Imputing for Missing Data in the ARMS Household Section: A Multivariate Imputation Approach

Christopher Burns and Daniel Prager[1]

USDA-Economic Research Service

christopher.burns@ers.usda.gov, daniel.prager@ers.usda.gov

Sujit Ghosh and Barry Goodwin[2]

North Carolina State University

sghosh2@ncsu.edu, bkgoodwi@ncsu.edu

Selected Paper prepared for presentation at the 2015 Agricultural & Applied Economics Associations and Western Agricultural Economics Association Annual Meeting, San Francisco, CA, July 26-28

[1]Christopher Burns and Daniel Prager are research economists at the Economic Research Service. The views expressed are the authors and should not be attributed to ERS or USDA.

[2]Sujit Ghosh is a professor in the Department of Statistics at North Carolina State University and the Deputy Director at SAMSI. Barry Goodwin is the William Neal Reynolds professor in the Economics and Agricultural and Resource Economics Departments at North Carolina State University.

Imputing for Missing Data in the ARMS Household Section: A Multivariate Imputation Approach

Abstract

This study proposes a new method to impute for ordinal missing data found in the household section of the Agricultural Resource Management Survey (ARMS). We extend a multivariate imputation method known as Iterative Sequential Regression (ISR) and make use of cut points to transform these ordinal variables into continuous variables for imputation. The household section contains important economic information on the well-being of the farm operator's household, asking respondents for information on off-farm income, household expenditures, and off-farm debt and assets. Currently, the USDA's Economic Research Service (ERS) uses conditional mean imputation in the household section, a method known to bias the variance of imputed variables downward and to distort multivariate relationships. The new transformation of these variables allows them to be jointly modeled with other ARMS variables using a Gaussian copula. A conditional linear model for imputation is then built using correlation analysis and economic theory. Finally, we discuss a Monte Carlo study which will randomly poke holes in the ARMS data to test the robustness of our proposed method. This will allow us to assess how well the adapted ISR imputation method works in comparison with two other missing data strategies, conditional mean imputation and a complete case analysis.

Key words: ARMS, missing data, ordinal data, ISR, imputation, Markov Chain Monte Carlo
JEL codes: C18, C15, C55

The Agricultural Resource Management Survey (ARMS) is a complex survey jointly administered by the U.S. Department of Agriculture's (USDA) National Agricultural Statistical Service (NASS) and Economic Research Service (ERS). ARMS[3] is the USDA 's primary source of information on the financial condition, production practices, and resource use of the nation 's farm households. In a given year it typically contains between 20,000-30,000 records of 1,000-2,000 variables. The survey is conducted in three phases and uses a stratified, multiframe, probability-weighted sampling design to ensure that it is a representative sample of the U.S. agricultural sector.

As with all large, complex surveys, ARMS suffers from non-response resulting in missing data for many key variables. Missing data in large complex surveys presents unique challenges for statistical and economic data analysis. Despite recent advances in survey imputation methodology, imputation in large-scale surveys will likely never be a trivial task. Each survey has its own unique characteristics, making it difficult to apply methods developed on other datasets.

This study is focused on the third phase of ARMS (the cost and returns report), which asks respondents a variety of questions about the financial characteristics of the farm operation and the household. Within the cost and returns report, the ARMS household (HH) section asks respondents for information on off-farm income, household expenditures, and off-farm debts and assets. This data is used for a wide variety of purposes, including calculating total off-farm income, estimating economic models on off-farm labor participation, and understanding the expenditure patterns of farm households.

Given the wide use of data from the HH section, item and section non-response is a serious problem for economists and policymakers. Because most statistical packages are designed for rectangular data arrays, missing data are often dropped from analysis. To avoid losing observations ERS has traditionally used a stratum-based conditional mean imputation for variables in the HH section. This method preserves the means and totals of the data, but has several drawbacks. Conditional mean imputation will bias the variance of the imputed variable downward and distort multivariate relationships between variables in the data (Little and Rubin, 2014). The loss of variability in the data is particularly problematic for economic modeling, where multivariate rela-

[3]The ARMS data is widely accessed by researchers, academics and policymakers to study economic, production and environmental issues surrounding U.S. agriculture. It is also widely used to study the efficacy of USDA programs. For more information on the uses of ARMS see: `http://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/uses-and-publications.aspx`

tionships may gain or lose significance. Additionally, this method treats all missing values in the household section as positive imputed values and does not allow for zero values.

Recent methodological advances now allow NASS to impute for missing data in the ARMS survey using multivariate imputation methods. This method, known as Iterative Sequential Regression (ISR) (Robbins, Ghosh, and Habiger, 2013) was specifically designed to impute for continuous/mixed variables in the ARMS data. While this method is a vast improvement over the previous conditional mean methods, ISR cannot currently handle ordinal missing data such as those found in the ARMS HH section. This study proposes extending the current ISR methodology by transforming these ordinal variables to continuous variables using a cut point method. This transformation allows for linear conditional models for imputation to be modeled using a Gaussian copula. We discuss two possible transformation methods, a likelihood-based method and a non-parametric method that uses the Anderson-Darling statistic. We then discuss how we will test these transformation methods and the resulting imputations using a simulation study.

The rest of the paper proceeds as follows: we first briefly review the missing data literature and then discuss the current state of imputation for ARMS. Next, we examine the missingness in the HH variables and look at common distributions found in the observed data. Following that, we describe a new method for transforming these variables so that a linear conditional model for imputation can be built. We follow this with a description of the ISR methodology. We conclude by discussing how a simulation study will test the imputation methodology and how the new imputed data may affect economic analysis for the U.S. agricultural sector.

Missing data in complex surveys

The ARMS data exhibit many characteristics found in large, complex surveys. This includes high dimensionality, different variable types, extreme skewness and non-montone missingness patterns. These characteristics make the development of a robust imputation methodology very challenging. When faced with missing data, three strategies are generally available: direct analysis of the incomplete data, imputation, and weighting the complete cases (Little, 1988).

Ad-hoc missing data methods such as list-wise deletion or complete case analysis are still commonly used by researchers because most statistical software platforms require a rectangular

data array. The drawback of complete-case analysis is that it can lead to biased results because relationships between variables are not preserved. It can also lead to reduced statistical power due to the reduction in sample size (Little and Rubin, 2014). Hot-deck imputation replaces missing data with observed values from pre-determined "donor cells". These donor cells are found in the complete cases within the survey. This method has been widely used in large surveys, including the Current Population Survey (Little, 1988). Hot-decking can reduce the mean-square error for univariate statistics and helps preserve the covariance structure in the data. However, the hot-deck approach is also known to underestimate the variance of estimates. It is also difficult to explicitly describe the data generating mechanism for missing values using this approach (He et al., 2009).

Multivariate models for imputation have focused on two methodologies: the multivariate normal model (Schafer, 1997) and the fully conditional specification. The fully conditional specification (FCS) is typically preferred due to the flexibility in modeling. However, it has been shown that FCS can result in over-specification of the joint distribution, resulting in imputations being sampled from an "incoherent" joint density (Van Buuren et al., 2006). When this occurs convergence of the Markov chain is no longer guaranteed and imputations may be very poor.

A popular method for imputation for missing data in surveys is multiple imputation (Rubin, 1996, 2004). Examples of surveys that use multiple imputation to impute for missing data include the National Medical Expenditure Survey (Rubin, 2003) and the National Health Interview Survey (Schenker et al., 2006). Multiple imputation has many advantages over more ad-hoc imputation methods. Because multiple imputed values are created for each missing value it preserves the uncertainty in the imputed data. By including both within and between imputation error, this method gives researchers more statistical power to detect relationships in multivariate modeling.

Iterative Sequential Regression (ISR) is a multivariate imputation method designed specifically to impute for mixed/continuous variables in ARMS (Robbins, Ghosh, and Habiger, 2013). The ISR method has been shown to avoid the over-specification issues of FCS because it uses a joint model specification by estimating the conditional distributions. ISR first transforms the mixed/continuous variables using skew-normal or log density transformations to ensure approximate normality. This transformation allows for the use of a Gaussian copula to model the joint distributions with conditional linear models. The resulting multivariate densities are then estimated using Gibbs sampling.

Missing ordinal data is problematic for researchers because current imputation software is only designed to impute for categorical or discrete data. Often, ordinal data is simply treated as a multinomial distributed variable, but this method is inappropriate because it ignores the information provided by the ordering. While there has been research into imputation for categorical and discrete missing data (Kropko et al., 2014), so far there has been little work done in the area of multivariate imputation for ordinal missing data.

There are currently several readily available software packages for imputing missing data in large datasets. These software packages include: the R package MICE (Multiple Imputation using Chained Equations) (van Buuren and Groothuis-Oudshoorn, 2011), Stata's MI(Multiple Imputation) package (Royston, 2004), and SRMI (Sequential Regression Multiple Imputation), which is run with a SAS (SAS Instititue, Cary NC) callable routine called IVEWare (Raghunathan et al., 2002). Unfortunately, none of these packages are currently able to handle ordinal missing data in a robust and satisfactory manner.

ARMS imputation methodology

A recent external review of ARMS highlighted the need to explore new imputation methodologies (National Research Council 2008) citing the failure of conditional mean imputation to satisfy the needs of data users. This review, as well as other studies (Briggeman, Koenig, and Moss, 2012; H. Kuethe, Briggeman, D. Paulson, and L. Katchova, 2014), also identified differences between ARMS estimates of debt and administrative data.

ERS and NASS currently run their own imputation procedures for subsets of variables in ARMS. NASS imputes for variables related to farm expenses and production, such as total sales, total expenses and total acres harvested for each commodity. ERS imputes for variables related to farm and household well-being, such as farm debt, operator wages and food expenditures. Before NASS updated its procedure in 2015, both agencies used a conditional mean imputation approach for imputation. This approach uses a stratum-based form of conditional mean imputation. A donor pool is created for a missing value by collecting all the positive values for a farm observed in the same sales class, farm type and region. The imputation is then taken as the mean of the donor pool.

In 2011, development began on a new multivariate imputation method to replace conditional mean imputation at NASS. The resulting methodology, known as ISR, was the result of a joint National Institute of Statistical Sciences (NISS)/NASS project to improve NASS's ability to impute for missing data in ARMS. This project first examined the impact of conditional mean imputation in the ARMS data and recommended improvements on the existing method. These methods included multivariate imputation approaches (Miller, Robbins, and Habiger, 2010; Habiger, Robbins, and Ghosh, 2010). In 2015, the ISR method will be used by NASS to impute for about 80 continuous/mixed variables in ARMS.

ERS imputation methodology for farm debt

A recent change to the imputation methodology in the Farm Debt section of ARMS highlights how conditional mean imputation may be insufficient. Each year the USDA collects detailed information on farm operator loans through the ARMS. This information includes interest rates, loan term, origination date, type of loan, purpose of loan and type of financing. The data collected is then used for the farm sector balance sheet, which contains information on key financial statistics such as debt-to-asset ratios, debt repayment capacity and liquidity. Information on debt is subject to non-response. ERS has historically used a generalized cell mean imputation for missing farm debt.

To address issues associated with conditional mean imputation with the farm debt table, ERS implemented a multivariate imputation methodology in 2012. Using the SAS callable imputation program IVEware (Raghunathan et al., 2002), conditional linear models can be specified for each variable being imputed. IVEware is flexible, allowing the user to impute for continuous, categorical and mixed variables (i.e. variables with both continuous and discrete values). Models for predicting the imputed values were built using economic theory and included variables such as operator age, acres, government payments, property taxes and region. Diagnostics were performed with the new imputed data, such as comparing the distributions for each of the imputed variables with the original data.

A recent study by (Morehart, Milkove, and Xu, 2014) used Sequential Regression Multiple Imputation (SRMI) to impute for on-farm debt in the 2012 ARMS. They show the resulting estimates of total on-farm debt are $27 billion greater under this imputation method, bringing them closer

to administrative data estimates. Preliminary analysis also shows this moved almost 2% more farm businesses into a critical zone based on their debt-to-asset measures. In 2013, this imputation method was found to add $31 billion to ARMS farm debt estimates, bringing total farm debt to $196 billion.

Missing ordinal data in HH section of ARMS

Previous work on imputing for missing variables in the ARMS HH section by Ahearn et al. (2011) examined the advantages of using SRMI over conditional mean imputation. Using a multivariate model, this study imputed for two HH variables: off-farm wages, and private retirement income and disability payments. They found evidence that a conditional mean imputation approach may bias the mean income of farm operator households when compared to a multivariate approach, such as SRMI. However, they were not able to generalize their findings and reach a definitive conclusion about which imputation method was better.

The ARMS HH section contains approximately 45 value-coded variables. Respondents are asked questions about their off-farm income, household expenses, debt and assets by filling in a value code between 1 and 34. Negative value codes are allowed for variables that can contain negative values, such as previous year farm operating income. These value codes are mapped to dollar amounts which range from zero to $10 million or more, as shown in table 1.

Starting in 2012, the HH section saw a large increase in refusals, likely due to changes in the survey administration. As seen in figure 1, section refusal rates of 20.4% and 25.6% were observed in 2012 and 2013, respectively. This sharply contrasts with the average section refusal rate of the preceding three years, at 5.4%. The general upward trend in section refusal rates is troubling for researchers who use ARMS.

A major factor in the increase in section refusals rates is changes in the ARMS questionnaire design. Before 2012, ARMS was mostly enumerated in person, with a shorter 'core' survey mailed to select households. In 2012, ARMS moved to an 'all-mail' survey format, where surveys were initially mailed to each respondent's household and enumerators followed up in person to households which had not responded. This may have contributed to an increased number of missing values in the household section. Another potential reason for the section refusals increase may

include the prevailing issues surrounding privacy and personal information. This problem has also been seen in other household surveys over the last few decades (Abraham, Maitland, and Bianchi, 2006).

Further analysis reveals that even among surveys with usable HH sections, item refusal rates observed after 2012 were much higher. For example, the item refusal rate for the HH question about 'total off-farm wages for the household' totaled 30% in 2013 as compared to 12.6% in 2011. Table 2 shows the item refusal rate and cumulative response rate for selected HH variables. Once the survey response rate, section and item response rates have been accounted for, the cumulative response rates for some variables were as low as 32%. For a variable such as financial assets, this is particular problematic because a household could have significant off-farm assets not accounted for in the survey. This might lead a researcher to conclude that the household's level of well-being is lower than in reality.

Distributional Characteristics of ARMS HH variables

In this section we discuss the three most common types of marginal distributions observed for the HH variables. The histograms displayed below show the frequency distribution by value code. Value codes range between 1 and 34 for questions which accept positive values and -34 to 34 for questions which accept both positive and negative values. The value code 1 corresponds to a zero dollar amount and it is assumed that households which incorrectly enter "0" intend a zero dollar amount.

The first commonly observed distribution has a large mass at zero, as well as a significant mass of positive values. The positive portion of the distribution seen in figure 2 appears almost normal in shape, with fewer observations at the higher and lower value codes. However, because the bins get wider in dollar terms as one moves to higher value codes, this picture of the distribution is misleading. The underlying distribution of data observed on the dollar scale is much more right skewed, as shown in figure 3. This makes sense as the higher value codes correspond to very large off-farm wages (e.g. millions of dollars), so we only observe a few data points at these levels. This right-skewed distribution is observed for several of the off-farm income variables.

The second common distribution has a much smaller number of zeroes, and large mass of

positive values with a unimodal shape. This is best illustrated by the variable food expenses, shown in figure 4. Again, because the value code bins correspond to an increasing range of dollar values at higher values, this distribution is also right skewed when observed on the dollars scale.

The third common distribution has large mass of zeroes, and distinctly different distributions for the positive and negative values. Overall, it has a highly non-symmetric shape. This type of distribution is best illustrated by the variable net farm income previous year, shown in figure 5. The mass of positive observed values is shown to be much greater than the mass of negative observed values.

In summary, the observed portions of the HH variables do not appear approximately normal. This is an issue when building linear regression models for imputation, and thus a transformation is warranted.

Transformation of HH variables

In order to apply multivariate models for imputation, we use a linear regression framework. Regression allows us to take advantage of the numerous predictors available in a high-dimensional ARMS dataset and build conditional models. Some predictors may also contain missing values. In order to apply regression methods, we require joint normality assumptions to be met. Given that the household section contains ordinal data, much of which is highly skewed, this will require a transformation for each of the variables we wish to impute. This section explores two potential methods for transforming the HH variables, one that uses parametric methods and another which does not.

Adjusting for Ordinal Variables

The HH section contains variables which take on ordered value codes. Our goal is to obtain a marginal distribution that is normal, so that a joint multivariate model for imputation can be used. Previous work by Robbins et al. (2013) used a parametric class of distributions to transform the data, including a log skew normal and log normal. Unfortunately, these distributional forms will not work with ordinal data. However, we can take advantage of the existing structure of the value codes, to identify a distribution for the latent variable that is segmented to construct the

ordinal variable. We describe two methods of estimating the underlying distribution of the latent variable.

Before proceeding, we briefly define some notation. Let

- Y_j be the j^{th} untransformed ordinal HH variable
- X_j be the j^{th} transformed continuous HH variable
- $\hat{X}_j$ be the j^{th} imputed continuous HH variable
- $\mathbf{Z}$ = $(Z_1, \ldots, Z_q)$ be a block of covariates that are fully observed
- $\mathbf{X}$ = $(X_1, \ldots, X_p)$ and χ = $(\mathbf{X}, \mathbf{Z})$ be a complete block of data with missing values

where $j = 1, \ldots, p$ index ARMS HH variables, $i = 1, \ldots, N$ index observations, and $k = 1, \ldots, m$ index value codes.

Likelihood Based Estimation

We begin by examining a method for transforming the HH ordinal variables that uses maximum likelihood techniques. One advantage of this method is that when the parametric form of the distribution is known, it is consistent and efficient. Suppose the HH ordinal variable takes on value codes $1, 2, \ldots, m$ and it is also known that $Y = k$ if $c_{k-1} < U < c_k$ for $k = 1, \ldots, m$, where c_k represents a known cut point and U represents the unobserved latent variable. The objective is to identify the distribution of the latent variable U using the observed values of Y and the known cut points. We assume that the latent variable has a cumulative distribution F, where $F(c_0) = 0$ and $F(c_m) = 1$. Notice that $P[Y = k] = F(c_k) - F(c_{k-1})$ for $k = 1, \ldots, m$. Assuming a completely observed sample of data, $Y_1, \ldots, Y_N$, the likelihood function of F is then

$$\mathcal{L}(F) = \prod_{k=1}^{m} [F(c_k) - F(c_{k-1})]^{n_k} \tag{1}$$

where $n_k = \sum_{i=1}^{N} I(Y_i = k)$, and $I(.)$ is an indicator function that equals one when the argument is true and zero otherwise. The maximum likelihood estimate of F at each c_k is then given by

$$\hat{F}(c_k) = \sum_{l=1}^{k} \frac{n_l}{N} = \sum_{i=1}^{N} \frac{I(Y_i \leq k)}{N} \tag{2}$$

Clearly, the function $F(.)$ cannot be estimated except perhaps at the cut points $c_1, \ldots, c_m$. Thus, a parametric form must be assumed before we can proceed with the estimation. Previous work with ARMS restricted F to a robust class of distributions, such as the log skew normal, log normal, or more general mixture of log normal families. Assuming a suitable class of parametric family $\{F_\theta(.) : \theta \in \Theta\}$, the likelihood function of θ is given by

$$\mathcal{L}(\theta) = \prod_{k=1}^{m} [F_\theta(c_k) - F_\theta(c_{k-1})]^{n_k} \tag{3}$$

The estimate of θ is then obtained by maximizing the log likelihood given in equation 3. The MLE of θ is given by

$$\hat{\theta} = arg\, max_{\theta \in \Theta}\, log\mathcal{L}(\theta) \tag{4}$$

Notice that once $\hat{\theta}$ is obtained, we have the estimate of $\hat{F}(U) = F_{\hat{\theta}}(U)$.

Anderson-Darling Method of Estimation

The maximum likelihood method usually requires numerical non-linear optimization methods, which can be computationally unstable. An alternative to the maximum likelihood method of transformation for the HH variables is based on a version of the Anderson-Darling statistic (Anderson and Darling, 1952). The Andersen-Darling statistic is typically used to determine whether a given sample of data comes from a certain probability distribution. We propose this statistic for our estimation for two reasons. First, it has the advantage of not assuming a parametric distribution for each HH variable. Second, the estimate of θ can be achieved through quadratic programming methods, which can ensure unique and efficient solutions. The estimation of θ is achieved by

$$\hat{\theta} = arg\, max_{\theta \in \Theta} \sum_{k=1}^{m} \frac{(\hat{F}(c_k) - \tilde{F}_\theta(c_k))^2}{(\hat{F}(c_k) + \epsilon)(1 - \hat{F}(c_k) + \epsilon)} \tag{5}$$

where $\epsilon = \frac{3}{8n}$ and $\hat{F}$ denotes the empirical likelihood estimate of F given in equation 2 and $\tilde{F}_\theta(U) = \sum_{j=1}^{m} \theta_j B_{jm}(U)$. One potential class of basis functions is a smooth class of distributions (e.g. a mixture of Beta's) where a set of mixing weights is estimated using a weighted least squares criterion subject to linear inequality constraints (Turnbull and Ghosh, 2014). Assuming $F_\theta(x) = \sum_{j=1}^{m} \theta_j B_{jm}(x)$, where $B_{jN}(x)$ is the cdf of a $Beta(j, N)$ distribution, with suitable linear

constraints on θ this optimization problem can be solved with quadratic programming. The basis functions $B_{jm}(.)$ are chosen suitably (e.g. cdf's of Betas or integrated B-splines) so that $F_\theta(c_0) = 0$ and $F_\theta(c_m) = 1$ for any $\theta = (\theta_1, \ldots, \theta_m) \in \{\theta_j \geq 0, \sum_{j=1}^m \theta_j = 1\}$.

After we estimate θ using one or both of the methods above, it will be possible to transform the HH variables to a continuous distribution for imputation. After imputation, we can then transform the data back to the ordinal scale using the same cut points. Let $X_j = T(Y_j)$ represent the transformed HH variable, then

$$X_j = T(Y_j) = \Phi^{-1}(F_{\hat{\theta}}((c_{Y_j} + c_{Y_j+1})/2)) \tag{6}$$

The next steps of this project will explore both the Anderson-Darling and likelihood-based methods of estimation. We'll choose the best method based on the robustness of the estimates and whether the parameters can be estimated efficiently.

Variable Groups and Model Selection

In total, 45 variables in the 2013 ARMS HH section require imputation. These variables can be categorized into five groups based on economic relationships. These five groups are listed in table 3. We also highlight some of the key HH variables commonly used for economic analysis in table 4. Many of these variables are key components in calculating total off-farm income, which ERS uses to calculate total farm household income. While some variables may change in survey years, we anticipate the vast majority of those in 2013 ARMS will be asked on a annual basis.

Using economic expertise and an analysis of correlation using Kendall's Tau, a lengthy list of predictors will be developed for each group. As a starting point, we use Kendall's Tau to measure correlation between each HH variable and other variables contained in ARMS. [4] As would be expected by economic theory, off-farm labor supply (i.e. hours worked off-farm) is highly predictive of off-farm wages, salaries and tips for both the operator and spouse. These correlations are shown in table 5. Because off-farm labor hours is highly correlated with off-farm wages, this variable will be included in a linear regression model that imputes for off-farm wages.

[4] We use Kendall's tau to measure correlation because a Pearsons correlation coefficient does not perform well with non-continuous variables.

Additional variables that may be included in a typical regression model could include operator age, farm type, education, farm assets, net farm income, total farm expenditures, total government payments, and distance to nearest city. We may also be able to use data from outside ARMS to supplement for variables that are highly predictive of HH variables. For example, the Census Bureau's American Community Survey (ACS) has county-level estimates of median household income.

Modeling Details

After choosing the appropriate predictors using economic theory and statistical techniques we build conditional linear models to impute for each HH variable. Our imputation methodology is run jointly on a block of variables. We assume that this block contains a group of fully observed covariates, $\mathbf{Z}$, and a group of p HH variables, all of which have missing values on some records.

Model Assumptions

Our model assumes that after the appropriate marginal transformation is performed, the X_j's are multivariate normal conditional on the all other covariates in the model. Once the transformation is performed on the HH variable of interest, the resulting joint model is similar to a Gaussian copula (Nelsen, 2006). This Gaussian copula enables us to impute for the missing HH variables by using a sequence of conditional linear models. For these model assumptions to be valid, pairwise scatterplots of variables should appear to have a bivariate normal relationship. We use both scatterplots and normal Q-Q plots as a check for the Gaussian copula assumptions, following transformation of the HH variables.

Iterative Sequential Regression

ISR was developed specifically to impute for mixed/continuous variables in ARMS. The ISR method has two phases: transformation and imputation. After the variables have been successfully transformed to ensure approximate normality, the second phase of ISR uses a form of data augmentation (Tanner and Wong, 1987). Data augmentation uses an Markov Chain Monte Carlo

algorithm to iteratively draw imputations from a predictive model. This method is appropriate because it allows the user to jointly model all of the variables that require imputation.

ISR constructs a joint model for the ARMS HH variables based on the properties of joint distributions. Using the fact that a joint distribution can be expressed as a production of conditionals, we model the joint distribution of the p HH variables as

$$P(X_1, \ldots, X_p|\mathbf{Z}) = \prod_{j=1}^{p} P(X_j|\mathbf{Z}, X_1, \ldots, X_{j-1}) \tag{7}$$

The conditional distribution on the right-hand side allows the researcher to build a model for imputation. We assume a linear conditional form

$$X_j = \alpha_{j0} + \boldsymbol{\alpha_j} \boldsymbol{Z} + \sum_{d \neq j}^{p} \beta_{jd} X_d + \sigma_j \epsilon_j \tag{8}$$

for $j = 1, \ldots, p$ where $\epsilon_j \sim$ N(0,1), $\boldsymbol{\alpha_j}$ represents a vector of regression parameters for the fully observed covariates, and $\boldsymbol{\beta_j}$ represents a vector of regression parameters for variables with missingness. This linear model can be modified to include a particular set of covariates for each HH variable, making ISR very flexible.

In order to estimate the parameters and impute for the missing data, ISR relies on a Bayesian model that places distributional assumptions on the regression parameters. After setting the appropriate distributional assumptions with priors, the MCMC method is started. ISR then proceeds by iterating over two steps.

The first step (I step) requires ISR to draw sample imputations. In the second step (P step) ISR draws from sample parameter values. We define the set of regression parameters as $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta} = (\theta_{\mathbf{1}}, \ldots, \theta_{\mathbf{p}})$ and $\theta_{\mathbf{j}} = (\boldsymbol{\alpha_j}, \boldsymbol{\beta_j}, \sigma_{\mathbf{j}}^{\mathbf{2}})$. The MCMC then uses Gibbs sampling to generate samples from the posterior distributions of $\boldsymbol{\Theta}$, and missing values of X's.

The ISR method obtains estimated parameters and imputations for each HH variable. Because we do observe some values in each of the HH variables, the imputed values are created as

$$\hat{X_j} = \begin{cases} X_j & if\ X_j\ is\ observed \\ \hat{\mu_j} + \hat{\sigma_j}\epsilon_j & if\ X_j\ is\ missing \end{cases} \tag{9}$$

where $\hat{\mu_j} = E[X_j|\mathbf{Z}, \hat{\mathbf{X}}_\mathbf{1}, \ldots, \hat{\mathbf{X}}_{\mathbf{j-1}}, \tilde{\theta}_\mathbf{j}]$ and $\hat{\sigma}_j = sd[X_j|\mathbf{Z}, \hat{\mathbf{X}}_\mathbf{1}, \ldots, \hat{\mathbf{X}}_{\mathbf{j-1}}, \tilde{\theta}_\mathbf{j}]$.

The first iteration of ISR generates initial values for each parameter and imputation, $\mathbf{\Theta^{(0)}}$ and $\chi^{(\mathbf{0})}$. The ISR method then iterates by alternating back and forth between the I-step and P-step. This process continues until the MCMC converges on the posterior distributions of the parameters and imputations. ISR has been shown to be superior to sequential regression because it preserves the covariance structure of the data, in addition to preserving the marginal variable characteristics (Habiger, Robbins, and Ghosh, 2010).

Testing the new imputation methodology

Once the proposed methodology has been successfully implemented using R, we will conduct diagnostics of the imputation method by examining scatterplots of the multivariate distributions for the data. We will examine scatterplots for each of the imputed HH variables to see how the relationships with other variables in ARMS change depending on the imputation method. This is a widely used method to assess how well the imputation method is performing (Abayomi, Gelman, and Levy, 2008). We will also compare the marginal distributions for the HH variables to see how the mean, mode, variance and quantiles change between the different imputation methodologies. One important question we hope to address in this study is whether systematic bias is being introduced via conditional mean imputation.

After diagnostics have been performed, we will conduct a simulation study to assess how well the new imputation procedure does. Successful imputation will be based on randomly poking holes in the observed data and then assessing how well the new imputation methodology fills them. The random holes will be generated using a missing-at-random (MAR) missing data generating process. We can compare how well the new imputation methodology works as compared to conditional mean and complete-case analysis. Several different metrics can measure how well the new imputation method works. One possible metric is the mean absolute deviation (MAD) across estimated variables. Another is the Brier Score, defined as

$$BS = \frac{1}{N} \sum_i \sum_j (p_{ij} - y_{ij})^2, \tag{10}$$

which compares squared differences between imputed (p) and reported (y) values calculated over B imputations for $i = 1, \ldots, N$ observations within the dataset.

Next Steps

Once the new imputation methodology has been tested, many economic research questions can be examined with the new data. Several studies have already examined the economic significance of the new ISR imputations on selected mixed/continuous variables in ARMS. A recent study by Robbins and White (2011) examined the effect of ISR imputation on estimates of farm commodity payments. They found the new imputed values show that commodity payments are shifting to households with higher net farm income, consistent with other studies. Robbins and White (2014) found that under the ISR imputation method an additional dollar of direct payments increases land value by about $2.69 more per acre than when using conditional mean imputation. They also show that ISR imputations for direct payments outperform other imputation methods. Because the ISR methodology should preserve the relationships between variables in the ARMS data, we would hope to find differences in marginal effects using econometric models.

One important area we intend to explore is how household financial stress changes under the new imputation method. For example, the new imputation method may show that farm households have more off-farm debt than previously thought. Additionally, we will be interested to know how the overall well-being of the U.S. farm household changes under the new imputation methodology. To examine this, we will estimate financial ratios such as the debt-to-asset ratio, debt repayment capacity utilization, working capital and return on assets for both the new and old imputation methods. We can then compare these financial measures to see if farm households financial stress is greater or lower. Further analysis will break these measures out by categories such as farm sales, commodity specialization, region, and land tenure. We can then conduct hypotheses tests to determine if there are significant changes in farm financial stress in these categories.

Conclusion

This study proposes a new method to impute for ordinal variables found in the HH section of ARMS. The household section contains important economic information on farm operator well-being, asking respondents for information on off-farm income, household expenditures, and off-farm debt and assets. We extend a multivariate imputation method known as Iterative Sequential Regression (ISR) to accommodate the ordinal variables found in the household section. This transformation method allows these variables to be jointly modeled with other ARMS variables using a Gaussian copula. A conditional linear model for imputation is then built using correlation analysis and economic theory.

In our future work, we will show that ISR is a better alternative than current imputation methods for the HH section, and gives a more accurate picture of farm operator well-being. After running appropriate diagnostics on the data, we will use a simulation study that randomly "pokes holes" in the observed ARMS data. This will allow us to compare the new imputation methodology to conditional mean and complete cases analysis.

A new imputation method for the HH section is needed because ARMS is widely used for academic research and policymaking. The current ERS methodology uses an outdated conditional mean approach that does not meet important criteria for data analysis. In particular, this approach has been shown to bias the variance of imputed variables downward and distort multivariate relationships in the data. This is potentially a serious issue, as missing data in the household section could give biased measures of total household income or financial stress. Using a better imputation methodology, we hope to not only give researchers and policymakers better estimates, but also contribute to the discussion of the efficacy of U.S. agricultural programs. Given that current measures of farm operator financial stress are based on ARMS data that uses older imputation methods, there are important policy questions about whether current programs are sufficient to help farmers mitigate financial risk. This new imputation method will not only have an impact on farm financial measures, but a wide range of economic models that make use of the HH data.

References

Abayomi, K., A. Gelman, and M. Levy (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 57*(3), 273–291.

Abraham, K. G., A. Maitland, and S. M. Bianchi (2006). Nonresponse in the american time use survey who is missing from the data and how much does it matter? *Public Opinion Quarterly 70*(5), 676–703.

Ahearn, M., D. Banker, D. M. Clay, and D. Milkove (2011). Comparative survey imputation methods for farm household income. *American journal of agricultural economics*, aaq167.

Anderson, T. W. and D. A. Darling (1952). Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, 193–212.

Briggeman, B. C., S. R. Koenig, and C. B. Moss (2012). U.s. farm debt: the role of arms. *Agricultural Finance Review 72*, 254–261.

H. Kuethe, T., B. Briggeman, N. D. Paulson, and A. L. Katchova (2014). A comparison of data collected through farm management associations and the agricultural resource management survey. *Agricultural Finance Review 74*(4), 492–500.

Habiger, J. D., M. Robbins, and S. Ghosh (2010). An assessment of imputation methods for the usdas agricultural resource management survey. *JSM Proceedings: Section on Survey Research Methods*.

He, Y., A. M. Zaslavsky, M. Landrum, D. Harrington, and P. Catalano (2009). Multiple imputation in a large-scale complex survey: a practical guide. *Statistical methods in medical research*.

Kropko, J., B. Goodrich, A. Gelman, and J. Hill (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis 22*(4), 497–519.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association 83*(404), 1198–1202.

Little, R. J. and D. B. Rubin (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Miller, D., M. Robbins, and J. Habiger (2010). Examining the challenges of missing data analysis in phase three of the agricultural resource management survey. *JSM Proceedings, Section on Survey Research Methods, Alexandria, VA: American Statistical Association*.

Morehart, M., D. Milkove, and Y. Xu (2014). Multivariate farm debt imputation in the agricultural resource management survey (arms). In *2014 Annual Meeting, July 27-29, 2014, Minneapolis, Minnesota*, Number 169401. Agricultural and Applied Economics Association.

Nelsen, R. B. (2006). *An Introduction to Copulas* (2nd ed.). Springer-Verlag New York.

Raghunathan, T. E., P. W. Solenberger, and J. Van Hoewyk (2002). Iveware: Imputation and variance estimation software. *Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan*.

Robbins, M. W., S. K. Ghosh, and J. D. Habiger (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.

Robbins, M. W. and T. K. White (2011). Farm commodity payments and imputation in the agricultural resource management survey. *American journal of agricultural economics*, aaq166.

Robbins, M. W. and T. K. White (2014). Direct payments, cash rents, land values, and the effects of imputation in us farm-level data. *Agricultural and Resource Economics Review 43*(3).

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal 4*, 227–241.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*(434), 473–489.

Rubin, D. B. (2003). Nested multiple imputation of nmes via partially incompatible mcmc. *Statistica Neerlandica 57*(1), 3–18.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, Volume 81. John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schenker, N., T. E. Raghunathan, P.-L. Chiu, D. M. Makuc, G. Zhang, and A. J. Cohen (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association 101*(475), 924–933.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association 82*(398), 528–540.

Turnbull, B. C. and S. K. Ghosh (2014). Unimodal density estimation using bernstein polynomials. *Computational Statistics & Data Analysis 72*, 13–29.

Van Buuren, S., J. P. Brand, C. Groothuis-Oudshoorn, and D. B. Rubin (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation 76*(12), 1049–1064.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software 45*(3). Open Access.

Figure 1: HH section refusal rate (ARMS 2003-2013)

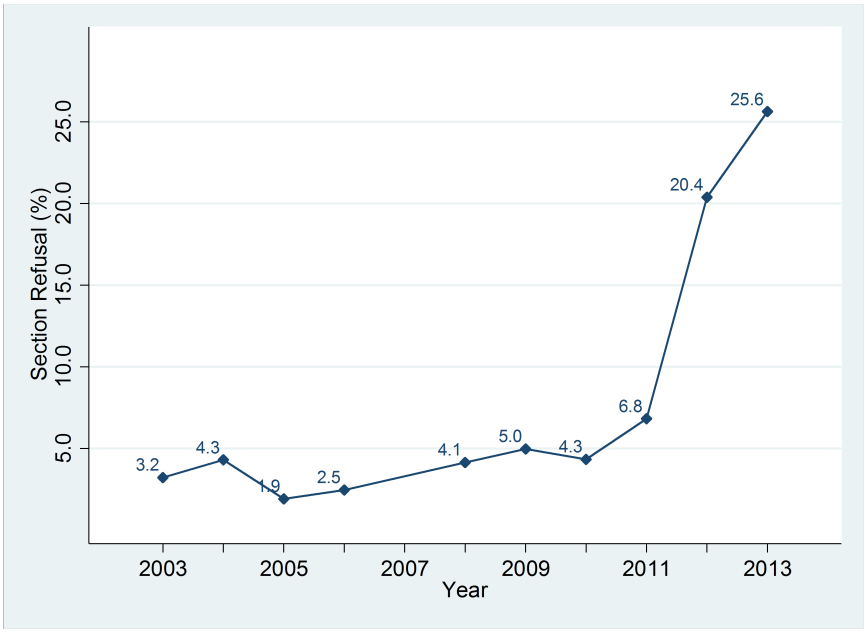

Figure 2: Histogram of Operator Off-farm wages (ARMS 2013)

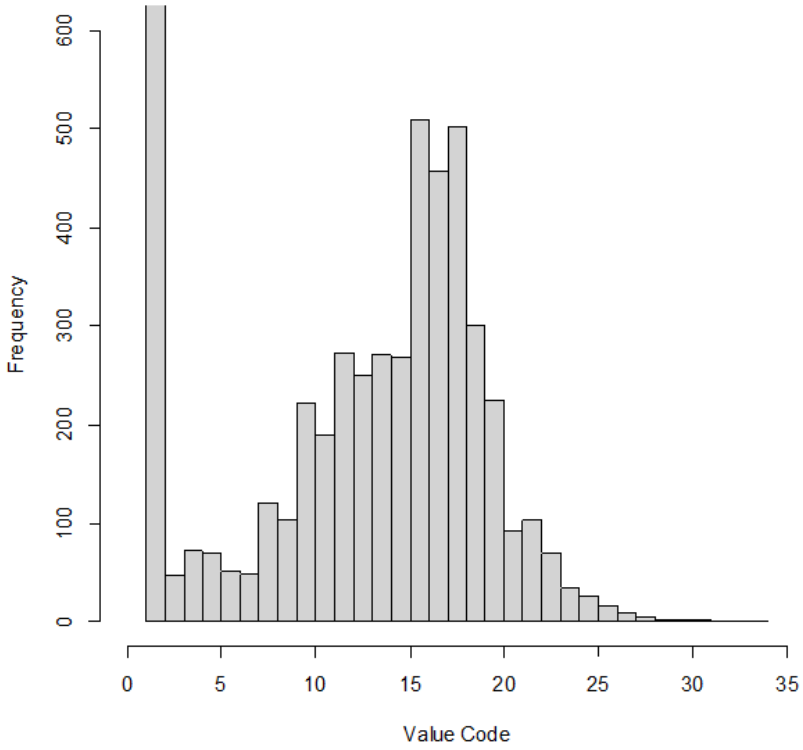

Figure 3: Histogram of Operator Off-farm wages (ARMS 2013)

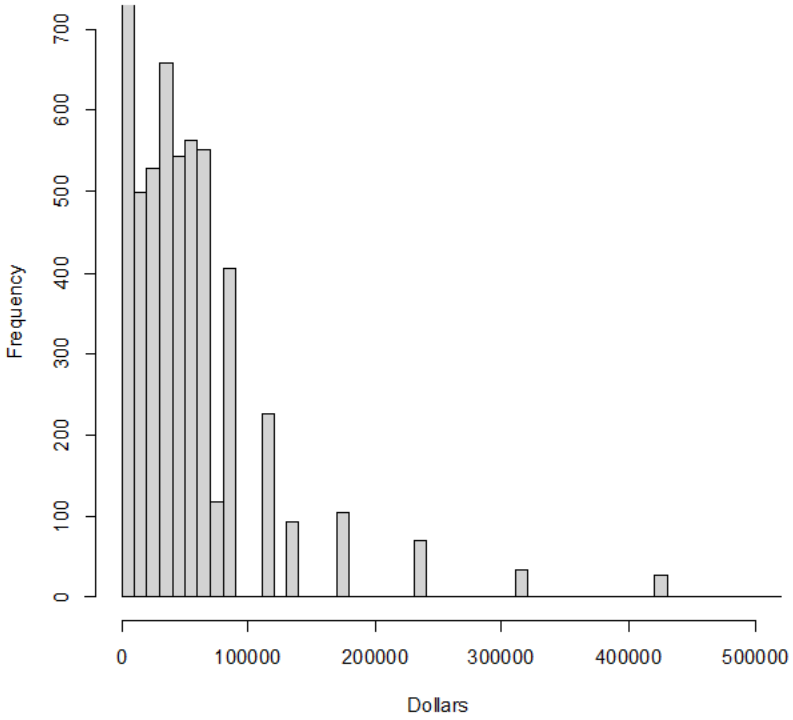

Figure 4: Histogram of Household Food Expenses (ARMS 2013)

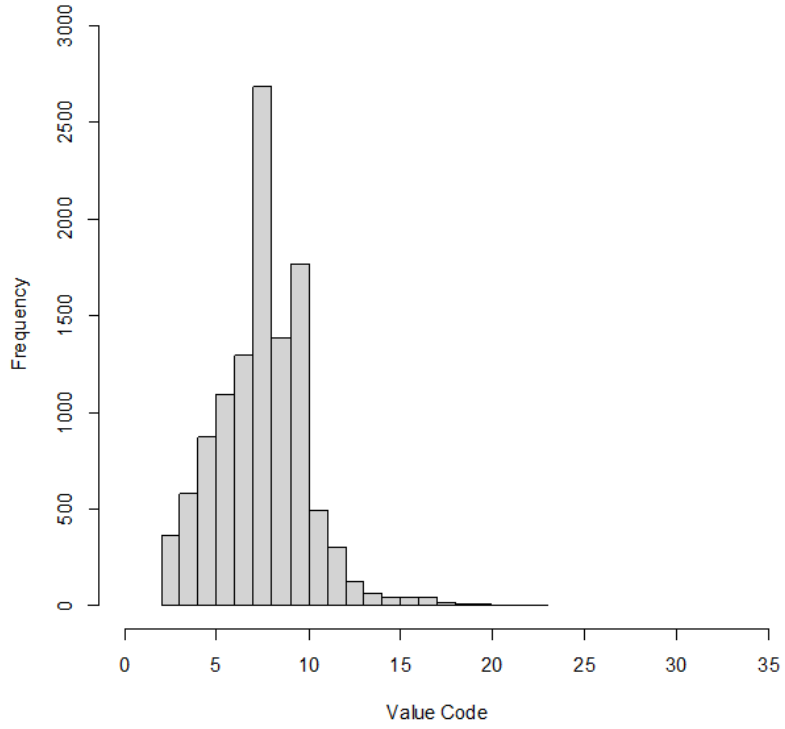

Figure 5: Histogram of Previous Year Net Farm Income (ARMS 2013)

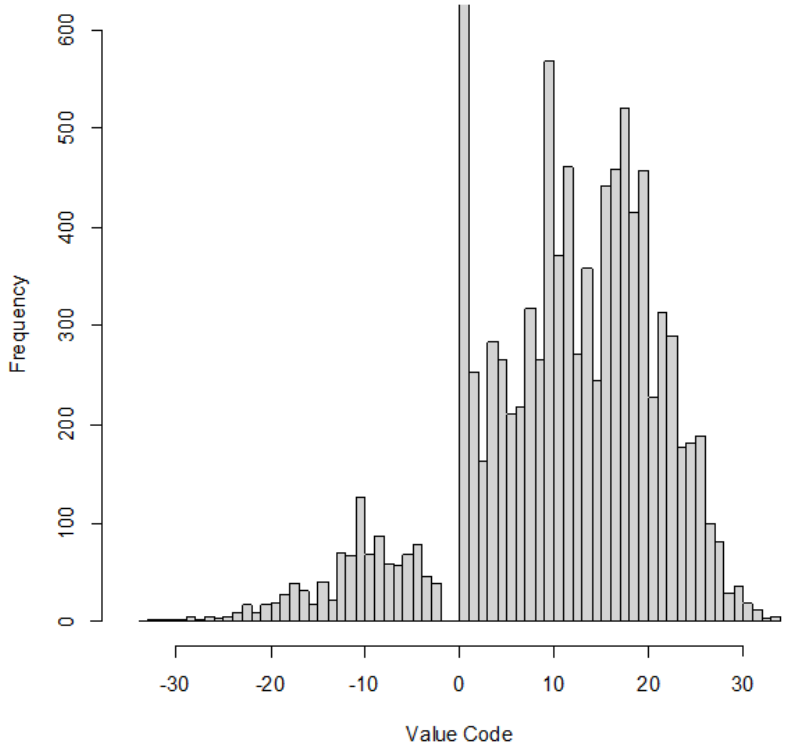

Table 1: Selected Value Codes for HH variables (ARMS 2013)

Dollar Range	Value Code
None	01
\$1 - \$499	02
\$500 - \$999	03
⋮	⋮
\$10,000 - \$14,999	10
\$15,000 - \$19,999	11
⋮	⋮
\$100,000 - \$124,999	20
\$125,000 - \$149,999	21
⋮	⋮
\$7,500,000 - \$9,999,999	33
\$10,000,000 and over	34

Table 2: Item refusal rates for selected HH variables (ARMS 2013)

Variable	Description	Item Refusal % (of usable surveys)	Item Cumulative Response %
R969	Interest Income	9.0%	35.4%
R1105	Food Expense	7.8%	35.9%
R953	Financial Assets	16.7%	32.4%
R988	Mortgage Debt	2.4%	38.0%

Table 3: Groups for ARMS HH variables requiring imputation (ARMS 2013)

Group	Description	# Requiring imputation
1	Off-farm Income	17
2	Household Spending	12
3	Off-farm Assets	7
4	Off-farm Debt	6
5	Previous Year Income & Expenses	3

Table 4: Selected ARMS HH variables requiring imputation (ARMS 2013)

Name	Description	Group
R950	Operator off-farm wages, salaries and tips	1
R956	Operator net cash income from operating another farm or ranch	1
R958	Operator net cash income from any other business	1
R1108	Health and dental insurance costs	2
R985	Asset value of operator dwelling (if not owned by operation)	3
R988	Mortgage on operator's dwelling (if not owned by operation)	4
R989	Mortgage on other real estate and other personal homes	4
R1114	Net operating income for operation in previous year	5

Table 5: Kendall's Tau Correlations for Selected HH variables (ARMS 2013)

Variable Description	Most Kendall's Tau	Correlated	Variables	
Operator off-farm wages (R950)	Operator's off-farm hours (Jan - Mar)	Operator's Major occupation	ERS Typology	Off-Farm Income Previous Year
	0.69	0.44	-0.40	0.36
Spouse off-farm wages (R951)	Spouse's off-farm hours (Jan - Mar)	Does Operator have spouse?	Spouse: Spanish, Hispanic, Latino?	Spouse's highest level of education
	0.71	-0.54	0.49	0.47