Implications of Missing Data Imputation for Agricultural Household Surveys: An Application to Technology Adoption

Haluk Gedikoglu
Assistant Professor of Agricultural Economics
Cooperative Research Programs
Lincoln University of Missouri
GedikogluH@lincolnu.edu

and

Joe L. Parcell
Professor
Department of Agricultural and Applied Economics
University of Missouri

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association's 2012 AAEA Annual Meeting, Seattle, Washington, August 12-14, 2012

Abstract

Missing data is a problem that occurs frequently in survey data. Missing data results in biased estimates and reduced efficiency for regression estimates. The objective of the current study is to analyze the impact of missing-data imputation, using multiple-imputation methods, on regression estimates for agricultural household surveys. The current study also analyzes the impact of multiple-imputation on regression results, when all the variables in the regression have missing observations. Finally, the current study compares the impact of univariate multiple imputation with multivariate normal multiple imputation, when some of the missing variables have discrete distribution. The results of the current study show that multivariate-normal multiple imputation performs better than univariate multiple imputation model, and overall both methods improve the efficiency of regression estimates.

Key Words: Missing Data, Multiple Imputation, Bayesian Inference, Household Surveys

1. Introduction

Missing data is a problem that occurs frequently in survey data. Missing data can cause biased estimates and reduce the efficiency of regression estimates (Rubin, 1987; Schafer, 1997). The loss of significant amount of observations can cause even simulation based regressions, such as seemingly unrelated multivariate probit model, not to reach a convergence due to lack of enough observations (Gedikoglu, 2008). The standard procedure on common econometrics and statistical software, such as Stata®, is to use observations those do not have any missing value, which is called listwise (casewise) deletion (Schafer, 1997). This can lead to a loss of significant number of observations for Agricultural Household Surveys. For example, the current study has a loss of 44% of the data due to missingness.

Over time, different methods have been used to handle missing data. A simple method of placing mean, predictive mean matching, single imputation, and multiple imputation are examples of missing data handling methods (Little and Rubin, 2002)[1]. Simple imputation treats imputed values as knows in the analysis, hence understates the variance of the estimates and overstates precision, which results in confidence intervals and significance tests to be too optimistic (Little and Rubin, 2002). Multiple imputation method addresses this problem by creating multiple imputations and taking into account the sampling variability due to the missing data (between-imputation variability) (Little and Rubin, 2002; Schafer, 1997).

Although statistical literature has been developed for missing data imputation, the agricultural economics literature has not used and got benefit from this literature, yet. There have been very few studies in agricultural economics that explored some parts of missing data imputation. The studies aimed to measure the impact of certain imputation techniques on the univariate missing variable (Robbins and White, 2011; Ahearn et. al., 2011). However, these

[1] Little and Rubin (2002) provides a detailed review on each method.

studies used complete data sets with randomly generated missing data. Hence, these studies do not address the complexities that the actual household surveys face, such us multiple missing variables and the correlation among these variables. An important aspect of multiple imputation literature that is lacking is how multiple imputation performs when all the variables in the data set have missing observations. Although the multiple imputation theory is developed, many of the practical problems are unanswered (Schafer, 1997). Another important question that has not been answered in the literature is whether or not to use multivariate multiple imputation instead of individual univariate imputations, when some of the missing variables have discrete distributions.

The objective of this paper is to analyze the impact of multiple imputation on Agricultural Household Surveys, when all the variables in the dataset have missing values. The current study also analyzes the impact of using multivariate normal multiple imputation, when some variables have discrete distribution and compares the results with univariate multiple imputation. In the next section, we provide information on missing-data patterns and missing-data mechanism. Multivariate normal imputation method is introduced next. Followingly, the univariate imputation methods are presented. The paper continuous with the results part and final conclusions are made.

2. Missing-Data Patterns

The missing data pattern is an important component of multiple imputation, which impact the choice of the multiple imputation method (Schafer, 1997). The missing data can occur in different patterns. We explain these patterns using an example from Enders (2010). Consider a 3 x 3 data matrix $X = (X_1, X_2, X_3)$ with 3 variables and 4 observations. An indicator matrix R can be formed based on R_{ij} = 1 if variable X_j is observed (complete) in observation i and otherwise:

$$R_1 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad R_2 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad R_3 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

where R_1 is an example for univariate missing-data pattern, R_2 is called monotone missing-data pattern, and R_3 is an general (nonmonotone) multivariate missing-data pattern (Enders, 2010). The univariate missing-data pattern R_1 can be imputed using the univariate multiple-imputation methods, based on the distribution of the variable with missing observations. For monotone missing-value pattern R_2, X_3 is at least as observed X_2 and X_2 is at least as observed as X_1. In this case multivariate multiple-imputation can be formulated as a sequence of independent univariate (conditional) imputation tasks, which allows a flexible imputation model (based on the distribution of each variable) and simplifies the imputation task (Rubin, 1987). However, in general it is very difficult to obtain monotone missing-value pattern. When the data shows the general missing-data pattern R_3, multivariate-normal (MVN) multiple-imputation model can be applied if the variables have continuous distribution (Rubin, 1987). Schafer (1999) shows that MVN model can be applied even when the variables have discrete distribution.

Table 1 shows the information about the missing data for the current study. All the variables in the current data set have some missing observations (percent of missing data is nonzero for all variables). Hence, a multivariate missing-data pattern exists in the current data set. A further analysis using Stata®'s data pattern function shows that there exists no monotone-data pattern. For that reason either a MVN multiple imputation method or a univariate multiple imputation method for each variable separately can be used.

3. Missing-Data Mechanisms

Missing data mechanism defines the distribution of the missing data for the data set and can be thought of the reason for the missingness (Rubin, 1987). The commonly used three missing data mechanisms are: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Denoting the observed part of n x p data matrix X, with n observations and p variables, by X_{obs} and missing part by X_{mis}, so that $X = (X_{obs}, X_{mis})$. For MAR, the probability that an observation is missing may depend on observed observation X_{obs} , but not on missing observations X_{mis}. This can be formally represented for a probability model as $\Pr\left(R_{np}|X_{obs}, X_{mis}, \varphi\right) = \Pr\left(R_{np}|X_{obs}, \varphi\right)$, where R n x p indicator matrix as and φ is the underlying vector of parameters of the missingness mechanism. The special case of MAR is MCAR. In this case observed data is a simple random sample of all potentially observable data values (Schafer, 1997). MCAR can be represented as $\Pr\left(R_{np}|X_{obs}, X_{mis}, \varphi\right) = \Pr\left(R_{np}|\varphi\right)$. Finally, MNAR is the case, if the probability of missing data on a variable is related to the missing values of itself and observed values of other variables.

When the data is generated as MCAR, then no bias would result from ignoring missing observation (Schafer, 1997). However, there will be still loss in efficiency of estimates, as not all the observations would be used. On the other hand, if MAR exists in the data set, ignoring missing observations will cause biased and inefficient estimates. The MAR restriction is not testable as the value of the missing data are unknown (Schafer, 1997). In general MAR is a better assumption than MCAR for most of the surveys, as MAR is more restrictive assumption. Lastly, MNAR would cause missingness mechanism to be accounted for in the model to obtain valid results, similar to Heckman’s selection model (Little and Rubin, 2002).

A missingness data mechanism is said to be ignorable if (a) the data set is MAR and (b) parameters for the missing data-generating process φ are unrelated to the regression parameters that we want to estimate. In this case there is no need to model data generating process for missing data (Schafer, 1997). In general, MAR and ignorability are often treated treated as equivalent under the condition that (b) for ignorability is almost always satisfied (Allison, 2002). Common statistical software, such as Stata®, assume MAR as it is difficult to test ignorability formall as MAR mechanism is distinguished from MNAR only through the missing data that is not observed (Schafer, 1997). In the current study we will assume MAR.

4. Multiple-Imputation Models

Multiple imputation methods, both multivariate and univariate, are based on simulation from a Bayesian posterior predictive distribution of missing data (Rubin, 1987: Schafer, 1997). The univariate imputation method uses noniterative techniques for simulation from the posterior predictive distribution of missing data, whereas multivariate methods use an iterative Markow Chain Monte Carlo (MCMC) technique (Rubin, 1987). Multiple imputation consists of three steps: imputation step, completed-data analysis step, and the pooling step. During the imputation step, M imputations (completed datasets) are generated under the chosen imputation model. The econometric model is performed separately on each imputation m=1,2,…,M in the completed − data analysis step. In the current study, a univarite probit model is used to represent the adoption of Roundup Ready® corn. Lastly, during the pooling step, the results obtained from M completed-data analyses are combined into a single multiple-imputation based estimation results. Below is the detailed description for each step.

4.1 Imputation Step

M imputations are generated under the chosen imputation model. The imputation model can be a univariate model or a multivariate model based on the number of variables to be imputed and the correlation among the variables. In the current study both univariate and multivariate models are used to evaluate the differences. In the current study there are three types of data: binary, ordered categorized, and continuous. Although multivaritate normal imputation is originally developed for imputing continues variables, studies show that multivariate normal model can be used for non-continous variables, given that imputed observation are again converted to categorized form after the imputation (Schafer, 1997; Lee and Carlin, 2010). For example, for binary variables, values smaller than 0.5 can be converted to 0, and other are converted to 1. Another option would be to use a logit based univariate multiple imputation for binary variables, ordered logit for ranked categorized variables, and linear regression based univariate multiple imputation for continuous variables. The disadvantage of this process is ignoring the correlation among imputed variables (Schafer, 1997). We provide information first on multivariate normal multiple imputation, then on univariate multiple imputation methods.

4.1.1 Multivariate Normal Multiple Regression

Multivariate normal (MVN) multiple regression model uses Data Augmentation, which is an iterative MCMC method, to impute missing values (Rubin, 1987). Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{N}$ be random sample from a *p*-variate normal distribution, representing the *p* imputation variables that have missing observations for observations *i*=1,…,*N*. The multivariate normal regression can be represented as:

$$\mathbf{x}_i = \mathbf{\Theta}'\mathbf{z}_i + \boldsymbol{\epsilon}_i,\ \ i = 1,\ldots,N$$

where $\boldsymbol{z}_i$ is a *q* x 1 vector of independent variables for observation *i*, $\mathbf{\Theta}$ is a *q* x *p* matrix of regression coefficients, and $\boldsymbol{\epsilon}_i$ is a *p* x 1 vector of random errors from a p-variate normal distribution with mean zero and a *p* x *p* variance-covariance matrix $\mathbf{\Sigma}$. $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ are referred as the model parameters. Next we provide the information on data augmentation.

4.1.1.1 Data Augmentation

Data augmentation consists of two steps, an I step (imputation step) and a P step (posterior step), which are preformed at each iteration *t* = 0,1,…,*T* (Schafer, 1997). Consider the partition of $\boldsymbol{x} = \left(\boldsymbol{x'}_{i(0)}, \boldsymbol{x'}_{i(m)}\right)$ corresponding to observed and missing values of imputation variables in observation *i*. At iteration *t* of the I step, the missing values in $\mathbf{x}_i$ are replaced with draws from the conditional posterior distribution of $\mathbf{x}_{i(m)}^{(t+1)}$ given observed data and the current values of model parameters $\mathbf{\Theta}^{(t)}$ and $\mathbf{\Sigma}^{(t)}$ independently for each observation (Little and Rubin, 2002). Following Little and Rubin (2002), in the current study, *T* is set as 100 following (Little and Rubin, 2002). Next, during the P step new values of model parameters $\mathbf{\Theta}^{(t+1)}$ and $\mathbf{\Sigma}^{(t+1)}$ are drawn from their conditional posterior distribution given observed data and data imputed in the previous I step $\mathbf{x}_{i(m)}^{(t+1)}$. These procedures can be represented as (Schafer, 1997; Little and Rubin, 2002):

I step: $$\mathbf{x}_{i(m)}^{(t+1)} \sim P\left(\boldsymbol{x}_{i(m)} \middle| \boldsymbol{z}_i, \boldsymbol{x}_{i(0)}, \mathbf{\Theta}^{(t)}, \mathbf{\Sigma}^{(t)}\right), i = 1, \dots, N$$

P step: $$\mathbf{\Sigma}^{(t+1)} \sim P\left(\mathbf{\Sigma} \middle| \boldsymbol{z}_i, \boldsymbol{x}_{i(0)}, \mathbf{x}_{i(m)}^{(t+1)}\right)$$

$$\mathbf{\Theta}^{(t+1)} \sim P\left(\mathbf{\Theta} \middle| \boldsymbol{z}_i, \boldsymbol{x}_{i(0)}, \mathbf{x}_{i(m)}^{(t+1)}\right)$$

the I and P steps are repeated until the MCMC sequence {($\mathbf{X}_m^{(t)}$, $\mathbf{\Theta}^{(t)}, \mathbf{\Sigma}^{(t)}$) : *t* = 1,2,…,*T*}, where $\mathbf{X}_m^{(t)}$ denotes all values imputed at iteration *t*, converges to the stationary distribution $P(\mathbf{X}_m, \mathbf{\Theta}, \mathbf{\Sigma} | \boldsymbol{Z}, \boldsymbol{X}_0,)$. The functional form of the conditional posterior distribution in the I and P steps above depends on the distribution of the data and a prior distribution of the model parameters. We use an improper uniform prior distribution for $\mathbf{\Theta}$, to reflect the uncertainty about $\mathbf{\Theta}$, and an inverted Wishart distribution $\mathrm{W}_p^{-1}(\lambda, \Lambda)$ for $\mathbf{\Sigma}$ (Rubin, 1987). In frequentist theory, Wishard distribution appears as the sampling distribution for the sample covariance matrix. The parameters λ and $\varLambda$ are called degrees of freedom and scale, respectively (Johnson and Wichern, 2002). The prior joint density function can be represented as:

$$f(\mathbf{\Theta}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\left(\frac{\lambda+p+1}{2}\right)} \exp\left(-\frac{1}{2}\mathrm{tr}\varLambda^{-1}\mathbf{\Sigma}^{-1}\right)$$

Wishard prior distribution is a natural conjugate to the multivariate normal likelihood function, which makes Bayesian inference to be conducted easily (Johnson and Wichern, 2002). Using the Baye's rule $P(\mathbf{\Theta}, \mathbf{\Sigma}|\boldsymbol{X}) \propto L(\mathbf{\Theta}, \mathbf{\Sigma}|\boldsymbol{X}) f(\mathbf{\Theta}, \mathbf{\Sigma})$, where $L(\mathbf{\Theta}, \mathbf{\Sigma}|\boldsymbol{X})$ is the standard multivariate normal likelihood function, the I and P steps become (Schafer, 1997):

I step: $\mathbf{x}_{i(m)}^{(t+1)} \sim N_{pi}\left(\boldsymbol{x}_{i(m)} \middle| \boldsymbol{\mu}_{m.o}^{(t)}, \mathbf{\Sigma}_{mm.o}^{(t)}\right), i = 1, \dots, N$

P step: $\mathbf{\Sigma}^{(t+1)} \sim W^{-1}(\boldsymbol{\varLambda}_{*}^{(t+1)}, \lambda_{*})$

$$\mathrm{vec}\left(\mathbf{\Theta}^{(t+1)}\right) \sim N_{pq}\left(\mathrm{vec}\left(\widehat{\mathbf{\Theta}}^{(t+1)}\right), \mathbf{\Sigma}^{(t+1)} \otimes (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\right)$$

where *pi* is the number of imputation variables containing missing values in observation *i*, $\otimes$ is the Kronecker product, and vec(.) is the vectorization of a matrix into a column vector. Submatrices $\boldsymbol{\mu}_{m.o}^{(t)}$ and $\mathbf{\Sigma}_{mm.o}^{(t)}$ are the mean and variance of the conditional distribution of $\boldsymbol{x}_{i(m)}$given $\boldsymbol{x}_{i(0)}$ based on $\mathbf{x}_i \sim N_p(\mathbf{\Theta}^{(\mathrm{t})'}\boldsymbol{z}_i, \mathbf{\Sigma}^{(\mathrm{t})})$. The matrix $\widehat{\mathbf{\Theta}}^{(t+1)} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}^{(t+1)}$ is the ordinary least squares estimate of regression coefficients based on the augmented data $\boldsymbol{X}^{(t+1)} = (\mathbf{X}_o, \mathbf{X}_m^{(t+1)})$ from iteration *t*. The posterior scale matrix $\varLambda_*^{(t+1)}$ and the posterior degrees of freedom for the inverted Wishart distribution λ_* are defined as (Johnson and Wichern, 2002):

$$\varLambda_*^{(t+1)} = \left\{\varLambda^{-1} + \left(\boldsymbol{X}^{(t+1)} - \boldsymbol{Z}\widehat{\mathbf{\Theta}}^{(t+1)}\right)'\left(\boldsymbol{X}^{(t+1)} - \boldsymbol{Z}\widehat{\mathbf{\Theta}}^{(t+1)}\right)\right\}^{-1}$$

$$\lambda_* = \lambda + N - q$$

Values for the degrees of freedom and the scale parameter are determined based on the requested prior distribution for $\mathbf{\Theta}$. For the uniform prior distribution for $\mathbf{\Theta}$, the values are λ = -(*p*+1) and

$\mathit{\Lambda}^{-1} = \ \mathbf{0}_{pxp}$, where $\mathbf{0}_{pxp}$ is a zero matrix (Johnson and Wichern, 2002). In the current study, to reflect uncertainty about model parameters, noninformative uniform prior distribution is used.

4.1.1.2 Expectation-Maximization Algorithm

The initial values $\boldsymbol{\Theta}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$for the Data Augmentation above are obtained from the Expectation-Maximization (EM) algorithm (Schafer, 1997). The EM algorithm iterates the expectation step (E step) and maximization step (M step) to maximize the log-likelihood function. The observed-data likelihood function is (Schafer, 1997).:

$$l_l(\boldsymbol{\Theta},\boldsymbol{\Sigma}|\boldsymbol{X_o}) = \sum_{s=1}^{S}\sum_{i\epsilon I(s)}\left\{-0.5\ln(|\boldsymbol{\Sigma}_{\mathbf{s}}|) - 0.5(\mathbf{x}_{i(o)} - \boldsymbol{\Theta}'_{(\mathrm{s})}\mathbf{z}_i)'\boldsymbol{\Sigma}^{-\mathbf{1}}{}_{\mathrm{s}}(\mathbf{x}_{i(o)} - \boldsymbol{\Theta}'_{(\mathrm{s})}\mathbf{z}_i)\right\}$$

where *S* is the number of unique missing-value patterns in the full-data, *I(s)* is the set of observations from the same missing-value pattern *s*, and $\boldsymbol{\Theta_s}$ and $\boldsymbol{\Sigma_s}$ are the submatrices of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ that correspond to the imputation variables, which are observed in pattern s. In the current data set *S* is 87. Using the prior joint density function, and the log-likelihood function above, the log-posterior function is obtained as (Schafer, 1997):

$$l_p(\boldsymbol{\Theta},\boldsymbol{\Sigma}|\boldsymbol{X_o}) = l_l(\boldsymbol{\Theta},\boldsymbol{\Sigma}|\boldsymbol{X_o}) + \ln\{f(\boldsymbol{\Theta},\boldsymbol{\Sigma})\} - \frac{\lambda+p+1}{2}\ln\,(|\boldsymbol{\Sigma}|) - \frac{1}{2}\mathrm{tr}(\mathit{\Lambda}^{-1}\boldsymbol{\Sigma}^{-1})$$

The E step and M step are processed as using the sufficient statistics for the multivariate normal distribution. Let $T_1 = \sum_{i=1}^{N}\boldsymbol{z}_i\boldsymbol{x'}_i$ and $T_2 = \sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x'}_i$ denote the sufficient statistics for the multivariate normal model. The submatrices $\boldsymbol{\Theta}_{i(\mathrm{s})}$and $\boldsymbol{\Theta}_{i(\mathrm{m})}$ of $\boldsymbol{\Theta}$, and the submatrices $\boldsymbol{\Sigma}_{i(\mathrm{mm})}$, $\boldsymbol{\Sigma}_{i(\mathrm{mo})}$, and $\boldsymbol{\Sigma}_{i(\mathrm{oo})}$ of $\boldsymbol{\Sigma}$ corresponding to the observed and missing column of $\boldsymbol{x}_i$. Let *O(s)* and *M(s)* correspond to the column indexes of the observed and missing parts of $\boldsymbol{x}_i$ for each missing-values pattern *s* (Little and Rubin, 2002; Rubin, 1987).

E Step: The expectations $\mathrm{E}(T_1)$ and $\mathrm{E}(T_2)$ are computed with respect to the conditional distribution $P\left(\mathbf{X}_m\middle|\boldsymbol{\Theta}^{(\mathrm{t})},\boldsymbol{\Sigma}^{(\mathrm{t})},\boldsymbol{X}_0\right)$ using the following relations (Little and Rubin, 2002; Rubin, 1987):

$$E\left(x_{ij}\middle|\boldsymbol{\Theta}^{(\mathrm{t})},\boldsymbol{\Sigma}^{(\mathrm{t})},\boldsymbol{X}_0\right)=\begin{cases}x_{ij},\text{for } j\epsilon\ O(s)\\ x_{ij}^{*},\text{for } j\epsilon\ M(s)\end{cases}$$

and

$$E\left(x_{ij}x_{il}\middle|\boldsymbol{\Theta}^{(\mathrm{t})},\boldsymbol{\Sigma}^{(\mathrm{t})},\boldsymbol{X}_0\right)=\begin{cases}x_{ij}x_{il},\text{for } j,l\ \epsilon\ O(s)\\ x_{ij}^{*}x_{il},\text{for } j\epsilon\ M(s),l\ \epsilon\ O(s)\\ c_{ij}+x_{ij}^{*}x_{il}^{*},\text{for } j,l\ \epsilon\ M(s)\end{cases}$$

where x_{ij}^{*} is the jth element of the vector $\boldsymbol{\Theta}'_{i(\mathrm{m})}\boldsymbol{z}_i+\boldsymbol{\Sigma}_{i(\mathrm{mo})}\boldsymbol{\Sigma}_{i(\mathrm{oo})}^{-1}(\mathbf{x}_{i(o)}-\boldsymbol{\Theta}'_{(\mathrm{s})}\mathbf{z}_i)$, and c_{ij} is the element of the matrix $\boldsymbol{\Sigma}_{i(\mathrm{mm})}-\boldsymbol{\Sigma}_{i(\mathrm{mo})}\boldsymbol{\Sigma}_{i(\mathrm{oo})}^{-1}\ \boldsymbol{\Sigma}'_{i(\mathrm{oo})}$ (Little and Rubin, 2002; Rubin, 1987).

M step: During the M step, the model parameters are updated using the computed expectations of the sufficient statistics:

$$\boldsymbol{\Theta}^{(t+1)}=(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathrm{E}(T_1)$$

$$\boldsymbol{\Sigma}^{(\mathrm{t}+1)}=\frac{\mathbf{1}}{N+\lambda+p+1}\{\mathrm{E}(T_2)-\mathrm{E}(T_1)'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathrm{E}(T_1)+\Lambda^{-1}\}$$

The EM iterates between the E step and M step until the maximum relative difference between the two successive values of all parameters is less than the specified tolerance (in this paper it is 1e-5) (Little and Rubin, 2002; Rubin, 1987).

4.1.1.3 Convergence of Data Augmentation

Data augmentation (DA) procedure is iterated until the MCMC sequence {($\mathbf{X}_m^{(t)}$, $\boldsymbol{\Theta}^{(t)},\boldsymbol{\Sigma}^{(t)}$) : *t* = 1,2,…,*T*} converges to a stationary distribution (Schafer, 1997). Unlike optimization based EM procedure, the DA procedure does not have a simple stopping rule that guarantees the convergence of the MCMC sequence to a stationary distribution (Schafer, 1997). Hence, the question is how long to iterate to achieve the convergence. Another issue is the serial dependence known to exist among MCMC draws (Schafer, 1997). Suppose that after an initial burn-in period *b*, the sequence {($\mathbf{X}_m^{(b+t)}$) : *t* = 1,2,…,*T*} can be regarded as an approximate sample from $\Pr(\mathbf{X}_m|\mathbf{X}_o)$ (Schafer, 1997, Little and Rubin, 2002). To achieve the independence among imputations, a chain can be sampled. In order to achieve the independence, the number of iterations *k* should be determined such that $\mathbf{X}_m^{(t)}$ and $\mathbf{X}_m^{(t+k)}$ are approximately independent (Schafer, 1997, Little and Rubin, 2002).. Then imputations can be obtained as the chain values of $\mathbf{X}_m$ from iterations *b*, *b+k*, *b+2k,…,b+Mk*, where *M* is the number of imputations. Hence, *b* is the number of iterations necessary for the chain to achieve stationary *k* is the number of iterations between imputations necessary to achieve independent values of the chain.

We use the worst linear function (WLF), developed by Schafer (1997) to detect the convergence and autocorrelation for the DA. WLF corresponds to the linear combination of parameter estimates where the coefficients are chosen such that this function has the highest asymptotic rate of missing information. Schafer (1997) shows that when the observed-data posterior distribution is nearly normal, WLF is among the slowest to approach stationarity. WLF can be calculated as (Schafer, 1997):

$$\omega(\boldsymbol{\theta}) = \widehat{v'}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)$$

where $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ are column vectors of the unique model parameters and their respective EM estimates; $\widehat{v'}\left(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\right)$, where $\boldsymbol{\theta}^{(t)} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{(t-1)}$are the estimates from the last and one before the last iterations of the EM algorithm. Below figure 1 shows the convergence of the DA and figure 2 shows the lag between iterations to analyze the autocorrelation for DA for the current study. Figure 1 shows no apparent trend, hence the convergence is reached for DA. Figure 2 shows that the autocorrelation dies out after the 1st lag. Hence, setting burn-in period *b* to be 100 and burn-between period *k* to be also 100 provided the convergence and no-serial correlation for the current study.

Figure 1: Convergence of DA

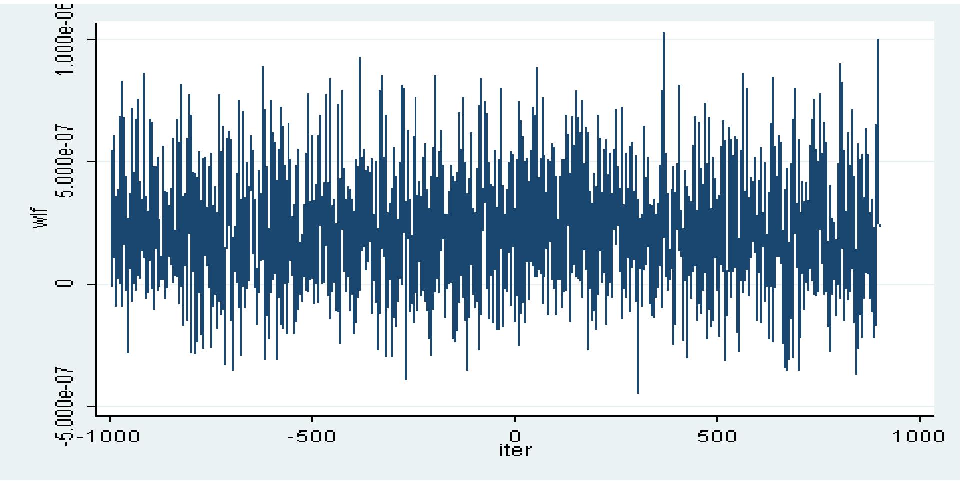
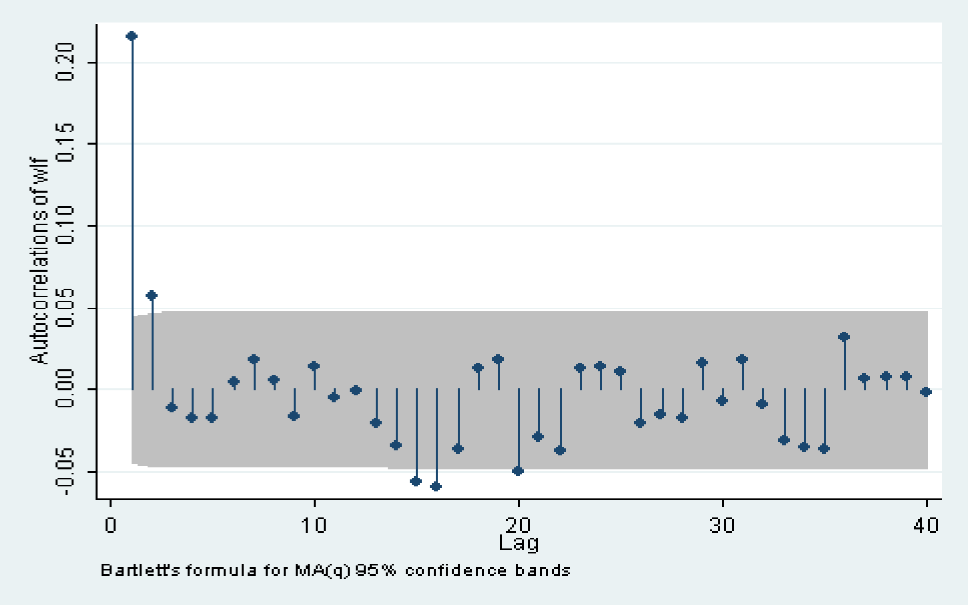


Figure 2: Autocorrelation of DA Draws



4.1.2 Univariate Multiple Imputation

The current data set have three types of variables: binary, ordered categorized, and continuous. Hence, we will use three univariate multiple imputation methods; logit, ordered logit, and linear regression, for the corresponding variables separately. Univariate multiple imputation methods address the specific distribution for each variable, but ignore the correlation among the variables. Below, we provide the information for each univariate multiple imputation method.

Logistic regression model is a parametric model that assumes an underlying logistic distribution for a binary missing variable. Imputation method for the logistic model is based on asymptotic approximation of posterior predictive distribution of missing data. Actual posterior distribution of logictic model parameters does not have a simple form. Hence, a large-sample normal approximation to the posterior distribution is used. A binary univarite variable that contains missing values to be imputed can be represented as $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$, which follows a logistic model (van Buuren, 2007; Rubin, 1987):

$$\Pr\left(x_i = 1 | \boldsymbol{z}_i\right) = \exp\left(\boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\beta}\right)/1 + \exp\left(\boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\beta}\right)$$

where $\boldsymbol{z}_{\boldsymbol{i}} = \left(z_{i1}, z_{i2}, \ldots, z_{iq}\right)'$ records values of predictors of $\mathbf{x}$ for observation i, $\boldsymbol{\beta}$ is the q x 1 vector of unknown regression coefficients. Consider the partition of $\boldsymbol{x} = (\boldsymbol{x'}_0, \boldsymbol{x'}_m)$ into two vectors containing the complete $(\boldsymbol{x'}_0)$ and incomplete $(\boldsymbol{x'}_m)$ observations. A similar partition is done for $\boldsymbol{Z} = (\boldsymbol{Z'}_0, \boldsymbol{Z'}_m)$ into $n_0\ x\ q$ and $n_1\ x\ q$ submatrices. The following steps are followed to fill in $\boldsymbol{x'}_m$ using multiple imputation (van Buuren, 2007; Rubin, 1987):

I. Fit a logistic model (1) to the observed data $(\boldsymbol{x}_0, \boldsymbol{z}_0)$ to obtain maximum likelihood estimates, $\widehat{\boldsymbol{\beta}}$, and their asymptotic sampling variance, $\widehat{\mathbf{U}}$.

II. New parameters $\boldsymbol{\beta}_*$ are *simulated* from the large-sample normal approximation, $N(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{U}})$, to its posterior distribution assuming the noninformative prior $\Pr\left(\boldsymbol{\beta}\right) \propto$ const.

III. One set of imputed values, $\boldsymbol{x}^1{}_m$, is obtained by simulating from the logistic distribution $\Pr\left(x_i = 1 | \boldsymbol{z}_i\right) = \exp\left(\boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\beta}\right)/1 + \exp\left(\boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\beta}\right)$ for every missing observation i_m.

IV. The steps 2 and 3 are repeated to obtain *M* sets of imputed values,

$$\boldsymbol{x}^1{}_m, \boldsymbol{x}^2{}_m, \ldots, \boldsymbol{x}^M{}_m.$$

Steps 2 and 3 correspond to only approximate draws from the posterior predictive distribution of missing data Pr ($\mathbf{x}_m|\mathbf{x}_o, \boldsymbol{Z}_o$), because $\boldsymbol{\beta}_*$ is drawn from the asymptotic approximation to its posterior distribution (Rubin, 1987).

For the ranked discrete missing variables, the ordered logistic model is used. The model can be represented for an ordered *K* categorized variable as $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$**,** as (van Buuren, 2007):

$$\text{Pr } (x_i = k|\boldsymbol{z}_i) = \text{Pr } (\gamma_{k-1} < \boldsymbol{z}_i'\boldsymbol{\beta} + u < \gamma_k) = \frac{1}{1+\exp(\boldsymbol{z}_i'\boldsymbol{\beta}-\gamma_k)} - \frac{1}{1+\exp(\boldsymbol{z}_i'\boldsymbol{\beta}-\gamma_{k-1})}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K-1})'$ are the unknown cutpoints. The steps I through IV above for the logistic model are conducted for the ordered logistic model to obtain the *M* set of imputed values.

For the continuous missing variables, linear regression model is used to obtain the imputed values. The linear regression model for a continuous variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ can be represented as (Gelman et al, 2004):

$$x_i|\boldsymbol{z_i} \sim N(\boldsymbol{z_i}'\boldsymbol{\beta}, \sigma^2)$$

The steps I through IV above for the logistic model are followed to obtain the *M* set of imputed values for the linear regression model (Little and Rubin, 2002). for the ordered logistic model to obtain the *M* set of imputed values.

4.2 Completed-data Analysis Step

In the current study, our ultimate objective is to analyze the adoption of Roundup Ready® corn. The adoption decision of farmers can be represented as binary variable *y* (Greene, 2008):

$$\Pr(\ y_i = 1|\boldsymbol{x}_i) = \exp(\boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{q})/1 + \exp(\boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{q})$$

where $y_i = 1$ if the farmer adopts Roundup Ready® corn and $y_i = 0$ if the farmer does not adopt Roundup Ready® corn. $\boldsymbol{q}$ is the vector of coefficients in interest, in the completed-data analysis, to be estimated. This model is performed separately on each set of imputed data (completed data) *m*=1,2,…,*M*.

4.3 Pooling Step

The results obtained from *M* completed-data analyses are combined into a single multiple-imputation based estimation results. Let $\{(\widehat{\boldsymbol{q}}_{\boldsymbol{\imath}}, \widehat{\mathbf{U}_{\mathrm{ı}}}) \colon \mathrm{i} = 1,2, \dots, M\}$ be the completed-data estimates of **q** and the respective variance covariance estimates **U** from *M* imputed datasets. The multiple imputation estimate of **q** is $\overline{\boldsymbol{q}}_M = \frac{1}{M}\sum_{i=1}^{M} \widehat{\boldsymbol{q}}_{\boldsymbol{\imath}}$. The variance-covariance estimate of $\overline{\boldsymbol{q}}_M$(total variance) is $\mathbf{T} = \overline{\mathbf{U}} + \left(1 + \frac{1}{M}\right)\mathbf{B}$, where $\overline{\mathbf{U}} = \frac{1}{M}\sum_{i=1}^{M} \widehat{\boldsymbol{U}_{\boldsymbol{\imath}}}/M$ is the within-imputation variance-covariance matrix and $\mathbf{B} = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{q}_{\boldsymbol{i}} - \overline{\boldsymbol{q}}_M)(\boldsymbol{q}_{\boldsymbol{i}} - \overline{\boldsymbol{q}}_M)'/(M-1)$ is the between-imputation variance-covariance matrix.

5. Data

The data for the current study is obtained through a mail survey of 2995 farm operations in Iowa and Missouri in spring 2011. The questions were designed to learn farmers' adoption of new technologies and how the farmer's and the farm's characteristics impacted the adoption decision. The survey was sent out to a test group of 100 farmers and was revised before developing the final survey instrument. The final survey was sent out with a cover letter and a postage paid return envelope. A reminder postcard was sent after two weeks. The effective response rate for the survey was 21 percent. Before calculating the response rate, the farmers that had stopped farming, farmers that returned the survey due to not being the farm operator, and undeliverable surveys were subtracted from the original number of surveys that were sent out. The effective rate is the number of returned surveys divided by the adjusted number of surveys sent, times 100.

Table 1 provides information about the type of the data, whether the variable is binary, categorized, or continuous. The current data set have both three types of variables. Also, both number and percentage of missing observations are reported. In general, the percentage of missing observations is low for most of the variables, except for spouse’s education and off-farm income. Even if the percentage of missing observations is low for most of the variables, percentage of complete observations, which is used in most statistical programs for the regression analysis, is 56 percent. Hence, 44 percent of the observation would not be used in the regression analysis using no imputation methods. In table 1, also reports that when MVN imputation is used, the missing observations are imputed leading the number of complete observation to be 472.

6. Results

To see the impact of imputation methods on the variables, mean and the standard deviation for the imputed variables are compared between no imputation (m=0), the 5th imputation and the 10th imputation for MVN and univarite multiple imputation methods. Table 2 reports the results for mean and table 3 reports the results for standard deviation. Overall, the multiple imputation methods do not cause a significant variation for both mean and standard deviation for the variable. The results vary for discrete variables only in the second or the third digit for most of the variables. Even for the spouse education variable, which has the highest percentage of missing observation 25 percent, both multiple imputation methods provided the mean and standard deviation very close to no imputation values.

Table 4 provides the comparison of probit regression results between no imputation case and MVN multiple-imputation with *M* is set as 10. The hypothesis that all the regression coefficients except the constant term is rejected for both regressions with the p-values of 0.000. Hence, both the no-imputation and MVN imputation regressions are significant. For the individual variables in the regression, two of the variables that were not significant in the no-imputation case became significant in the multiple-imputation case, e.g. age and land rented in. On the other hand, the university / extension variable was significant in no-imputation regression, but it is not significant in the MVN imputed regression. It is important to see that all the variable estimates have *lower* standard error in the MVN imputed regression than no-imputed regression. Hence, MVN imputation increased the efficiency of the estimates and corrected for the biased estimates. For the cases of age and land rented in variables, no-imputation regression underestimated these variables and it over estimated the university / extension variable.

Table 4 provides the comparison of probit regression results between no imputation case and univariate multiple-imputation method with *M* is set as 10. The hypothesis that all the regression coefficients except the constant term is rejected for the univarite-imputed regression with the p-values of 0.000. Hence, in addition to no-imputation regression being significant, univariate-imputed regression is also significant. Similar to no-imputed regression, age and land rented in variables are not statistically significant in the univariate-imputed regression at 10 percent significance level (even if they have lower p-values in the univariate-imputed regression). However, university / extension variable is not significant in the univariate-imputed regression. Hence, the univariate-imputed regression shows that the university / extension variable is overstated in the no-imputation regression. Also, comparison of the standard errors of the variables estimates show that all the estimates have *lower* standard errors in the univaria-imputed regression than no-imputation regression.

The comparison of the standard errors between the MVN multiple imputation regression and the univariate multiple imputation regression shows that all the estimates have lower standard errors in the MVN imputation regression than the univariate imputed regression. Hence, MVN imputed regression results are more efficient than univariate imputed regression. Table 6 shows the impact of missing variables on the variable estimates for MVN multiple imputed regression and univariate imputed regression. The relative variance increase (RVI), which is the increase in variance of a variable due to missing information, are relatively low for all variables both MVN imputed regression and univariate imputed regression, with the exception for spouse off-farm income and education variables for MVN imputed regression. This is expected as these two variables had the highest percentage of missing observations. Same results are also valid for the fraction of missing information (FMI), which shows the ratio of information lost due to missing data to the total information that would be present if there were no missing data. Lastly, the relative efficiency in table 6 is helpful in deciding the number of total imputation *M*, which is 10 in the current study. Relative efficiency shows the ratio of the variance of an estimator with *M* is set as 10 to the variance if *M* was infinite. The results show that the relative efficiencies are very high for both MVN and univariated imputed regressions. Hence, choice is setting *M* in the current study is justified.

7. Conclusion

The current study analyzed the impact of missing data and multiple imputation methods on an agricultural household survey. Current study shows that there can be significant amount of information lost in household surveys due to non-response. Even the individual percentages of missing data was low for the variables of interest in the current study, overall 44 percent of observations could not be used in a standard regression due to non-response, when no imputation methods are used.

The current study also analyzed the impact of using multivariate normal multiple imputation method when some of the variables have discrete distribution and the results are compared to univariate multiple imputation method, which provides imputation based on the distribution of each variable. Our results showed that the regression estimates had lower standard error for multivariate normal imputed regression than the univariate imputed regression. Hence, multivariate normal method is preferred to univariate method, even when the variables have non continuous distributions. Overall, multiple imputation methods provided estimates with lower standard error than no imputation regression. Hence, use of multiple imputation methods can improve the efficiency of regression results.

8. References

Ahearn, M., B. David, D. M. Clay, and D. Milkove. 2011. “Comparative Survey Imputation Methods for Farm Household Income.” *American journal of Agricultural Economics*, 93(2): 613-618.

Enders, C.K. 2010. “Applied Missing Data Analysis.” Gilford Press, New York.

Gedikoglu, H. 2008. “Adoption of Nutrient Management Practices.” Ph.D. Dissertation, University of Missouri.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis. 2nd ed. London: Chapman & Hall/CRC.

Greene, W. H. 2008. Econometric Analysis. New York: Prentice-Hall Inc.

Johnson, R.A., and D.W. Wichern. 2002. Applied Multivariate Statistical Analysis. New York: Prentice-Hall Inc.

Lee, K.J., and J.B**.** Carlin. 2010. “Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation.” *American Journal of Epidemiology* (5):624-632.

Little, R. J. A., and D. B. Rubin. 2002. Statistical Analysis with Missing Data. 2nd ed. Hoboken, NJ: Wiley.

Robbins, M.W., and T.K. White. 2011. “Farm Commodity Payments and Imputation in the Agricultural Resource Management Survey.” *American journal of Agricultural Economics*, 93(2): 606-612.

Rubin, D. B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Schafer, J. L. 1997. Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman & Hall/CRC.

van Buuren, S. 2007. “Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification.” *Statistical Methods in Medical Research* 16: 219–242.

Table 1. Number of Missing Observations and Imputed

Variables	Type	Complete Obs.	Missing Obs.	Percentage Missing	Imputed MVN	Total MVN	Imputed Univ.	Total Univ.
Roundup Ready Corn	Binary	453	19	4%	19	472	1	452
Age	Continuous	460	12	3%	12	472	3	463
Owned Land	Continuous	466	6	1%	6	472	2	468
Land Rented Out	Continuous	461	11	2%	11	472	6	467
Land Rented In	Continuous	459	13	3%	13	472	6	465
State	Binary	467	5	1%	5	472	1	468
Farm Sales	Categorized	456	16	3%	16	472	4	460
Non-Family Labor	Binary	463	9	2%	9	472	2	465
Environmental Perceptions								
Water Quality	Categorized	463	9	2%	9	472	2	465
Air Quality	Categorized	443	29	6%	29	472	8	451
Global Warming	Categorized	463	9	2%	9	472	1	464
Sources of Information								
Other Farmers	Categorized	452	20	4%	20	472	2	455
Non-farming Neighbors	Categorized	450	22	5%	22	472	1	451
Banks	Categorized	449	23	5%	23	472	1	450
Contractors	Categorized	449	23	5%	23	472	0	449
University / Extension	Categorized	453	19	4%	19	472	0	453
USDA	Categorized	447	25	5%	25	472	0	447
Other Government Org.	Categorized	446	26	6%	26	472	0	446
Off-Farm Income								
Farm Operator	Categorized	458	14	3%	14	472	0	458
Spouse	Categorized	373	99	21%	99	472	6	379
Education								
Farm Operator	Categorized	436	36	8%	36	472	4	440
Spouse	Categorized	352	120	25%	120	472	12	364
Total Animal Units	Continuous	462	10	2%	10	472	3	465

Table 2. Comparison of the Mean Between No Imputation and Multiple Imputations

Variables	**No Imputation**	**MVN Multiple Imputation**		**Univariate Multiple Imputation**	
	m = 0	**m = 5**	**m=10**	**m =5**	**m= 10**
Roundup Ready Corn	0.466	0.464	0.466	0.453	0.455
Age	53	53	53	53	53
Owned Land	235	234	234	236	234
Land Rented Out	20	20	20	21	19
Land Rented In	170	167	166	168	167
State	0.490	0.489	0.489	0.493	0.493
Farm Sales	3.171	3.167	3.172	3.160	3.169
Non-Family Labor	0.283	0.282	0.284	0.283	0.282
Environmental Perceptions					
Water Quality	3.994	3.994	3.994	3.991	3.983
Air Quality	4.115	4.104	4.113	4.097	4.116
Global Warming	2.544	2.541	2.541	2.542	2.534
Sources of Information					
Other Farmers	2.573	2.571	2.574	2.560	2.586
Non-farming Neighbors	1.718	1.716	1.716	1.714	1.714
Banks	1.866	1.864	1.864	1.851	1.837
Contractors	1.490	1.490	1.490	1.498	1.482
University / Extension	2.210	2.210	2.210	2.215	2.197
USDA	2.145	2.145	2.145	2.158	2.151
Other Government Org.	1.794	1.794	1.794	1.812	1.798
Off-Farm Income					
Farm Operator	2.614	2.614	2.614	2.621	2.602
Spouse	2.842	2.842	2.842	2.821	2.777
Education					
Farm Operator	2.489	2.502	2.498	2.491	2.472
Spouse	2.744	2.758	2.742	2.680	2.646
Total Animal Units	187	187	187	189	185

Table 3. Comparison of the Standard Deviation Between No Imputation and Multiple Imputations

Variables	**No Imputation**	**MVN Multiple Imputation**		**Univariate Multiple Imputation**	
	m = 0	**m = 5**	**m = 10**	**m = 5**	**m= 10**
Roundup Ready Corn	0.499	0.499	0.501	0.500	0.498
Age	11	11	11	11	11
Owned Land	256	255	255	256	256
Land Rented Out	102	102	102	102	102
Land Rented In	339	337	339	341	342
State	0.500	0.503	0.502	0.500	0.500
Farm Sales	1.502	1.498	1.489	1.498	1.498
Non-Family Labor	0.451	0.449	0.452	0.450	0.451
Environmental Perceptions					
Water Quality	1.191	1.192	1.195	1.190	1.190
Air Quality	1.080	1.077	1.083	1.098	1.078
Global Warming	1.357	1.361	1.358	1.358	1.358
Sources of Information					
Other Farmers	1.164	1.166	1.167	1.162	1.160
Non-farming Neighbors	1.020	1.040	1.024	1.019	1.019
Banks	1.118	1.136	1.145	1.118	1.118
Contractors	0.869	0.865	0.879	0.869	0.869
University / Extension	1.178	1.179	1.177	1.178	1.178
USDA	1.192	1.201	1.206	1.192	1.192
Other Government Org.	1.040	1.045	1.033	1.040	1.040
Off-Farm Income					
Farm Operator	1.158	1.180	1.162	1.158	1.158
Spouse	1.248	1.260	1.268	1.248	1.250
Education					
Farm Operator	1.608	1.643	1.620	1.609	1.613
Spouse	1.453	1.540	1.472	1.461	1.454
Total Animal Units	362	361	362	362	362

Table 4. Regression Results for Roundup Ready Corn

Variables	**No-Imputation**			**Multivariate Normal Imputation**				
	Coeff.	**Std. Err.**	**p-Value**	**Coeff.**	**Std. Err.**	**p-Value**	**DOF**	**% Inc. S.E.**
Age	0.008	0.010	0.390	0.023	0.006	0.000	3332	4%
Owned Land	-0.001	0.000	0.182	0.000	0.000	0.349	9474	2%
Land Rented Out	0.001	0.001	0.451	0.000	0.001	0.820	10896	2%
Land Rented In	0.001	0.000	0.424	0.001	0.000	0.088	3826	4%
State	-0.995	0.207	0.000	-0.640	0.148	0.000	8818	2%
Farm Sales	0.366	0.108	0.001	0.323	0.073	0.000	2903	4%
Non-Family Labor	-0.041	0.222	0.854	-0.154	0.165	0.351	14121	2%
Environmental Perceptions								
Water Quality	0.052	0.104	0.614	-0.052	0.069	0.457	4457	3%
Air Quality	-0.093	0.105	0.377	0.022	0.080	0.786	969	8%
Global Warming	-0.198	0.081	0.014	-0.126	0.056	0.023	5654	3%
Sources of Information								
Other Farmers	0.086	0.096	0.373	-0.077	0.070	0.271	2488	5%
Non-farming Neighbors	0.054	0.105	0.610	0.065	0.085	0.446	1105	7%
Banks	0.094	0.107	0.382	0.114	0.083	0.168	3378	4%
Contractors	0.088	0.143	0.537	-0.088	0.104	0.394	2373	5%
University / Extension	-0.205	0.119	0.085	-0.048	0.081	0.552	4986	3%
USDA	0.237	0.118	0.044	0.247	0.092	0.008	1986	5%
Other Government Org.	0.179	0.119	0.131	0.054	0.089	0.545	4506	3%
Off-Farm Income								
Farm Operator	0.038	0.114	0.739	0.008	0.081	0.920	3118	4%
Spouse	0.001	0.110	0.994	-0.018	0.084	0.826	444	12%
Education								
Farm Operator	0.004	0.078	0.956	-0.038	0.055	0.485	2051	5%
Spouse	0.074	0.083	0.378	0.021	0.069	0.763	355	14%
Total Animal Units	0.000	0.000	0.715	0.000	0.000	0.683	9528	2%
Constant	-1.910	0.877	0.029	-1.945	0.608	0.001	3592	4%

Table 5. Regression Results for Roundup Ready Corn

Variables	**No-Imputation**			**Univariate Imputation**				
	Coeff.	**Std. Err.**	**p-Value**	**Coeff.**	**Std. Err.**	**p-Value**	**DOF**	**% Inc. S.E.**
Age	0.008	0.010	0.390	0.010	0.008	0.233	300561	0%
Owned Land	-0.001	0.000	0.182	-0.001	0.000	0.195	13901	1%
Land Rented Out	0.001	0.001	0.451	0.002	0.001	0.211	1386	4%
Land Rented In	0.000	0.000	0.424	0.000	0.000	0.123	6991	2%
State	-0.995	0.207	0.000	-0.870	0.181	0.000	120161	0%
Farm Sales	0.366	0.108	0.001	0.303	0.090	0.001	83813	1%
Non-Family Labor	-0.041	0.222	0.854	0.012	0.192	0.950	59573	1%
Environmental Perceptions								
Water Quality	0.052	0.104	0.614	0.059	0.089	0.506	24990	1%
Air Quality	-0.093	0.105	0.377	-0.034	0.095	0.717	5623	2%
Global Warming	-0.198	0.081	0.014	-0.176	0.068	0.010	114456	0%
Sources of Information								
Other Farmers	0.086	0.096	0.373	0.020	0.086	0.812	34299	1%
Non-farming Neighbors	0.054	0.105	0.610	0.051	0.096	0.598	31997	1%
Banks	0.094	0.107	0.382	0.149	0.093	0.111	18213	1%
Contractors	0.088	0.143	0.537	-0.010	0.125	0.934	75602	1%
University / Extension	-0.205	0.119	0.085	-0.079	0.095	0.409	9467098	0%
USDA	0.237	0.118	0.044	0.232	0.103	0.024	989302	0%
Other Government Org.	0.179	0.119	0.131	0.091	0.102	0.372	27383	1%
Off-Farm Income								
Farm Operator	0.038	0.114	0.739	0.062	0.097	0.523	23185	1%
Spouse	0.001	0.110	0.994	-0.013	0.095	0.889	10483	2%
Education								
Farm Operator	0.004	0.078	0.956	-0.023	0.067	0.734	10164	2%
Spouse	0.074	0.083	0.378	0.051	0.074	0.493	8455	2%
Total Animal Units	0.000	0.000	0.715	0.000	0.000	0.884	32138	1%
Constant	-1.910	0.877	0.029	-2.009	0.739	0.007	228952	0%

Table 6. Impact of Missing Observations on Variable Estimates

	MVN Imputation			Univariate Imputation		
	RVI	**FMI**	**Rel. Eff.**	**RVI**	**FMI**	**Rel. Eff.**
Age	0.068	0.064	0.997	0.006	0.005	0.999
Owned Land	0.050	0.048	0.998	0.026	0.026	0.997
Land Rented Out	0.028	0.027	0.999	0.088	0.082	0.992
Land Rented In	0.130	0.117	0.994	0.037	0.036	0.996
State	0.042	0.040	0.998	0.009	0.009	0.999
Farm Sales	0.076	0.071	0.996	0.010	0.010	0.999
Non-Family Labor	0.047	0.045	0.998	0.012	0.012	0.999
Environmental Perceptions						
Water Quality	0.088	0.081	0.996	0.019	0.019	0.998
Air Quality	0.088	0.081	0.996	0.042	0.040	0.996
Global Warming	0.120	0.109	0.995	0.009	0.009	0.999
Sources of Information						
Other Farmers	0.119	0.107	0.995	0.016	0.016	0.998
Non-farming Neighbors	0.104	0.095	0.995	0.017	0.017	0.998
Banks	0.162	0.141	0.993	0.023	0.022	0.998
Contractors	0.106	0.096	0.995	0.011	0.011	0.999
University / Extension	0.076	0.072	0.996	0.001	0.001	0.999
USDA	0.106	0.097	0.995	0.003	0.003	0.999
Other Government Org.	0.071	0.066	0.997	0.018	0.018	0.998
Off-Farm Income						
Farm Operator	0.124	0.111	0.994	0.020	0.020	0.998
Spouse	0.244	0.200	0.990	0.030	0.029	0.997
Education						
Farm Operator	0.196	0.166	0.992	0.031	0.030	0.997
Spouse	0.358	0.269	0.987	0.034	0.033	0.997
Total Animal Units	0.023	0.023	0.999	0.017	0.017	0.998
Constant	0.088	0.081	0.996	0.006	0.006	0.999