Comparison of Modern Imputation Methodologies on Complex Data from Agricultural Operations

Darcy Miller
Andrew Dau
Jonathan Lisic

National Agricultural Statistics Service
1400 Independence Avenue SW
Washington, DC 20250

Abstract

The National Agricultural Statistics Service (NASS), in conjunction with the Economic Research Service (ERS), conducts the Agricultural Resource Management Survey (ARMS) to study the well-being of farm households. ARMS has a complex survey design, collects personal household information, and has a burden exceeding ninety minutes. Due to item nonresponse, some of the ARMS data are missing. Prior to 2015, a complete data set for use by NASS was formed by a mixture of machine imputation (conditional mean) and manual imputation. Recently, Sequential Iterative Regression (ISR), a multivariate imputation methodology, was implemented through a cooperative agreement between the National Institute of Statistical Sciences and NASS (see Robbins, et al. 2013), and it better preserves relationships and the distribution of the data. ISR has been noted to be a blend of the popular data augmentation and fully conditional specification (FCS) methods, allowing for some flexibility of the conditional models while providing a valid joint distribution. IVEware, a product of the University of Michigan, utilizes Sequential Regression Multiple Imputation (SRMI; see Ragunathan et al. 2001), which utilizes FCS methodology for mixed data types and allows for more flexibility in terms of variable type and the incorporation of edit logic. In this study, the empirical performance of IVEware and ISR for use in ARMS are compared through simulation.

Introduction

The Agricultural Resource Management Survey (ARMS) is an annual survey administered by the National Agricultural Statistics Service (NASS) in three phases. ARMS is an invaluable source of information on the current state of agriculture. Farm production expenditures are published annually in conjunction with the United States Department of Agriculture (USDA) Economic Research Service's (ERS) related releases, such as reports on net farm income. Data users of the ARMS survey include Congress, USDA, NASS, ERS, Bureau of Economic Analysis, researchers, and agri-business officials. Some data users develop forecasts for personal business that affect food supply and prices. Others use ARMS III data as part of an analysis to establish and review policy or to assess standards in an array of areas, such as food production, rural economies, bioenergy, and the environment.

To fully assess the link between policy, operation profitability, and operator household financial health, the third phase (ARMS III) survey instrument is long and complex. Moreover, it asks detailed characteristics and financial information about the farming operation, field practices, and the operator's household. NASS has taken steps to increase awareness of the benefits of the survey and to reduce respondent burden. Many operators, however, still may not find utility in responding due to the magnitude of perceived personal costs (time, privacy, etc.), confusing questionnaire design, or possibly anti-government sentiment (Dillman, 2007). Item nonresponse can be over 50% in the responding units for some items requested on the questionnaire (Miller and O'Connor, 2012).

Despite the difficulty in obtaining full responses from all operations that are sampled, the need remains from ARMS III data users to perform effective multivariate statistical analysis with the high-dimensional, mixed-type data. With item responses missing, care must be taken to ensure that estimates are not biased and subsequent inferences are valid. The potential magnitude of the bias increases as the proportion of missing responses increases. Although the calibration of the sampling weights is useful in mitigating bias in many estimated totals (Earp, et. al., 2008), the potential disturbance of the true variation and of the complex relationships among the items of ARMS III is of concern. To mitigate bias of estimates and to maintain the integrity of relationships in ARMS III data in the presence of item nonresponse, NASS uses imputation.

In 2009, NASS entered into a cooperative agreement with the National Institute of Statistical Sciences to update the imputation methodology used for ARMS III to better reflect the multivariate nature of the data. The review of potential options focused on two classes of modern methods that showed the greatest potential for application to ARMS III data: 1) data augmentation procedures for multivariate normal data (DA) and 2) a class of methods that build an imputation model via fully conditional specification (FCS). Schafer (1997) describes DA in detail, and implementations of the methodology can be found in the software package NORM (Schafer 1999) and within the SAS procedure MI (Yuan 2014). The DA method was applied to ARMS III data (following data transformations) in Robbins and White (2011) and was shown to markedly out-perform the ARMS III conditional mean imputation method. However, one drawback of this method is that the imputer is forced to assume that there exists a relationship among all variables in the model. With the high dimensionality of the ARMS III dataset, this was an unreasonable assumption. Van Buuren et al. (2006) describes the benefits and drawbacks of FCS in detail. FCS develops a conditional model for each variable by conditioning on all other variables, which allows the imputer flexibility to impute where the joint distribution is not explicitly defined (e.g. mixed categorical and continuous data). However, the joint distribution is implicitly assumed. By conditioning on all other variables, the joint distribution is over-specified; consequentially, the Markov chain may diverge. FCS appears in several widely-used imputation algorithms, including MICE (Van Buuren and Oudshoorn 1999); SRMI, which is built into the well-known IVEware package (Raghunathan et al. 2001); mi (Gelman and Hill 2011); and the SAS MI procedure (Yuan 2014). Methodology to update the ARMS III imputation methodology was developed by a cross-sector (NASS/ERS and academia), cross-discipline (statisticians and economists) team over the course of the two-year agreement. It is called Iterative Sequential Regression (ISR) and is described by Robbins et al as a blend of FCS and DA (Robbins, et al, 2013). ISR is a regression-based technique that allows for flexibility in the selection of conditional models while providing a valid joint distribution. Parameter estimates and imputation are obtained using a Markov chain Monte Carlo sampling method (Robbins, et al, 2011). ISR performed better than the conditional mean method and in terms of sample means, variances, and covariances. Further analysis concluded that ISR maintained the distribution of the data better than the conditional mean method. NASS further developed the ISR methodology and research program into an operational program implemented in 2015.

Agricultural Resource Management Survey (ARMS)

The ARMS is administered in three phases. The first phase is a screening phase for in-scope and in-business farms as well as presence of the targeted commodities for that year, which changes from year-to-year. The second phase asks for detailed field-level data for the targeted commodity for that year. The third phase (ARMS III) is a multi-mode, dual frame survey conducted annually in all states except Alaska and Hawaii. The sample consists of approximately 35,000 farms and ranches and is selected from NASS's list frame that attempts to cover all agricultural establishments within the U.S. and an area frame which compensates for the incompleteness of the list frame. The survey questionnaire is mailed to the entire sample, but additional modes of data collection include web and face-to-face. The questionnaire contains over 800 items for the respondent to complete and for NASS to process after data collection.

Based on data collected from the ARMS III, NASS publishes estimates of farm production expenditures for the U.S. (except Alaska and Hawaii) in addition to five regions. The regional estimates are broken down by some of the leading cash receipt states and then all other states within the region. Farm production expenditures are also estimated for eight economic sales classes and for two farm type categories. In addition to farm production expenditures, the ARMS III collects data on production practices and costs of production for one to three targeted crop and livestock commodities each year, selected on a rotational basis. The production practices and cost of production data for these designated commodities are collected in the top producing states while the farm production expenditures data are collected in all states (except Alaska and Hawaii).

Processing for the ARMS III survey is conducted in phases (See Figure 1). In the first phase, a computer edit checks the consistency of the data and verifies that data values fall within a certain range. A statistician reviews all questionnaire items flagged as errors, and either manually imputes data or marks the items to be imputed as non-zero by a separate statistical imputation process near the end of data collection. In 2014, the number of items that were eligible to be flagged for the statistical imputation process was eighty. Typically, manual imputation is performed when the statistician has knowledge about the questionnaire item for that operation. Other variables may be flagged for imputation by ERS using a separate imputation process for their use. After all of the items on the records have either a value to pass the edit or have been flagged for statistical imputation, statistical imputation (the second phase) is run. Once the data are imputed, the third phase begins with a run through an edit. After the edit has flagged any new errors, the statistician can make changes to fields to resolve an edit flag. The sampling weights are then calibrated; the final phase of editing and imputation is outlier analysis. Weights are changed and re-calibrated as necessary.

After all of the phases are complete, the data are summarized and NASS produces a report that includes estimated totals of farm production expenses. The dataset used for the summary is passed to ERS for further multivariate analysis.

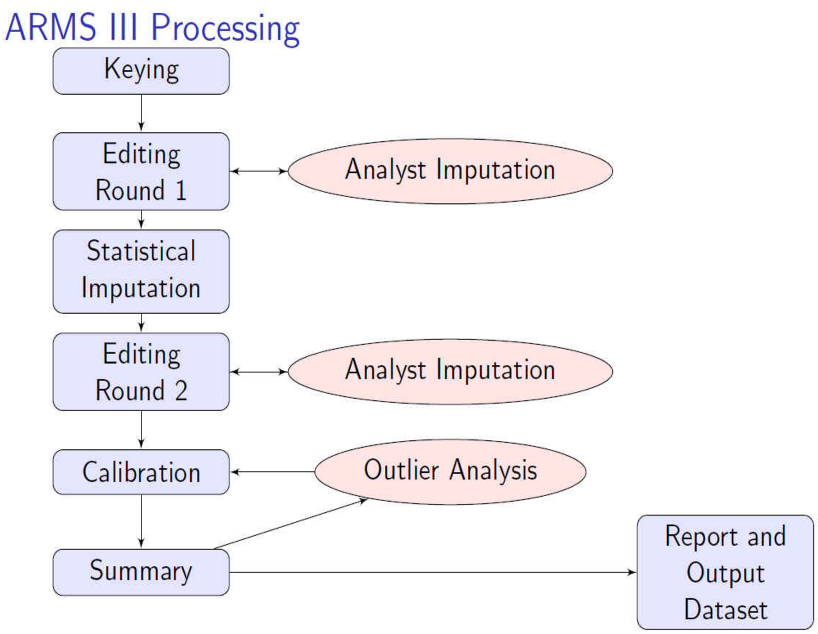


Figure 1. Processing of ARMS III

Iterative Sequential Regression (ISR)

Three parts to the ISR procedure are overviewed here. For more technical details and theory, see Robbins, et al., 2013.

1. Transformations
2. Model Selection
3. Generation of Imputations

Transformation techniques are used to handle the semi-continuous nature of the ARMS III dataset. The density of many ARMS III items can be described as a mixture of a skewed distribution and a point mass at zero. In ISR all missing values are assumed to be non-zero; this is determined through the edit and manual imputation processes.

Then zero portions of the variables are set to missing and the non-zero portion of the variables are transformed to be normal using one of a suite of transformations available in the procedure: log, log skew normal, log kernel density, and log empirical density. The transformations can be specified within a parameter file, or a default can be used where the transformation used is determined by the number of non-zero observations available.

Model structure in ISR allows the imputation procedure to run jointly on a group of variables, while allowing select variables (both imputed and fully observed) to span across blocks by defining each variable's role in the model. Model Groups are selected and some Model Groups are run together. Model Groups that run together are in the same Imputation Group. Descriptions of the roles are in table 1 and a diagram of the dependencies are in figure 2. The program will drop potential covariates from the models where the number of pairwise non-zero values are too few or a covariate may lead to a poorly conditioned covariance matrix.

Table 1. Description of variable roles in the models for ISR

Variable Role	Description
Global Covariates	Fully observed and used as a covariate in all of the imputations
Require Imputation	Require imputation by NASS and are not used as a covariate in imputations outside of its Model Group
Group Covariates	Fully observed and only used as covariates within the assigned Model Group
Global Contributors	Require imputation by NASS or are later imputed by ERS and are used to inform imputations for other variables that need imputation within its Impute Group

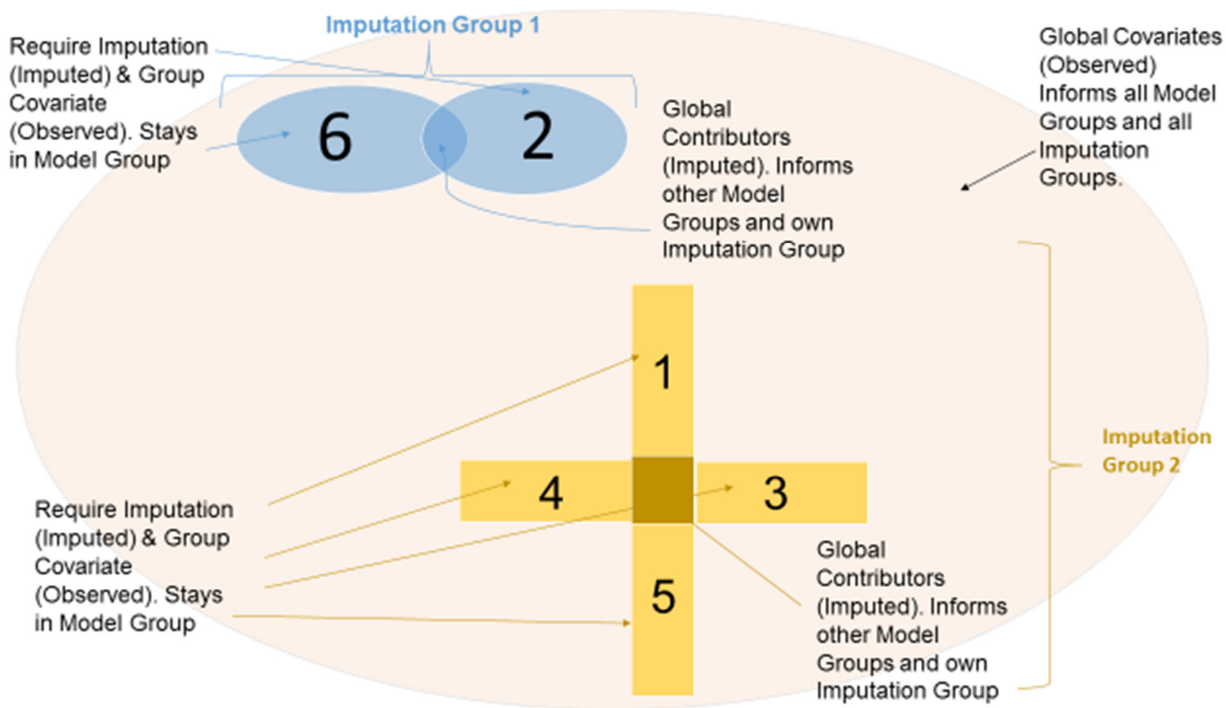


Figure 2. Diagram of model with variable roles. Large dark numbers within each shape denote the Model Group number.

Imputations are generated via MCMC sampling from the joint distribution of the variables requiring imputation conditioned on the fully observed covariates. The procedure is initialized using a sequential regression and may be regarded as an application specific example of the initialization step of the SRMI technique of Raguhnathan, et al.

(2001). ISR executes Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) and iterative draws of parameters from the posterior distribution and imputations from the conditional distributions. The technique used falls into a general class of methods known as data augmentation (DA, Tanner and Wong, 1987). The transformation and modeling ensures that the series of conditional models are jointly normal. Because a valid density is formed, established theory (Tierney 1994) assures convergence of the chain. Hence, ISR has theoretical justification and some flexibility in selection of certain conditional distributions. However, ISR is not constructed to retain theoretical justification where data that are not continuous or semi-continuous in nature need to be imputed.

Moving into production, generalizations to allow ISR's application to NASS's ARMS III data over survey years were made through the integration of a series of parameter files. The program is written in R and C, which are hosted on a Linux server, and a SAS interface was written to move necessary files and to run the program from a Windows environment.

IVEware

IVEware is software created by researchers at the Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, to produce single or multiple imputations using SRMI as described in Raghunathan, et al., 2001. SRMI is a popular and well understood methodology; a brief overview of the process follows to allow for comparison to ISR. Full technical detail can be found in the previously noted paper.

As with ISR, the joint conditional distribution can be factored as a series of univariate conditional distributions. A Gibbs sampling algorithm (Geman and Geman 1984; Gelfand and Smith 1990) is developed. After initialization, iterative draws of parameters and imputations can be made, where each conditional model may be linear or non-linear (e.g. generalized logit) in nature and a diffuse prior is used for the parameters.

IVEware is available as stand-alone program, or it can be run in SAS (SAS callable). Several modules are available to not only do imputation but to also conduct analysis of the data. For this study, the IMPUTE module was used. The IMPUTE module not only defines the model but also contains a host of other features that may be appealing to NASS. Some of the features of the IMPUTE model are defined below (see IVEware manual for full details).

Within the IMPUTE module, the type of regression used can be determined by defining the variable type. Variable types that can be imputed include continuous, binary, categorical (polytomous with more than two categories), counts, and semi-continuous. All variables in the dataset are potentially used in each conditional model, unless indicated in the transfer statement. Hence, variables may not take on all of the roles allowed in the ISR program; therefore, some of the relationships preserved by the conditional models may not be preserved using IVEware. The imputer has options to utilize statements for model selection, such as step-wise regression, minimum R-squared, and maximum number of predictors. The user also has features to incorporate some types of edits, such as restrictions on variables to be imputed based on the value of another variables and bounded imputations. The user may opt to transform the data before imputing.

IVEware is free, user-friendly, and easy to apply on a variety of data sources. Empirically, FCS methods, like those implemented in IVEware, have produced reasonable results (see Ragunathan, et al., 2001; Van Buuren et al., 2006; White and Reiter, 2008) with a high degree of variable flexibility and other desirable features for implementation by a statistical agency. However, the user accepts that convergence may not be reached due to a potential lack of a valid joint distribution.

Methods

The final goal of this work is to compare the performances of ISR and IVEware using selected fully observed study variables from 2013 ARMS III, simulating missingness, and then imputing using both methods. This is a progress report toward that goal.

Here progress on the following two steps is discussed.

1. Conduct a simulation study of IVEware using simulated missing values for ARMS IIII 2013 data and assess the performance of IVEware using analysis measures to be described.

2. Impute ARMS III data as in an operational setting using IVEware and ISR and compare the estimates based on the imputed data from the two approaches. .

1. Simulation Study Description.

Some of the variables are strongly correlated while others are weakly correlated. The study variables are listed in table 2.

Table 2. List of variables used in simulation study. *denotes variables with imposed missingness

Variable	Variable Description	Variable Type
FARMTYPE*	Type of farm	Categorical (Crop = 1, Livestock = 2)
EOY BREEDING LIVESTOCK VALUE (P864)*	End of year livestock value	Semi-Continuous
FERTSEXP*	Fertilizer expenses for the year	Semi-Continuous
LVSTKEXP	Livestock related expenses	Semi-Continuous
SEEDSEXP	Seed expenses	Semi-Continuous
EOY CROP VALUES (P889)	End of year crop value	Semi-Continuous
CROPLAND ACRES (P63)	Acres of cropland on the operation	Semi-Continuous
TOTAL ACRES (P26)	Total acres on the operation	Continuous
Region	Region	Categorical
GVCLS	Gross total value of the operation	Categorical

Three variables, FARMTYPE, EOY LIVESTOCK VALUE, and FERTSEXP, were selected to impose missingness. FARMTYPE is a categorical variable for which the quality of the survey collected measurement is considered to be high. FERTSEXP is a value an operator has available on tax forms, so it is considered reliable and relatively error free. EOY LIVESTOCK VALUE is a value that NASS is considering for imputation in the future.

Missing was induced under three missingness models: (1)Missing Completely At Random (MCAR), (2) Missing at Random (MAR), and (3) Missing Not At Random (MNAR). Data are MCAR if the probability of missingness is unrelated to the value of the observation or the value of other variables in the dataset. If the missingness depends on other variables in the dataset but is unrelated to the value of the observation, data are MAR. Data are MNAR if other variables in the data set do not explain the missingness and the pattern of missingness is related to the value of the missing variable itself. Both the ISR and IVEware methodologies studied here are constructed to handle data that are MAR and thus would be appropriate for MCAR data as well, but not MNAR data.

For ARMS III, missing items eligible for imputation have been through an edit process that determines whether the value is zero or non-zero. Hence, in this first stage of the study, missing values are only imposed for non-zero values. From the population of fully observed respondents, 250 datasets were created with missing values under each of the missingness models. For each of the three selected variables and under each of the missingness models, approximately 30% of the nonzero values were removed from each dataset.

We apply two IVEware imputation strategies to ARMS III simulated datasets: IVEtrans and IVEuntrans. Descriptions are in table 3.

Table 3. Descriptions of imputation strategies for simulation study

Imputation Strategy	Description
IVEuntrans	IVEware applied without transformation of continuous variables.
IVEtrans	IVEware applied with non-zero indicator created, zeros set to missing values (but replaced), and nonzero values transformed using log transformation.

For both of the IVEware applications, a minimum Rsquare was used for variable selection and bounds close to the 99^{th} percentile of the observed values in the population were used for FERTSEXP and EOY LIVESTOCK VALUE.

For analysis, estimates were compared using the dataset containing all of the observed values to estimates from the imputed datasets. The differences in the means (proportion in the case of FARMTYPE) were examined.

2. Operational Application

For reference year 2013, ARMS III data were imputed using three imputation strategies: (1) transformed IVEware (IVEtrans), (2) untransformed IVEware (IVEuntrans), and (3) ISR.ISR (See table 4).

Table 4. Descriptions of imputation strategies for operation application

Imputation Strategy	Description
ISR	Six model groups. Two imputation groups. Variable roles are covariate, require imputation, global covariate, global contributor. ISR transformed data with non-zero indicator created, zeros set to missing values (but replaced), and nonzero values transformed using default described.
IVEuntrans	IVEware applied without transformation of continuous variables. Used two imputation groups.
IVEtrans	IVEware applied with non-zero indicator created, zeros set to missing values (but replaced), and nonzero values transformed using log transformation. Used two imputation groups.

Approximately 150 variables out of over 800 variables collected on the ARMS III questionnaire were imputed for each year's dataset. . The models change some from year to year. The models used for IVEware were as similar as possible to ISR in terms of eligible covariates. Due to ISR's flexibility to allow multiple model groups to be run together but only some of the variables to be shared between model groups (see previous sections of this paper), we could not match this precisely.

Results

1. Simulation Study

We consider the differences in the mean estimates from the imputed datasets relative to the true mean from the observed data for FERSTEXP and EOY LIVESTOCK VALUE. Estimates made using single imputations.

$$\text{Relative Difference} = \frac{\text{Imputed Mean} - \text{True Mean}}{\text{True Mean}}$$

Positive values of the relative difference indicate larger mean estimates in the imputed dataset than the dataset with the true values, and negative values indicate smaller mean estimates in the imputed dataset than in the dataset with the true values.

In the case of FARMTYPE, the differences in the total number of crop farm estimates from the imputed datasets relative to the true number of crop farms from the observed data was evaluated for each simulated dataset. Estimates were made using imputations.

$$\text{Relative Difference} = \frac{\text{Imputed Total} - \text{True Total}}{\text{True Total}}$$

Positive values of the relative difference indicate more crop farms in the imputed dataset than the dataset with the true values, and negative values indicate less crop farms in the imputed dataset than in the dataset with the true values.

The following figures (Figure 3, Figure 4, and Figure 5) plot the differences in the mean estimates relative to the true mean. Each plotted point represents this relative difference for a simulated dataset. The red line indicates a relative difference of 0 and points scattered across the plot randomly above and below this line indicate no bias in the estimate. However, observing a tendency for points to be above or below the red line indicate positive and negative bias, respectively.

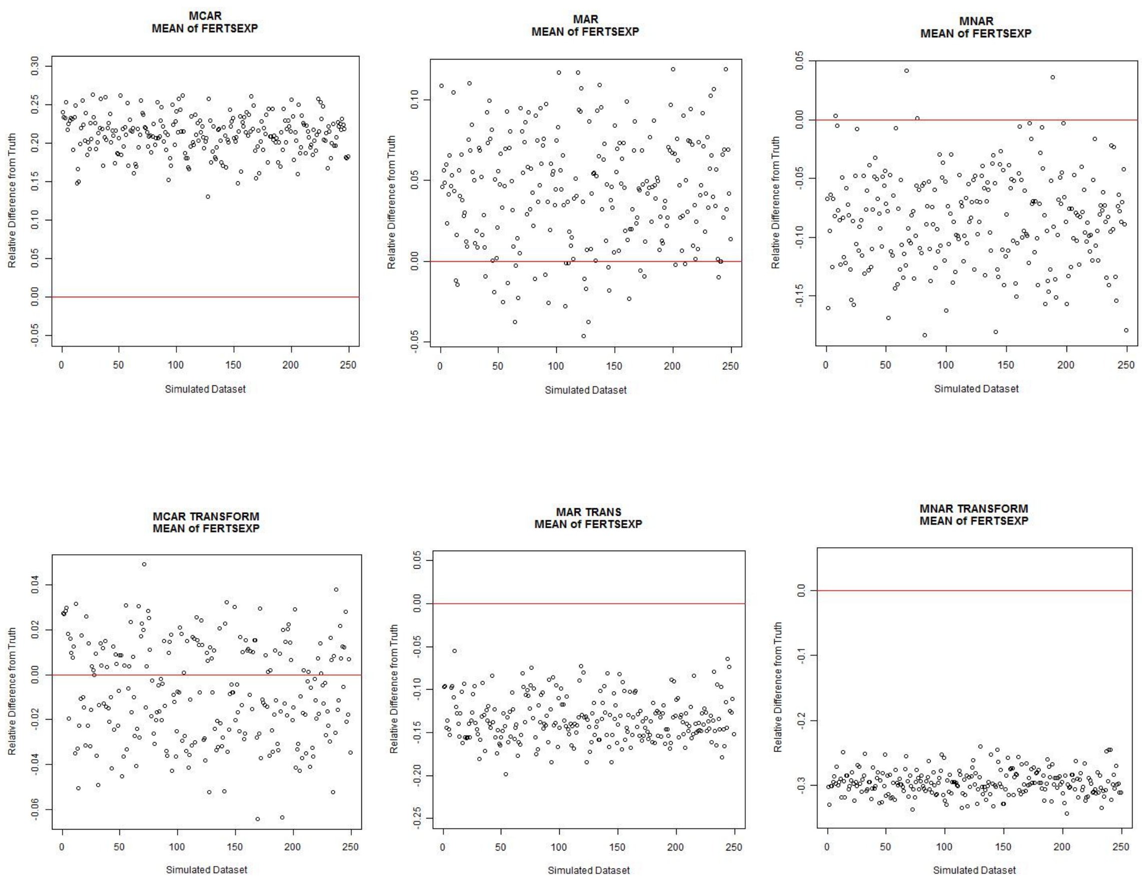
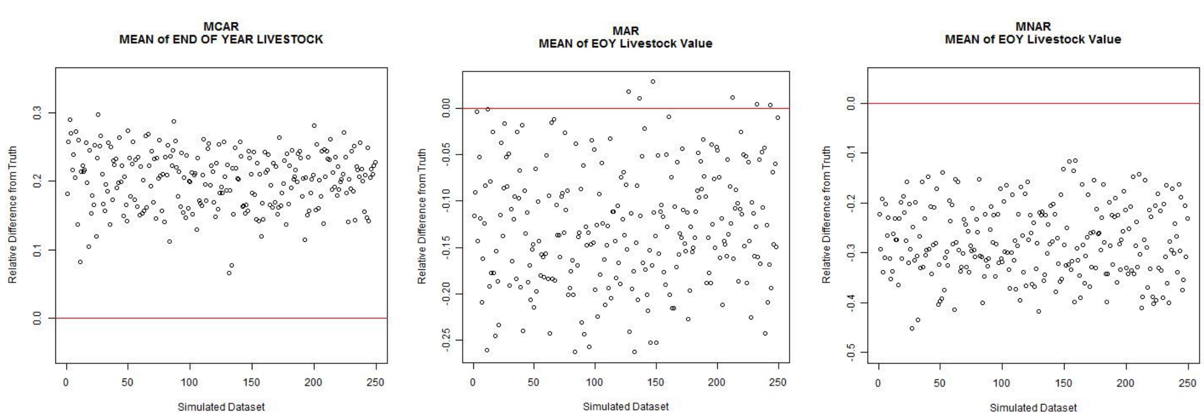


Figure 3. Plot of relative differences in means for FERTSEXP. The first row of plots represents the relative differences using IVEuntrans method under each missingness condition and the second row contains the plots for IVEtrans.



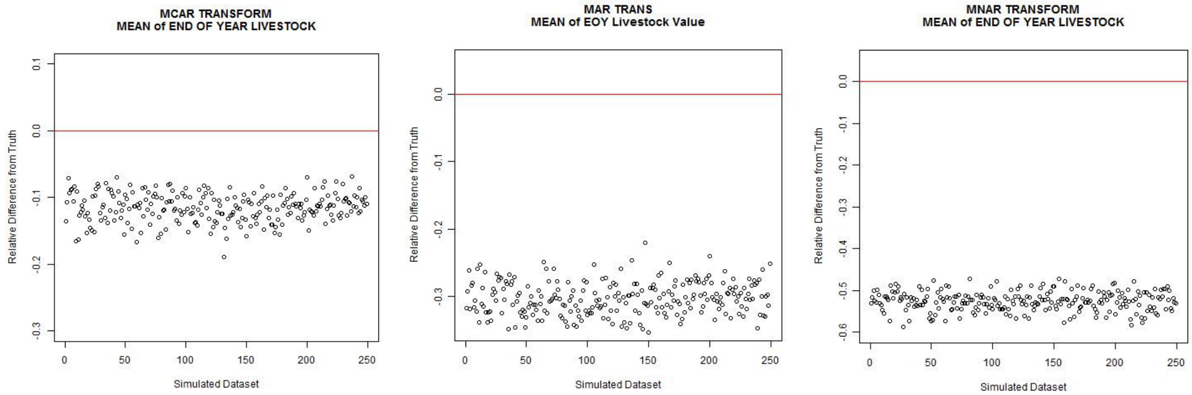


Figure 4. Plot of relative differences in means for EOY Livestock Value. The first row of plots represents the relative differences using IVEuntrans method under each missingness condition and the second row contains the plots for IVEtrans.

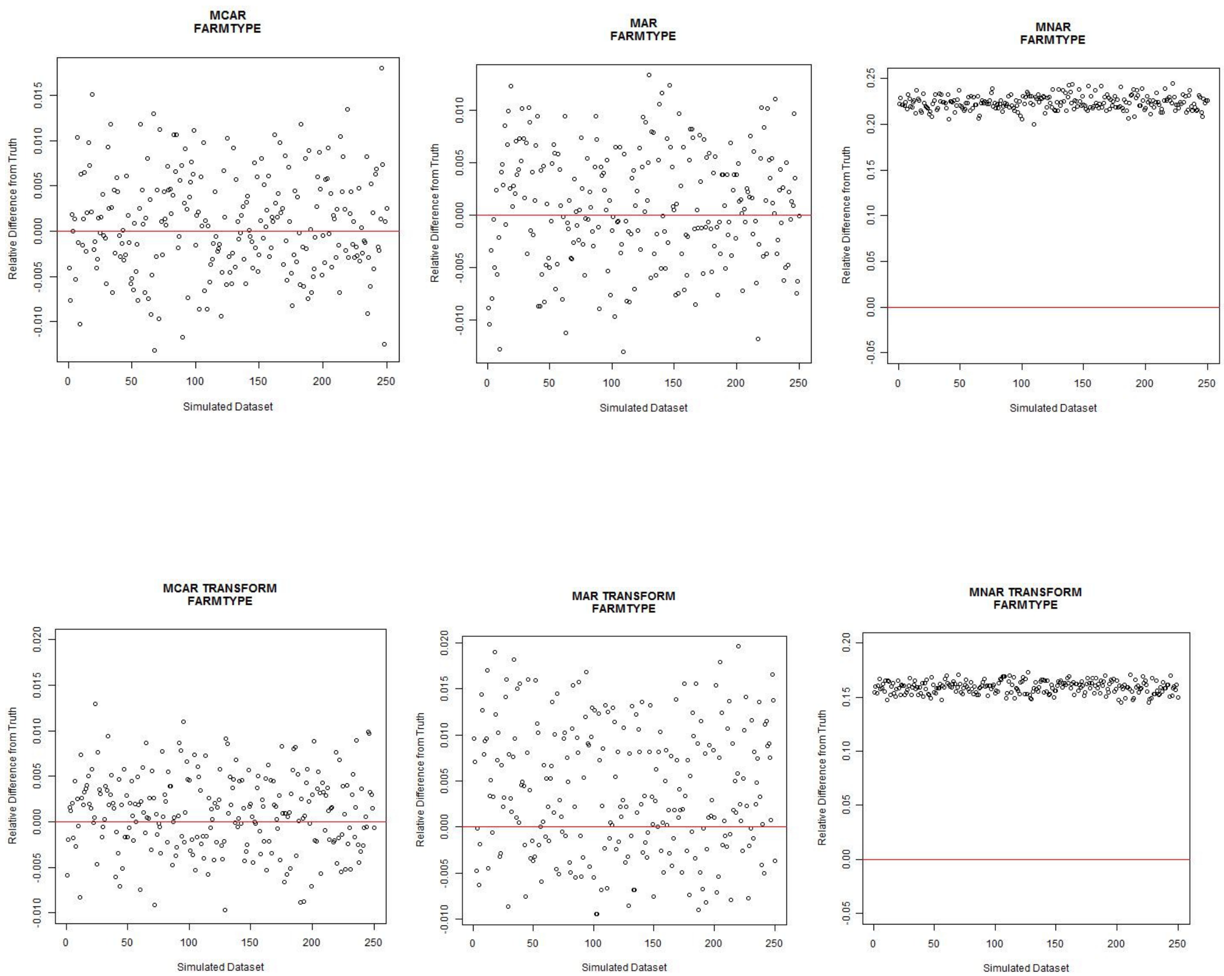


Figure 5. Plot of relative differences in means for FARMTYPE. The first row of plots represents the relative differences using IVEuntrans method under each missingness condition and the second row contains the plots for IVEtrans.

In the case of MAR, where IVEware is intended for use, we see some bias in the case of FERSTEXP and EOY LIVESTOCK VALUE. Although no sensitivity analysis was conducted for the models used, it may be that case that utilizing a different set of covariates for imputation and producing the missing value introduced bias. When the data are MNAR, we expect to see bias in the estimate. Note that the direction of the bias corresponds with the model used to produce the missingness (deleting values) under MNAR. For example, to create MNAR data for FARMTYPE, the probability of being missing was larger for livestock farms, so we see that bias in the proportion of crop farms is positive. We also see that the estimates can be biased even in the case of MCAR for some of our imputation strategy/variable combinations.

Overall, the performance of IVEware was best for FARMTYPE, the categorical variable in the study. If we consider the case where data are MAR (most often assumed in practice), the not transforming the variables before imputation performed the best in terms of bias.

2. Operational Application

Running the IVEware imputation process on the ARMS III dataset revealed that IVEware did not always impute within the bounds set by the programmer. Imputed values outside the set bounds were noted in the log produced by IVEware; this occurred infrequently. Also, ISR was applied using five hundred iterations while IVEware was applied using ten iterations. Five hundred iterations was determined to be used for ISR applications from review of convergence diagnostics over applications of ISR to several years of ARMS III data. Even with some modifications to the workhorses (macros) of the IVEware software, IVEware failed to complete more than ten iterations on a consistent basis. Examining the mean estimates using a number of iterations between three and ten revealed little change across iterations. Moreover, recommendations from the literature suggest no more than ten iterations for applications with moderate amounts of missingness, which is the case for most of the ARMS III variables. So, ten iterations for the IVEware models were used. Therefore, the run time for IVEware was significantly less than ISR (one hour versus eight hours). If results are similar, this would be a positive aspect of IVEware.

Conclusion

Ultimately, the goals are to assess the cost and benefits of using each software with different datatypes. In this report of the progress to answer this question, we conclude a couple of points from our application of IVEware in our simulation study: 1) IVEware performed the best in terms of bias with our categorical variable in the study and 2) IVEware tended to perform best under MAR (the assumed case in practice) when data are not transformed. Developing ISR into a program that can be applied to surveys other than ARMS III would take considerable effort. Would making those revisions and applying ISR to the different types of data frequently collected by NASS be better? Further evaluations are being conducted. However, at the present time, it would be better to move to IVEware.

References

Allison, P. (2001). “Missing Data”. Thousand Oaks, CA: Sage Publications.

Barboza, W., Miller, D. and Cruze, N. (2014). “Assessing the Impact of a New Imputation Methodology for the Agricultural Resource Management Survey”. United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Paris, France, 28-30, April 2014.

Dillman, D. A. (2007). “Mail and Internet Surveys: The Tailored Design Method”. Hoboken, NJ: John Wiley & Sons, Inc. 2nd ed.

Earp, M., McCarthy, J., Schauer, N. and Kott, P. (2009). “Assessing the Effect of Calibration on Nonresponse Bias in the 2006 ARMS Phase III Sample Using the Census 2002 Data”. Proceedings from the 2009 Joint Statistical Meetings.

Gefland, A. E. and Smith, A. F. M. (1990), “Sampling-Based Approaches to Calculating Marginal Densities,” Journalof the American Statistical Association, 85, 398-409.

Gelman, A.and Hill, J. (2011), “ Opening Windows to the Black Box”. Journal of Statistical Software, 40.

Geman, D. and Geman, S. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Reconstruction of Images". IEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

Little, R. J. A. (1988). "Missing-Data Adjustments in Large Surveys". Journal of the Business & Economic Statistics, 6, 287-296.

Little, R. J. A. and Rubin, D. B. (2002). "Stastitical Analysis with Missing Data", New Jersey: John Wiley & Sons, 2nd ed.

Miller, D., Robbins, M., and Habiger, J. (2010). "Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey". Proceedings of the 2010 Joint Statistical Meetings, pages 816-829.

Miller, D. and O'Connor, T. (2012). "Item Response Rates for the Agricultural Resource Management Survey III in 2006 and 2007". National Agricultural Statistics Service Research Report, RDD-11-07.

Miller, D. and Dau, A. (2015). "Capturing Additional Variability Introduced by Imputation within the Agricultural Resource Management Survey". 2015 Joint Statistical Meetings Proceedings.

National Agricultural Statistics Service (2014). "Farm Production Expenditures Methodology and Quality Measures".
http://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Farm_Production_Expenditures/08_2014/FPEXPQM.pdf

National Agricultural Statistics Service (2014). "Farm Production Expenditures 2013 Summary".
http://usda.mannlib.cornell.edu/usda/nass/FarmProdEx//2010s/2014/FarmProdEx-08-01-2014.pdf

Raghunathan, T.E., Lepkowski, J.M., Hoewyk, J.V. and Solenberger, P. (2001). "A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models". Survey Methodology, 27, 85-95.

Robbins, M., Ghosh, S., and Habiger, J. (2010). "Innovative Imputation Techniques Designed for the Agricultural Resource Management Survey". Proceedings of the 2010 Joint Statistical Meetings, pages 634-641.

Robbins, M., Ghosh, S., Goodwin, B., Habiger, J., Kosler, J., Miller, D., and White, K. (2011), "ARMSimpute: A Computation Algorithm for Imputation in ARMS III." Tech. re., National Institute of Statistical Sciences/National Agricultural Statistics Service.

Robbins, M. W. and White, T. K. (2011). "Farm Commodity Payments and Imputation in the agricultural Resource Management Survey". The American Journal of Agricultural Economics, 93, 606-612.

Robbins, M., Gosh, S., and Habiger, J. (2013). "Imputation in high-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey". Journal of the American Statistical Association, 108:501, 81-95, DOI: 10.1080/01621459.2012.734158.

Rubin, D. B. (1987) "Multiple Imputation in Surveys". A Phenomenological Bayesian Approach to Nonresponse". In Proceedings of the Survey Research Methods Section of the American Statistical Association, American Statistical Asssociatiaon, p.20-34.

Rubin, D. B. (1987). "Multiple Imputation for Nonresponse in Surveys". New York, New York: John Wiley & Sons.Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data, Chapman & Hall/CRC.

Schafer, J.L. (1999). NORM users' guide (Version 2). University Park: The Methodology Center, Penn

State. Retrieved from http://methodology.psu.edu

Schafer, J. L. and Graham, J. W. (2002). "Multiple Imputation for Missing Data: Our View of the State of the Art," Pyschological Methods, 6, 147-177.

Schenker, N., Ragunathan, T. E. , Chiu, P.L., Makuc, D. M. , Zhang, G., and Cohen, A.J. (2006). "Multiple Imputation of Missing Income Data in the National Health Interview survey" Journal of the American Statistical Association, 101, 924-933.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distriutions by Data Augmentation (with Discussion)". Journal of the American Statistical Association, 82, 528-550.

Tierney, L. (1994). "Markov Chains for Exploring Posterior Districtutions (with Discussion)". Annals of Statisticas, 22, 1049-1064.

Van Buuren , S., Brand , J. P.L., Groothuis-Oudshoorn, C. G.M., and Rubin, D.B. (2006). "Fully conditional specification in multivariate imputation". Journal of Statistical Computation and Simulation, 76:12, 1049-1064, DOI: 10.1080/10629360600810434

Van Buuren, S. and Oudshoorn, C. G. M. (1999). Flexible Multivariate Impuation by MICE, TNO Preventie en Gezondheid, Leiden, for associated software see http://www.multiple-impuation.com

Vizcarra, B. and Sukasih, A. (2013). "Comparing SAS PROC MI and IVEware Callable Software". 2013 SouthEast SAS Users Group Conference Proceedings.

von Hippel, P. T. (2007). "Regression with Missing Y's: An Improved Strategy for Analyzing Multiply Imputed Data" Sociological Methodology, 37, 1-54.

von Hippel, P.T. (2013). "Should a Normal Imputation Model be Modified to Impute Skewed Variables?" Sociological Methods & Research, vol 42: 105 – 138.

de Waal, T., Pannekoek, J., and Scholtus, S. (2011). "Handbook of Statistical Data Editing and Imputation". Wiley Handbooks in Survey Methodology. John Wiley & Sons, Inc.

White, T.K., and Reiter, J.P. (2008). "Multiple Imputation in the Annual Survey of Manufacturers". 2007 Research Conference Papers. Federal Committee on Statistical Methodology, Office of Management and Budget, Washington D.C.

Yuan, Y. (2014). Sensitivity Analysis in Multiple Imputation for Missing Data. In Proceedings of the SAS Global Forum 2014 Conference, Washington D.C.