Multiple Imputation by Chained Equations: An Overview of Conceptual and Operational Aspects and Software: Review

Zincir Denklemleriyle Çoklu Değer Atama: Kavramsal ve İşletimsel Unsurlar ile Yazılıma Genel Bir Bakış

Jennifer L. BEAUMONT,[a,b]
Hakan DEMİRTAŞ[a]

[a]Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, IL 60612, USA
[b]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, U.S.A.

Geliş Tarihi/*Received:* 06.11.2012
Kabul Tarihi/*Accepted:* 09.12.2012

Yazışma Adresi/*Correspondence:*
Hakan DEMİRTAŞ
University of Illinois at Chicago, School of Public Health, Division of Epidemiology and Biostatistics (MC 923), 1603 West Taylor Street, Room 950 Chicago, IL, 60612-4336, U.S.A.
demirtas@uic.edu

ABSTRACT Missing data are prevalent in nearly all areas of research. It is of utmost importance that appropriate methods are used to obtain statistically valid inferences in the presence of missing data. Multiple imputation relies on the creation of multiple sets of plausible values for the missing data. We provide background on missing data and its consequences, and describe the fundamental concepts underlying multiple imputation. We then describe a flexible implementation of multiple imputation, called multiple imputation by chained equations, that allows the analyst to specify a separate conditional regression for each variable with missing data. We conclude with an overview of software options for multiple imputation by chained equations.

Key Words: Missing data; multiple imputation; chained equations

ÖZET Hemen hemen bütün çalışma alanlarında kayıp veriler ortaya çıkmaktadır. Kayıp veri durumunda, istatistiksel olarak geçerli çıkarsamalar yapabilmek için uygun yöntemlerin kullanılması son derece önemlidir. Çoklu değer atama, kayıp veriler yerine geçebilecek uygun değerleri içeren veri setleri oluşturma temeline dayanır. Bu çalışmada, kayıp verilere ilişkin genel bilgiler ve neden olacağı sonuçların yanı sıra, çoklu değer atamanın temelini oluşturan kavramlara değinilmiştir. Kayıp veri içeren her değişken için ayrı bir regresyon denklemi tanımlamaya olanak sağlayan ve bu anlamda esnek bir yöntem olan zincir denklemleri ile çoklu değer atamadan bahsedilmiştir. Zincir denklemleri ile çoklu değer atama için kullanılan yazılımlar incelenmiştir.

Anahtar Kelimeler: Kayip veriler; çoklu değer atama; zincir denklemleri

Turkiye Klinikleri J Biostat 2013;5(1):29-36

Missing data are ubiquitous in statistical practice. Missing data can arise due to laboratory instrument failure, participant withdrawal or loss-to-follow-up in longitudinal studies, participant refusal to answer particular survey questions, or even intentionally by study design. Many statistical techniques used in data analysis require complete data and are not equipped to handle missing data. Some common software implementations of these statistical analysis procedures drop all cases with missing values for any of the variables included in the analysis. For analyses involving a large number of variables, such as a complex logistic regression model, missing data may occur in the outcome variable and in any of the potentially numerous explanatory variables and confounders included in the model. Where a scattered missing data pattern such as this may occur, case-

wise deletion of subjects with missing data on any variable can substantially reduce power by limiting the sample size available for analysis. Furthermore, the particular subset of observations with complete data may not be representative of the original sample and may lead to biased estimates and incorrect conclusions.

Arbitrary missing data patterns where data may be missing for any variable on any subject are very common, though there are some special missing data patterns that can afford simpler treatment. The special case of univariate missing data occurs when only one variable in a data set has missing data. A monotone missing data pattern occurs when the variables under study, $Y_1,...,Y_p$ can be ordered in a such a way that if Y_j is missing for a subject then $Y_{j+1},...,Y_p$ are missing as well, while $Y_1,...,Y_{j-1}$ are observed. The monotone missing data pattern is commonly seen in longitudinal studies where missing data are due to attrition or study drop-out.

Missing data mechanisms can be classified into three main types that were first described by Rubin[1] and later refined in Little and Rubin's seminal text on missing data.[2] Missing data are considered missing completely at random, or MCAR, if the probability of missingness is completely independent of both the observed and unobserved data. That is, every data element has equal probability of being missing. To describe this mathematically, we must first define $R = R_1,..., R_p$ as vectors of missingness indicators that take the value 1 if Y is observed and 0 if Y is missing. MCAR implies that $P(R \mid Y) = P(R)$. It is quite rare in practice for data to be missing completely randomly. For an example of MCAR data, imagine a printer that failed to print two pages of a multi-page survey for a random 10% of surveys. The probability of missing data is completely unrelated to the survey respondents and their characteristics. Data that are missing by design are sometimes MCAR. For example, a portion of a very long survey may be broken into blocks and survey participants randomly assigned to complete only one block of questions, rather than the whole survey.

If the probability of missingness depends on observed data, then the missing data are missing at random, or MAR. That is, $P(R \mid Y) = P(R \mid Y_{obs})$, where Y_{obs} is the portion of Y that is observed. If, however, the probability of missingness depends on the unobserved values, then the missing data are called missing not at random, or MNAR, and $P(R \mid Y)$ cannot be reduced to a simpler form that does not require specific assumptions about the missing data mechanism. In a cross-sectional study examining correlates of physical functioning, for example, perhaps cholesterol level is one such potential correlate under investigation and 15% of study participants are missing data for their cholesterol level. These missing data may be MNAR if people with low cholesterol levels are more likely to have missing cholesterol data, perhaps because they are generally younger and healthier and less likely to have ever had cholesterol testing conducted. Note that if data are collected on age and general health status, it may then be reasonable to consider the missing cholesterol data MAR conditional on those observed variables. This is an important point that is leveraged by many statistical techniques for handing missing data that will be described later. It is important to note that the distinction between MAR and MNAR is essentially untestable in most circumstances because the data required to evaluate the distinction are, by definition, missing. Researchers must rely on external data if possible and their understanding of the constructs under study to judge the reasonableness of the MAR assumption.

All statistical analysis techniques make some assumptions about the nature of missing data. It is up to the researcher to recognize them and determine whether those assumptions are reasonable. The simplistic technique mentioned previously, throwing out cases with missing data and analyzing only complete cases, assumes that missing data are MCAR.[3] In many cases, this may not be a reasonable assumption, in addition to leading to a potentially substantial loss of power due to reduced sample size. Single imputation techniques, such as substituting missing data with a mean value or regression prediction, offer some improvements over simply ignoring cases with missing data but they also rely on assumptions that are very likely untenable and do not account for the uncertainty in

the imputation of the missing data.[4] After missing values are "filled in" using some form of single imputation, the analysis then carries on as if this were the complete data set. However, it is not a recreation of the complete data since some of the values are not real data but "guesses," and there is some level of uncertainty in these guesses that should be incorporated into the statistical analysis to avoid over-stating the precision in the form of p-values that are too small and confidence intervals that are too narrow.

Some statistical analysis techniques can handle missing data directly via direct maximum likelihood methods, including some software for multilevel models and structural equation modeling. For example, multilevel models, such as those implemented in SAS PROC MIXED, are able to use all of the available data and provide appropriate inferences even in the presence of missing values in the outcome variable under the assumption of MAR. However, these multilevel models do not accommodate missing values in the explanatory variables.

MULTIPLE IMPUTATION

Rubin developed the beginnings of multiple imputation in 1977[5] and published his seminal book on the topic in 1987.[6] He reviewed the progress of multiple imputation nearly two decades later in 1996[7] and many others in the field have published reviews and instructive tutorials.[3,8,9] At first, multiple imputation was conceived of in the context of large-scale sample surveys—a situation where a single data set may be analyzed by a large number of users, some of whom may not be granted access to the full data set (e.g., detailed confidential data such as protected health information). It has now achieved broader application in many fields. In the most basic sense, multiple imputations are simply multiple sets of plausible values for the missing values. Different multiple imputation methods rely on different techniques to create the plausible values as will be described below. The use of multiple sets of plausible values allows the analyst to incorporate uncertainty about the imputed values. Once imputations have been created, analysis of multiply imputed data requires little to no specialized statistical knowledge as analyses can be conducted using any statistical package and complete-data technique. Subsequently, results are combined to a single inferential summary by Rubin's rules.[6]

The foundation of multiple imputation has a Bayesian motivation, although it can be carried out by frequentist arguments.[6] If we define Q as the vector of parameters of interest, then the actual posterior distribution of Q is the complete-data posterior distribution of Q, averaged over the repeated imputations. The repeated imputations, in turn, are draws from the posterior predictive distribution of the missing data given the observed data:

$$P(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})dY_{mis}$$

where $Y = (Y_{obs}, Y_{mis})$ is the complete data where Y_{obs} is the portion of Y that is observed and Y_{mis} is the portion of Y that is missing. Where R=0, we have no information for $P(Y_j^{mis}|Y_{-j}, R)$, but the ignorability, or MAR, assumption allows us to define $P(Y_j^{mis}|Y_{-j}, R = 0) = P(Y_j^{mis}|Y_{-j}, R = 1)$.

The basic result led to the rules for estimating the posterior mean and variance of Q, which are frequently referred to as "Rubin's Rules".[6] Specifically, the posterior mean of Q is the average of the repeated complete-data posterior means of Q:

$$E(Q|Y_{obs}) = E[E(Q|Y_{obs}, Y_{mis})|Y_{obs}],$$

which is estimated by

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i$$

where $\hat{Q}_1, \ldots, \hat{Q}_m$ are the values of the complete data statistics calculated on the m completed data sets. The posterior variance of Q is the average of the repeated complete-data variances of Q, plus the variance of the repeated complete-data posterior means of Q:

$$V(Q|Y_{obs}) = E[V(Q|Y_{obs}, Y_{mis})|Y_{obs}] + V[E(Q|Y_{obs}, Y_{mis})|Y_{obs}],$$

which is estimated by

$$T_m = \bar{U}_m + \frac{m+1}{m}B_m,$$

where the within imputation variability is

$$\bar{U}_m = \frac{1}{m}\sum_{i=1}^{m} U_i$$

and the between imputation variability is

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q}_m)(\hat{Q}_i - \bar{Q}_m)'.$$

Confidence intervals can then be calculated as $\bar{Q}_m \pm t_{df}\sqrt{T_m}$, where

$$df = (m-1)\left(1 + \frac{m\bar{U}_m}{(m+1)B_m}\right)^2.$$

We have not yet addressed how the multiple imputations are generated. Imputations are repeatedly drawn from the posterior predictive distribution of the missing values under a specific model. A common implementation of multiple imputation relies on specification of a joint model for the full data set, *Y*. Most traditionally, the multivariate normal distribution has been used. Each iteration then alternates between making draws from the posterior distribution of the parameters and imputing the missing values based on these parameter draws. The imputation model should include as many variables as possible, including all variables that will be included in analysis models as outcomes, predictors, or covariates, as well as clustering and stratification factors, predictors of missing data, and interactions of scientific relevance.[10] Most multiple imputation methods assume that the missing data mechanism is MAR and generous inclusion of variables in the imputation model will improve the plausibility of the MAR assumption. Furthermore, it is important to include variables whose relationships one is interested in testing because failure to do so will result in those associations being biased towards zero at the time of analysis.

As mentioned before, some statistical analysis techniques can handle missing data directly via direct maximum likelihood methods, including some software for multilevel models and structural equation modeling. In cases where the MAR assumption is reasonable, these methods may be more efficient than multiple imputation. However, the ability to include additional variables in the imputation model can make the MAR assumption more tenable under multiple imputation. The basic implementation of all multiple imputation methods assumes MAR but multiple imputation can generalize to MNAR scenarios. Specific assumptions about departures from MAR must be specified, for example, imputed values for a particular variable are reduced by 20% from the values derived from the observed data. Sensitivity analyses evaluating the impact of different MNAR assumptions should be performed, although it is important to remember that no MNAR assumption is testable using the data available except for simple situations. For some imputation methods that can accommodate MNAR type of missingness see Demirtas and Schafer[11] and Demirtas.[12]

After the multiple imputed data sets are created, it is important to conduct some diagnostic checking. Of particular use is an examination of the distribution of the imputed and observed data. Some degree of difference between the distributions is to be expected and is acceptable, especially if missing data are not MCAR, but differences between observed and imputed data distributions should seem reasonable. The goal of all multiple imputation procedures is not to recreate the individual missing values but to provide a means to make statistically valid inferences.

MULTIPLE IMPUTATION BY CHAINED EQUATIONS

Real data rarely conform to the multivariate normal model. Procedures for other joint models have been developed, such as the log-linear and general location models,[8] but there are many data sets for which a definable joint model may not apply. Multiple imputation by chained equations (MICE), on the other hand, is flexible enough for nearly any data structure. MICE is also called fully conditional specification[13] or sequential regressions.[14] Rather than relying on a single joint model for the data, the analyst specifies a separate imputation model for each variable with missing data. MICE is much more flexible than joint modeling as one can easily specify imputation models that do not fit within any known joint multivariate distribution. Impu-

tations can also incorporate complexities in the data such as skip patterns, summed scores, interactions, and variable bounds. A tutorial for multiple imputation using chained equations has been published by White et al.[10] and is an excellent resource for analysts, as is van Buuren's latest book on the topic of imputation.[15]

A conditional density, $P(Y_j|Y_{-j}, \theta_j)$, is defined for each Y_j and used to impute Y_j^{mis} given Y_{-j}. Note that since there is a separate conditional density for each Y_j, the θ_j are specific to that particular conditional density. Each iteration of the chained equation approach effectively draws θ_j from its distribution conditional on the observed portion of Y_j and Y_{-j}, which may contain observed and previously imputed values. Then Y_j is sampled from its distribution conditional on this sampled θ_j and the observed and imputed Y. Once the algorithm has cycled through θ_1, Y_1, θ_2, Y_2,..., θ_p, Y_p, that is one full iteration and the process starts again. To begin the first iteration, initial values are generated by randomly selected observed values.[10] Certain orderings of the variables may be more efficient than others. For example, ordering monotone missing data according to the decreasing proportion of observed data will result in instant convergence. However, in general cases the order should have little impact if a large number of iterations are performed.

These conditional densities can be measurement scale-appropriate models for different types of variables. Continuous variables may be modeled using linear regression. If continuous variables are non-normal, transformations can be applied or predictive mean matching can be used to generate the imputed values. In predictive mean matching, imputed values are sampled from a subset of values in the observed data that are "close" to the predicted mean for the missing value.[10] Predictive mean matching will not perform well if the imputations necessitate an extrapolation beyond the range of the observed values. Binary variables may use a logistic regression model for imputation. Nominal, or unordered categorical, variables can be modeled using multinomial logistic regression; while ordinal, or ordered categorical, variables can use multinomial logistic regression or ordinal logistic regression based on the proportional odds assumption. An overdispersed Poisson regression model may be a desirable option for counts or ordinal data with many categories. Longitudinal data can be straightforwardly imputed by treating the different time points as different variables. Censored time-to-event variables can be imputed by using time, log(time) and the censoring indicator all as separate variables or by using the cumulative hazard function for each individual with the censoring indicator. MICE using all linear regressions is equivalent to imputation under the joint multivariate normal model. Perfect prediction can be a problem in imputation models for categorical variables. For example, if a 2x2 table representing the association between two variables contains a zero cell. This leads to infinite parameter estimates and problems in estimation.[10]

Plots of parameter estimates versus the iteration number for each of the *m* parallel streams can be examined to evaluate convergence of the method. There should be mixing of streams and no apparent trend after convergence. In many cases 5 to 10 iterations are sufficient to achieve convergence but analysts should examine plots and, if computing time is not a limiting factor, increase the maximum number of iterations. The suggestions for the number of imputed data sets was traditionally m = 3 to 5. These suggestions were related to the concept of the fraction of missing information (FMI), which is defined as B/(U+B), or the between imputation variance divided by the total variance. If the fraction of missing information is ≤ 0.25 then m=5 imputations is generally sufficient to achieve loss of efficiency of 5 percent or less.[10] However, Graham et al.[16] recommended at least 20 imputations to minimize power loss and White et al.[10] recommended at least as many imputations as the percentage of incomplete cases in the data set. As suggested for multiple imputation in general, after creation of the multiple imputed data sets by chained equations, one should examine the imputed data to ensure the imputations are plausible.

One limitation of MICE that is frequently highlighted is its lack of theoretical basis. Separate

conditional distributions may be incompatible in the sense that they may not correspond to a unique, well-defined joint distribution. Nevertheless, positing a joint distribution may be a daunting task, especially when different types of variables are involved. Even with this theoretical shortcoming, MICE has been found to perform well in simulation studies.[17-20] One consequence of this potential incompatibility is that the results of an analysis may depend on the order of the variables in the imputation but so far this appears to have little impact in practice.[17] Comparisons between multiple imputation with multivariate joint modeling and MICE have been mixed,[13,21,22] and it is likely that the relative performance of the two methods depends upon the level and type of missing data and degree of departure from multivariate normality. More work examining the performance of MICE under different conditions is needed to identify whether there are any situations where the method breaks down and should not be used.

SOFTWARE OVERVIEW

A recent compilation of papers provides details on several software options for the implementation of multiple imputation in general and MICE in particular.[23-27] We provide an overview of the software options for MICE here.

The **R** package *mice 2.9* implements multiple imputation by chained equations with many flexible options.[24] The package includes useful functions for examining the number and pattern of missing values and creating graphical displays. A basic call to the `mice()` function with no changes to defaults will produce five multiply imputed data sets using the software's default method for each variable type (Table 1). Other models can be explicitly specified if desired. For numeric variables additional options include linear regression (Bayesian and non-Bayesian), a two-level linear model, and unconditional mean imputation. For categorical variables, a linear discriminant analysis can be specified. For any type of variable, a random sample from the observed data can be used, and this is the method used by the program to obtain initial starting values. The order in which variables are imputed can be specified by the user, the seed value can be defined so that analyses are repeatable, and the number of imputations can be changed.

TABLE 1: Default imputation models.

Variable type	*mice* (R)	*ice* (Stata)
Numeric	Predictive mean matching	Linear regression
Dichotomous	Logistic regression	Logistic regression
Nominal	Multinomial logit	Multinomial logit
Ordinal	Ordered logit	Multinomial logit

A square matrix of zeros and ones is used to specify the predictors to be used for each imputation model. By default, all variables are used to predict all others but a custom matrix can be inputted. While general advice for imputation models is to include as many variables as possible to make the MAR assumption more plausible, there are many situations where one may need to reduce the number of predictors. Collinearity and computational problems can arise when there are many variables in a model. There may be little gain in explained variance after the best 15-25 predictors are included anyway.[24] *mice 2.9* includes a function to aid in selection of predictors that will automatically create a custom predictor matrix. It also includes a capability for passive imputation, which maintains consistency between different transformations of the same data. This is useful in the context of variable transformations, interaction terms, summed scores, or other derived scores. The plot function can then be used to plot, for example, means of the imputed values against the iteration number to evaluate convergence.

After the multiple imputed data sets have been created, **R**'s many graphics capabilities can be used to examine the distributions of the original and imputed data for diagnostic checks. The `stripplot` and `xyplot` functions use different colors to differentiate between the observed and imputed values. The **R** package *mice 2.9* includes a wrapper function, `with.mids()`, that can be used to conduct any analysis on a set of multiply imputed data. Finally, the function `pool()` will pool the re-

sults of the multiple analyses providing estimates, standard errors, confidence intervals, t-tests, the fraction of missing information, and the proportion of the total variance attributable to the missing data. A `pool.r.squared()` function also exists for pooling estimates of R^2 using the Fisher *z* transformation, while `pool.compare()` can be used to compare two nested models.

Another **R** package, *mi*, also implements MICE and includes many additional features for diagnostic checks and an optional graphical user interface.[25] The implementation, options, and defaults for *mi* are different from those for *mice*; we do not go into details here but do summarize some of the unique features. *mi* includes Bayesian versions of the general linear models to automatically handle computational problems with perfect prediction. Special transformations are applied for semi-continuous variables. To deal with collinearity, *mi* will screen for perfectly correlated variables and automatically exclude one from the imputation models and also uses reshuffling noise to deal with additive collinearity. Multiple diagnostic plots are available for the imputation models including overlaid histograms of the observed, imputed and completed data, binned residuals plotted against expected values, and scatterplots of observed or imputed values versus predicted values.

Multiple imputation by chained equations is implemented by **Stata** in *ice*.[26] The default imputation models for different variable types are listed in Table 1. Other methods that are also available for user-specification are predictive mean matching, ordinal logistic regression, interval censored regression for a continuous variable, negative binomial regression for a count variable, and a method that estimates regression coefficients in a bootstrap sample. The interval censored regression may be useful in applications where, for example, age is recorded only in age groups (e.g., <18, 19-29, 30-39, 40-49, etc.). In addition to the imputation model for each variable, analysts can specify the number of imputations, the random seed, and the pool-size for predictive mean matching. Conditional imputation can be implemented easily, allowing a variable to be imputed only for subsets of subjects, for example, imputing number of cigarettes smoked for smokers only or imputing number of pregnancies for women only. By default, *ice* places the variables to be imputed in order of increasing missingness. Because monotone missing data patterns converge instantly, using the monotone option in *ice* will result in only one iteration per imputation stream. A user-written program called *mim* is used to implement Rubin's rules for combining estimates.

The multiple imputation procedure for **SAS**, *proc mi*, implements multiple imputation by joint modeling and by MICE, using the FCS statement.[28] Linear regression is the default method for all continuous variables and the discriminant function method is the default for all classification variables. Other methods that can be specified by the user include predictive mean matching, logistic regression and ordinal logistic regression. *proc mianalyze* reads in datasets of parameters and standard errors and produces the pooled estimates. Prior to the release of version 9.3, **SAS** implemented sequential regressions for monotone missing data patterns only[27] and *IVEware* was developed as a **SAS** callable application that implements MICE via four modules: `IMPUTE`, `DESCRIBE`, `REGRESS`, and `SASMOD`.[29]

SPSS implements MICE in its *MULTIPLE IMPUTATION* procedure and analysis procedures recognize multiply imputed data sets and automatically produce pooled estimates. Multiply imputed data from other software can be imported into **SPSS** and analyzed in this way if it is in the proper format.[24]

FINAL REMARKS

Missing data are a prevalent and persistent problem throughout research and the application of principled techniques has not yet become widespread to the extent that it should. In the field of quality of life research, for example, a review of randomized controlled trials published in major medical journals found that in a sample of 61 trials reporting on quality of life outcomes, 90% reported some missing data but only two studies used mul-

tiple imputation while the others used less statistically rigorous techniques such as last value carried forward or complete case analysis.[30] Multiple imputation was first developed in the 1970s and provides a principled, statistically valid means to make inferences in the presence of missing data.

Multiple imputation by chained equations is a flexible approach to multiple imputation that allows for a wide range of variable-specific imputation models and avoids the daunting task of specifying a joint distribution for the data. Although sometimes computationally intensive, the methods are straightforward to apply and are now included in many statistical software packages. Further research is warranted to continue to investigate the performance of MICE under different missing data circumstances and methods for handling MNAR data. The current availability of software to implement the flexible method of multiple imputation by chained equations will hopefully lead to an increasing trend in appropriate analyses in the presence of missing data.

REFERENCES

1. Rubin DB. Inference and missing data. Biometrika 1976;63(3):581-92.
2. Little, RJ, Rubin DB. Statistical Analysis with Missing Data. 1st ed. New York: Wiley; 1987. p.1-278.
3. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8(1):3-15.
4. Shen C, Weissfeld L. Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. Biostatistics 2005;6(2):333-47.
5. Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association 1977;72(359):538-43.
6. Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1st ed. New York: John Wiley; 1987. p.1-253.
7. Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association 1996;91(434): 473-89.
8. Schafer JL. Analysis of Incomplete Multivariate Data. 1st ed. London: Chapman & Hall; 1997. p.1-430.
9. Kenward MG, Carpenter J. Multiple imputation: current perspectives. Stat Methods Med Res 2007;16(3):199-218.
10. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011; 30(4): 377-99.
11. Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. Stat Med 2003; 22(16):2553-75.
12. Demirtas H. Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. Stat Med 2005; 24(15):2345-63.
13. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007; 16(3):219-42.
14. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001;27(1):85-95.
15. Van Buuren S. Flexible Imputation of Missing Data. 1st ed. New York: CRC Press; 2012. p.1-247.
16. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci 2007;8(3):206-13.
17. Van Buuren, S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation & Simulation 2006;76(12):1049-64.
18. Horton N J, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 2001;55(3):244-54.
19. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007;61(1):79-90.
20. Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. Stat Methods Med Res 2007;16(3):243-58.
21. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol 2010;171(5):624-32.
22. Demırtas H, Freels SA, Yucel RM. Plausibility of multivariate normality assumption when multiply imputing non-Guassian continuous outcomes: a simulation assessment. Journal of Statistical Computation and Simulation 2008;78(1):69-84.
23. Yucel RM. State of the multiple imputation software. J Stat Softw 2011;45(1). doi:pii: v45/i01.
24. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw 2011;45(3):1-66.
25. Su Y-S, Gelman A, Hill J, Yajima M. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. J Stat Softw 2001;45(2):1-30.
26. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. J Stat Softw 2011;45(4):1-20.
27. Yuan Y. Multiple imputation using SAS software. J Stat Softw 2011;45(6):1-25.
28. INC, SI. SAS/STAT(R) 12.1 User's Guide. Cary, NC: SAS Institute Inc; 2012. p.1-65.
29. Raghunathan TE, Solenberger PW, Van Hoewyk J. IVEware: Imputation and Variance Estimation Software User Guide. Michigan: Survey Methodology Program Survey Research Center, Institute for Social Research University of Michigan; 2002. p.1-73.
30. Fielding S, Maclennan G, Cook JA, Ramsay CR. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. Trials 2008;9:51.