Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Yucel, J Biomet Biostat 2012, 3:1
DOI: 10.4172/2155-6180.1000e105

Editorial Open Access

Dissemination of Missing Data Techniques in Medical, Biomedical and Social Research

Recai M. Yucel*

Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, State University of New York

Keywords: Missing data; Multiple imputation; Software

Among all state-of-the-art techniques in the statistical methodology, perhaps one of most readily available topic to the researchers in medical, biomedical and social research is the missing data. This is partly due to its natural containment of not only the theoretical development but also the computational algorithms. The efforts for producing software for public use accompanying the underlying methodology are often much more involved then the development of the method itself. Then why would the methodologists assume the responsibility to achieve this often rewardless task? Aside from achieving professional reputation, the underlying motivation is **dissemination**. Dissemination efforts as true scientific novelty merely aims to contribute validity of the results as well as to improve methods underlying the implemented methodology.

Aside from all the professional benefits to the developers, dissemination of the missing data methods undoubtedly serves the greater scientific community, ranging from social sciences to medicine and from life sciences to population-based sciences. From my limited experience, I have observed that the subject-matter studies often employ primitive and sometimes unprincipled methods for dealing with missing data even though the employed methods for data collection can be state-of-the-art, comprehensive and costly. Missing data literature repeatedly demonstrated the extreme adversities facing the data analysis when no sound action taken to address the problem of missing data. One of the most striking examples of unacceptable behavior of "unprincipled" missing-data techniques was demonstrated by Schafer JL and Graham JW [1]. In a very simple bivariate data example, they demonstrated that ignoring missing data can lead to coverage rates that are far below than nominal rates, and in some cases zero coverage rate. As the data structures become more complex (e.g. multilevel studies, longitudinal clinical trials) the adverse effects of "naive" methods can be extreme and can easily become a threat to the validity of the underlying inferences.

Fortunately, there is a growing awareness in the subject-matter fields about the missing-data methods. One overall observation is that the "complete-case-only" analysis is decreasingly in use. While researchers may become increasingly aware of its adverse side-effects, they still do not seem to be aware of under what circumstances their complete-case-only analysis would be valid. We recently investigated the use of missing data techniques in three high impact medical journals. Initial investigation revealed that the penetration of "principled" missing data techniques is nowhere near the scientific expectations. I use the term "principled" rather loosely as the search criteria included any terms that would be indicative of any modern missing data technique. There even seems to be complete misuse of the underlying assumption of missingness mechanism of missing completely at random. It is clear that there is an obvious disconnect in communication between statistical methodologists and subject-matter researchers.

Statisticians should necessarily assume responsibility in closing the gap in communication so the ill-practice is not a threat to the validity of inferences intensively used in diverse set of fields, ranging from health policy to social sciences and to lab-based fields. Publications that communicate what we develop and implement via software to the greater scientific community. In addition to the availability of software, there is also a great documentation purely aimed to benefit both sophisticated and/or moderately sophisticated consumers of such software tools. As a good sign of increasing communication, more and more statisticians are contributing to the subject-matter journals on the subject of missing data with special focus on commonly misunderstood concepts such as missing completely at random versus missing at random and multiple imputation.

A recently-edited issue appeared in Journal of Statistical Software aims to achieve the goal of greater dissemination of not only how to use the state-of-the-art missing data software but also the underlying techniques and assumptions [2]. Manuscripts are organized following the underlying "imputation" philosophy implemented by the respective software. First group shares the common theme of variable-by-variable approach (also referred as chained imputation models). This approach is particularly useful in problems with a set of incompletely-observed variables with diverse set of measurement scales (e.g., continuous, categorical, count and semi-continuous) and in problems complicated by common survey practices including skip patterns and truncation. First paper in this group is by Su YS, Gelman A, Hill J, Yajima M [3]. Their software implements flexible imputation techniques via chained imputation models and diagnostic tools that allow users to assess plausibility of the assumed imputation models. Specifically, their package **mi** features flexible choice of predictors, models, and transformations for chained imputation models; binned residual plots for checking the fit of the conditional distributions used for imputation; and plots for comparing the distributions of observed and imputed data in one and two dimensions. Bayesian models are also used to construct more stable estimates when data are sparse and supported by a prior knowledge.

An increasingly popular approach to producing multiple imputations in settings pertaining to variables that are of varying natures and measured with restrictions is illustrated by Buuren SV, Groothuis-Oudshoorn K [4]. They present the most recent version of their *R* [5] package called **mice** which imputes incomplete values by fully conditional specification. This package offers many practical solutions including predictor selection, passive imputation and automatic pooling to combine estimates from the multiply imputed datasets.

***Corresponding author:** Recai M. Yucel is Associate Professor of Biostatistics at the Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, State University of New York, E-mail: ryucel@albany.edu

Received December 23, 2011; **Accepted** December 23, 2011; **Published** December 26, 2011

Citation: Yucel RM (2012) Dissemination of Missing Data Techniques in Medical, Biomedical and Social Research. J Biomet Biostat 3:e105. doi:10.4172/2155-6180.1000e105

These features are also extended to the multilevel continuous data. Finally, this version adds a capability of multilevel **MI** and interactive use with *SPSS*. The third contribution presents an implementation of a similar approach in *Stata*. Manuscript by Royston P, White IR [6] describes **ice** which is the *Stata* module of the approach using the fully automatic pooling to produce multiple imputation. Royston and White [6] illustrate this fully-integrated module in *Stata* using real data from an observational study in ovarian cancer.

Joint modeling approach follows the variable-by-variable approach. Carpenter and his colleagues [7] describe a comprehensive module called REALCOM-IMPUTE of the multilevel model fitting software MLwiN [7]. Variables subject to missing values are modeled under a multivariate latent normal model with random-effects, which is used as a basis to approximate the underlying posterior predictive distribution. The authors use Markov chain Monte Carlo (MCMC) simulation techniques to fit the imputation models and thus draw the multiple imputations. The software also allows for weights to account for sampling design both at level 1 and level 2. A variety of variables can be imputed: continuous, ordinal or nominal. Users can further analyze the imputed datasets under multilevel models and combine estimates using MI rules defined by Rubin [8].

Another increasingly popular package is **PROC MI** and **PROC MIANALYZE** procedures of *SAS*. Yuan [9] illustrates how to conduct MI inference in *SAS*. **PROC MI** implements three major techniques one can adopt to produce multiple imputations. Specific choice of these techniques depends on the missingness pattern and the type of imputed variable. For the problems with monotone patterns of missingness (i.e. a variable missing implies that all subsequent variables to be missing), one can choose from the following three methods depending on the type of the variable(s) to be imputed: matching (using propensity score or predictive mean) or MCMC which draws imputations from a multivariate normal if the underlying variables are continuous. If they are categorical, one can choose logistic regression or discriminant-function-based method to match. For the arbitrary patterns of missingness, one would have to approximate the underlying posterior predictive distribution using a multivariate normal distribution with a set of priors provided by **PROC MI** (e.g., ridge or Jeffreys prior).

Another popular software mostly among social and political researchers is **Amelia** by Honaker J, King G, Blackwell M [10]. **Amelia** integrates two important computational tools EM and bootstrap to produce multiple imputations [11]. It implements a new computationally-improved EM-bootstrapping algorithm as an alternative to MCMC-based solutions. The imputation model still relies on a joint model, but the underlying sampling from the posterior predictive distribution is fundamentally different. Because the computations are centered around maximum likelihood (or posterior mode) estimates and it merely uses a re-sampling-based algorithm, it provides a computational efficiency. It also includes features to accurately impute cross-sectional datasets, individual time series, or sets of time series for different cross-sections. Finally, it allows users to facilitate graphical diagnostics for the imputed datasets.

Software development is one of the significant keys to dissemination of the statistical methods. Without it, the greater scientific community simply does not have the means to access to state-of-the-art techniques. Due to high prevalence of missing data in research problems relying on empirical evidence, it is critical for the statistical community to provide objective and open source for missing data software. This special issue aims to provide exactly this, and it is my hope to see updates to this special issue to provide statistical and substantive literatures with the up-to-date documentation of software. The diversity of the contributions to this special issue provides an impression about the progress of the last decade in the software development in the multiple imputation.

It should be noted that there are many other multiple imputation software products. Some of the most commonly-used software include R packages **aregImpute, norm, cat, mix** [12,13] for a variety of techniques to create multiple imputations in continuous, categorical or mixture of continuous and categorical datasets. Another useful R package for imputing continuous variables in clustered or longitudinal designs is **pan** [14]. There is also a very important package in the form of *SAS* macro for multiple imputation using a sequences of regression models. This *SAS*-callable program is called **IveWare** written by iveware and very similar to the R [5] and *Stata* implementation of **mice** and **ice**.

The implemented methodology of MI has so far focused on the improved computational algorithms geared towards relatively simpler data designs. In other words, software development for MI is just starting. There are many problems for which the greater scientific community is looking for principled and ready-to-use tools. Some examples include extensions of variable-by-variable-based methodology to clustered designs, multilevel datasets, incorporation of non-ignorable mechanisms. I believe that the software development will be greatly helped by open-source forums such as R as it provides a great forum for steady improvements via users' feedback and constructive criticism.

References

1. Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7: 147-177.
2. Yucel RM (2011) State of the multiple imputation software. J Stat Softw 45: 1-7.
3. Su YS, Gelman A, Hill J, Yajima M (2011) Multiple imputation with diagnostics (mi) in R: opening windows into the black box. J Stat Softw 45: 1-31.
4. Buuren SV, Groothuis-Oudshoorn K (2011) mice: Multivariate Imputa-tion by Chained Equations in R. J Stat Softw 45: 1-67.
5. R Development Core Team (2011) R: A language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria.
6. Royston P, White IR (2011) Multiple imputation by chained equations (mice): implementation in stata. J Stat Softw 45: 1-20.
7. Carpenter JR, Goldstein H, Kenward MG (2011) REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. J Stat Softw 45: 1-14.
8. Rubin DB (1987) Multiple Imputation for Nonresponse in Surveys, New York: John Wiley and Sons.
9. Yuan Y (2011) Multiple Imputation Using SAS software. J Stat Softw 45: 1-25.
10. Honaker J, King G, Blackwell M (2011) AMELIA II: A pro-gram for missing data. J Stat Softw 45: 1-47.
11. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society Ser B 39: 1-38.
12. Harrel F (2010) Multiple imputation using additive regression, bootstrap-ping, and predictive mean matching. J Stat Softw
13. Schafer JL (2000) Multiple imputation of incomplete multivariate normal data. The Pennsylvania State University, PA, USA.
14. Schafer JL, Yucel R (2002) Computational strategies for multivariate linear mixed-effects models with missing values. J Comput Graph Stat 11: 437-457.