Research Article

Statistics in Medicine

Received 6 November 2014, Accepted 9 February 2016 Published online 7 March 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6926

Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models

Irina Bondarenko[a] **and Trivellore Raghunathan**[b*†]

Multiple imputation has become a popular approach for analyzing incomplete data. Many software packages are available to multiply impute the missing values and to analyze the resulting completed data sets. However, diagnostic tools to check the validity of the imputations are limited, and the majority of the currently available methods need considerable knowledge of the imputation model. In many practical settings, however, the imputer and the analyst may be different individuals or from different organizations, and the analyst model may or may not be congenial to the model used by the imputer. This article develops and evaluates a set of graphical and numerical diagnostic tools for two practical purposes: (i) for an analyst to determine whether the imputations are reasonable under his/her model assumptions without actually knowing the imputation model assumptions; and (ii) for an imputer to fine tune the imputation model by checking the key characteristics of the observed and imputed values. The tools are based on the numerical and graphical comparisons of the distributions of the observed and imputed values conditional on the propensity of response. The methodology is illustrated using simulated data sets created under a variety of scenarios. The examples focus on continuous and binary variables, but the principles can be used to extend methods for other types of variables. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: multiple imputation; propensity score; diagnostics; congeniality

1. Introduction

Multiple imputation (MI) is a general purpose approach for analyzing data with missing values where the missing set of values are replaced by several sets of plausible values. These values are generated as draws, typically from a predictive distribution of the missing set conditional on the observed set of values. Each plausible or imputed set, when combined with the observed set of values, results in a completed data set. Each completed data set is then analyzed separately, and then the inferential statistics (such as point estimates, covariance matrices, test statistics, or p-values) are combined across the completed data sets to construct the MI inference [1–5].

Data with missing values may be complex with several types of variables, (such as continuous, ordinal, nominal, count, and semi-continuous), involve skip patterns (for example, some variables are not applicable to a particular group of subjects), and restrictions (such as bounds for plausible lab values). Availability of MI software to handle such complexities has made the MI procedure attractive. During the last two decades, procedures for MI of missing values have been incorporated into many popular statistical software packages such as SAS, R, Stata [6–9].

Despite this extensive development of the imputation software, the tools to diagnose the validity of the imputations are limited, and the majority of currently available tools assume that the imputation model is known. In most practical applications of MI, however, the imputer and the analyst are different individuals or may work at different organizations. Hence, the analyst may have limited or no knowledge of the

[a] *Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[b] *Survey Research Center, Institute for Social Research and Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbot, MI 48106, U.S.A.*
**Correspondence to: Trivellore Raghunathan, Survey Research Center, Institute for Social Research and Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbot, MI 48106, U.S.A.*
[†] *E-mail: teraghu@umich.edu*

imputation model. This may lead to a scenario where the imputation and analyst models make different assumptions with respect to data generating process and, hence, in broad terms uncongenial [10].

There are different kinds of uncongeniality having varying consequences on the point and interval estimates. For example, if the analyst model is a submodel of the imputer model, then the point estimates are typically unbiased but may yield wider interval estimates. On the other hand, if the imputer model is a submodel of the analyst model, then the point estimates may be biased. Another type of uncongeniality occurs when the model assumptions made by the imputer and analyst are the same but the analyst uses suboptimal estimation procedure (for example, the method of moments instead of the maximum likelihood) [11, 12]. Note, that we always assume that the analyst model is the correct model for performing the repeated sampling calculations.

This paper develops a set of diagnostic tools with two broad objectives:

(1) To assist an analyst to determine whether the imputations are reasonable under his/her model, without actually knowing the exact imputation model. That is, the analyst can diagnose uncongeniality.
(2) To assist an imputer to fine tune the imputation model through checking whether the observed and imputed values exhibit similar characteristics, and thus capturing important features to be preserved in the imputation process.

An imputer can use several standard model building techniques such as regression diagnostics, posterior predictive checks, and variable selection methods. These tools allow the modeler to discern important features and structures in the data and then to incorporate them in the imputation model. Some examples include inclusion of nonlinear or interaction terms, variable transformation, etc. The goal of the second objective listed earlier is slightly different from these standard procedures. That is, the goal is to check whether, conditional on the model chosen, the observed and imputed values exhibit similar characteristics. After all, the goal of the imputation is to create a plausible completed data set from the population. If the observed and imputed values are not exhibiting similar characteristics, then the imputation model needs further refinement.

Recently, several imputation diagnostic tools have been proposed [13–16]. These tools proceed in two stages: (i) initial comparison of the marginal distributions of the imputed and observed values and (ii) evaluation of variables flagged based on the first stage with the knowledge of the imputation model. The initial comparison of the marginal distributions of the observed and imputed values is useful to identify variables that need future evaluation. Existing statistical packages provide a variety of strategies for comparison of marginal distributions of imputed and observed values, including comparison of histograms, density plots, $q-q$ plots, or descriptive statistics of the imputed and observed. Various numeric tests and rules of thumb have been suggested to identify variables with differences in the marginal distributions of observed and imputed values. For example, flagging variables with an absolute difference in means between the observed and imputed values greater than 2 standard deviations, or with a ratio of variances of the observed and imputed values that is less than 0.5 or greater than 2 [16]. As an alternative, compare variances of point estimates and correlation coefficients from multiply imputed data and available cases [14]. Kolmogorov–Smirnov (KS) test can also be used to flag variables with significant differences in the marginal distributions of imputed and observed values [13]. In such comparisons, attention should be paid to the extreme values produced by imputations, because they can indicate a potential problem with the imputations and a need for model improvement [15].

It is important to note that the marginal distributions of the observed and imputed values are expected to be similar only when the data are missing completely at random. Under the missing at random (MAR) mechanism, the differences in the marginal distributions of the observed and imputed values are not necessarily an indicative of problems with the imputation model. Consider a bivariate example with missing values in Y with X fully observed. Suppose that the probability of missing Y values is positively (or negatively) correlated with X. Depending upon the correlation between Y and X, the marginal distributions of the observed and imputed values will differ even when the imputations are created under the correctly specified model. Thus, the initial evaluation described earlier may result in false alarms about the imputation model. Thus, many of these procedures are mostly useful for filtering a large set to smaller subset for further inspection.

The second stage of MI diagnostics includes more elaborate tools developed to assist in the evaluation of certain conditional, weighted distributions of observed and imputed values. For example, one can compare complete-case, MI, and weighted analyses [14]. This approach, however, is applicable only for a certain missing data patterns. An alternative is to test the fit of the imputation model to the observed data by comparing distributions of the residuals for imputed and observed values, conditional on their

predicted values under the imputation model [13]. If the imputation model is a good fit, then the pattern of residuals should be random with no difference between the observed and imputed values. A sensitivity analysis based on the posterior predictive checking under an imputation model or its surrogate based on the subset of variables may also be useful [14]. For example, suppose that the completed data through imputation, D_C, are used to generate several copies or replicates $D_C^{*(r)}, r = 1, 2, \ldots$ through posterior predictive check mechanism. If the original imputations are valid, then the estimates based on the completed data through imputation, D_C, should be similar to the estimates calculated from $D_C^{*(r)}$ (under the correctly specified imputation model). However, considerable knowledge of the imputation model is necessary to implement many of these methods.

As emphasized earlier, the imputation models may not be available when the analyst and imputer work in different organizations. Sometimes, internal variables (such as paradata or administrative data) may be used in the imputation process which may not be released as a part of the completed data sets. In such situations, it is necessary to determine whether the imputations are valid under the analyst model or, in other words, to assess whether the imputation model is congenial with the model posited by the analyst. The notion of congeniality of the imputation and analyst models is an important issue. The techniques proposed in this paper can be used to assess congeniality of the imputations for the analyst model without knowing the imputation model.

The rest of the paper is organized into the following five sections. Section 2 develops the proposed diagnostic methods. Section 3 outlines a set of graphical tools to implement the method for continuous and binary variables. Section 4 uses simulated data sets to illustrate and evaluate the proposed approaches for assessing the validity of the imputations. How these tools can be used for detecting uncongeniality is also discussed in Section 4. Section 5 proposes numerical summaries or test procedures and evaluates them using simulated data sets. Section 6 concludes with discussions, limitations, and future work. Throughout the paper, the missing data are assumed to be MAR.

2. Proposed method

Consider a data set with missing values that has n observations and p variables, Y_v, $v = 1, 2, \ldots, p$ and the goal is to assess imputation for the variable Y_v. Without loss of generality, denote the set of observed values of Y_v as $y_{obs,v} = \{y_{sv}, s = 1, 2, \ldots, n_v\}$ and the set of missing values as $y_{mis,v} = \{y_{sv}, s = n_v+1, n_v+2, \ldots, n\}$. Let $R_{sv} = 1, s = 1, 2, \ldots, n_v$ and $R_{sv} = 0, s = n_v + 1, n_v + 2, \ldots, n$ be the response indicator. Let $Y_{obs,-v}$ denote all the observed set of values of variables $Y_i, i = 1, 2, \ldots, v-1, v+1, \ldots, p$ across all subjects. If values of Y_v are MAR, then, by definition, the conditional distribution of $y_{obs,v}$ given $Y_{obs,-v}$ should be the same to that of $y_{mis,v}$ given $Y_{obs,-v}$. That is, $Pr(y_{obs,v}|Y_{obs,-v}) = Pr(y_{mis,v}|Y_{obs,-v})$

Let $e_{obs,-v} = Pr(R_v = 1|Y_{obs,-v})$ be the actual response propensity for variable Y_v as a function of the observed data, $Y_{obs,-v}$. The propensity score $e_{obs,-v}$ is an efficient summary of the covariates $Y_{obs,-v}$ [17].Thus, under the MAR assumption, distributions of observed $y_{obs,v}$ and missing $y_{mis,v}$ must be the same, conditional on $e_{obs,-v}$.

Missing values $y_{mis,v}$ are unknown quantities and replaced in the course of imputations with a set of imputed values, $y_{mis,v}^{(l)} = \{y_{sv}^{(l)}, s = n_v + 1, n_v + 2, \ldots, n\}$, where $l = 1, 2, \ldots, M$. If the imputations are reasonable under the MAR assumption, then the observed set $y_{obs,v}$ and the imputed set $y_{mis,v}^{(l)}$ should have similar distributions conditional on the propensity score $e_{obs,-v}$ or equivalently, $Pr(y_{obs,v}|e_{obs,-v}) \sim Pr(y_{mis,v}^{(l)}|e_{obs,-v})$

The actual response propensity, $e_{obs,-v}$ is not known, but it can be estimated for each subject. One option is to estimate the propensity scores in the presence of missing data by conditioning on the response indicators as well as the observed covariates [18]. The second option is to use the following approach to estimate $e_{obs,-v}$.

(1) Let $Y_{mis,-v}^{(l)}, l = 1, 2, \ldots, M$ denote the M sets of imputed values for the missing values in all the variables except Y_v. Analyze each completed data set to build a response propensity model for Y_v using the standard logistic, probit, or any other regression model with binary outcome variable, R_v, and the completed data $Y_i, i = 1, 2, \ldots, v-1, v+1, \ldots, p$ as predictors. One could also use nonparametric regression models, such as CART or semiparametric models such as generalized additive models, to estimate the observed data propensity scores. The choice of the model for propensity of response is guided by the best fit model that ensures balance of all covariates, that is, $Pr(Y_{obs,-v}|e_{obs,-v}, R_v = 1) = Pr(Y_{obs,-v}|e_{obs,-v}, R_v = 0)$. The extent of balancing can be determined using the methods described in [19,20]. Standard model building tools can be used to inspect

whether interaction terms or transformations are needed and to assess the goodness of fit statistics (for example, in Hosmer and Lemeshaw [21]). After a satisfactory fit of the model in each completed data is achieved, obtain the values of the estimated propensities, $\widehat{Pr}(R_v = 1|Y_{obs,-v}, Y^{(l)}_{mis,-v})$.

(2) For sufficiently large M, approximate

$$\hat{e}_{obs,-v} = \widehat{Pr}\left(R_v = 1|Y_{obs,-v}\right) \approx \sum_{l=1}^{M} \widehat{Pr}\left(R_v = 1|Y_{obs,-v}, Y^{(l)}_{mis,-v}\right)/M.$$

3. Specific diagnostics

3.1. Continuous variables

The next step of comparing the imputed and observed values for a variable Y_v, conditional on the estimated propensity scores, can be performed in a number of ways. For example, H strata could be created from the propensity scores with n_h and m_h observed and imputed values in stratum $h = 1, 2, \ldots, H$. Analysis of variance (ANOVA) technique may be used with stratum ($H - 1$ degrees of freedom), indicator for observed/imputed (1 degree of freedom), and their interactions ($H - 1$ degrees of freedom) as factors. Under the correctly specified imputation model, the mean squares for both the missingness indicator and the interaction effect should be small. A large between-stratum sum of squares indicates significant departure from missing completely at random assumption. Because the imputed and observed values are correlated, the derivation of the sampling distribution even under the correctly specified imputation model is complex. Hence, it is analytically difficult to define large or small using the significance testing framework. In Section 5, we describe some empirically derived decision rules to reject the imputations that are calibrated (the exact and nominal levels are similar) and good power detect problems with the imputation.

Because the goal is to diagnose the problem with the imputations, a set of graphical tools (similar to the residual plots in a regression analysis) might be more useful. Two useful diagnostic plots are described to visualize the differences in the conditional distributions: *scatter plot diagnostics* and *residual density diagnostics*.

3.1.1. Scatter plot diagnostics. Plot values of Y_v versus the estimated propensity of response $\hat{e}_{obs,-v}$. Use different colors or symbols to identify observed and imputed values in Y_v. Systematic differences in the patterns of scatter across observed and imputed values, for a given value of propensity of response, indicate problems with the imputation model. Adding separate local regression (LOESS) curves to the scatter for imputed and observed groups may help further to visualize the differences.

3.1.2. Residual density diagnostics. First, regress Y_v on the estimated propensity score $\hat{e}_{obs,-v}$. Next, create histograms or kernel density plots of the residuals, separately for the observed and imputed values in Y_v. Differences in shape as well as the location of the residual densities indicate problems with the imputations.

3.2. Diagnostics for binary variables

For a binary variable Y_v, the analog of ANOVA method, described earlier for the continuous variable, may be developed. As before, create H strata-based propensity scores and fit a logistic regression model with Y_v as the outcome (imputed or observed), $H - 1$ dummy variables for strata, 1 dummy variable for observed/imputed and their products (interaction). We could use the deviance statistic as a measure, and under the correctly specified model, the deviance for both missing indicator and the interaction effect should be small. As in the continuous case, a set of graphical diagnostics tools may be more useful.

The scatter plot diagnostics can be applied to any type of variable: continuous or categorical. For binary variables, the average of the outcome values across imputations gives a better picture (because the response is either 1 or 0). Adding LOESS curves plotted on the top of the scatter make this plot, generally, informative. For a categorical variable, we may use the frequency distribution averaged across imputations.

The residual density diagnostic described earlier cannot be directly used for a binary variable because the residuals (given that the observed data is either 1 or 0) are not very informative and is even more problematic with nominal categorical outcome. Deviance residuals may be used, but the simulation study described in the next section did not show much promise for such residuals.

To develop an analog of residual density diagnostic for a binary variable Y_v, we modify the method described for the continuous variable as follows. Let p_v denote the actual conditional probability $Pr(Y_v = 1|Y_{-v})$. Then p_v is a balancing score for the covariates Y_{-v} across the two populations defined by the outcome Y_v. In particular, the distributions of Y_{-v}, conditional on p_v, are independent of Y_v [17]. This ideas were first proposed in [22] for logistic regression models and extended to the ordinal outcome in [23].

Note that, as discussed earlier, conditional on $e_{obs,-v}$, the $Y_{obs,-v}$ is independent of R_v. Thus, it follows from the properties of these two scores $e_{obs,-v}$ and p_v, that under MAR, conditional on $e_{obs,-v}$ and p_v, the Y_{-v} is independent of both Y_v and R_v. That is,

$$Pr\left(Y_{-v}|e_{obs,-v}, p_v\right) \perp\!\!\!\perp R_v, Y_v$$

A simple proof follows by noting that

$$Pr\left(Y_{-v}, Y_v, R_v|e_{obs,-v}, p_v\right) = Pr\left(Y_{-v}|e_{obs,-v}, p_v\right) \cdot Pr\left(Y_v, R_v|e_{obs,-v}, p_v, Y_{-v}\right). \quad (1)$$

Because, under MAR and the property of the propensity scores,

$$Pr\left(R_v, Y_v|e_{obs,-v}, p_v, Y_{-v}\right) = Pr\left(Y_v|e_{obs,-v}, p_v\right) \cdot f\left(R_v|e_{obs,-v}, p_v\right),$$

the right-hand side of Equation (1) reduces to

$$Pr\left(Y_{-v}|e_{obs,-v}, p_v\right) \cdot Pr\left(Y_v|e_{obs,-v}, p_v, Y_{-v}\right) \cdot Pr\left(R_v|e_{obs,-v}, p_v, Y_{-v}, Y_v\right)$$

The aforementioned result implies that the distribution of the covariates conditional on both propensities should be similar for the two populations defined by Y_v regardless of whether Y_v is imputed or observed. Specifically,

$$Pr\left(Y_{-v}|e_{obs,-v}, p_v, R_v, Y_v = 0\right) \sim Pr\left(Y_{-v}|e_{obs,-v}, p_v, R_v, Y_v = 1\right).$$

Note that matching on both propensity scores is akin to residual density diagnostics for the continuous variable. Estimation of $\widehat{p}_v$ poses a challenge because Y_v is not observed for $R_v = 1$, and using the imputed values makes the estimate very much dependent on the imputation model. The suggestion is to estimate $\widehat{p}_v$ using $Y_{obs,v}$ (that is, subset the completed data to $R_v = 0$), conditional on the observed values $Y_{obs,-v}$. Specifically, fit, for example, a logistic regression (or any other regression) model for

$$\widehat{p}_v^{(l)} = \widehat{Pr}\left(Y_v = 1|Y_{obs,-v}, Y_{mis,-v}^{(l)}, R_v = 0\right),$$

to obtain the maximum likelihood estimate of the regression coefficients, $\widehat{\beta}_l$. Note that the complete-case maximum likelihood estimates are the correct estimates for the regression model under MAR. Average $\widehat{\beta}_l$ across imputations ($\bar{\beta} = \sum_l \widehat{\beta}_l / M$) and then apply the resulting logistic regression equation (with the average $\bar{\beta}$ as the regression coefficient) to estimate the propensity score $\widehat{p}_v$ for the whole sample.

In the simulation study (described in the next section), the following strategy to develop residual density diagnostic for binary variables proved to be highly useful to diagnose problems with the imputations of binary variables. First, conduct a principal component analyses on $Y_{-v}^{(l)}$ and extract the principal components $Z_{-v}^{(l)}$. Next, regress these principal component $Z_{-v}^{(l)}$ on p_v, $\widehat{e}_{obs,-v}$ and their interaction. Finally, separate the data by the missingness indicator R_v, create two sets of the create histograms or kernel density plots of the residuals, grouping subjects by the value of outcome Y_v within each set (a total of four kernel density plots for the cross classification $R_v \times Y_v$). Differences in distributions of the residuals for the two values of Y_v for subjects with $R_v = 1$ or for subjects with missing value of Y_v ($R_v = 0$) indicate problem with the imputations.

4. Simulation study

To evaluate these diagnostic tools, we conducted a simulation study involving the following four steps: generation of complete data; setting some values to missing; multiply imputing under the correct and incorrect models; and finally applying the diagnostic tools developed in this paper. The simulated data contained both continuous and binary variables.

4.1. Simulations for continuous variables

First, the complete data of size 1000 were generated from normal model with mean function given by (2) and variance 1. Variables x_1, x_2, x_3 were independently drawn from the standard normal distribution and are assumed to be fully observed.

$$\mathrm{E}(y|x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_1^2 + \alpha_5 x_1 \cdot x_2 \,. \tag{2}$$

We considered two scenarios by choosing the values of α. The null scenario or the correctly specified model where $\alpha_4 = \alpha_5 = 0$, or the non-null scenarios with varying levels of mis-specification by various choices of $(\alpha_3, \alpha_4, \alpha_5)$. The mis-specification scenarios correspond to omitting a linear, quadratic or interaction term.

Second, some values for y were deleted. The missing data mechanism was defined by a logistic regression model (3). The regression coefficients were chosen to yield the response rate as 0.5 (50% missing values on the average).

$$\mathrm{logit}(\Pr(y = .)) = 1 - x_1 + x_2 - x_3 \,. \tag{3}$$

At the third step, the missing values of y were multiply imputed using the sequential regression approach as implemented in IVEWARE [6,24] package. For this simple case with missing values only in y, this is equivalent to drawing values from the posterior predictive distribution under the normal model with non-informative prior for the parameters. The imputation models adopted for these simulations did not include interaction, or a quadratic term (misspecified models but all variables included), and in one scenario did not include x_3 term (omitted variable in the model). Five multiply imputed data sets were produced under all scenarios.

Lastly, scatter plot and residual density diagnostic were applied to the imputed data sets to assess the validity of the imputations. The results are given in Figure 1. It shows that the two proposed diagnostics and, for comparison purposes, also the expected value-based diagnostics, which requires the knowledge of imputation model [13].

Red and black dots indicate imputed and observed values, respectively. The first row corresponds to the null scenario, where the imputation and data generating models included only linear terms for x_1, x_2, x_3, and y with no interactions or square terms. All three methods, as expected, show no difference in patterns of red (imputed) and black (observed) values.

In the second row, the imputation model is misspecified by omitting x_3. Scatter plot diagnostic shows differences in the observed and imputed values of y for a given values of the propensity scores. The differences in the patterns are made clearer by the addition of LOESS curves to the scatter plot. The residual density plot shows less variance in the residual values for imputed y's than for the observed values. However, the expected value-based plot shown in the third column fails to diagnose the problem. In fact, if a variable is erroneously omitted from the imputations, the diagnostics based on the imputation model are not valid, and hence, the expected value-based diagnostic cannot be used to detect omitted variable from the imputation model.

In the third row, the data generating model includes both the linear and quadratic terms for x_3, but the imputation model includes only a linear term. For this scenario, all three methods clearly detect substantial differences in patterns of the observed and imputed values. For example, the residuals for the observed values (in black) show a long left tail, whereas the residuals for the imputed values (in red) show much more compact and symmetric distribution. This difference in in the residual densities indicates a need for a nonlinear transformation. We tested this aspect through several replicates of the simulation study.

The final row examines a scenario where the interaction term is part of the data generation model but not included in the imputation model. Here, the residual density plot and the expected value-based diagnostics clearly indicate differences in patterns between observed and imputed values, whereas the scatter plot diagnostics fail to show substantial differences where the response propensities overlap. That is, the residual density and the scatter plot diagnostics call for different conclusions.

A more detailed investigation reveals that the residual density and scatter plot diagnostics target different aspects of misspecification of the imputation model. To investigate further, consider two data sets A and B with the following two regression models for y involving interaction between x_1 and x_2:

$$\mathrm{E}(y|x) = 1 + x_1 + x_2 + x_3 + x_1 \cdot x_2 \,, \tag{4}$$

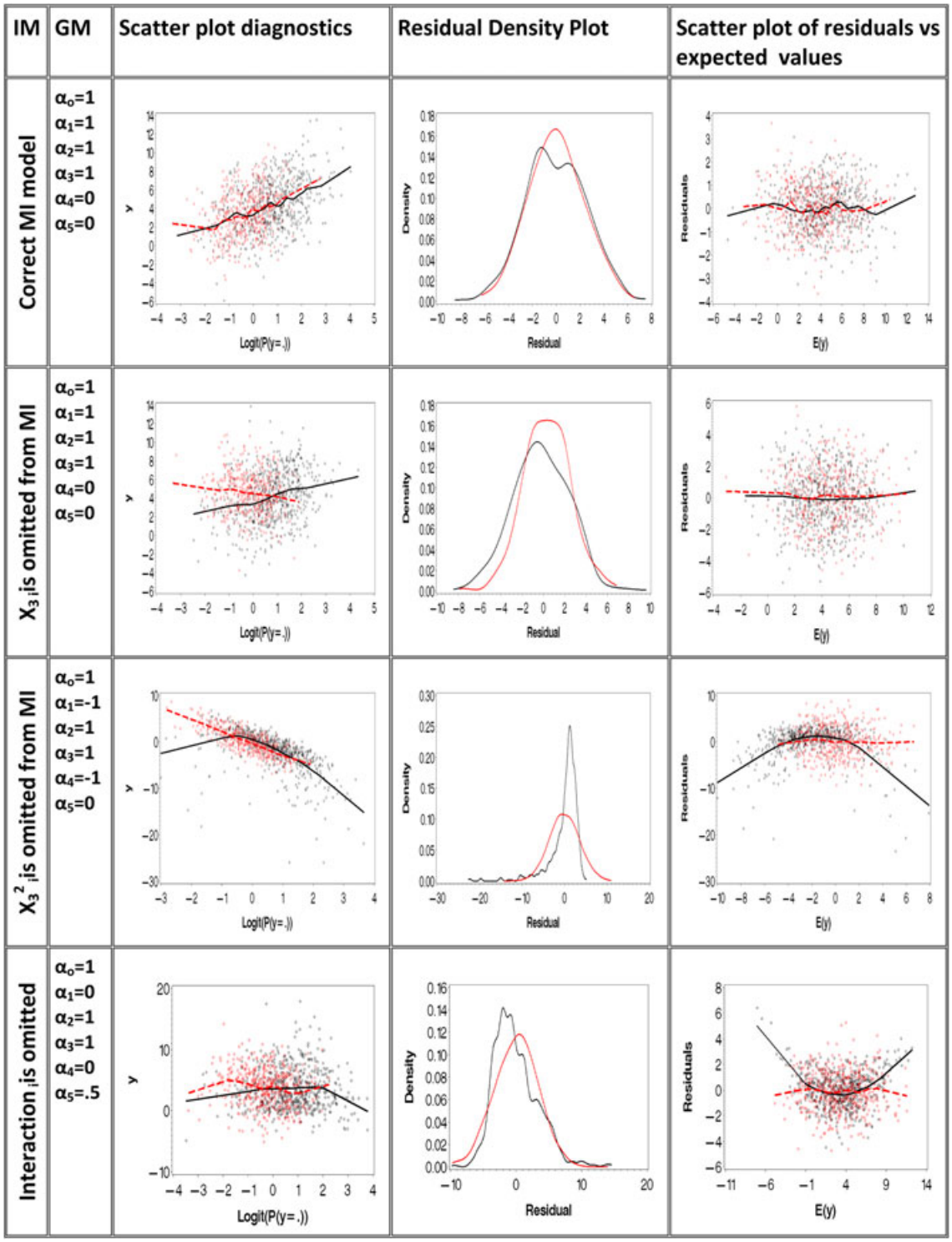

Figure 1. Diagnostics for continuous variables.

$$E(y|x) = 1 + x_1 + x_2 + x_3 - x_1 \cdot x_2 \,, \tag{5}$$

As before, some values of y were set to missing in correspondence using the logistic model (3) and then multiply imputed both data sets assuming that values of y depend on the linear combination of x_1, x_2, and x_3 but ignoring the interaction term $x_1 \cdot x_2$.

After the imputations, the marginal and conditional distributions of the original values of y that were set to missing and the corresponding imputed values were compared. The first and second columns of

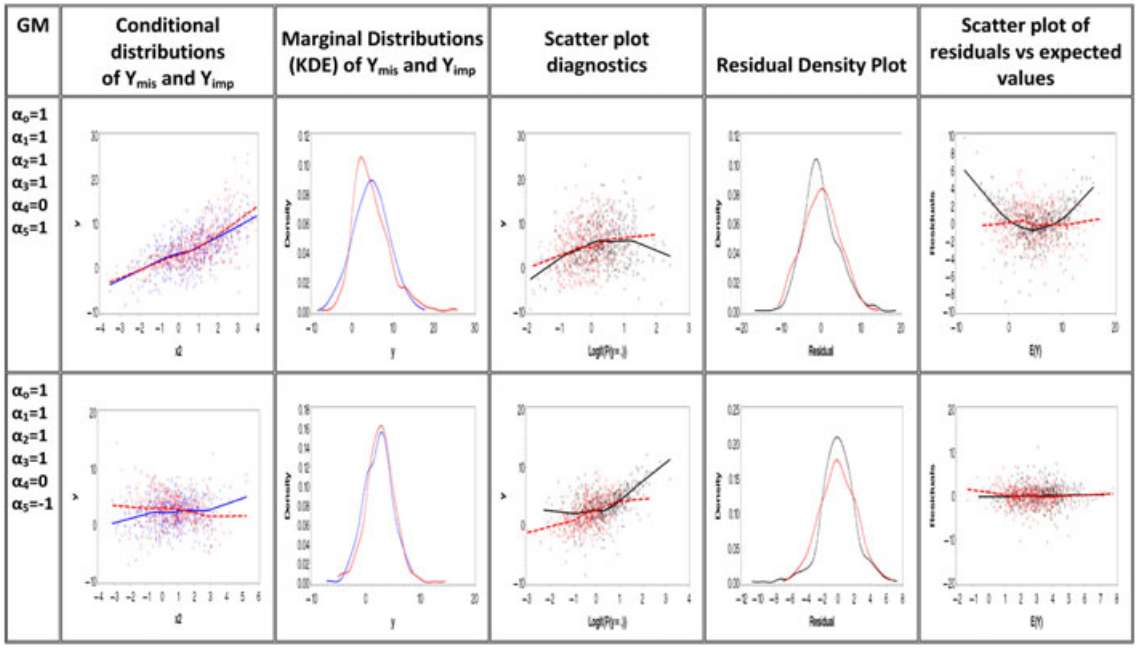

Figure 2. Targets for scatter plot and residual density diagnostics.

graphs in Figure 2 show the conditional and marginal distributions of the original (red) and imputed (blue) values of y. The imputations for data set A (4) preserved conditional distributions of y given x_3 but distorted the marginal distribution of the residuals. The imputations of data set B (5) retained the marginal distributions of y but distorted the relationship of y and x_3.

The plots comparing the marginal distributions of the original and imputed values of y are shown in the second and third columns of Figure 2. The proposed diagnostic tools as well as expected value-based diagnostics were applied to check whether they are equally capable to detect flaws in the imputations of data sets A and B. As shown in the last three columns of Figure 2, the scatter plot is effective in recognizing problems with the conditional distributions of imputed values, and the residual density plot targets the marginal distributions of the residuals. If conditional distributions are preserved, but the marginal distributions of the residuals are distorted by imputations (as in data set A), then the residual density plots and the expected value-based diagnostic are helpful to diagnose the problem. If marginal distributions are similar, but the conditional distributions are distorted (data set B), then the scatter plots are useful. Thus, it is important to apply both residual density diagnostics and the scatter plot diagnostics, as they detect different types of misspecfications.

4.2. Simulations for binary variable

The simulation study for a binary variable consisted of the same four steps as for the continuous variable. Binary variable, y, was generated by the logistic regression model (6).

$$\text{logit}(P(y=1|x)) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_1^2 + \alpha_5 x_1 \cdot x_2 \,, \tag{6}$$

As in the continuous case, values x_1, x_2, x_3 are independent draws from standard normal distributions and fully observed. Missing mechanism for y was imposed by Equation (3).

The values of y were set to be missing and then imputed using IVEWARE. Again, for this simple case, the imputations are the draws from the posterior predictive distribution under the binomial-logistic regression model with non-informative prior for the regression coefficients with the posterior distribution of the parameters being approximated by a multivariate normal distribution with the maximum likelihood estimate as the mean and the inverse of the observed information matrix as the covariance matrix. Imputation models were prone to various degrees of misspecification, including omitted linear, quadratic, or interaction terms. The results are illustrated in Figure 3.

The first row corresponds to the correct imputation where the imputation model is consistent with the data generating model. The scatter plot diagnostic, in the first column, shows that observed (black)

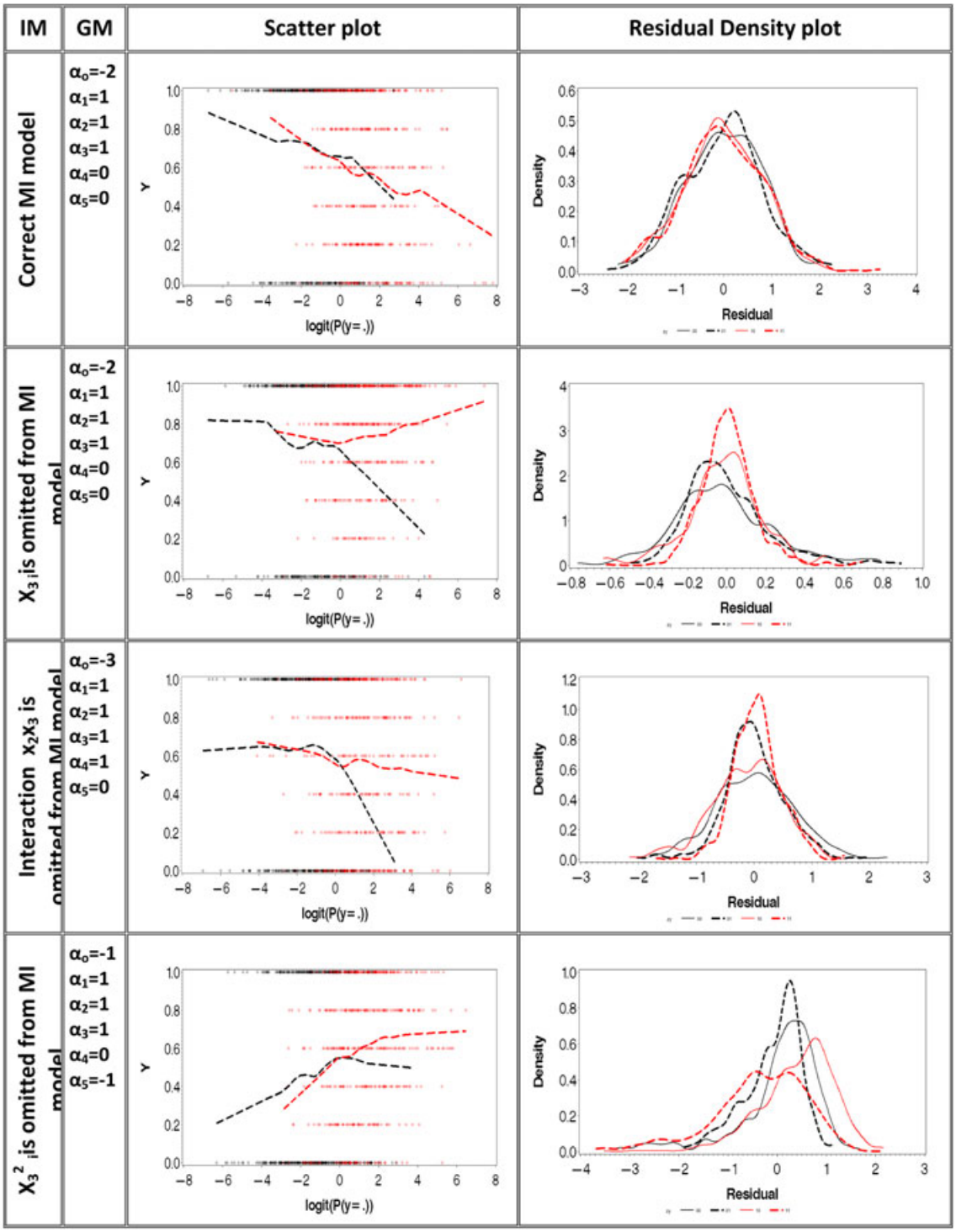

Figure 3. Diagnostics for binary variables.

and imputed (red) values are very similar. Residual density diagnostic shows a set of four kernal density curves. Again, the observed values are depicted by the black lines and imputed by the red lines. Solid lines of both colors represent observations with $y = 0$, and the dashed lines represent $y = 1$. The kernel densities in the first row are very similar across all four subgroups. Thus, both diagnostics confirm the validity of the imputations.

The second row of graphs corresponds to the omitted variable scenario. LOESS curves are clearly different for the observed and imputed values of y. For the third row, a case of ignored interaction, the

scatter plot diagnostic, clearly diagnoses a problem with the imputations. On the other hand, the residual density diagnostic are not very informative.

The last row shows a scenario where a square term for x_3 is present in the data generation model but is omitted from the imputation model. Here, the scatter plot diagnostic is rather ambiguous. LOESS curves for the observed and imputed values of y coincide in the area of the most dense overlap of propensities for respondents and nonrespondents, and differ on the sides. However, residual plot clearly shows the difference in distributions for nonrespondents for whom y was imputed to be 0 and for those with y imputed as 1. The last two scenarios indicate discrepancy between conclusions from scatter plot diagnostic and residual density diagnostic. As for the continuous variable, it is important to check both diagnostic tools to evaluate the imputations.

4.3. Assessment of uncongeniality

An analyst can detect uncongeniality by constructing scatter plot diagnostic and residual density diagnostic using the residuals from the analyst model. Suppose that the analyst using a regression model that involves Y_v and has some imputed values. Suppose that the analyst model was applied to the imputed data set l yielding the residuals $r^{(l)}$. If the imputations are 'agreeable' with the analyst model, then under MAR assumption, in each imputed data set l distribution of the residuals $r^{(l)}$, conditional on the estimated propensity of response, must be similar for $R_v = 1$ and $R_v = 0$ groups.

To create residual density diagnostic, first regress $r^{(l)}$ on the propensity score $\widehat{e}_{obs,-v}$ and construct the residuals $u^{(l)}$ from this model. Next, or each imputation l generate the kernel density plots for $u^{(l)}$ by R_v. If imputation model is congenial to the analyst model, then the marginal distributions of the residuals $u^{(l)}$ must be similar between the two groups, respondents, and nonrespondents.

To further elaborate on this approach, we simulated data for the congenial and uncongenial scenarios. Specifically, an analyst is interested in a regression model for y as a linear function of x_1, x_3 and their product. The analyst believes that interaction between x_1 and x_3 is important. However, the analyst has no knowledge if the interaction was included in the imputation model and wants to assess if the imputations are valid under the posited model.

We simulated the data under the analyst model, reset some values of y to missing, and then multiply imputed the two data sets. One set was imputed incorporating interaction $x_1 \cdot x_3$ into the imputation model. The second set was imputed ignoring the interaction. Figure 4a shows that the scatter plot and residual density diagnostics for the congenial models were imputation model included in the interaction. Neither plot shows any differences between distributions for the residuals based on the observed and imputed values. However, the plots in Figure 4b, corresponding to the imputations omitting the interaction term, show differences between the observed and imputed values. These differences allow the analyst to conclude that imputations are not reasonable under the analyst model.

We conducted several other simulation studies varying the missingness rate, alternative models for both regression model for the outcome as well as missing data mechanism. Across all these simulation studies, the proposed diagnostic procedure detected modest to severe discrepancies. When the residual variance in the outcome regression model is large compared with systematic bias because of underfitting, the residual diagnostic plots were better in detecting the problems than the scatter plot diagnostic. Generally, together, they practically identified problems when they exist.

5. Numerical tests

Though the emphasis of this paper has been on graphical approach, we also propose two test procedures that mirror the graphical diagnostic tools. These are useful to formally test the validity of the imputations using the significance testing framework.

As indicated earlier, a numeric test analogous to the scatter plot diagnostics can be performed applying the ANOVA technique. A similar approach has been used to ensure balance of covariates on the estimated propensity score [17]. To apply the ANOVA technique to diagnose problems, the estimated probabilities of response $\widehat{e}_{obs,-v}$ are grouped into H strata. The ANOVA model includes stratum ($H-1$ degrees of freedom), indicator for observed/imputed (1 degree of freedom), and their interactions ($H-1$ degrees of freedom) as factors. It has been shown that with five strata, 90% of the bias is removed [19,25] Thus, we use $H = 5$ as default guidance for this test. The ANOVA test basically uses a F-statistic that compares the full model with the null model that drops the missing data indicator and stratum by indicator interaction term.

As a numerical analog of kernel density estimate (KDE) diagnostics, we propose using the KS test

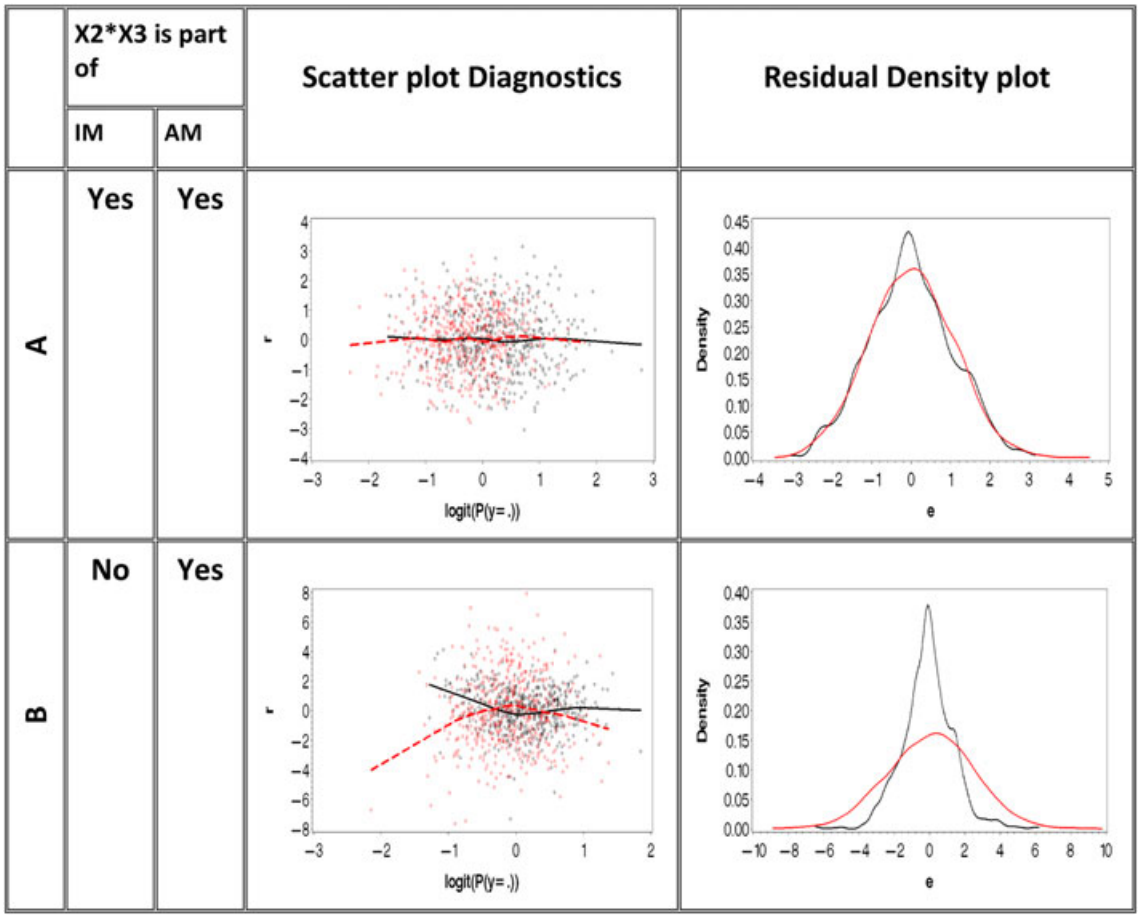

Figure 4. Use of the scatter plot and residual density diagnostic to check congeniality.

for equality of the distribution of the residuals for the imputed and observed units conditional on the estimated propensity of response. Define the residuals $\boldsymbol{r}_{sv} = y_{sv} - \hat{y}_{sv}$, where $\hat{y}_{sv}$ is the prediction from the regression of Y_v on $\widehat{e}_{obs,-v}$. Compute the KS statistic comparing the residuals for $R_v = 1$ with $R_v = 0$. Thus, the ANOVA test would formally test the equality of the location, and the KS test examines equality of the full residual distribution.

As indicated earlier, because of correlation between the imputed and observed values, the standard sampling distributions for the ANOVA and KS procedures cannot be applied. It is difficult to analytically derive the actual sampling distributions. Instead, we conducted a simulation study by generating data under both null and non-null scenarios, applied the test procedures on each multiply imputed data set, and determined the cut points under the null scenarios that resulted in the exact and nominal level to be the same. We then used the same cut points for the non-null scenarios to determine the power of these tests.

We generated 500 data sets under the simulation scenarios described in Section 4. We also added modest and severe misspecification. Table I provides the true and imputation models used in the simulation.

We calculated the number of rejections of the null hypothesis for each test in every imputed data set. Based on these empirical distributions, we established a rule based on the number data sets in which the hypothesis should rejected (under the null model) to ensure that type I error is close to the nominal level (we chose 0.05). For both the tests, the following two simple rules of rejecting the imputation model worked best:

(1) Rule 1: Reject the imputation model if the ANOVA test is rejected in at least two of the five imputed data sets.
(2) Rule 2: Reject if at least KS is rejected in at least one of the five imputed data sets or ANOVA test is rejected in at least two of the five imputed data sets.

Table I. True and imputation models used in determining the exact and power of ANOVA and Kolmogorov–Smirnov procedures. Residual variance is 1 for all seven true models.

Model number	Type of model	True model	Imputation model
1	Null	$E(y) = 1 + x_1 + x_2 + x_3$	$y \sim N\left(\alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3, \sigma^2\right)$
2	Omitted	$E(y) = 1 + x_1 + x_2 + x_3$	$y \sim N\left(\alpha_o + \alpha_1 X_1 + \alpha_2 x_2, \sigma^2\right)$
3	Variable	$E(y) = 1 + x_1 + x_2 + 0.5x_3$	Same as above
4	Interaction	$E(y) = 1 + x_1 + x_2 + x_3 + x_2 x_3$	$y \sim N\left(\alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_2 x_3, \sigma^2\right)$
5	Omitted	$E(y) = 1 + x_1 + x_2 + x_3 + 0.5x_2 x_3$	Same as above
6	Square term	$E(y) = 1 + x_1 + x_2 + x_3 + x_3^2$	$y \sim N\left(\alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3, \sigma^2\right)$
7	Omitted	$E(y) = 1 + x_1 + x_2 + x_3 + 0.5x_3^2$	Same as above

Table II. The exact level and power (both in %) based on the two rules for rejecting the imputation model.

Model	Rule 1	Rule 2
1	5.2	6.2
2	98.8	99.4
3	50.4	64.4
4	37.8	87.6
5	36.4	72.8
6	86.4	100
7	13.8	72.2

Table II provides the exact level (under the null) and power (under the alternative) based on the 500 simulated data sets.

The results in Table II suggest that though using just the ANOVA test may be sufficient to detect large departures from the true model, but using both KS and ANOVA tests add considerable power to detect relatively modest differences between the true and imputation models.

6. Conclusion

The major emphasis of this paper is to propose a set of graphical diagnostic tools that adopt propensity score methods to assist in assessing suitability of MI for an analyst without the full knowledge of the imputation model. Additionally, the proposed diagnostics can be used by an imputer to check the validity of the imputations from the working model and refine it, if necessary. The central theme is the comparison of the imputed and observed values conditional on the estimated propensity of response even with missing values in other variables.

Two proposed diagnostic tools help to evaluate different features of imputed values. The scatter plot diagnostics assist in comparison of the conditional distributions, whereas the residual density plots are useful in comparison of means as well as second and third moments of the residual distributions. It is important to address both issues when evaluating MI. We also proposed and evaluated some simple rules for rejecting the imputation model using the significance testing framework.

Although the focus was on binary and continuous variables, the method could be extended to count or semi-continuous variables. For example, in the case of semi-continuous variables, the diagnostics can be carried out in two parts. The first part uses the binary variable diagnostic tools to assess the validity of the imputations of zero/non-zero status. Next, conditional on the binary variable being non-zero, one can use the continuous variable approach to assess the validity of imputed continuous values. The methodology can be easily implemented in standard software packages such as R or Stata and has already been used in a complex setting [26] where they use limited techniques in the non-peer reviewed technical report [27]. This report, considered mainly continuous variables, did not provide theoretical underpinnings, refinements for binary variable, and the numerical tests. The present article also carries out a more thorough evaluation of the proposed tools.

The proposed approach can be used by an analyst who was not involved in the imputation process and seeks to assess congeniality of the imputation and analyst models. If the diagnostic procedures indicate problems with the imputation, then it may be prudent for the analyst to ignore the imputed values and adopt alternative approaches such as MI analysis by re-imputing the missing values in just the variables in the analyst model, the maximum likelihood, or the fully Bayesian analysis. A comparison of the results with original and re-imputed values might be useful to quantify sensitivity of inferences to model misspecification. Such analysis across the data sets might be useful to assess the impact of uncongeniality on MI inferences.

There are number of limitations that can be addressed with further research. First, the approach assumes that the data set includes all the variables used in the imputation process. In some applications, the imputer may use some internally available variables, and it is possible that our diagnostics may indicate problems where there are none. However, as long as all variables used in the imputations are used in propensity score estimation, the inferences should be valid. Second limitation is the assumption that the data are MAR. The problems identified by the procedures may be because of nonignorable missing data mechanism rather than problems with the imputation model. The third limitation is that while focusing on variable Y_v, the estimated propensity score is constructed by averaging over the imputations of all other variables Y_{-v}. If the models imputing those variables are severely misspecified, then the estimated propensity score may be affected. One solution is to apply the diagnostics sequentially and iteratively for all the variables until all the models show reasonable fit across all variables. Nevertheless, further work is needed to address these limitations.

References

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New Tork, New York, 1987.
2. Li KH, Raghunathan TE, Rubin D. Large sample significance levels from multiply-imputed data using momentbased statistics and an f-reference distribution. *Journal of American Statistical Association* 1991; **86**:1065–1073.
3. Li KH, Meng X, Raghunathan T, Rubin DB. Significance levels from repeated p-values with multiply imputed data. *Statistica Sinica* 1991; **1**:65–92.
4. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999; **86**(4):948–955.
5. XL Meng X, Rubin DB. Performing likelihood ratio tests with multiply imputed data sets. *Biometrika* 1995; **79**:103–111.
6. Raghunathan TE, Lepkowski JM, Hoewyk JV, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**(1):85–95.
7. Berglund P, Heeringa SH. *Multiple Imputation of Missing Data Using SAS*. SAS Institute: Cary, North Carolina, 2014.
8. van Buuren S. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC: Boca Raton, Florida, 2012.
9. Stata. *Multiple-Imputation Reference Manual*. Stata Press: College Station, Texas, 2015.
10. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**(4):538–558.
11. Kim JK, Brick JM, Fuller WA, Kalton G. On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006; **68**(3):509–521.
12. Robins JM, Wang N. Inference for imputation estimators. *Biometrika* 2000; **87**(1):113–124.
13. Aboyomi K, Gelman A, Levy M. Diagnostics for multiple imputations. *Applied Statistics* 2008; **57**(3):273–291.
14. He Y, Zaslavsky AM, Harrington DP, Catalano P, Landrum MB. Multiple imputation in a large-scale complex survey: A practical guide. *Statistical Methods in Medical Research* 2010; **19**(6):653–670.
15. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research* 2011; **20**(1):40–49.
16. Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology* 2009; **169**(9):1133–1139.
17. Rosenbaum P, Rubin D. The central role of the propensity score in the observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
18. Ralph B, D'Agostino J, Rubin DB. Estimating and using propensity scores with partially missing data. *JASA* 2000; **95**(451):749–759.
19. Rosenbaum P, Rubin D. Reducing bias in observational studies using subclassification on the propensity score. *Biometrika* 1984; **79**(387):516–524.
20. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. *The Indian Journal of Statistics, Series A* 1973; **35**(4):417–446.
21. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. Wiley: New York, 1989.
22. Rubin DB. Graphical methods for assessing logistic regression models: comment. *Journal of the American Statistical Association* 1984; **79**:79–80.
23. Boonstra P, Bondarenko I, Park S, Voconas P, Mukherjee B. Propensity score-based diagnostics for categorical response regression models. *Statistics in Medecine* 2014; **33**(3):455–469.
24. Raghuthan TE, Solenberger PW, Van Hoewyk JH. *IVEware: Imputation and Variance Estimation Software, Survery Methodology Program*. Institute for Social Research, University of Michigan: Ann Arbor, Michigan, 48106, USA, 2002. (Download: www.iverware.org).

25. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhy: The Indian Journal of Statistics, Series A* 1973; **35**:417–446.
26. Robbins MW, Ghosh SK, Habiger JD. Imputation in high-dimensional economic data as applied to the agricultural resource management survey. *Journal of the American Statistical Association* 2013; **108**(501):81–95.
27. Raghunathan TE, Bondarenko I. Diagnostics for multiple impuation. *SSRN* Technical Report, Socail Science Research Network Rochester, New York, 2007. http://ssrn.com/abstract=1031750 [accessed on 21 November 2007].