Published in final edited form as:

Stat Methods Med Res. 2018 June ; 27(6): 1683–1694. doi:10.1177/0962280216667763.

Multiple Imputation with non-additively related variables: Joint-modeling and approximations

Soeun Kim¹, Thomas R. Belin², Catherine A. Sugar²

¹Department of Biostatistics, University of Texas Health Science Center

²Department of Biostatistics, University of California Los Angeles

Abstract

This paper investigates multiple imputation methods for regression models with interacting continuous and binary predictors when continuous variable may be missing. Usual implementations for parametric multiple imputation assume a multivariate normal structure for the variables, which is not satisfied for a binary variable nor its interaction with a continuous variable. To accommodate interactions, missing covariates are multiply imputed from conditional distribution in a manner consistent with the joint model. Alternative imputation methods under multivariate normal assumptions are also considered as candidate approximations and evaluated in a simulation study. The results suggest that the joint modeling procedure performs generally well across a wide range of scenarios, and so do the approximation methods that incorporate interactions in the model appropriately by stratification. It is critical to include interactions in the imputation model as failure to do so may result in low coverage and bias. We apply the joint modeling approach and approximation methods in the study of childhood trauma with gender trauma interaction.

Keywords

Multiple Imputation; missing covariate; interaction; binary predictor; multivariate normal assumption; joint modeling

1. Introduction

Multiple imputation(MI)^{1–3} is one of the most widely used methods for inference in the presence of incomplete data. Availability of MI tools in standard software packages^{4–6} has made it popular for practitioners to fill in missing observations with several plausible values and then analyze the resulting data sets with standard methods for complete data.

Theoretical insights underscore the importance of a multivariate perspective on generating multiple imputations⁷. Examples include incorporating the outcome when predicting for a missing covariate, and including auxiliary variables to improve precision of imputations⁸. Similarly, when there are categorical predictors or non-additive relationships, e.g.

Reprints and permission: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Soeun Kim, Department of Biostatistics, University of Texas Health Science Center, 1200 Pressler St., Houston, TX 77030, USA, Soeun.Kim@uth.tmc.edu.

interactions or transformations in the model, the relationship between the variables would not be strictly multivariate normal as is assumed in most MI implementations under parametric imputation models. This paper aims to develop MI methods that carefully account for such structure, using as an illustration the case when the underlying substantive model involves an interaction between binary and continuous variables.

The importance of using appropriate methods for dealing with missing covariates has long been discussed in the literature⁹. The scenario with continuous predictors was handled in Kim et al.¹⁰, and this paper is specifically concerned with missing covariates for regression models with interacting continuous and binary predictors in the presence of interactions. This setting arises in a number of contexts, and one example where this would be applicable includes a randomized clinical trial (RCT) in which the treatment effect is related to a continuous covariate. In RCTs that compare a treatment and a control group, it is often the case that the treatment assignment is fully observed but other continuous predictors e.g. potential treatment modifiers, are incomplete, implying that the interaction variable among the two is also missing. With non-additive relationships among variables, obtaining valid inference for the underlying parameters for multivariate distribution of the data poses special challenges.

In this paper we propose a general method for imputing missing values of the missing covariate and interactions by deriving their conditional distribution under a joint modeling (JM) framework. We also investigate MI strategies that can be implemented with readily available software using a multivariate normal model. Building on a regression setting with two predictors and their interaction, the conditional distribution for the missing covariate is derived and used in a simulation study to evaluate the properties of estimators and to compare several approximation methods. This paper is concerned with the case in which there is a completely observed discrete predictor and a continuous predictor which may be missing. In an interaction model, when covariates are missing, the corresponding interaction term related to the missing covariates is also missing. Even if the continuous predictors are normally distributed, the presence of categorical predictors or non-additive relationships will make the joint distribution under substantive model not multivariate normal. Note, however, in the case of interaction between binary predictor and continuous predictor, the stochastic dependence disappears when conditioning on the binary variable, so imputations can be produced after stratification.

Studies have considered the extent to which the multivariate normal approximation is robust for imputation of incomplete binary data, noting that it can give good performance when adjusted appropriately, for example by adaptive rounding, or by including it in the imputation model as a dummy variable. However, there are inherent limitations given the imperfection of a normal approximation to a Bernoulli distribution^{4;11–14}. With a completely observed binary predictor, there is no need for imputation of binary variables, but the binary variable and the interaction term must still be incorporated appropriately in the imputation model. We use the derived conditional distribution to handle the binary variable and the interaction in a manner consistent with the joint model.

Several approximation methods are also considered. An alternative strategy involves including the binary variable as a dummy variable ignoring its non-normality and including the interaction term as just another variable in the imputation model under multivariate normal assumption. With a completely observed binary indicator, the interaction is already known for the observations when indicator has value zero, and the analyst may or may not fill in the zero values prior to imputation. Another alternative strategy is to impute the missing covariate and recalculate the interaction to preserve the multiplicative relationship. Finally, one can perform stratified MI, filling in values separately for the two subgroups defined by the value of the binary variable.

The following sections present and evaluate imputation methods that accommodate an interaction in the setting with a completely observed binary predictor and missing continuous predictor, along with simulation evaluations of different procedures, a real data application, and discussion. Specifically, Section 2 derives the conditional distribution for a missing covariate and uses a Markov Chain Monte Carlo (MCMC) imputation procedure based on this conditional distribution for a regression setting with a partially missing continuous predictor and a binary predictor and their interaction. In Section 3, we present the imputation procedures based on candidate approximations, and perform simulation studies in section 4 in order to evaluate and compare the performance of imputation procedures based on JM and approximation methods. Section 5 presents an application of the procedures to a study of childhood trauma¹⁵. Discussion and recommendations are given in Section 6.

2. Methods Based on Joint Modeling Approach

In this section, we derive the conditional distribution of a missing covariate for a regression setting with a completely observed binary variable and a partially observed continuous variable and their interaction, and present the corresponding MCMC imputation procedure.

2.1 Derivation of the Conditional Distribution

We assume that X_1 is a binary variable that follows a Bernoulli distribution, X_2 is a continuous variable that follows a normal distribution, and Y has a linear regression relationship with X_1 , X_2 and their interaction $X_{12} = X_1 X_2$. Parameters of X_2 may change conditionally on the value of X_1 . We define the following notation for means and variances of each group:

$$\mu_{20} = E(X_2 | X_1 = 0), \sigma_{20}^2 = \text{Var}(X_2 | X_1 = 0)$$

$$\mu_{21} = E(X_2 | X_1 = 1), \sigma_{21}^2 = \text{Var}(X_2 | X_1 = 1)$$

Specifically, the joint distribution is assumed to be

$$X_1 \sim \text{Bernoulli}(p)$$

$$X_2 | X_1 \sim N(\mu_{X_1}, \sigma_{X_1}^2) \quad (1)$$

where μ_{X_1} and $\sigma_{X_1}^2$ can be expressed as

$$\mu_{X_1} = (1 - X_1)\mu_{20} + X_1\mu_{21}$$

$$\sigma_{X_1}^2 = (1 - X_1)\sigma_{20}^2 + X_1\sigma_{21}^2$$

since X_1 is a binary variable and takes values 0 or 1. We assume Y has a regression relationship with the predictors,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{12} + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2) \quad (2)$$

or alternatively,

$$[Y | X_1, X_2] \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{12}, \sigma_\epsilon^2).$$

Suppose Y and X_1 are completely observed but some values of X_2 and therefore X_{12} are missing. The conditional distribution of X_2 given X_1 and Y can be derived using Bayes theorem. We wish to impute these missing values using draws from the conditional distribution $[X_2, X_{12} | Y, X_1]$, and to calculate $X_{12} = X_1 X_2$ using the observed X_1 and the drawn value of X_2 . By Bayes theorem,

$$P(X_2 | X_1, Y) \propto P(Y | X_1, X_2) P(X_2 | X_1)$$

Invoking (1), the relationship becomes

$$P(X_2 | Y, X_1) \propto P(Y | X_1, X_2) P(X_2 | X_1) \propto \exp \left[-\frac{1}{2} \left(X_2^2 \left(\frac{1}{\sigma_{X_1}^2} + \frac{(\beta_2 + \beta_3 X_1)^2}{\sigma_\epsilon^2} \right) - 2X_2 \left(\frac{\mu_{X_1}}{\sigma_{X_1}^2} + \frac{(Y - \beta_0 - \beta_1 X_1)(\beta_2 + \beta_3 X_1)}{\sigma_\epsilon^2} \right) \right) \right]$$

Therefore the conditional distribution of $[X_2 | X_1, Y]$ remains normal, although the expressions for the mean and variance are complicated, with common denominators for the mean and the variance. Specifically, the conditional distribution is $[X_2 | X_1, Y] \sim N(E(X_2 | X_1, Y), \text{Var}(X_2 | X_1, Y))$, where

$$E(X_2|X_1, Y) = \frac{\frac{(1-X_1)\mu_{20} + X_1\mu_{21}}{(1-X_1)\sigma_{20}^2 + X_1\sigma_{21}^2} + \frac{(\beta_2 + \beta_3 X_1)(Y - \beta_0 - \beta_1 X_1)}{\sigma_\epsilon^2}}{\frac{1}{(1-X_1)\sigma_{20}^2 + X_1\sigma_{21}^2} + \frac{(\beta_2 + \beta_3 X_1)^2}{\sigma_\epsilon^2}} \quad (3)$$

$$\text{Var}(X_2|X_1, Y) = \frac{1}{\frac{1}{(1-X_1)\sigma_{20}^2 + X_1\sigma_{21}^2} + \frac{(\beta_2 + \beta_3 X_1)^2}{\sigma_\epsilon^2}}$$

Using the values of X_2 drawn from this distribution, the interaction term X_{12} can be calculated by multiplying the observed X_1 and the drawn X_2 . It can be shown that this is equivalent to drawing pairs from the joint conditional distribution of X_2 and X_{12} given the observed X_1 and Y .

2.2 MCMC Imputation Procedure for the Joint Modeling Procedure

The MCMC imputation procedure is based on the draws from the conditional distribution in (3). In our implementation, the initial cycle is based on the completely observed cases, which is then followed by cycles in which all observed and imputed cases from the previous step are used. The MCMC algorithm draws variance parameters given the observed data, followed by the mean parameters given the drawn variance parameters. The regression error variance and regression coefficients are drawn next, after which in the final step, a missing covariate value is drawn from the conditional distribution.

Algorithm:

Based on n cases $Y, X_1, X_2^{(t)}$ and to complete the specification needed to derive the step of an MCMC procedure, we assume flat prior distributions for location parameters and variance parameters following an Inverse-Gamma distribution. Letting

$$\bar{X}_{20}^{(t)} = \frac{1}{(1-\hat{p})n} \sum_{i=1}^n I_{X_1=1} \bar{X}_{2i}^{(t)}, \text{ and } S_{20}^{(t)} = \frac{1}{(1-\hat{p})n-1} \sum_{i=1}^n I_{X_1=1} (X_{2i}^{(t)} - \bar{X}_{20}^{(t)})^2,$$

1. Draw $(\sigma_{20}^2)^{(t+1)}, (\sigma_{21}^2)^{(t+1)}$ from:

$$\begin{aligned} (\sigma_{20}^2)^{(t+1)} & \left| X_1, X_2^{(t)}, Y \sim ((1-\hat{p})n-1) \frac{s_{20}^2}{\chi_{(1-\hat{p})n-1}^2} \right. \\ (\sigma_{21}^2)^{(t+1)} & \left| X_1, X_2^{(t)}, Y \sim (\hat{p}n-1) \frac{s_{21}^2}{\chi_{\hat{p}n-1}^2} \right. \end{aligned}$$

where $s_{20}^2 = \text{Var}(X_{20}^{(t)})$, $s_{21}^2 = \text{Var}(X_{21}^{(t)})$ and $\hat{p}$ is the probability that $X_1 = 1$.

2. Draw

$$\mu_{20}^{(t+1)} \Big| X_1, X_2^{(t)}, Y, (\sigma_{20}^2)^{(t+1)} \sim N \left(\bar{X}_{20}^{(t)}, \frac{1}{(1-\hat{p})n-1} (\sigma_{20}^2)^{(t+1)} \right)$$

$$\mu_{21}^{(t+1)} \Big| X_1, X_2^{(t)}, Y, (\sigma_{21}^2)^{(t+1)} \sim N \left(\bar{X}_{21}^{(t)}, \frac{1}{\hat{p}n-1} (\sigma_{21}^2)^{(t+1)} \right)$$

3. Draw $(\sigma_e^2)^{(t+1)} \mid X_1, X_2^{(t)}, Y \sim \text{Scaled-Inverse } \chi_{n-4}^2$

$$(\sigma_e^2)^{(t+1)} \sim (n-4) \frac{(s^2)^{(t)}}{\chi_{n-4}^2}$$

where $(s^2)^{(t)} = \text{Var}(Y - b_0^{(t)} - b_1^{(t)}X_1 - b_2^{(t)}X_2^{(t)} - b_3^{(t)}X_{12}^{(t)})$, and $b_0^{(t)}, b_1^{(t)}, b_2^{(t)}, b_3^{(t)}$ are fitted values of regression coefficients.

4. Draw $\beta^{(t+1)} \mid X_1, X_2^{(t)}, Y, \sigma_e^{2(t+1)} \sim$

$$MVN(\widehat{\beta}^{(t+1)}, \sigma_e^{2(t+1)}(X'X)^{-1}),$$

where the rows of X_i are denoted by $(1, X_{1i}, X_{2i}^{(t)}, X_{12i}^{(t)})$, and $\widehat{\beta}^{(t)} = (b_0^{(t)}, b_1^{(t)}, b_2^{(t)}, b_3^{(t)})^T$.

5. Draw $X_2^{(t+1)}, X_{12}^{(t+1)} \mid X_1, Y, \mu^{(t+1)}, \Sigma^{(t+1)}, \beta^{(t+1)}, \sigma_e^{2(t+1)}$ jointly by drawing $X_2^{(t+1)}$ from the following conditional distribution and calculating $[X_{12}^{(t+1)} \mid X_1, X_2^{(t+1)}] = X_1 \times X_2^{(t+1)}$.

$$[X_2^{(t+1)} \mid X_1, Y] \sim N(E(X_2^{(t+1)} \mid X_1, Y), \text{Var}(X_2^{(t+1)}, X_{12}^{(t+1)} \mid X_1, Y))$$

$$E(X_2^{(t+1)} \mid X_1, Y) =$$

$$\frac{(1-X_1)\mu_{20}^{(t+1)} + X_1\mu_{21}^{(t+1)}}{(1-X_1)\sigma_{20}^{2(t+1)} + X_1\sigma_{21}^{2(t+1)}} + \frac{(\beta_2^{(t+1)} + \beta_3^{(t+1)}X_1)(Y - \beta_0^{(t+1)} - \beta_1^{(t+1)}X_1)}{\sigma_e^{2(t+1)}} \\ \frac{1}{(1-X_1)\sigma_{20}^{2(t+1)} + X_1\sigma_{21}^{2(t+1)}} + \frac{(\beta_2^{(t+1)} + \beta_3^{(t+1)}X_1)^2}{\sigma_e^{2(t+1)}}$$

$$\text{Var}(X_2^{(t+1)} | X_1, Y) = \frac{1}{\frac{1}{(1-X_1)\sigma_{20}^{2(t+1)} + X_1\sigma_{21}^{2(t+1)}} + \frac{(\beta_2^{(t+1)} + \beta_3^{(t+1)}X_1)^2}{\sigma_e^{2(t+1)}}}$$

6. Iterate steps 1–5 until convergence in distribution.

With m imputations, m completed data sets are created and analyzed using standard complete data methods, then combined using Rubin's rules².

3. Imputation Procedures Based on Approximation Methods

In this section we describe approaches making use of available software for MI under a multivariate normal assumption. Although the marginal distribution of a binary variable does not conform to a normality assumption, the fact that the binary variable is fully observed means that its primary relevance is the role it plays in conditional distributions for imputing other variables. In that context, all the available information is used in the process of imputation including its contribution to the interaction effect.

We consider two main approximate approaches to imputing interaction terms. The first, referred to as Passive Imputation (PSI)^{13;14;16;17} respects the constraint that the interaction term is the product of its component terms. In PSI one simply imputes the value of the missing covariate, X_2 based on Y and X_1 and then calculates the interaction term as X_{12} . The second approach treats the interaction term as "just another variable (JAV)", ignoring the multiplicative constraint. Simulation studies have shown that under the assumption of multivariate normality, the unconstrained JAV approach has good coverage properties for the case when the interaction is the product of two continuous variables^{13;16;17}. We will employ a similar approach for the case when one of the interaction components is binary although the violation of the multivariate normal assumption is more severe, making it unclear whether the JAV approach will perform as well in this context.

In the special case when one of the components of the interaction is an indicator variable, two other factors may be considered. First, the interaction term X_{12} is known to be zero whenever the indicator, $X_1 = 0$, and is equal to the value of X_2 when $X_1 = 1$. One thus has the option of filling in $X_{12} = 0$ for cases with $X_1 = 0$ and X_2 missing before performing the remaining imputation and analyses. Second, the data can be stratified so that the imputation model is different depending on the value of the binary variable. This correctly accommodates the interaction structure, (in analogy with using stratified regression to allow for interaction of one categorical predictor with all other covariates) but may result in a loss of efficiency since only a subset of the sample is used to impute each of the missing observations.

These considerations yield eight alternative MI methods for approximating the JM approach using standard multivariate normal MI approaches. Specifically, we consider constrained (PSI) vs unconstrained (JAV), filling in or not filling in 0's and stratifying or not stratifying the imputation. These approaches are denoted as follows:

1. JAV : Just another variable^{13;14} (also known as Impute all Unconstrained¹⁷) - Impute X_2 and X_{12} under a multivariate normal model, but without requiring that the imputed value of X_1 , X_2 be equal to be the product of X_1 and the imputed value of X_2 .
2. S-JAV : Stratified JAV, also known as Impute All Stratified (IAS)¹⁷ - Impute X_2 and X_{12} separately for two subgroups resulting from the binary variable X_1 .
3. PSI : Passive imputation, also known as Impute and Calculate Unstratified - Impute X_2 only and calculate $X_1 \times X_2$.
4. S-PSI : Stratified PSI, also known as Impute and Calculate, Stratified (ICS)¹⁷ - Impute X_2 only and calculate X_{12} separately for two subgroups resulting from the binary variable X_1 .
5. JAV0: JAV with interactions filled in with zero for $X_1 = 0$.
6. S-JAV0: S-JAV with interactions filled in with zero for $X_1 = 0$.
7. PSI0: PSI with interactions filled in with zero for $X_1 = 0$.
8. S-PSI0: S-PSI with interactions filled in with zero for $X_1 = 0$.

JAV imputes X_2 and X_{12} with all four variables (Y , X_1 , X_2 , X_{12}) in the imputation model. S-JAV stratifies and performs imputation separately for two groups resulting from $X_1 = 0$ vs $X_1 = 1$. The imputation model for PSI consists of three variables (Y , X_1 , X_2), and X_{12} is calculated based on imputed values of X_2 and observed indicator X_1 . Similarly, S-PSI constrains the interaction by imputing X_2 and calculating X_{12} based on stratified data. Methods 5–8 are variations of 1–4 but with zeros filled in as ‘known’ values of interaction term for all the cases with $X_1 = 0$. In summary, JAV and S-JAV do not necessarily constrain the interaction to be a product of the two variables, but PSI and S-PSI preserve the relationship between variables by calculation.

4. A Comparison of Imputation Procedures based on Joint Modeling Approach and Approximation Methods

The regression relationship (2) implies that Y is a mixture of normals with two components according to the value of X_1 . We consider a wide range of scenarios allowing different mixture proportions as well as different mean and variance assumptions. The overall sample size values, the nature of the missing data mechanism, and the percentage of missing covariate values are also varied in the simulation to evaluate the performance of procedures across a range of conditions. The data are assumed to be either missing completely at random (MCAR) or missing at random (MAR) where the MAR mechanism is generated by allowing the percentage of missing values to be different for the two subgroups dependent on the binary variable. In the MAR condition, the missingness percentage is assumed to be twice as large for the group with $X_1 = 1$ compared to the group with $X_1 = 0$, with overall missingness percentage being either 20 percent or 50 percent. True regression coefficients are varied in the interest of creating qualitatively different mixture distributions for Y , with the error variance taken to be either 1 or 16 to reflect degrees of overlap between mixture

components in the distribution of Y . The resulting scenarios are illustrated by histograms in Supplementary materials. In summary, the following are the factors varied for the simulation:

- Proportion of cases with $X_1 = 0$: 0.3 or 0.5
- Means, variances of $X_2|X_1$: (0,1)/(0,1), (0,1)/(1,1), (1,1)/(1,1)
- Coefficients in regression of Y on X_1, X_2, X_{12} : $\beta_0, \beta_1, \beta_2, \beta_3 = (1,1,1,1), (1,3,1,1), (1,10,1,1), (1,1,1,2)$
- Error Variance = 1 or 16
- Mechanism : MCAR, MAR
- Sample size: 100, 500, 1000
- Percent missing: 20, 50

A factorial design on these parameters results in 576 scenarios, each of which is replicated 200 times. Bias, confidence interval(CI) width and coverage of nominal 95% CI's for regression coefficients are recorded to evaluate the performance of the various procedures. The JM procedure was implemented in R, and the eight multivariate normal approximation methods were implemented using SAS PROC MI and MIANALYZE. For each replicate data set the corresponding twenty imputations were drawn, followed by the regression analyses of the imputed data sets with the results then combined to form a single inference.

Table 1 presents the results for confidence interval coverage for each of the procedures averaged across all scenarios. Tables 2 and 3 provide summaries of the bias and confidence interval widths, respectively. The coverage probabilities for JM approach range from 0.944 to 0.953 for the four parameters $\beta_0, \beta_1, \beta_2$, and β_3 , with mean bias ranging from -0.009 to 0.006 , and mean CI widths between 1.13 and 1.66. Among the other procedures, JAV, S-JAV and S-PSI tend to yield close to nominal coverage. PSI had coverage ranging from 0.762 to 0.943, suggesting that it can result in lower coverage compared to other methods. As expected, PSI0 yields similar results as PSI because filling in the interactions with zeros for the $X_1 = 0$ group does not substantially affect the results, since X_{12} is not included in the imputation model and the interaction is calculated after imputation. For the stratified analyses, the SAS PROC MI procedure excludes the variable X_{12} in the $X_1 = 0$ subgroup from the analyses since all its observed values are identical, which often happens because $X_{12} = 0$ for all cases with $X_1 = 0$. This makes some of these candidate procedures to be equivalent to one another. For example, S-JAV is equivalent to S-PSI for the $X_1 = 0$ group, although not necessarily for $X_1 = 1$, and therefore the two procedures yield similar results. Similarly, S-PSI0 is equivalent to S-PSI when $X_1 = 0$, and S-JAV and S-JAV0 yield the same results as expected. Surprisingly, the only difference made by imputing zeros for interactions is for JAV0, which results in lower mean coverage and higher mean bias across the scenarios.

PSI and PSI0 have largest mean bias for each coefficient across all methods, with mean values between -0.282 and 0.162 . Overall, CI widths tend to be similar across the methods, except for some cases where JAV and JAV0 results in slightly greater widths. JAV yields

larger bias than JM, but greater widths explain the good coverage probabilities for JAV. In summary, the imputation procedure based on draws from the conditional distribution of missing covariate under the JM approach has good performance overall, as do procedures that take into account the interaction term in the imputation model through stratification.

Figure 1 presents confidence interval coverage for all the scenarios considered in the simulation study based on each of the 576 scenarios. An ideal portrait would feature coverage at the nominal 95% level across all scenarios; deficits in coverage across multiple scenarios which shared characteristics appear as gaps in the pictures. The coverage plots illustrate that JM, S-JAV, S-PSI, S-JAV0, S-PSI0 are robust in terms of various factors considered in the simulation and results in stable nominal coverage across the scenarios. PSI and JAV0 have fluctuations in coverage probabilities, and JAV tends to work well in most scenarios but not all of the time. Investigation of the results broken down by various simulation factors suggest that JM, S-JAV, S-PSI, S-JAV0, S-PSI0 tend to be generally robust to factors considered for simulation, whereas JAV, JAV0, PSI, and PSI0 yields below nominal coverage for some scenarios. Generally, the factors that seemed to affect the results were: error variance = 1 (consistent with more pronounced bimodality in the mixture distributions for Y), larger percent missing data, and larger sample size (where bias can be expected to be a larger proportion of mean squared error) lead to lower coverage. In conclusion, the results suggest that JM and stratified procedures are generally robust across all parameters $\beta_0 - \beta_3$ for all the scenarios considered. However, PSI, PSI0, and JAV0 are sensitive to the scenarios, leading to lower confidence interval coverage.

5. An Application to the Study of Childhood Trauma

Next, we present an application using the Core Data Set from the National Child Traumatic Stress Network (NCTSN)¹⁵. The data set contains records of 14,088 subject participants seen at one of 56 participating clinics in the United States. The UCLA Post Traumatic Stress Disorder Reaction Index (PTSD-RI)^{18;19}, an instrument that measures post-traumatic-stress symptoms for each child, is an important clinical indicator and a key outcome in the study. We are particularly interested in a subgroup of children between the ages of 7 and 18 who have confirmed trauma exposures on the Networks Trauma History Profile form, an instrument that records for each year of childhood the presence or absence of any of twenty possible trauma types²⁰, and who also have complete PTSD-RI scores. A total of 6,291 youths, comprised of 2,794 boys and 3,497 girls, are included in this analysis. Similarly to Layne et al.²¹, who investigate the total number of traumas experienced by an individual, we calculate the average number of traumas per year (labeled here as *Num_Traumas/Yr*) experienced by a child as a measure of trauma exposure.

It was of substantive interest to NCTSN investigators to summarize evidence regarding any gender difference in the regression relationship between the PTSD-RI score and *Num_Traumas/Yr*, i.e., whether there is an interaction between gender and *Num_Traumas/Yr*. A subject could have experienced multiple types of trauma within a year, so the number of trauma per year could range between 0 and 20, although most individuals do not report more than a few trauma types in a given year. In our analysis, the mean value of *Num_Traumas/Yr* was 0.960, and the standard deviation was 1.06. The outcome variable,

PTSD-RI score, and the binary predictor, gender, are fully observed, but the Trauma History Profile was not complete for 47% of the sample. We perform MI for the missing covariate and the interaction term and fit the regression model:

$$PTSD - RI = \beta_0 + \beta_1 Gender + \beta_2 Num_Traumas/Yr + \beta_3 Gender \times Num_Traumas/Yr$$

where gender is coded 1 for boys and 0 for girls. The JM approach derived in section 2, as well as the candidate approximation methods and complete case analyses was applied to the data.

- JM: Imputation with joint modeling approach using the conditional distribution derived in section 2.
- JAV: The binary variable gender is included in the imputation model as a dummy variable, and the interaction is imputed without forcing a constraint to preserve the relationship of the interaction as a product of the predictors.
- S-JAV: Impute with all variables in the imputation model, separately for males and females without constraining the interaction to be a product.
- PSI: Impute missing *Num_Traumas/Yr* only without the interaction term in the imputation model and calculate the interaction, constraining interaction to be a product.
- S-PSI: Impute *Num_Traumas/Yr* only, separately for boys and girls, and calculate the interaction by taking the product.
- JAV0 : Fill in interactions with the value '0' for girls prior to performing JAV
- S-JAV0: Stratified JAV with zeros imputed for girls
- PSI0: Fill in interactions with the value '0' for girls prior to performing PSI
- S-PSI0: Stratified PSI with zeros imputed for girls
- CC: Complete Case Analysis, i.e. use only the complete cases, deleting observations that contain missing data.
- CC0: Complete Case Analysis with zeros imputed for girls, i.e. fill in zeros first, then use only the complete cases, deleting observations that contain missing data.

The results are presented in Table 4, where the parameter estimates in the model can be compared across procedures. The results point to a differential effect of *Num_Traumas/Yr* by gender. Specifically, there is a widening gender gap (girls worse off than boys) as *Num_Traumas/Yr* increases, although both genders show an increasing relationship between experience and symptoms. The signs of the parameter estimates are in the same direction across the procedures with comparable magnitudes for the most part. The gender×trauma interaction is significant at 0.05 level ($p=0.02$) for JM. S-JAV, S-PSI and JAV0 give a marginally significant ($p=0.09$) interaction effect, and JAV, PSI, PSI0 and CC yield larger p -values with smaller magnitude of the coefficient for interaction term. As reflected in the

simulation results in the previous section, PSI/PSI0 tends to behave differently from other procedures, with results deviating the most from the JM.

Without the full data available to cross check the results, it is not possible to formally assess the performance in comparison with the true parameter values. Simulation results from the previous sections suggest that JM is generally robust to various factors considered, and we also noted that procedures that take into account interaction term by stratification such as S-JAV and S-PSI also yield good coverage probabilities. The results from this analysis that involves gender×trauma interaction seem to be in agreement with what was observed in the simulation results: JM picks up the significance of the gender trauma interaction ($p=0.02$) while other procedures give slightly weaker significance. In addition, this example illustrates that complete case analysis can lead to underestimation of the interaction effect that may exist in the relationship as suggested by results from JM.

Figure 2 is a plot that overlays the estimated regression lines for all the methods. The fitted equation for JM is: $PTSD = 26.65 - 3.77G + 2.11T - 0.97G \times T$, yielding two lines, $PTSD = 26.65 + 2.11Num_Traumas/Yr$ for girls, and $PTSD = 22.88 + 1.14Num_Traumas/Yr$ for boys. In the overlay plot, alternative procedures lie in between these two regression lines from JM, illustrating that these methods might underestimate the interaction relationship.

6. Discussion

MI deals with missing values by filling in multiple draws from the posterior predictive distribution. This process is implemented in various available software, assuming a joint multivariate normal distribution among the variables. When interactions are involved in regression settings, multivariate normality is not satisfied and care is required to obtain appropriate imputation method. This paper specifically accommodates interactions in MI for regression relationships with a binary and a missing continuous predictor using a procedure based on JM approach. Several approaches are assessed in a detailed simulation study. The results suggest that the use of established methods under multivariate normality can be adapted to appropriately take into account the interactions in the imputation model with minimal bias and reasonable coverage properties.

As expected, the JM approach performs well across all the scenarios when we have a completely observed binary predictor and a missing continuous predictor in the model. Approximations under the multivariate normal assumption can perform well with nominal coverage and comparable bias and CI widths compared to JM if the interaction is correctly specified in the imputation model. All procedures are generally robust to the scenarios we considered, for MCAR and MAR mechanisms, and with small or large sample sizes or percentages of missing data.

Failure to include in the imputation model the interaction term that belongs in the substantive analysis model may lead to large bias with below nominal coverage. Filling in the interaction to be zero for all the cases when the binary indicator equals zero seems intuitively sensible but does not always yield good coverage, so it is recommended to be cautious in constraining the interaction term.

This paper illustrates the importance of accounting for interactions in the imputation model for imputing missing continuous covariates when we have a fully observed binary predictor in these regression relationships. In the context we consider, the JM approach used in this paper is the "gold standard" for imputation in the presence of interactions. Stratifying and then performing imputation separately for two groups implicitly takes into account the interaction relationship of the stratification variable and other covariates, and generally leads to good coverage and performs well in a variety of settings considered. In the case with interaction between two continuous variables, stratification requires some degree of categorization of one of the continuous variables, which might give rise to questions about the degree to which exchangeability assumptions are met. When more interactions are involved in the substantive model, obtaining draws from conditional distributions may require approximations and iterative methods, an issue that deserves attention in future research. Especially when the binary variable is an indicator for a treatment that has been assigned at random, stratifying and imputing separately seems to be an effective way to account for interactions as such an approach not only implicitly accounts for interaction effects but also does so in a way that ensures that imputations depend only on individuals who received the same treatment.

Overall, it is not surprising that JM has better performance than alternatives in scenarios where data are generated consistent with the joint model. The application where JM was the only approach under which interaction effect was found to reach conventional levels of statistical significance, suggest that the advantage of JM over other methods can meaningfully influence the interpretation of study results. It is perhaps more surprising that some of the approximation methods perform as well as they do. Here, the stratified methods (S-JAV, S-JAV0, S-PSI0) performed comparably in simulations to JM, reflecting the intuitive perspective that stratifying on a binary variable and performing separate analyses is conceptually close to the idea of fitting a joint model with interactions, the main distinction having to do with residual variance assumptions in the regression part of the model.

This paper evaluated several approximations to the conditional distribution method under a joint modeling framework. MI with chained equations (MICE)^{22;23} and other related sequential regression imputation methods^{13;14;24} are iterative procedures that offer flexibility in dealing with various types of variables, but do not reflect a joint modeling approach for imputations, and are not examined in this paper.

The investigation is restricted to the simplest non-trivial example with two predictors, one continuous and one binary. Realistic incomplete data settings often include a larger number of variables, and we believe the research presented in this paper is a starting point for the study of imputation in settings with more covariates and their interactions and other transformations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH-funded projects P50-HL105188, P30-MH058017, P50-MH066286, R01-HD061404, and UL1-TR000124, and by the Center for Mental Health Services (CMHS), Substance Abuse and Mental Health Services Administration (SAMHSA), US Department of Health and Human Services (USDHHS) through a cooperative agreement (3U79SM054284-10S1) to the UCLA/Duke University National Center for Child Traumatic Stress. The views, policies, and opinions expressed are those of the authors and do not necessarily reflect those of CMHS, SAMHSA, or USDHHS. We would also like to acknowledge the 56 centers within the National Child Traumatic Stress Network (NCTSN) that have contributed data to the Core Data Set as well as the staff, children, youth, and families at NCTSN centers throughout the United States who have made this collaborative network possible.

References

1. Rubin DB. Multiple imputations in sample surveys Proceedings of the Survey Research Methods Section, American Statistical Association 1978;20–34.
2. Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons: New York, 1987.
3. Rubin DB. Multiple Imputation After 18+ Years. Journal of the American Statistical Association 1996; 91:473–489.
4. Horton NJ, Lipsitz SR and Parzen M. A Potential for Bias When Rounding in Multiple Imputation. The American Statistician 2003;57:229–232.
5. Schater JL. Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, 1997.
6. Harel O and Zhou X. Multiple imputation: review of theory, implementation and software. Stat Med 2007;26(16):3057–77. [PubMed: 17256804]
7. Carpenter J and Kenward M. Multiple Imputation and its Application. Wiley, 2013.
8. Collins LM, Schafer JL and Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6(4):330–51. [PubMed: 11778676]
9. Little RJA. Regression With Missing X's: A Review. Journal of the American Statistical Association 1992;87:1227–1237.
10. Kim SE, Sugar CA and Belin TR. Evaluating model-based imputation methods for missing covariates in regression models with interactions. Stat Med 2015;34(11):1876–88. [PubMed: 25630757]
11. Bernaards CA, Belin TR and Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. Stat Med 2007;26:1368–1382. [PubMed: 16810713]
12. Tang L, Song J, Belin TR and Untzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. Statistics in Medicine 2005;24:2111–2128. [PubMed: 15889392]
13. Seaman SR, Bartlett JW and White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Medical Research Methodology 2012;12:46 [PubMed: 22489953]
14. Bartlett JW, Seaman S, White I et al. Multiple imputation of covariates by fully conditional specification: Accomodating the substantive model. Stat Methods Med Res 2014; 24(4): 462–87. [PubMed: 24525487]
15. Pynoos RS, Fairbank JA, Steinberg AM, et al. The National Child Traumatic Stress Network: Collaborating to improve the standard of care. Professional Psychology: Research and Practice 2008; 39(4):389–395.
16. Von Hippell PT. How to impute interactions, squares and other transformed variables. Sociol Methodol 2009;39:265–291.
17. Kim SE. Multiple Imputation for Missing Covariates in Regression Models with Interactions PhD Thesis, University of California Los Angeles, USA, 2011.
18. Steinberg AM, Brymer M, Decker K and Pynoos RS. The UCLA PTSD Reaction Index. Current Psychiatry Reports 2004;6:96–100. [PubMed: 15038911]
19. Steinberg AM and Brymer MJ. The UCLA PTSD Reaction Index Encyclopedia of Psychological Trauma, John Wiley & Sons, Ltd, 2008.

20. Pynoos R, Steinberg A, Layne C, et al. Modeling constellations of trauma exposure in the National Child Traumatic Stress Network Core Data Set. *Psychological Trauma: Theory, Research, Practice, and Policy* 2014; 6:S9–S17.
21. Layne CM, Greeson JKP, Ostrowski S, et al. Cumulative trauma exposure and high risk behavior in adolescence: Findings from the National Child Traumatic Stress Network Core Data Set. *Psychological Trauma: Theory, Research, Practice, and Policy* 2014; 6:S40–S49.
22. van Buuren S Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007;16:219–42. [PubMed: 17621469]
23. van Buuren S and Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011;45:1–67.
24. Raghunathan TE, Lepkowski JM, Hoewyk J, et al. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* 2001;27:1

Figure 1.
Confidence Interval Coverage by each scenario
(Nominal coverage is 95%.)

Figure 2.
Plot of Results for Each Procedure by Gender: Alternative procedures lie between the two lines by JM

Table 1.

Confidence Interval Coverage Probability: Across Scenarios

Method	β_0		β_1		β_2		β_3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Joint Modeling (JM)	0.944	0.014	0.946	0.014	0.953	0.016	0.953	0.015
Just Another Variable (JAV)	0.946	0.016	0.943	0.019	0.938	0.037	0.947	0.026
Stratified JAV (S-JAV)	0.945	0.014	0.948	0.014	0.953	0.015	0.952	0.017
Passive Imputation (PSI)	0.943	0.028	0.883	0.172	0.925	0.067	0.762	0.305
Stratified PSI (S-PSI)	0.945	0.014	0.946	0.015	0.953	0.015	0.953	0.017
JAV zero imputed (JAV0)	0.931	0.045	0.935	0.041	0.890	0.124	0.947	0.029
S-JAV zero imputed (S-JAV0)	0.945	0.014	0.948	0.014	0.953	0.015	0.952	0.017
PSI zero imputed (PSI0)	0.943	0.028	0.883	0.172	0.925	0.067	0.762	0.305
S-PSI zero imputed (S-PSI0)	0.945	0.014	0.946	0.015	0.953	0.015	0.953	0.017

Values presented are means across scenarios.

Bold face indicates coverage lower than 0.90.

Table 2.

Bias: Across Scenarios

Method	β_0		β_1		β_2		β_3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Joint Modeling (JM)	-0.009	0.015	0.006	0.035	0.000	0.019	-0.006	0.041
Just Another Variable (JAV)	-0.028	0.062	0.015	0.070	0.061	0.084	-0.045	0.089
Stratified JAV (S-JAV)	-0.008	0.016	0.018	0.048	-0.001	0.019	-0.022	0.050
Passive Imputation (PSI)	-0.043	0.075	0.162	0.181	0.103	0.093	-0.282	0.181
Stratified PSI (S-PSI)	-0.008	0.016	0.011	0.039	-0.001	0.019	-0.014	0.044
JAV zero imputed (JAV0)	-0.027	0.040	-0.018	0.059	0.056	0.053	0.004	0.058
S-JAV zero imputed (S-JAV0)	-0.008	0.016	0.018	0.048	-0.001	0.019	-0.022	0.050
PSI zero imputed (PSI0)	-0.043	0.075	0.162	0.181	0.103	0.093	-0.282	0.181
S-PSI zero imputed (S-PSI0)	-0.008	0.016	0.011	0.039	-0.001	0.019	-0.014	0.044

Values presented are means across scenarios.

Table 3.

Confidence Interval Widths

Method	β_0		β_1		β_2		β_3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Joint Modeling (JM)	1.16	1.06	1.66	1.45	1.13	1.03	1.50	1.32
Just Another Variable (JAV)	1.20	1.06	1.70	1.47	1.16	1.02	1.57	1.36
Stratified-JAV (S-JAV)	1.17	1.06	1.68	1.46	1.14	1.02	1.52	1.34
Passive Imputation (PSI)	1.17	1.04	1.69	1.40	1.10	0.96	1.46	1.23
Stratified-PSI (S-PSI)	1.17	1.06	1.67	1.45	1.14	1.02	1.51	1.33
JAV zero imputed (JAV0)	1.18	1.08	1.71	1.50	1.20	1.09	1.64	1.45
S-JAV zero imputed (S-JAV0)	1.17	1.06	1.68	1.46	1.14	1.02	1.52	1.34
PSI zero imputed (PSI0)	1.17	1.04	1.69	1.40	1.10	0.96	1.46	1.23
S-PSI zero imputed (S-PSI0)	1.17	1.06	1.67	1.45	1.14	1.02	1.51	1.33

Values presented are means across scenarios.

Table 4.

Comparison of results for the interaction model for the study of childhood trauma

	Intercept	Gender	Trauma	Gender × Trauma	Gender × Trauma
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	<i>P</i>
Joint Modeling (JM)	26.65	−3.77	2.11	−0.97	0.02
Just Another Variable (JAV)	26.78	−3.99	1.98	−0.74	0.13
Stratified JAV (S-JAV)	26.75	−3.94	2.01	−0.80	0.09
Passive Imputation (PSI)	26.88	−4.22	1.87	−0.51	0.22
Stratified PSI (S-PSI)	26.75	−3.93	2.01	−0.81	0.09
JAV zero imputed (JAV0)	26.77	−3.93	1.99	−0.82	0.09
S-JAV zero imputed (S-JAV0)	26.75	−3.94	2.01	−0.80	0.09
PSI zero imputed (PSI0)	26.88	−4.22	1.87	−0.51	0.22
S-PSI zero imputed (S-PSI0)	26.75	−3.93	2.01	−0.81	0.09
Complete Case (CC)	26.52	−3.85	1.95	−0.75	0.12