Innovative Imputation Techniques Designed for the Agricultural Resource Management Survey

Michael W. Robbins * Sujit K. Ghosh † Joshua D. Habiger ‡

Abstract

The Agricultural Resource Management Survey (ARMS) is a high dimensional, complex economic survey which suffers from item non-response. Here, we introduce methods of varying complexity for imputation in this survey. The methods include stratified mean imputation, the approximate Bayesian bootstrap, and non-iterative and iterative sequential regression. The iterative sequential regression is a form of Markov chain Monte Carlo (MCMC) that is unique in that it allows for flexible selection of conditional distributions while utilizing joint modeling. Each of the regression procedures require data-driven transformations that allow for the implementation of a conditional multivariate normal model.

Key Words: Missing Data, Imputation, ARMS, Markov Chain Monte Carlo

1. Introduction

In this paper we consider imputation methods which are applicable to the US Department of Agriculture's (USDA) Agricultural Resource Management Survey (ARMS). The ARMS is a multi-phase survey which contains 35,000 records of 1,000-2,000 variables that is administered annually by NASS and the Economic Research Service (ERS), which are both subsidiaries of the US Department of Agriculture (USDA). The current imputation method which NASS uses on the ARMS is an out-dated form of mean imputation which distorts several data attributes. Our goal is to develop a procedure that will maintain all distributional characteristics of the complete data, had there been no missing values.

The ARMS is the USDA's primary source of information on the financial condition, production practices, and resource use of farms, as well as the economic well-being of the nation's farm households. The scope of the information collected in the ARMS is too large to be further paraphrased here — to quote National Research Council (2008), "No other source affords such a comprehensive view of the American farm." The ARMS data are indispensable to federal and private sector decision makers when considering policies and programs or business strategies relating to the farm sector.

The complete survey is administered in three phases, and here we concentrate on imputation in the third phase (ARMS III). ARMS III typically has 3-5 versions which are administered in total to about 35,000 farm operations over the contiguous United States.

The Panel to Review the USDA's Agricultural Resource Management Survey was established in 2006 and was chaired by Bruce Gardner; its findings are outlined in National Research Council (2008). This reference also provides a detailed overview of the ARMS, as well as the survey design and processing. The research discussed in this paper is the result of the Panel's recommendations.

*National Institute of Statistical Sciences, Research Triangle Park, NC 27709–4006

†Department of Statistics, North Carolina State University, Raleigh, NC 27695–8203

‡Department of Statistics, Oklahoma State University, Stillwater, OK 74078–1056

Miller et al. (2010) provide a good outline of the ARMS and its data characteristics as well as a discussion on particular survey aspects that make imputation in ARMS a particularly challenging problem. Here, we paraphrase these challenges. Due to the large number of ARMS data users, it is essential that no data characteristics (i.e., means, variances, covariances) be distorted by the imputation processes. The large number of variables within the survey make it particularly difficult to preserve all variable relationships throughout the imputation process. Likewise, it is difficult to preserve the confounding marginal structure of ARMS variables throughout the imputation process. For instance Miller et al. (2010) notes that most ARMS variables are mixed discrete/continuous in distribution. That is, these variables contain a portion of zeros and the remaining portion has a positive continuous density. A skew normal density (Azzalini, 1985) often fits the log of the positive portions. All values which require imputation are known to be positive.

We continue by introducing imputation methods which are applicable to ARMS. In Section 2 we outline methods that utilize stratification, including the current NASS method. In Section 3 we outline transformation techniques which will be required in order utilize regression methods. In Section 4 we outline a non-iterative regression technique which we call sequential regression. In Section 5 we introduce iterative sequential regression, which is a type of Markov chain Monte Carlo (MCMC). Section 6 offers some concluding thoughts.

2. Imputation via Stratification

The current NASS imputation procedure involves stratification. Hence, the imputation model used may be described as a 3-factor ANOVA table with interaction effects, where the three factors are: 1) Farm Type, 2) Region, and 3) Sales Class. The data are grouped into cells (or strata), where each cell contains all observations that have each the same value for each of the three factors. If a specific observation has a missing value for a specific variable, all observations of that variable in the corresponding cell with a positive and observed value make up the donor pool. NASS requires that a donor pool has 10 or more values, and if that requirement is not met, fallback groupings are used in order to broaden/merge the cells and to thereby expand the donor pool. See Banker (2007) for an ordered list of the fallback groups, as well as a more detailed description of the NASS and ERS imputation processes. Observed values that are determined to be outliers are excluded from the process.

2.1 Conditional Mean Imputation

The current NASS method employs conditional mean imputation. For this method, the impute for each missing value is taken as the mean of the values within the donor pool corresponding to that specific observation and variable.

The drawbacks of this method are numerous. Most noticeably, conditional mean imputation is well known to distort marginal variable characteristics, primarily by causing a downward bias in classical estimates of variance (see Little and Rubin, 2002; Schafer and Graham, 2002; Fichman and Cummings, 2003; Newman, 2003, among others).

2.2 Approximate Bayesian Bootstrap Imputation

The most obvious improvement over conditional mean imputation is a method that imputes with a random draw from a conditional distribution, as opposed to the mean of that distribution. Doing so should alleviate the downward bias in variance estimation. However, proper simulation from the true posterior distribution within each cell is infeasible, since the small number of observations within cells makes it difficult to determine appropriate distributional assumptions. It may be more feasible to impute using a draw from the observed values within that cell.

Approximate Bayesian bootstrap (ABB) imputation (Rubin and Schenker, 1986; Kim, 2002) accomplishes just that. For this method, donor pools are determined in the same fashion as in the current NASS method. Assume that the k^{th} cell corresponding to the j^{th} variable contains $n_{j,k}$ positive observed values and $m_{j,k}$ missing values. The set of positive values (the donor pool) is denoted $A_{j,k}$. ABB imputations are generated in two steps:

1. Draw a bootstrapped donor pool, $A^*_{j,k}$, by selecting $n_{j,k}$ values with replacement from $A_{j,k}$,
2. Draw imputations for the k^{th} cell of the j^{th} variable by selecting $m_{j,k}$ values with replacement from $A^*_{j,k}$.

ABB imputation is not thought to be proper in the Bayesian sense. Kim (2002) notes that this method induces bias into variances estimates found using MI. However, it does provide a simple method that should show certain improvements over the current mean imputation procedure.

3. Transformation Techniques

In order to integrate sophisticated multivariate models into the imputation scheme, we abandon the stratified approach and consider linear modeling. For our purposes, this will require normality assumptions, so we now consider transformation techniques that will achieve approximate joint normality.

3.1 Adjusting for the Mixed Variables

We adjust for the mixed nature of certain variables by using the following. Assume that Y_j, the j^{th} variable, represents a mixed-continuous variable. We break down Y_j into two variables, B_j and Y^*_j, where

$$B_j = \begin{cases} 1 & \text{if } Y_j > 0 \text{ or } Y_j = ?, \\ 0 & \text{if } Y_j = 0, \end{cases} \qquad \text{and} \qquad Y^*_j = \begin{cases} Y_j & \text{if } Y_j > 0, \\ ? & \text{if } Y_j = 0 \text{ or } Y_j = ?, \end{cases} \tag{1}$$

where a "?" represents a missing value. Any missing value of Y_j is known to be positive, thereby B_j is fully observed. In terms of the joint model, if Y_j is 0 then it is treated as being missing. An example of the creation of B_j and Y^*_j from Y_j is given in Table 1.

If Y_j is mixed and fully observed, we can still break the variable down in this fashion. Therefore, Y^*_j will have missing values whereas Y_j has none.

This technique for addressing the mixed nature of ARMS data results in a dataset where all variables with missing values have continuous distributions. Also, all information provided by observed zeros is still contained within the data (in the form of the B_j's).

Y_j	B_j	Y_j^*
0	0	?
9876	1	9876
0	0	?
?	1	?
0	0	?
?	1	?
12345	1	12345

Table 1: The process of breaking down a mixed variable (Y_j) into a fully-observed binary variable (B_j) and a positive continuous variable (Y_j^*).

3.2 Transformation of Positive Portions of Variables

We now consider the marginal distributions of the Y_j^*'s. As mentioned previously, the skew normal density often fits the log of the positive portions. A skew normal density contains three parameters: a location parameter (ξ), a scale parameter (ω) and a shape parameter (α). The j^{th} variable will have its own skew normal parameter set, which we denote $\{\xi_j, \omega_j, \alpha_j\}$. If these parameters are known, then skew normal data may easily be transformed into standard normal data. Let $F(y|\xi_j, \omega_j, \alpha_j), y \in \Re$ represent the cumulative density function (cdf) of the skew normal variate $\log Y_j$. If we define

$$T_j(y) = \Phi^{-1}(F(y|\xi_j, \omega_j, \alpha_j)) \tag{2}$$

then

$$T_j(\log Y_j) \sim N(0, 1).$$

where $\Phi(\cdot)$ represents the standard normal cdf. Since the values of ξ_j, ω_j, and α_j are unknown for each relevant j, we use MLEs found using available data. An inverse of this transformation may also be easily applied. We refer to the transformation in (2) as a "SN transformation".

For the j^{th} variable (which may or may not have missing values) we will consider one of three possible transformations to create the transformed variables X_j:

1. $X_j = Y_j^*$ (no transformation),
2. $X_j = \log Y_j^*$ (log transformation), (3)
3. $X_j = T_j(\log Y_j^*)$ (density transformation), (4)

where $T_j(\cdot)$ is defined in (2).

In the remaining procedures, we will impute for the missing values throughout the set of X_j's. Next, the resulting imputed vectors, which are denoted with $\hat{X}_j$, are untransformed, and values originally observed as zero are reset to zero.

4. Sequential Regression

One notable drawback of the stratified approach is that covariates must be categorical. Inclusion of additional covariates would likely result in having far too many empty cells. In order to incorporate more covariates (in particular, those which are continuous) into the imputation model, we must abandon the stratified approach and utilize regression techniques.

We continue with our specific notation which is in accordance with notation introduced in Section 3. Our imputation methods are run jointly on a block of variables. Of the variables in this block, we assume that r are mixed variables and have missing values. These are denoted $Y_1, \ldots, Y_r$. We also have q fully-observed mixed variables, denoted $Y_{r+1}, \ldots, Y_{r+q}$, and a set of fully observed discrete or continuous variables which are denoted $\boldsymbol{Z}$. We let $p = r + q$ represent the total number of mixed variables. Of course, as indicated at the end of Section 3, our methods will be applied to the corresponding $X_1, \ldots, X_p$. For our purposes, each of these X's has missing values, and thereby, in hopes of achieving a near-monotone missingness structure, they are indexed so that they are increasing in missingness (i.e., X_1 is the variable with the fewest missing values). We let $\boldsymbol{B} = \{B_1; \ldots; B_p\}$ and $\boldsymbol{\chi} = \{\boldsymbol{Z}; \boldsymbol{B}; X_1; \ldots; X_p\}$, and for completeness, we write $X_j = \{x_{1j}, \ldots, x_{nj}\}^t$ and $Y_j = \{y_{1j}, \ldots, y_{nj}\}^t$ for each j, where n represents the total number of observations.

We now introduce a class of regression procedures that will create imputations for the missing values in the p variables. These procedures are akin to the predictive mean matching technique analyzed in Horton and Lipsitz (2001) and the SRMI technique of Raghunathan et al. (2001) (the initialization step, to be specific). We will refer to these methods as *sequential regression* (SR). SR techniques are motivated by the fact that the joint distribution of $X_1, X_2, \ldots, X_p$ given $\boldsymbol{Z}$ and $\boldsymbol{B}$ can be factored into a sequence of conditional distributions as follows

$$\begin{aligned} P(X_1, X_2, \ldots, X_p | \boldsymbol{Z}, \boldsymbol{B}) = \\ P(X_1|\boldsymbol{Z}) \cdot P(X_2|\boldsymbol{Z}, \boldsymbol{B}, X_1) \cdots \\ P(X_p|\boldsymbol{Z}, \boldsymbol{B}, X_1, X_2, \ldots, X_{p-1}), \end{aligned} \tag{5}$$

where we use $P(\,\cdot\,)$ to denote a distribution function.

Letting $\hat{X}_j = \{\hat{x}_{1j}, \ldots, \hat{x}_{nj}\}^t$ represent the imputed version of X_j, sequential regression techniques will attempt to use $P(X_j|\boldsymbol{Z}, \boldsymbol{B}, \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{j-1})$ to create $\hat{X}_j$.

4.1 SR2* and SR3*

We let $\boldsymbol{B}_{-j} = \{B_1; \ldots; B_{j-1}; B_{j+1}; \ldots; B_p\}$, where the B_j's are defined in (1). We assume that, for $j = 1, \ldots, p$,

$$X_j = \beta_{j0} + \boldsymbol{Z}\boldsymbol{\alpha}_j + \boldsymbol{B}_{-j}\boldsymbol{\gamma}_j + \beta_{j1}X_1 + \ldots + \beta_{j,j-1}X_{j-1} + \sigma_j\boldsymbol{\epsilon}_j, \tag{6}$$

where $\boldsymbol{\alpha}_j$ and $\boldsymbol{\gamma}_j$ are vectors of coefficients and $\boldsymbol{\epsilon}_j$ is a length-n vector of IID standard normal variates. We let

$$\boldsymbol{\theta}_j = \{\beta_{j0}, \boldsymbol{\alpha}_j, \boldsymbol{\gamma}_j, \beta_{j1}, \ldots, \beta_{j,j-1}, \sigma_j\}, \qquad \text{and} \qquad \boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p\}.$$

We will find the imputed vector, $\hat{X}_j$, sequentially for $j = 1, \ldots, p$. The first step in imputing for X_j is to draw values of regression parameters that will be used to create the imputations. Assuming the model in (6), we let $\tilde{\boldsymbol{\theta}}_j$ represent a draw from the posterior distribution of $\boldsymbol{\theta}_j$ found using formulas of the form in Little and Rubin (2002), p. 114. The covariate matrix contains $X_1, \ldots, X_{j-1}$ (each of which have missing values), but the sequential nature of this procedure allows us to use the imputed versions of these variables instead. Since the response variable, X_j, also contains missing values, we include only observations which have an observed value of X_j when calculating the posterior distribution.

Sequentially for $j = 1, \ldots, p$, we create $\hat{X}_j$ by drawing from

$$\hat{x}_{ij} \sim P(x_{ij}|Z_i, \boldsymbol{B}_{i,-j}, \hat{x}_{i1}, \ldots, \hat{x}_{i,j-1}, \tilde{\boldsymbol{\theta}}_j),$$

whenever x_{ij} is missing. This is done by adding a randomly sampled error to the predicted mean found using (6) while assuming $\boldsymbol{\theta}_j = \tilde{\boldsymbol{\theta}}_j$. In the above, Z_i and $\boldsymbol{B}_{i,-j}$ represent the i^{th} row of $\boldsymbol{Z}$ and $\boldsymbol{B}_{-j}$ respectively. This process can be done while using the transformation in (3) or in (4), which yield the "SR2*" and "SR3*" methods respectively.

5. Iterative Sequential Regression

Most robust procedures (Spiess and Keller, 1999; Little and An, 2004; Von Hippel, 2007) follow the SR scheme we have outlined; however, in order to draw proper imputations using a SR technique, the missingness structure must be monotone. If the missingness is not monotone, it is possible, for example, that certain unit has a missing value for X_1 whereas $X_2, \ldots, X_{p-1}$ are observed. In this case, the imputed value of X_1 would be sampled from $P(X_1|\boldsymbol{Z}, \boldsymbol{B})$ when the SR technique is used. Doing so may disrupt the relationships (as gauged using the imputed dataset) between X_1 and X_j for $j = 2, \ldots, p$. In order to avoid such a disruption, we must sample X_1 from $P(X_1|\boldsymbol{Z}, \boldsymbol{B}, X_2, \ldots, X_p)$. Also, under non-monotone missingness it is difficult to obtain unbiased draws of regression parameters using the SR technique since the covariate matrix used to obtain such draws often contains imputed values (and as we just mentioned, these imputed values may be improperly sampled).

5.1 ISR2 and ISR3

We assume that the sequence of models seen in (6) holds true for $j = 1, \ldots, p$. We iteratively draw imputes and parameter estimates. Given starting values, we produce a sequence of completed datasets, $\boldsymbol{\chi}^{(t)} = \{\boldsymbol{Z}; \boldsymbol{B}; X_1^{(t)}; \ldots; X_p^{(t)}\}$, and a sequence of model parameters, $\boldsymbol{\Theta}^{(t)} = \{\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_p^{(t)}\}$ for $t \geq 0$. For each j, $X_j^{(t)}$ represents the value of X_j and $\boldsymbol{\theta}_j^{(t)}$ represents the value of $\boldsymbol{\theta}_j$ (at the t^{th} iteration). Like most MCMC techniques used for imputation, imputes and parameters are updated at each iteration via an imputation step (I step) and a parameter step (P step).

The I step samples $\boldsymbol{\chi}^{(t+1)}$ from:

$$\boldsymbol{\chi}^{(t+1)} \sim P\left(\boldsymbol{\chi} \left| \boldsymbol{\Theta}^{(t)}, \boldsymbol{\chi}_{\text{obs}} \right.\right),$$

where $\boldsymbol{\chi}_{\text{obs}}$ represents the observed values in $\boldsymbol{\chi}$. The P step samples $\boldsymbol{\Theta}^{(t+1)}$ from:

$$\boldsymbol{\Theta}^{(t+1)} \sim P\left(\boldsymbol{\Theta} \left| \boldsymbol{\chi}^{(t+1)} \right.\right).$$

The sequence of conditional models seen in (6) ensures that

$$P\left(x_{i1}, \ldots, x_{ip} \left| \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{\Theta}^{(t)} \right.\right) \tag{7}$$

is multivariate normal for $1 \leq i \leq n$ and for each $t \geq 0$. During the I step of the $(t+1)^{\text{th}}$ iteration, we calculate $X_j^{(t+1)} = \{x_{1j}^{(t+1)}, \ldots, x_{nj}^{(t+1)}\}$ sequentially for $j = 1, \ldots, p$ by sampling from the following density whenever x_{ij} is missing:

$$x_{ij}^{(t+1)} \sim P\left(x_{ij} \left| x_{i1}^{(t+1)}, \ldots, x_{i,j-1}^{(t+1)}, x_{i,j+1}^{(t)}, \ldots, x_{ip}^{(t)}, \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{\Theta}^{(t)} \right.\right).$$

This is done by first calculating the mean vector and covariance matrix of the expression in (7), and then using known equations for the conditional distributions of a multivariate normal density.

The P step of this procedure will closely resemble the parameter simulation process seen in the SR techniques above. For $j = 1, \ldots, p$, we calculate $\boldsymbol{\theta}_j^{(t+1)}$ by sampling from its posterior distribution via formulas of the form in Little and Rubin (2002), p. 114. We use all units of $X_j^{(t+1)}$ as the response variable, and the covariate matrix, in accordance with (6), includes $\boldsymbol{Z}, \boldsymbol{B}_{-j}, X_1^{(t+1)}, \ldots, X_{j-1}^{(t+1)}$.

We determine $\boldsymbol{\chi}^{(0)}$ and $\boldsymbol{\Theta}^{(0)}$ via the SR procedure outlined in Section 4. After a burn-in period (b) we return $\boldsymbol{\chi}^{(b)}$. We refer to this MCMC procedure as *Iterative Sequential Regression* (ISR). It may be implemented in conjunction with the transformations in (3) or (4) which yield the "ISR2" and "ISR3" methods respectively.

6. Conclusion

Both the current NASS method and the ABB method lack the multivariate sophistication required for a high dimensional dataset. These methods only utilize three covariates, and there are several highly informative covariates that go unused. Also, the methods do not allow the imputer to model variables with missing values on other variables with missing values, thereby implying that relationships between these variables will likely be distorted by the imputation process.

The SR methods should enable the imputer to capture the marginal characteristics of the data. Likewise, it will offer improvement over the NASS and ABB methods in terms of preserving variable relationships since it allows variables with missing values to be modeled on any of the fully observed covariates as well as other variables with missing values. However, its non-iterative nature implies that imputations found using this technique will still induce bias into variable relationships as long as those relationships are not sufficiently explained using the fully observed covariates.

The ISR technique allows for flexible selection of conditional distributions, which is an attribute of other popular MCMC techniques, such as MICE (Van Buuren and Oudshoorn, 1999), SRMI (Raghunathan et al., 2001), and `mi` (Su et al., 2010). ISR utilizes joint modeling, since conditional models of the form in (6) are used as opposed to the respective full conditional models. Joint modeling (which is an attribute of the data augmentation class of imputation procedures — see Little and Rubin 2002 and Schafer 1997 for an outline of such methodology) ensures that after a sufficient number of iterations, the imputes represent a draw from the posterior distribution of the complete data given the observed data.

REFERENCES

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Banker, D. (2007), "ARMS Phase III: Data Processing and Analysis," Tech. rep., Economic Research Service, prepared for presentation at the FAO Regional Consultation on Statistics for Farmer Income, Bangkok, Thailand, December 11-14, 2007.

Fichman, M. and Cummings, J. N. (2003), "Multiple Imputation for Missing Data: Making the Most of What You Know," *Organizational Research Methods*, 6, 282–308.

Horton, N. J. and Lipsitz, S. R. (2001), "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables," *The American Statistician*, 55, 244–254.

Kim, J. K. (2002), "A Note on Approximate Bayesian Bootstrap Imputation," *Biometrika*, 89, 470–477.

Little, R. J. A. and An, H. (2004), "Robust Likelihood-Based Analysis of Multivariate Data with Missing Values," *Statistica Sinica*, 14, 949–968.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical 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," in *JSM Proceedings*, Section on Survey Research Methods, Alexandria, VA: American Statistical Association.

National Research Council (2008), *Understanding American Agriculture: Challenges for the Agricultural Resource Management Survey*, Washington, D.C.: The National Academies Press.

Newman, D. (2003), "Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood and Multiple Imputation Techniques," *Organizational Research Methods*, 6, 328–362.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85–95.

Rubin, D. B. and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, New York, New York: Chapman and Hall/CRC.

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.

Spiess, M. and Keller, F. (1999), "A Mixed Approach and a Distribution Free Multiple Imputation Technique for the Estimation of Multivariate Probit Models with Missing Values," *British Journal of Mathematical and Statistical Psychology*, 52, 1–17.

Su, Y.-S., Gelman, A., Hill, J., and Yajima, M. (2010), "Multiple Imputation with Diagnostics (`mi`) in R: Opening Windows into the Black Box," *Journal of Statistical Software*, forthcoming.

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

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