator ($P<0.05$). In contrast, when the correct or overfitted imputation model is used to generate MIs, the distributions of the observed and imputed data conditional on the propensity score are similar, and the values of Y are independent of m ($P>0.05$). These results suggest we can flag potential problems with the underlying imputation model that is used to create imputations. Additionally, the propensity scores estimated from the correct or overfitted propensity model are proved to be reliable and will not affect the diagnostic results.

A recent study conducted by Abayomi, et al. (2008) considered diagnostics for imputations in three steps as described in the introduction. Our simulation study confirms

the limitation of their graphical methods. They can work well only for the extreme departures between observed and imputed values. However, as shown in Figure 4 and Figure 5, deviations can be expected under MAR and they do not necessarily indicate problems with the imputation model. Such deviation is due to the effect of other variables in the dataset on the probability of missingness.

The key property of our diagnostic method is that the adjustment for the propensity score is sufficient to remove the effects of all other covariates that contribute to the probability of missingness. Therefore, assuming MAR, our graphical display conditioning on the propensity score is more robust than the marginal distribution of the completed data or the conditional distribution given only one variable as described in Abayomi's research. In application to 2002 US Natality file, the results show the similarities of the conditional distributions given propensity score between observed and imputed gestational age and the insignificant effect of the missing indicator on gestational age. All of the results suggest the high quality of the imputations we create, that is, the missingness of gestational age can be totally explained by the propensity score.

In Natality file, gestation age is subject to two problems: missing data and implausible data. Therefore, the imputation for US Natality data is complicated by the uncertainty about which records need to be imputed due to implausible values (Parker and Schenker, 2007). We simplified this issue by setting the records with over two-week difference between computed and clinical estimate as missing data. Attempts have been made by Parker and Schenker (2007) to use multiple imputation technique for imputing missing and implausible gestational age values. Multiple imputation is an appropriate technique to handle missing data, which takes into account both the relationships among the variables

and the uncertainty added from the imputation, thus it can yield more valid statistical results relating to gestational age in future analytical studies. We use SRMI, which is an extension of MI in which the missing values of each variable are imputed conditionally on all the other variables in the data set and the types of regression models used depend on the type of variable being imputed. Moreover, it can incorporate restrictions to a relevant subpopulation for some variables and logical bounds for the imputed data. In addition to the imputation techniques, to improve the quality of imputations, our imputation model includes variables from 10 categories with respects to both the newborn and the maternal characteristics. Because of these efforts, our diagnostic method identifies the imputations we create with high quality.

The findings in this study contribute to the ongoing search to identify reasonable and reliable diagnostic techniques to check the quality of multiple imputation. An important assumption of these diagnostics is the missing at random. Nevertheless, in this study, the MAR assumption cannot be approved. Another limitation in this study is the nonlinear relationship between the values of the dependent variable and the propensity score. The scatter plots in Figure 6 show the curvilinear, rather than linear, relationship between Y and the propensity score. The future research can extend our method by using smoothing spline to model the relationship between Y and the propensity score. Furthermore, a quantitative test can be employed to numerically compare the conditional distribution of the observed and the imputed data given the propensity score.

Appendix

Table 1: Summary of mean function of Y, correct, overfitted and incorrect imputation model.

Mean Function μ_y =	Imputation Model, Y=		
	Correct	Overfitted	Incorrect
$1+X_1+2*X_2$	$\beta_0+\beta_1*X_1+\beta_2*X_2$ $[X_1,X_2]$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_{12}*X_{12}$ $[X_1,X_2,\ldots,X_{12}]$	$\beta_0+\beta_1*X_1$ $[X_1]$
$1+2*X_1+2*X_2+3*X_1X_2$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_1X_2$ $[X_1,X_2,X_1X_2]$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_{12}*X_{12}+\beta_{13}*X_1X_2$ $[X_1,X_2,\ldots,X_{12},X_1X_2]$	$\beta_0+\beta_1*X_1+\beta_2*X_2$ $[X_1,X_2]$
$1+2*X_1+2*X_2+3*X_1^2$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_1^2$ $[X_1,X_2,X_1^2]$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_{12}*X_{12}+\beta_{13}*X_1^2$ $[X_1,X_2,\ldots,X_{12},X_1^2]$	$\beta_0+\beta_1*X_1+\beta_2*X_2$ $[X_1,X_2]$

Table 2: Summary of mean function of Y, true propensity score function, percentage of missing of Y, correct and overfitted propensity model.

Mean Function μ_y =	True Propensity Function logit (P)=	M (%)	Propensity Model logit (P)=logit (m=1|X_1, X_2,…, X_p)	
			Correct	Overfitted
$1+X_1+2*X_2$	$-2+X_1+X_2$	15.8	$\beta_0+\beta_1*X_1+\beta_2*X_2$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_{12}*X_{12}$
	$-2+3*X_1+X_2$	26.0		
	X_1	48.6		
	$1+X_1+X_2$	64.0		
$1+2*X_1+2*X_2+3*X_1X_2$	$-2+2*X_1+2*X_2+3*X_1X_2$	23.6	$\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_1X_2$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_{12}*X_{12}+\beta_{13}*X_1X_2$
	$1+3*X_1-3*X_2-4*X_1X_2$	54.2		
	$1+2*X_1+3*X_2-4*X_1X_2$	63.6		
$1+2*X_1+2*X_2+3*X_1^2$	$-4+3*X_1+X_2+X_1^2$	16.2	$\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_1^2$	$\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_{12}*X_{12}+\beta_{13}*X_1^2$
	$-1+5*X_1+1*X_2-2*X_1^2$	33.8		
	$2+X_1+X_2-2*X_1^2$	59.0		

Table 3: Regression of Y (Y_{obs}+Y_{imp}) on the propensity score (p-score) and the missing indicator (m): three imputation models are fitted to impute missing data, and two propensity models are used to estimate p-scores.

			Imputation Model						Propensity Model	
			Model 1*: Correct		Model 2*: Overfitted		Model 3*: Incorrect		Model 4*: Overfitted	
Mean Function	M%	Variable	β	p-value	β	p-value	β	p-value	β	p-value
$Y=1+X_1+2*X_2+\varepsilon$	15.8%	p-score	11.76	<0.0001	11.71	<0.0001	9.57	<0.0001	10.42	<0.0001
		m	0.07	0.7456	0.10	0.6972	-1.33	0.0005	0.13	0.7153
	26.0%	p-score	4.56	<0.0001	4.52	<0.0001	3.68	<0.0001	4.23	<0.0001
		m	-0.09	0.7828	0.01	0.9738	-1.01	0.0050	0.07	0.8533
	48.6%	p-score	5.43	<0.0001	5.32	<0.0001	4.44	0.0007	4.78	<0.0001
		m	-0.02	0.9480	0.08	0.7777	0.10	0.7643	-0.05	0.8258
	64.0%	p-score	7.72	<0.0001	7.95	<0.0001	3.59	<0.0001	7.25	<0.0001
		m	0.04	0.8304	0.09	0.7344	-0.83	0.0101	-0.04	0.8343
$Y=1+2*X_1+2*X_2+3*X_1*X_2+\varepsilon$	23.6%	p-score	10.31	<0.0001	10.27	<0.0001	5.59	<0.0001	10.01	<0.0001
		m	0.15	0.7518	0.04	0.9363	-2.34	0.0044	0.15	0.7520
	54.2%	p-score	-4.35	<0.0001	-4.33	<0.0001	1.20	0.4018	-4.24	<0.0001
		m	-0.07	0.9005	0.05	0.9401	1.78	0.0279	0.00	0.9993
	63.6%	p-score	0.73	0.3903	0.74	0.3909	2.53	0.0102	0.81	0.3583
		m	0.05	0.9434	0.13	0.8560	1.84	0.0818	-0.06	0.9381
$Y=1+2*X_1+2*X_2+3*X_1^2+\varepsilon$	16.2%	p-score	13.54	<0.0001	13.31	<0.0001	2.58	0.0095	13.25	<0.0001
		m	-0.18	0.8152	-0.29	0.7128	-3.16	0.0001	-0.14	0.8564
	33.8%	p-score	5.33	<0.0001	5.25	<0.0001	3.92	0.0018	5.13	<0.0001
		m	0.08	0.9137	0.11	0.8747	0.09	0.9578	0.06	0.9308
	59.0%	p-score	-5.74	<0.0001	-5.84	<0.0001	-2.89	0.0640	-5.55	<0.0001
		m	-0.01	0.9844	-0.05	0.9347	3.57	<0.0001	0.05	0.9417

Model 1: Imputations are created by fitting correct imputation model. Propensity score is estimated by fitting correct model
Model 2: Imputations are created by fitting overfitted imputation model. Propensity score is estimated by fitting correct model
Model 3: Imputations are created by fitting incorrect imputation model. Propensity score is estimated by fitting correct model
Model 4: Propensity score is estimated by fitting overfitted propensity model. Imputations are created by fitting correct imputation model.

Table 4: Correlation matrix among all continuous variables.

	BIRWT	DMAGE	DFAGE	FMAPS	WTGAIN	NPREVIS	CIGAR	DRINK	MEDRK	NEWBN	LABOR	OBSTET	CONGN
BIRWT	1.000	**0.068	**0.040	**0.278	**0.175	**0.107	**-0.092	-0.007	**-0.101	**-0.163	**-0.074	**0.029	**-0.046
DMAGE		1.000	**0.758	0.003	**-0.072	**0.107	**-0.062	**0.020	**0.052	0.007	**0.013	*0.012	-0.006
DFAGE			1.000	-0.009	**-0.063	**0.062	**-0.042	**0.024	**0.042	0.006	0.009	0.005	0.000
FMAPS				1.000	**0.026	**0.067	-0.009	0.005	**-0.062	**-0.231	**-0.131	-0.010	**-0.104
WTGAIN					1.000	**0.092	**-0.022	0.008	**-0.015	**-0.016	**0.034	**0.036	**-0.018
NPREVIS						1.000	**-0.044	**-0.041	-0.003	**-0.048	**-0.025	**0.032	**-0.023
CIGAR							1.000	**0.090	**0.041	**0.017	**0.016	*0.012	0.004
DRINK								1.000	0.004	0.005	0.004	-0.001	-0.001
MEDRK									1.000	**0.181	**0.176	**0.196	**0.042
NEWBN										1.000	**0.196	**0.084	**0.117
LABOR											1.000	**0.176	**0.039
OBSTET												1.000	**0.026
CONGN													1.000

* P<0.01

** 0.01<P<0.05

Table 5: Spearman correlation coefficients and p-values.

Spearman Correlation Coefficients Prob >|r| under H_0: Rho=0					
	NLBNL	NLBND	NOTERM	DTOTORD	DLIVORD
NLBNL	1.0000	0.0691 <.0001	0.1451 <.0001	0.8756 <.0001	0.9937 <.0001
NLBND		1.0000	0.0425 <.0001	0.1492 <.0001	0.1691 <.0001
NOTERM			1.0000	0.5474 <.0001	0.1479 <.0001
DTOTORD				1.0000	0.8817 <.0001
DLIVORD					1.0000

Table 6: Categorical variables included in the imputation model.

Variable Name	Definition of Variable	Category	Definition of Categories	Freq	Percent
RESTATUS	Residents status	1	Residents	30106	74.75
		2	Intrastate nonresidents	9230	22.92
		3	Interstate nonresidents	884	2.19
		4	Foreign residents	54	0.13
PLDEL3	Place of delivery	1	In hospital	39909	99.09
		2	Not in a hospital	363	0.90
		.	Unknown or not stated	2	0.00
REGNRES	Region of residence	0	Foreign residents	54	0.13
		1	Northeast	6786	16.85
		2	Midwest	8767	21.77
		3	South	14813	36.78
		4	West	9853	24.46
CITRSPOP	Population size of city of residence	0	>=1,000,000	3500	8.69
		1	Place of 500,000 to 1,000,000	1821	4.52
		2	Place of 250,000 to 500,000	3085	7.66
		3	Place of 100,000 to 250,000	3776	9.38
		9	All other areas in the U.S.	28038	69.62
		z	Foreign residents	54	0.13
METRORES	Metropolitan	1	Metropolitan county	33209	82.46
		2	Nonmetropolitan county	7011	17.41
		z	Foreign residents	54	0.13
CNTRSPOP	Population size of county of residence	0	>=1,000,000	10243	25.43
		1	Place of 500,000 to 1,000,000	7528	18.69
		2	Place of 250,000 to 500,000	6113	15.18
		3	Place of 100,000 to 250,000	6330	15.72
		9	All other areas in the U.S.	10006	24.84
		z	Foreign residents	54	0.13
MRACE3	Race of Mother	1	White	31861	79.11
		2	Races other than White or Black	2554	6.34
		3	Black	5859	14.55
MEDUC6	Education of mother	1	0 – 8 years	2397	5.95
		2	9 – 11 years	6061	15.05
		3	12 years	12324	30.60
		4	13 – 15 years	8651	21.48
		5	16 years and over	10288	25.55
		.	Not stated	553	1.37
DMAR	Marital status of mother	1	Married	26768	66.46
		2	Unmarried	13506	33.54
		.	Unknown or not stated	0	0.00
MPLBIRR	Place of birth of mother	1	Native born	30712	76.26
		2	Foreign born	9476	23.53
		.	Unknown or not stated	86	0.21

Table 6 (cont.): Categorical variables included in the imputation model.

Variable Name	Definition of Variable	Category	Definition of Categories	Freq	Percent
ADEQUACY	Adequacy of care	1	Adequate	29464	73.16
		2	Intermediate	7179	17.83
		3	Inadequate	2018	5.01
		.	Unknown	1613	4.01
MPRE5	Month prenatal care began	1	1st trimester	33040	82.04
		2	2nd trimester	5017	12.46
		3	3rd trimester	1056	2.62
		4	No prenatal care	374	0.93
		.	Unknown or not stated	787	1.95
DFRACE4	Race of father	1	White	28186	69.99
		2	Races other White, Black or unknown	2107	5.23
		3	Black	4333	10.76
		.	Unknown or not stated	5648	14.02
CSEX	Sex	1	Male	20544	51.01
		2	Female	19730	48.99
DPLURAL	Plurality	1	Single	38940	96.69
		2	Twin	1267	3.15
		3	Triplet	62	0.15
		4	Quadruplet	5	0.01
		5	Quintuplet or higher	0	0.00
DELMETH5	Method of delivery	1	Vaginal (excludes vaginal after previous C-section)	28974	71.94
		2	Vaginal birth after previous C-section	589	1.46
		3	Primary C-section	6328	15.71
		4	Repeat C-section	4130	10.25
		.	Not stated	253	0.63
NLBNL	Number of live birth, now living	0	No live birth, now living	16166	40.14
		1	One live birth, now living	13303	33.03
		2	Two live births, now living	6574	16.32
		3	Three live births, now living	2585	6.42
		4	Four live births, now living	894	2.22
		5	Five live births, now living	349	0.87
		6	Six live births, now living	155	0.38
		7	Seven live births, now living	86	0.21
		8	Eight live births, now living	31	0.08
		9	Nine live births, now living	21	0.05
		10	Ten live births, now living	14	0.03
		11	Eleven live births, now living	8	0.02
		12	Twelve live births, now living	5	0.01
		13	Thirteen live births, now living	2	0
		.	Not stated	81	0.2

Table 6 (cont.): Categorical variables included in the imputation model.

Variable Name	Definition of Variable	Category	Definition of Categories	Freq	Percent
NLBND	Number of live births, now dead	0	No live birth, now dead	39515	98.12
		1	One live birth, now dead	549	1.36
		2	Two live births, now dead	73	0.18
		3	Three live births, now dead	15	0.04
		4	Four live births, now dead	7	0.02
		5	Five live births, now dead	1	0
		6	Six live births, now dead	1	0
		9	Nine live births, now dead	3	0.01
		.	Not stated	110	0.27
NOTERM	Number of other termination	0	No other termination	30572	75.91
		1	One other termination	6463	16.05
		2	Two other terminations	2062	5.12
		3	Three other terminations	682	1.69
		4	Four other terminations	226	0.56
		5	Five other terminations	76	0.19
		6	Six other terminations	44	0.11
		7	Seven other terminations	12	0.03
		8	Eight other terminations	9	0.02
		9	Nine other terminations	3	0.01
		10	Ten other terminations	3	0.01
		.	Not stated	121	0.3
MEDRK[1]	Total number of medical risks	0	No medical risk	27767	68.95
		1	One medical risk	9721	24.14
		2	Two medical risks	1975	4.9
		3	Three medical risks	399	0.99
		4	Four medical risks	58	0.14
		5	Five medical risks	15	0.04
		6	Six medical risks	4	0.01
		.	Not stated	335	0.83
OBSTET[2]	Total number of abnormal conditions	0	No abnormal condition	2816	6.99
		1	One abnormal condition	8062	20.02
		2	Two abnormal conditions	17251	42.83
		3	Three abnormal conditions	9686	24.05
		4	Four abnormal conditions	2063	5.12
		5	Five abnormal conditions	193	0.48
		6	Six abnormal conditions	11	0.03
		.	Not stated	192	0.48
CONGN[3]	Total number of congenital anomalies	0	No congenital anomaly	39276	97.52
		1	One congenital anomaly	341	0.85
		2	Two congenital anomalies	38	0.09
		3	Three congenital anomalies	6	0.01
		4	Four congenital anomalies	3	0.01
		.	Not stated	610	1.51

Table 6 (cont.): Categorical variables included in the imputation model.

Variable Name	Definition of Variable	Category	Definition of Categories	Freq	Percent
NEWBN[4]	Total number of newborn complications	0	No newborn complication	36994	91.86
		1	One newborn complication	2545	6.32
		2	Two newborn complications	341	0.85
		3	Three newborn complications	60	0.15
		4	Four newborn complications	8	0.02
		.	Not stated	326	0.81
LABOR[5]	Total number of labor complications	0	No labor complication	27159	67.44
		1	One labor complication	10118	25.12
		2	Two labor complications	2307	5.73
		3	Three labor complications	368	0.91
		4	Four labor complications	68	0.17
		5	Five labor complications	12	0.03
		6	Six labor complications	2	0
		.	Not stated	240	0.6

MEDRISK[1]: medical risk variables include anemia, cardiac disease, acute or chronic lung disease, etc.
OBSTETRIC[2]: obstetric procedures include amniocentesis, electronic fetal monitor, induction of labor, etc.
CONGNTL[3]: congenital anomalies include anencephalus, spina bifida, hydrocephalus, microcephalus, etc.
NEWBORN[4]: newborn complications include anemia, birth injury, fetal alcohol, etc.
LABCOMP[5]: labor complications include febrile, meconium, premature rupture of membrane, etc.

Table 7: Continuous, count or mixed variables included in the imputation model.

Variable Name	Definition of Variable	N	N Miss	Mean	Std Dev	Min	Max	Miss%
Continuous variable								
DBIRWT	Birth weight of child (gram)	40240	34	3297.12	604.87	227	6039	0.08%
DMAGE	Age of mother	40274	0	27.35	6.2	12	52	0.00%
DFAGE	Age of father	34936	5338	30.49	6.85	14	72	13.25%
FMAPS	Apgar score	31106	9168	8.91	0.75	0	10	22.76%
WTGAIN	Weight gain (lb)	32784	7490	30.89	13.8	0	98	18.60%
Count variable								
NPREVIS	Total number of prenatal visits	39239	1035	11.55	3.99	0	49	2.57%
Mixed variable								
CIGAR	Number of cigars/day	34318	5956	1.08	3.88	0	70	14.79%
DRINK	Number of drinks/week	34659	5615	0.04	0.71	0	84	13.94%

Table 8: Variables included in the propensity model and percent missing.

Variable	Definition of Variable	Missing %	Included
Categorical Variables			
RESTATUS	Residence status	-	Y
PLDEL3	Place of delivery	-	Y
REGNRES	Region of residence	-	Y
CITRSPOP	Population size of city of residence	-	Y
CNTRPOP	Population size of county of residence	-	Y
METRORES	Metropolitan of residence	-	Y
CSEX	Sex of child	-	Y
DPLURAL	Plurality	-	Y
MARCE3	Race of mother	-	Y
MEDUC6	Education of mother	1.37%	N
DMAR	Marital status of mother	-	Y
MPLBIRR	Place of birth of mother	0.21%	N
ADEQUACY	Adequacy of care	4.01%	N
MPRE5	Month prenatal care began	1.95%	N
DFRACE	Race of father	14.02%	N
DELMETH5	Method of delivery	0.63%	N
NLBNL	Number of live births, now living	0.20%	N
NLBND	Number of live births, now dead	0.27%	N
NOTERM	Number of other termination	0.30%	N
MEDRK	Total number of medical risks	0.83%	N
NEWBN	Total number of newborn complications	0.81%	N
LABOR	Total number of labor complications	0.60%	N
OBSTET	Total number of abnormal conditions	0.48%	N
CONGN	Total number of congenital anomalies	1.51%	N
Continuous Variables			
DBIRWT	Birth weight of child	0.08%	N
DMAGE	Age of mother	-	Y
DFAGE	Age of father	13.25%	N
FMAPS	Apgar score	22.76%	N
WTGAIN	Weight gain	18.60%	N
Count Variable			
NPREVIS	Total number of prenatal visits	2.57%	N
Mixed Variables			
CIGAR	Number of cigars/day	14.79%	N
DRINK	Number of drinks/week	13.94%	N

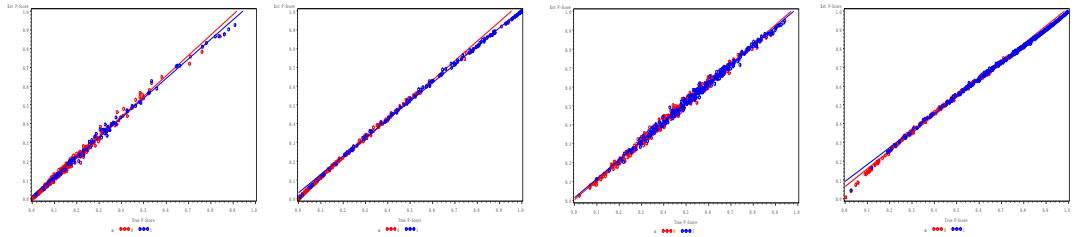
Table 9: Point estimates (Standard Errors) and p-values of linear regression coefficients for model of gestational age for complete cases, only variables without any missing values included in the propensity model.

Imputation	Propensity score		Missing indicator	
	β (SE)	p-value	β (SE)	p-value
1	-1.94 (0.26)	<0.0001	0.02 (0.03)	0.5193
2	-1.87 (0.26)	<0.0001	0.01 (0.03)	0.7955
3	-1.95 (0.26)	<0.0001	0.01 (0.03)	0.8120
4	- 1.99 (0.26)	<0.0001	0.02 (0.03)	0.4176
5	- 1.97 (0.26)	<0.0001	0.01 (0.03)	0.7483
Summary	- 1.95 (0.26)	<0.0001	0.01 (0.03)	0.6619

Figure 1: Scatter plots of true propensity score versus estimated propensity score. Propensity score is estimated by fitting correct model.

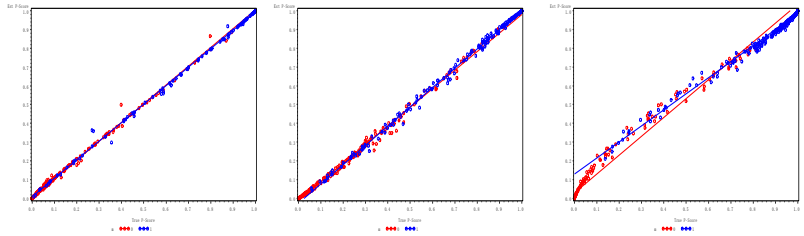
$Y=1+X_1+2*X_2+\varepsilon$

m%=15.8% m%=26.0% m%=48.6% m%=64.0%



$Y=1+2*X_1+2*X_2+3*X_1X_2+\varepsilon$

m%=23.6% m%=54.2% m%=63.6%



$Y=1+2*X_1+2*X_2+3*X_1^2+\varepsilon$

m%=16.2% m%=33.8% m%=59.0%

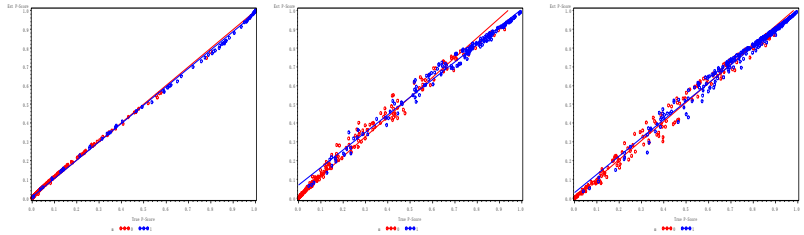
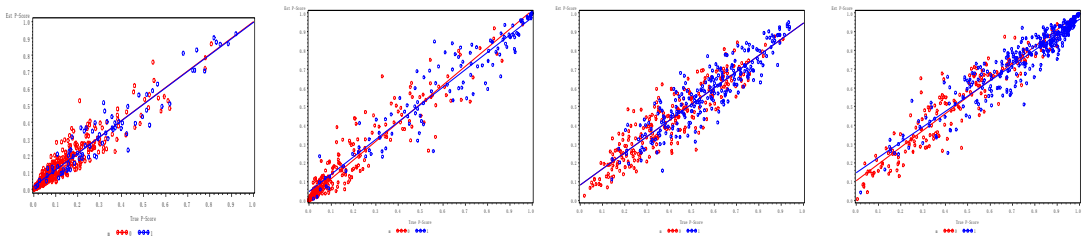


Figure 2: Scatter plots of true propensity score versus estimated propensity score. Propensity score is estimated by fitting overfitted model.

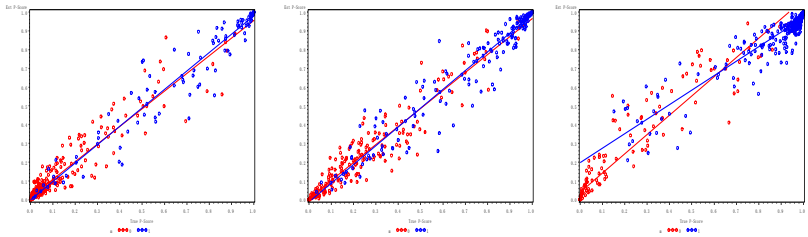
$Y=1+X_1+2*X_2+\varepsilon$

m%=15.8% m%=26.0% m%=48.6% m%=64.0%



$Y=1+2*X_1+2*X_2+3*X_1X_2+\varepsilon$

m%=23.6% m%=54.2% m%=63.6%



$Y=1+2*X_1+2*X_2+3*X_1^2+\varepsilon$

m%=16.2% m%=33.8% m%=59.0%

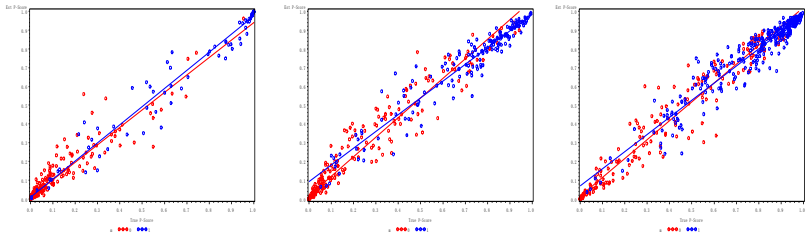
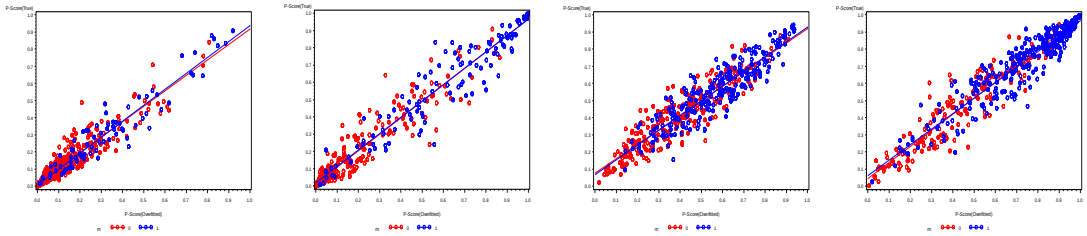


Figure 3: Scatter plots of propensity score from correct model versus propensity score from overfitted model.

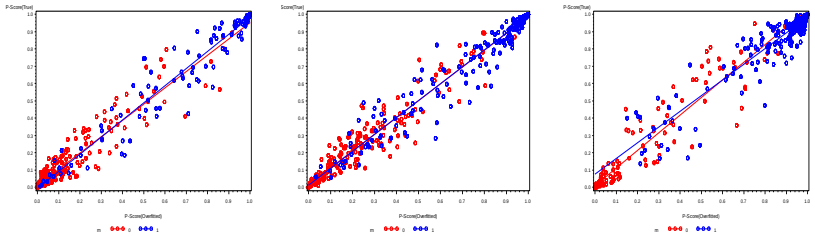
$Y=1+X_1+2*X_2+\varepsilon$

m%=15.8% m%=26.0% m%=48.6% m%=64.0%



$Y=1+2*X_1+2*X_2+3*X_1X_2+\varepsilon$

m%=23.6% m%=54.2% m%=63.6%



$Y=1+2*X_1+2*X_2+3*X_1^2+\varepsilon$

m%=16.2% m%=33.8% m%=59.0%

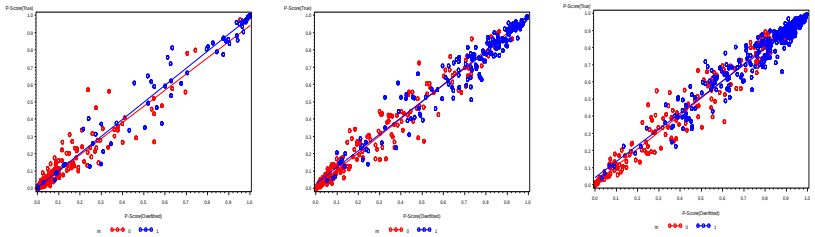
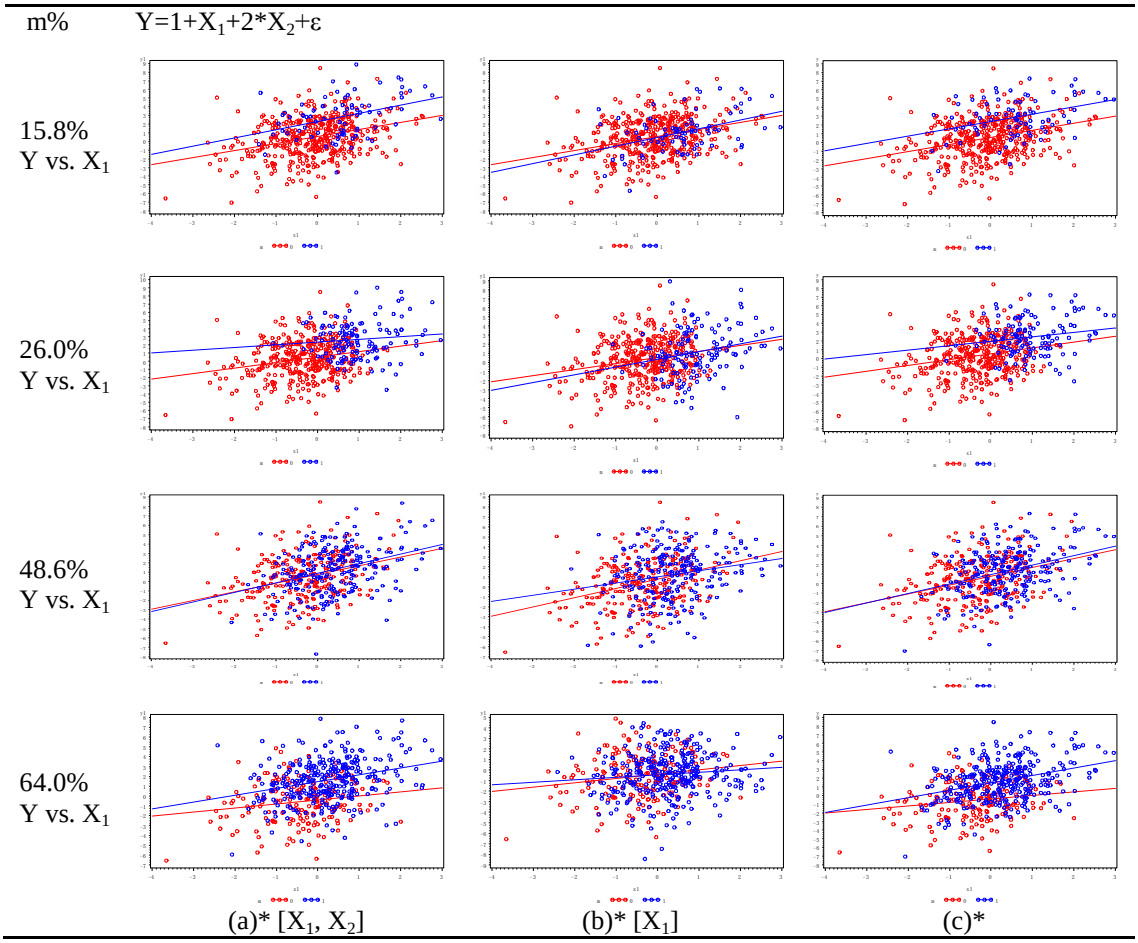


Figure 4: Distribution of completed Y vs. X_1.

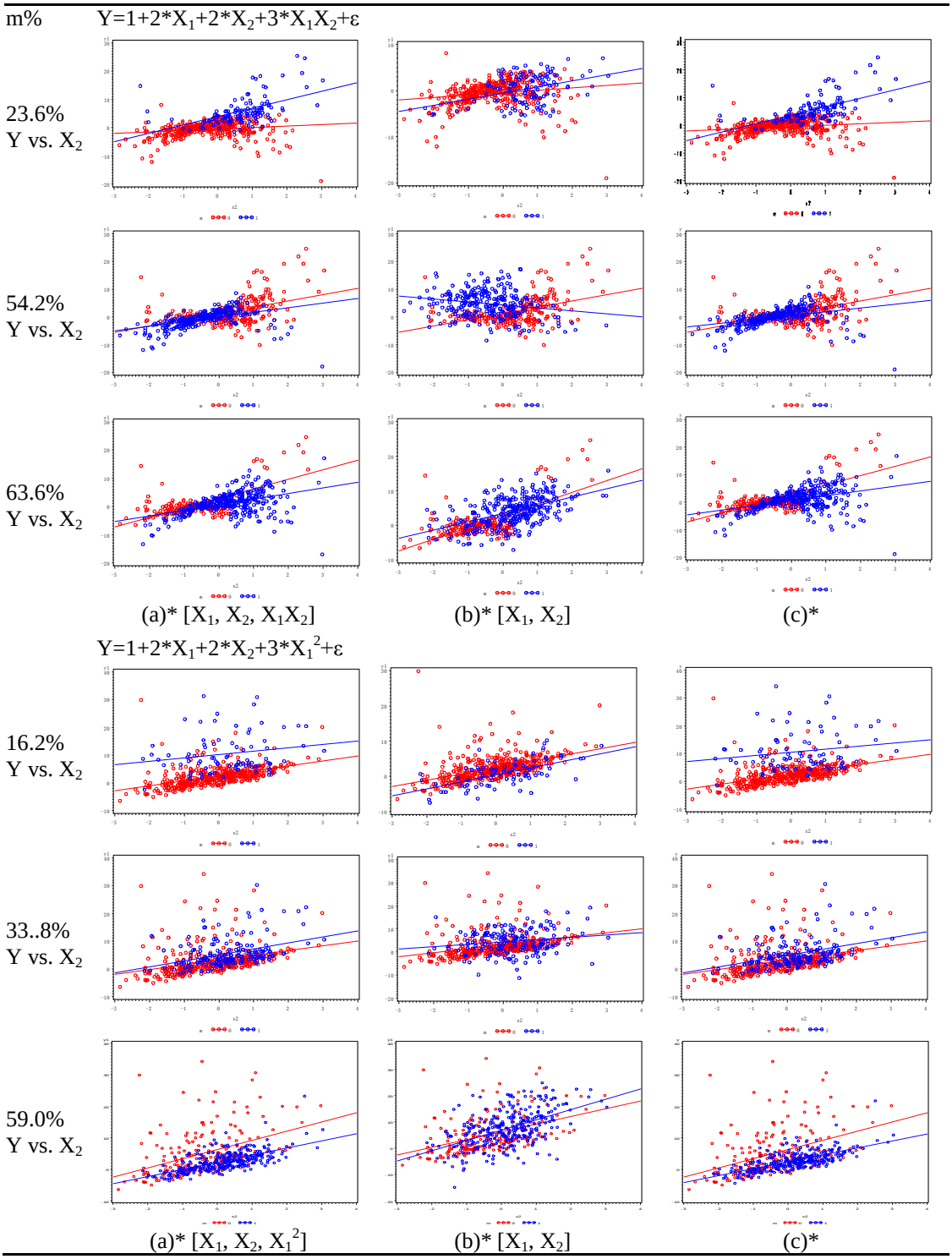


(a) The observed and imputed Ys are plotted versus X_1. Regression lines are produced with I=R operand to show the relationship between Ys and X_1. Correct imputation models are fitted to generate the imputations of Y.

(b) The observed and imputed Ys are plotted against X_1. Incorrect imputation models are fitted to create imputations of Y.

(c) The true Ys are plotted against X_1 at two levels of the missingness (red: m=0, blue: m=1).

Figure 4 (Cont.): Distribution of completed Y vs. X_2.

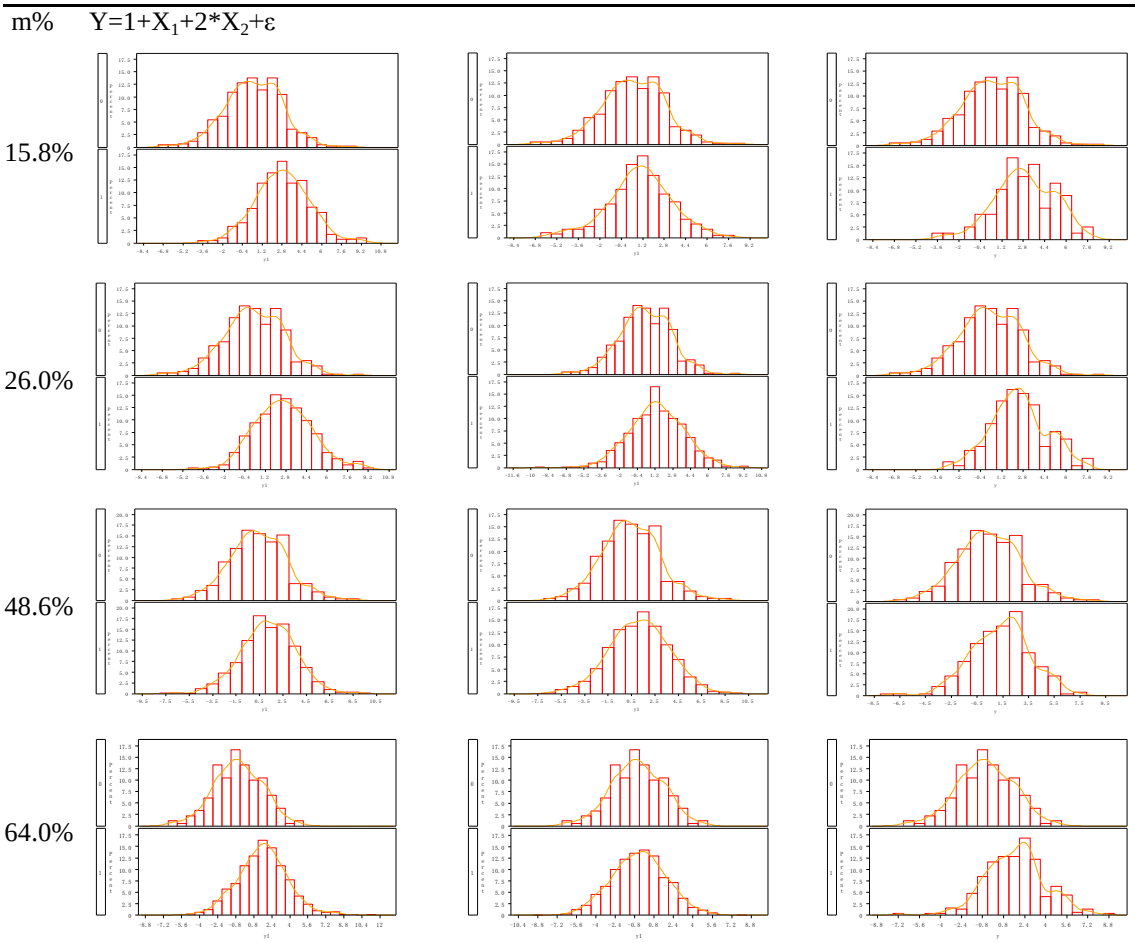


(a) The observed and imputed Ys are plotted versus X_1. Regression lines are produced with I=R operand to show the relationship between Ys and X_1. Correct imputation models are fitted to generate the imputations of Y.

(b) The observed and imputed Ys are plotted against X_1. Incorrect imputation models are fitted to create imputations of Y.

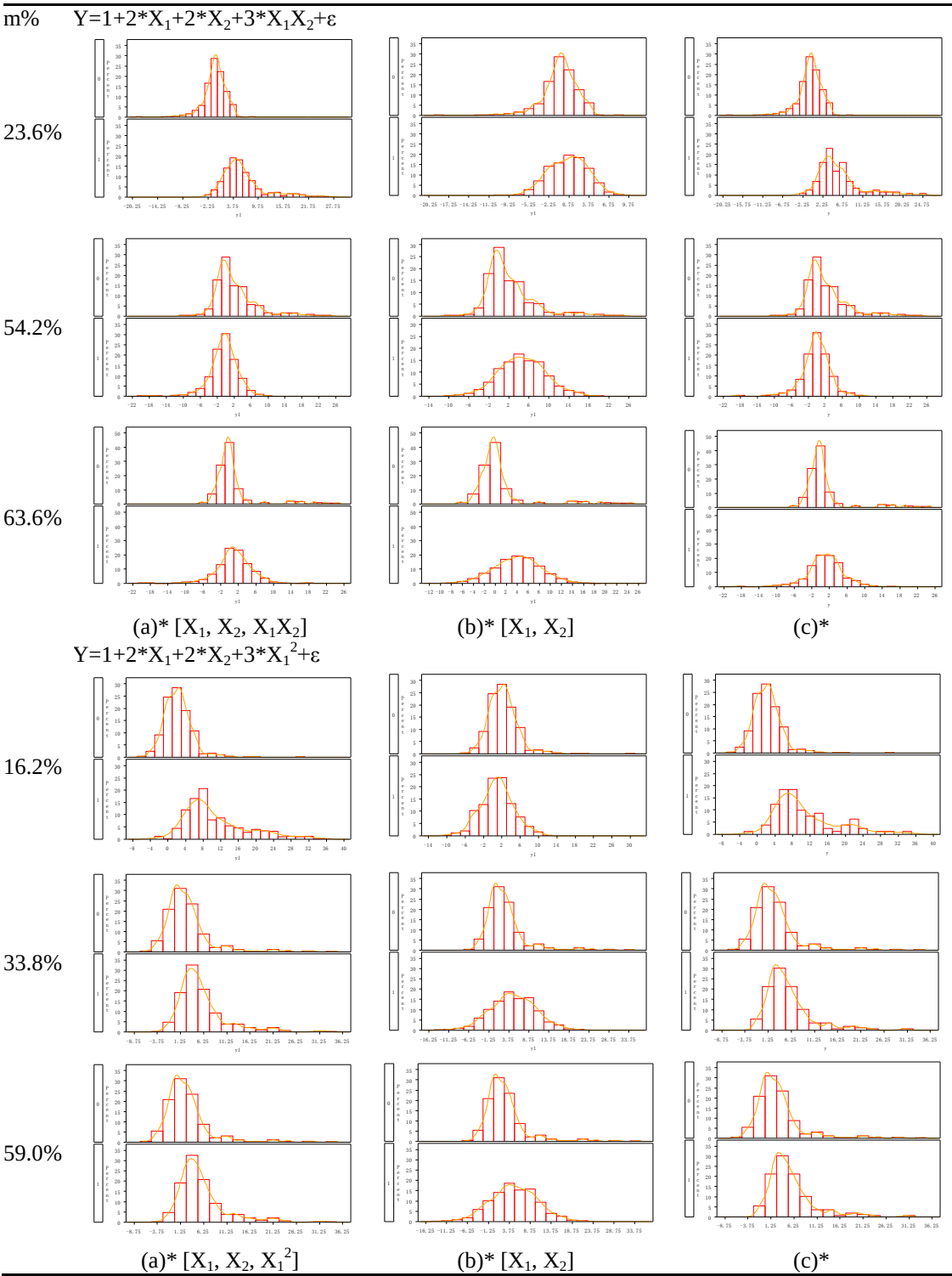
(c) The true Ys are plotted against X_1 at two levels of the missingness (red: m=0, blue: m=1).

Figure 5: Histograms of observed Y and imputed Y.



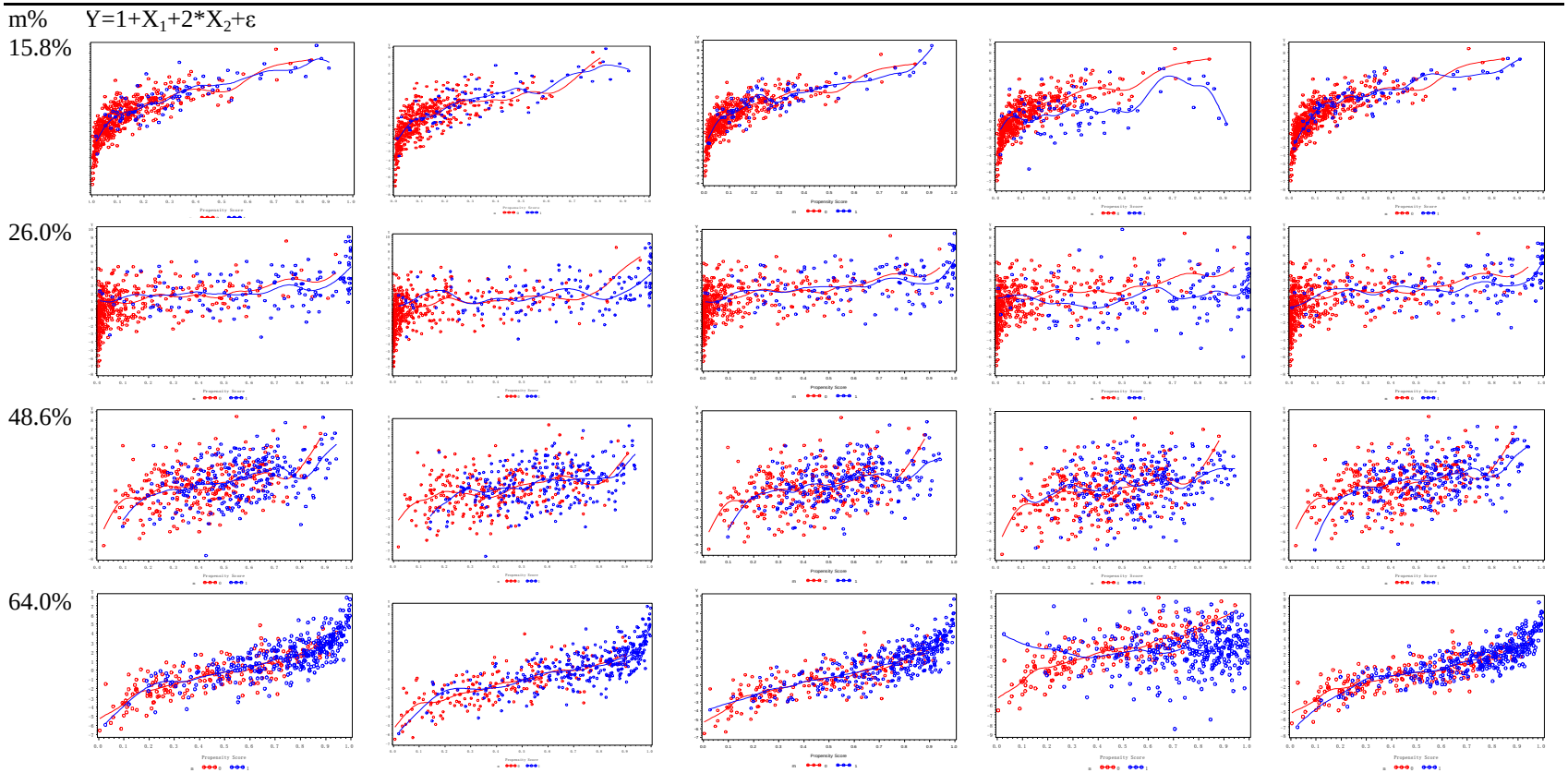
(a) Histogram with Kernel Curve is plotted to show the distribution of Ys (Top: Observed Ys; Bottom: Imputed Ys): correct imputation model is fitted to create imputations.
(b) Histogram of Y (Top: Observed Ys; Bottom: Imputed Ys): incorrect imputation model is fitted to create imputations.
(c) Histogram of Y at two levels of missingness (Top: m=0, Bottom: m=1).

Figure 5 (Cont.): Histograms of observed Y and imputed Y.



(a) Histogram with Kernel Curve is plotted to show the distribution of Ys (Top: Observed Ys; Bottom: Imputed Ys): correct imputation model is fitted to create imputations.
(b) Histogram of Y (Top: Observed Ys; Bottom: Imputed Ys): incorrect imputation model is fitted to create imputations.
(c) Histogram of Y at two levels of missingness (Top: m=0, Bottom: m=1).

Figure 6: Distribution of completed Y versus the propensity score.



(a) Correct imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}) with smooth curves plotted to indicate a possible nonlinear relationship between Ys and the propensity score.

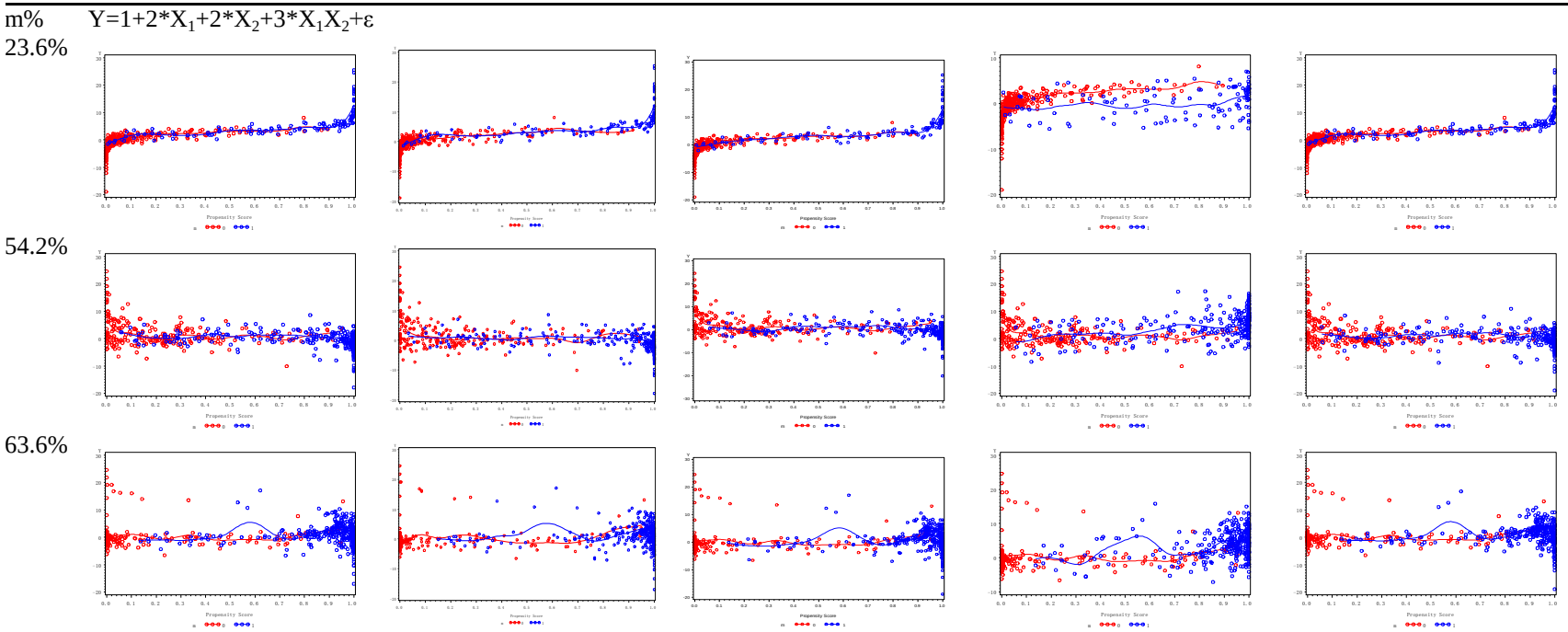
(b) Correct imputation model and overfitted propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(c) Overfitted imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(d) Incorrect imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(e) True Ys are plotted versus the propensity scores from correct model at two levels of the missingness (red: m=0, blue: m=1) with smooth curves plotted.

Figure 6 (cont.): Distribution of completed Y versus the propensity score.



(a) Correct imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}) with smooth curves plotted to indicate a possible nonlinear relationship between Ys and the propensity score.

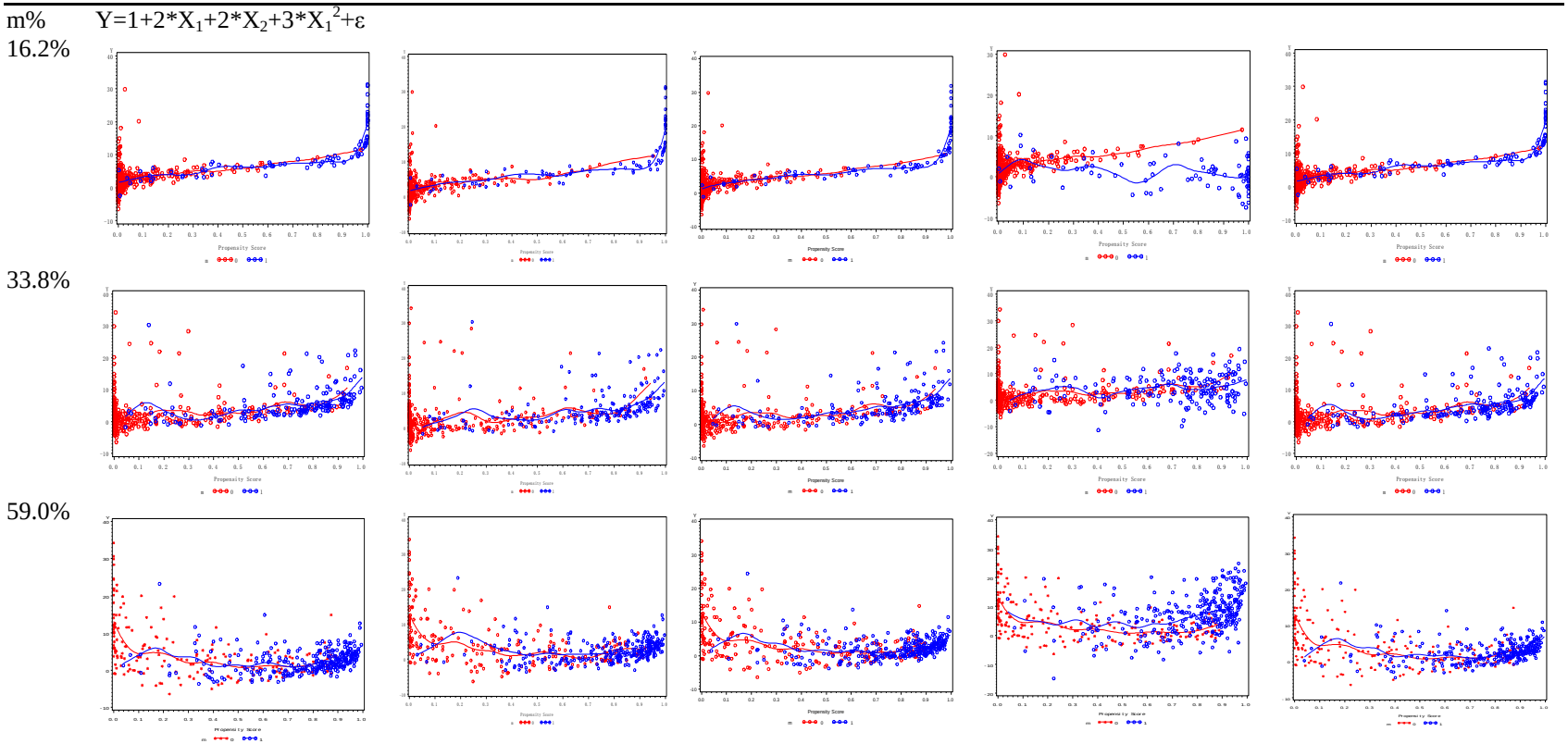
(b) Correct imputation model and overfitted propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(c) Overfitted imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(d) Incorrect imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(e) True Ys are plotted versus the propensity scores from correct model at two levels of the missingness (red: m=0, blue: m=1) with smooth curves plotted.

Figure 6 (Cont.): Distribution of completed Y versus the propensity score.



(a) Correct imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}) with smooth curves plotted to indicate a possible nonlinear relationship between Ys and the propensity score.

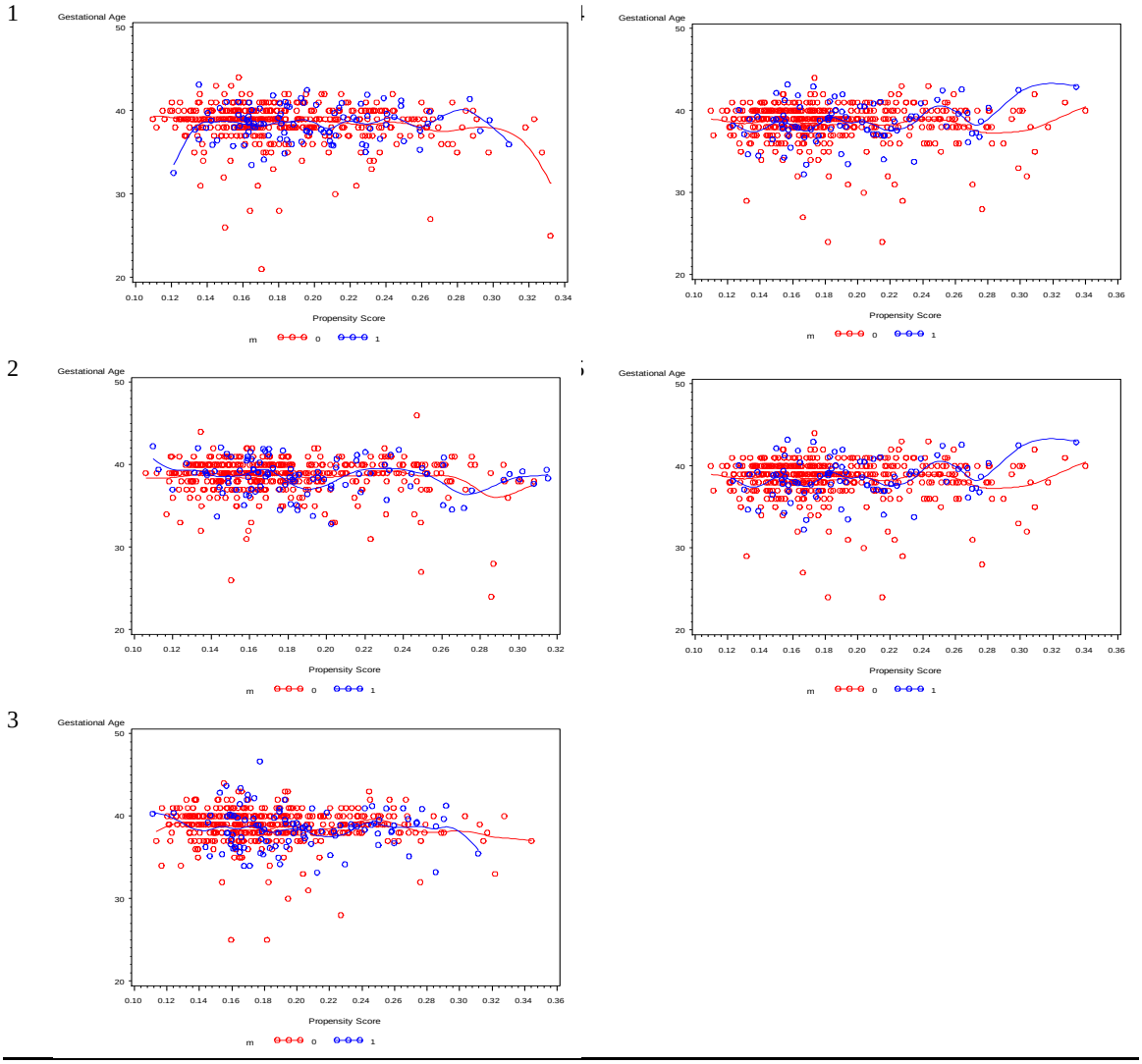
(b) Correct imputation model and overfitted propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(c) Overfitted imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(d) Incorrect imputation model and correct propensity model: observed and imputed Ys are plotted versus the propensity score (red: Y_{obs}, blue: Y_{imp}).

(e) True Ys are plotted versus the propensity scores from correct model at two levels of the missingness (red: m=0, blue: m=1) with smooth curves plotted.

Figure 7: Plots of gestational age (observed+imputed) versus propensity score: imputations are created by using Sequential Regression Multiple Imputation method. Only variables without missing value are included in the propensity model.



Bibliography

Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. Journal of the Royal Statistical Society, 57: 273-291.

Alexander, G.R. & Allen, M.C. (1996). Conceptualization, measurement, and use of gestational age. I. Clinical and Public Health Practice. J Perinatal , 16:53–9.

Barnard, J. & Meng, X. (1999). Application of multiple imputation in medical studies: From AIDS to NHANES. Statistical Methods in Medical Research, 8: 17-36.

Hammad, H.T. (2009). Thesis: Identification of factors that relate to gestational age in term and preterm babies using 2002 National Birth Data.

Little, R.J & Rubin, D.B. (1989). The analysis of social science data with missing values. Sociological Methods & Research, 18: 292-326.

Little, R.J. & Rubin,D.B. (2002). Statistical Analysis with missing data. 2nd edn. New York: john Wiley & Sons.

Little, R.J. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83: 1198-1202.

Martin, J.A., Hamilton, B.E., Sutton, P.D., Ventura, S.J., Menacker, F., & Munson, M.L. (2003). Births: Final data for 2002. National vital statistics reports, vol 52 no 10. Hyattsville, Maryland: National Center for Health Statistics.

National Center for Health Statistics. (2002). Natality 2002. Hyattsville, MD: National Center for Health Statistics.

Parker, J. & Schenker, N. (2007). Multiple imputation for national public-use datasets and its possible application for gestational age in United States Natality files. Paediatric & Perinatal Epidemiology, 21: 97-105.

Raghunathan, T.E., Lepkowski, J.M., Hoewyk, J.V., & Solenberger, P. (2001). a multivariate technique for multiple imputing missing values using a sequence of regression models. Suvery Methodology, 27: 85-95.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70: 41-55.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63: 581-592.

Sinharay, S., Stern, H.S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. Psychological methods, 6: 3317-329.