# An Assessment of Imputation Methods for the USDA's Agricultural Resource Management Survey

Joshua D. Habiger* Michael Robbins † Sujit Ghosh ‡

**Abstract**

This paper provides an assessment of imputations methods applicable to the Agricultural Resource Management Survey data. We find that both iterative and noniterative regression procedures perform better than the current NASS method and an approximate Bayesian bootstrap method. Both regression methods perform better when utilizing a log-skew-normal transformation over a log-normal transformation.

**Key Words:** Missing Data, ARMS, Multivariate Imputation

## 1. Introduction

For reasons outlined in [1], it is important that an imputation method not perturb the joint distribution structure of the Agricultural Resource Management Survey (ARMS) data. This paper assesses the performance of the "state-of-the-art" imputation methods in [2] relative to their ability to preserve the joint distribution structure of the ARMS data. The methods considered are: the sequential regression procedure with log-normal and log-skew normal transformations, referred to as SR2* and SR3*, respectively, the iterative sequential regression procedure with log-normal and log-skew-normal transformations, called ISR2 and ISR3, respectively, the approximate Bayesian bootstrap, or ABB, and the current conditional mean method used by the National Agricultural Statistics Service (NASS). See [2] for a detailed description of the methods.

Simulation studies suggest that, in terms of preserving the joint distribution structure, both the SR and ISR methods perform better when making use of the log-skew-normal transformation over the log-normal transformation. Both regression methods outperform the NASS and ABB methods, with the iterative sequential regression method performing the best.

This paper proceeds as follows. In section 2, we outline the simulation setup and metrics for assessing the performance of the methods. A summary of our findings is in section 3. Concluding remarks are in section 4.

## 2. Simulation Setup

Imputation methods are assessed with a simulation study, which is described as follows. First, a subset of fully observed representative ARMS variables, which we denote by $(X_1, X_2, ..., X_Q)$, where $X_q = (X_{q1}, X_{q2}, ..., X_{qN})^t$ for $q = 1, 2, ..., Q = 24$, was selected. Here, $N$ is greater than 20,000. To simulate an ARMS data set

*Oklahoma State University, Department of Statistics, Stillwater, OK 74078

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

‡Department of Statistics, North Carolina State University, Raleigh, NC 27695

with missing values, we "poke holes" in two of the variables: "Corn for grain acres harvested" and "Corn for grain total production", which we denote by $X_1$ and $X_2$, respectively. To allow for different missingness mechanisms, we let response indicator $R_{qn}$, which is 1 if $X_{qn}$ is observed and 0 otherwise, be a bernoulli random variable with

$$logit(\Pr(R_{qn}|x^*)) = \beta_0 + \beta_1 x^*_{1n} + \beta_2 x^*_{2n} + \beta_3 x^*_{3n}.$$

Here, the variable $x^*_3$ is the log skew-normal transformed version of "Gross Value of Sales". See [2]. To simulate a missing completely at random (MCAR) mechanism, we choose $\beta_0 = log(1/p - 1)$ for $p = .5$ and $p = .8$, and set the rest of the $\beta_i$'s to 0. Note that $\Pr(R_{qn} = 1|x^*_3) = \Pr(R_{qn} = 1)$ is .5 and .8, respectively. To simulate a missing at random (MAR) mechanism, we choose $(\beta_0, \beta_3) = (-1.5, .75)$ and $(\beta_0, \beta_3) = (0, 1)$. Again, the average response rate in our nonzero population, computed

$$\frac{1}{K_3} \sum_{\{n:x_{3n}\neq 0\}}^{N} \Pr(R_{qn}|x^*_{3n}),$$

where $K_q$ is the number of nonzero elements of $x_q$, is .8 and .5, respectively. The coefficients $\beta_1$ and $\beta_2$ are still set to 0. To simulate a not missing at random (NMAR) mechanism for $X_q$, we take $(\beta_0, \beta_q) = (0, 1)$ and $(\beta_0, \beta_q) = (-1.5, .75)$ for $q = 1, 2$, which again results in average response rates of .8 and .5, respectively.

We will consider several different metrics, generally denoted $\theta$, for summarizing the joint behavior of a data set. The performance of an imputation method is measured in terms of the relative change of each metric post imputation. That is, for replication $k$, each missing value of $x_{1k}$ and $x_{2k}$ is imputed for using any one of the methods in the previous section, which yields a completed data set $\hat{x}_k = (\hat{x}_{1k}, \hat{x}_{2k}, x_3, ..., x_Q)$. Then, the percent change in the metric is computed via

$$\% \text{ change in } \theta(\hat{x}_k) = 100 \left( \frac{\theta(\hat{x}_k) - \theta(x)}{\theta(x)} \right)$$

The metrics of interest are the the marginal means and variances of $x_1$ and $x_2$, and the covariance of $x_1$ and $x_2$. The marginal means and variance are computed on the positive portion of the variables via

$$\mu_q(x) = \frac{1}{K_q} \sum_{\{n:x_{qn}\neq 0\}} x_{qn}$$

and

$$\sigma^2_q(x) = \frac{1}{K_q} \sum_{\{n:x_{qn}\neq 0\}} (x_{qn} - \mu_q(x))^2,$$

respectively. The covariance of $x^*_1$ and $x^*_2$ is computed

$$Cov_{1,2}(x) = \frac{\sum_{\{n:x_{1n},x_{2n}\neq 0\}} (x^*_{1n} - \mu_1(x^*))(x^*_{2n} - \mu_2(x^*))}{K_{12}}$$

We are interested in maintaining the covariance between variables on the log scale, rather than on the original scale, because scatterplots suggest that relationships between variables are linear on the log scale.

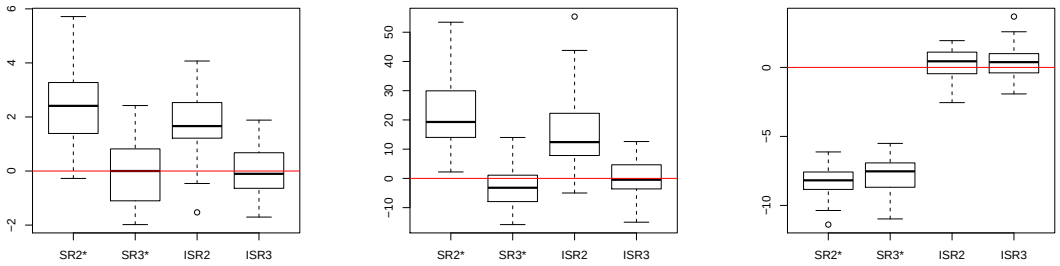


**Figure 1**: Box Plots of % change in (from left to right) mean, variance, and covariance when data are MCAR

## 3. Results

For the remainder of this paper, we discuss only simulation results when the average response rate is .5. Results are analogous when the response rate is .8. Results regarding marginal means and variances are presented for the "corn for grain acres harvested" variable.

We first compare the impact of the log-normal and log-skew-normal transformations on the performance of the regression procedures. We find that making use of the more robust log-skew-normal transformation allows for these methods to perform better. In general, whether data are MCAR, MAR, or NMAR, the SR3* method performs better than the SR2* method, and the ISR3 method performs better than the ISR2 for any metric considered. Figure 1 displays the percent change in mean, variance, and covariance of these 4 methods when data are MCAR. Indeed, we see that the SR3** method preserves the mean and variance while the SR2** method does not. Analagously, the ISR3 method preserves the mean and variance while the ISR2 method does not. Interestingly, we see that it is not important to make use of the appropriate transformation in order to preserve the covariance between variables.

Figure 1 also illustrates our finding that the ISR3 method tends to perform better than the SR3* method. In particular, we see that though the ISR3 and SR3* are comparable in terms of the preservation of the mean and variance, the ISR3 method will better preserve the covariance between variables. This result holds when data are MAR and NMAR as well.

We now compare the NASS and ABB methods to the SR3* and ISR3 methods for each metric and missingness mechanism. In terms of preserving the mean, all methods perform reasonably well when data are MCAR. See Figure 2. However, when data are not MCAR, all methods tend to overestimate the mean, especially when data are NMAR. This phenomenon is typically observed in our simulation studies when values on the right tail of the distribution are more likely observed, and this is the case here since $\beta_1$, $\beta_2$, and/or $\beta_3$ are taken to be positive. The SR and ISR methods can often capture some of this missingness mechanism, especially when imputation methods depend upon the same variable that was used to poke

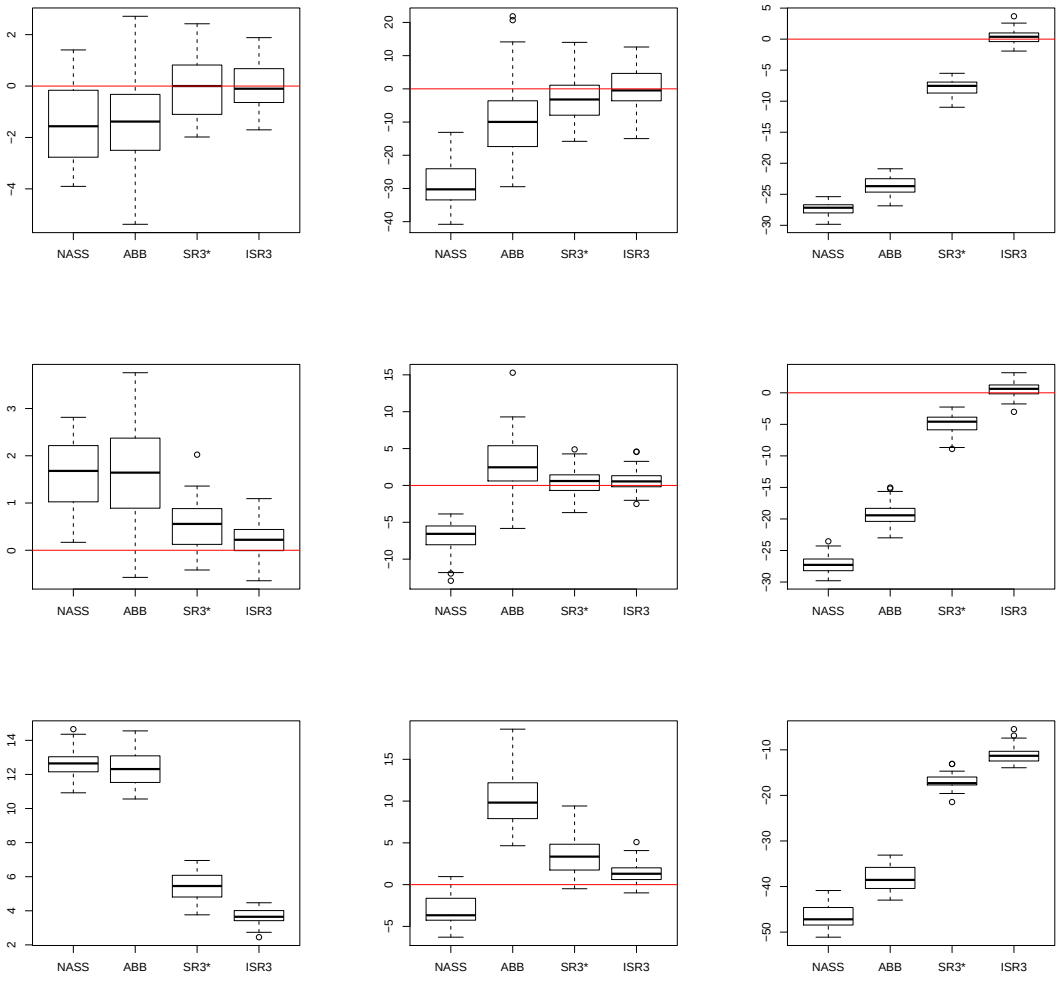


**Figure 2**: Box Plots of % change in mean, variance, and covariance (columns 1, 2, and 3) when data are MCAR, MAR, and NMAR (rows 1, 2, and 3).

holes in the data. For example, when data are MAR, we see that since SR3** and ISR3 methods use $x_3$ as a covariate, and $x_3$ also determines the probability that $x_{2n}$ or $x_{1n}$ is observed, both methods tend to preserve the mean. When data are NMAR, however, so that the response probability depends on $x_1$ or $x_2$, even ISR and SR methods will fail to preserve the mean. However, they do still perform significantly better than the NASS and ABB methods.

An analogous phenomenon is observed in the variance metric for this particular simulation study. That is, as we go from MCAR to MAR to NMAR, the estimate of the variance increases for any given method. This result is expected. Our variable "corn for grain acres harvested" is highly skewed right, and the extreme observation in the right tail have the highest probability of being observed under MAR and NMAR settings. Since the addition of extreme observations to a data set increases its sample variance, we see that imputation methods tend to overestimate the variance. Additional simulation studies indicate that this phenomenon will only occur for highly skewed right variables and for this particular type of missingness mechanism. When the missingness mechanism is defined so that these extreme values in the right tail of the distribution have the lowest probability of being observed, both the mean and the variance metrics tend to be negatively biased. Regardless, the ISR method performs the best, followed by the SR method, in terms of the preservation of the variance. The NASS and ABB method perform the worst.

In terms of the preservation of the covariance, the NASS and ABB methods tend to perform poorly regardless of the missingness mechanism. This is to be expected since these methods only exploit the relationship between the variable to be imputed and 'Gross Value of Sales'. Hence, we may only expect the relationship between two variables to be preserved if each variable is highly correlated with 'Gross Value of Sales', and this is not the case. The SR3* and ISR3 methods, however, allow for each variable to be imputed to depend upon multiple variables, including one another. We also find that the ISR method typically does a substantially better job at preserving the covariance between variables over the SR method, especially when data are MAR or NMAR.

## 4. Conclusions

In this paper, we assessed the performance of several imputation methods in terms of their ability to preserve the joint distribution structure of the ARMS data. We saw that ISR and SR regression methods perform best when making use of a log-skew-normal transformation rather than a skew-normal transformation. Also, we found that both the ISR and SR methods dominate the NASS and ABB methods, especially in terms of their ability to preserve the relationship between variables. The ISR method performs the best. It allows for the mean, variance, and covariance between variables to be preserved for most types of missingness mechanisms.

## References


[1] D. Miller, M. Robbins, and J. Habiger. Examining the challenges of missing data analysis in phase three of the agricultural resource management survey. *Proceeding of the Joint Statistical Meetings*, 2010.

[2] M. Robbins, S. Ghosh, and J. Habiger. State-of-art techniques for missing data analysis as applied to the agricultural resource management survey. *Proceeding of the Joint Statistical Meetings*, 2010.