Implications of Survey Sampling Design for Missing Data Imputation

Haluk Gedikoglu
Assistant Professor of Agricultural Economics
Cooperative Research Programs
Lincoln University of Missouri
Jefferson City, MO 65101
GedikogluH@lincolnu.edu

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 2013 AAEA & CAEA Joint Annual Meeting, Washington, DC, August 4-6, 2013

Abstract

Previous studies that analyzed multiple imputation using survey data did not take into account the survey sampling design. The objective of the current study is to analyze the impact of survey sampling design missing data imputation, using multivariate multiple imputation method. The results of the current study show that multiple imputation methods result in lower standard errors for regression analysis than the regression using only complete observation. Furthermore, the standard errors for all regression coefficients are found to be higher for multiple imputation with taking into account the survey sampling design than without taking into account the survey sampling design. Hence, sampling based estimation leads to more realistic standard errors.

Key Words: Multiple Imputation, Sampling Based Estimation, Missing Data

Although statistical theory has been developed for missing data imputation, the use of these methods has been relatively rare by agricultural economists (Robbins and White, 2011). Majority of the existing studies measured the impact of imputation techniques on the distribution of univariate missing variables using arbitrarily created missing data patterns (e.g., Robbins and White, 2011). Among the few studies that analyzed missing data imputation in agricultural economics, Robbins and White (2011), Ahearn et al (2011), and Moss and Mishra (2011) used ARMS data for their analysis on missing data imputation. These studies can be thought of simulation studies, as they generate missing data randomly from a complete dataset. Robbins and White (2011) analyze how the distribution of the variable "farm commodity payments received" changes between two imputation methods when some of the observations for this variable are randomly removed from the data. One of the imputation methods used their analysis is the method used by the USDA, which is the conditional mean imputation (a non-model-based estimation method). The second method is based on Data Augmentation (DA), which is a Markow Chain Monte Carlo method (Robbins and White, 2011). This paper applied DA to conduct a single imputation, rather than multiple imputation. Their results show that the method of imputation impacts the distribution of the variable imputed. As ARMS date set does not include missing observations, missing data for this study was created by random removal of observations from the dataset, this is referred to as Missing Completely at Random (MCAR). This is an important limitation of the Robbins and White (2011) research, because it is likely that there may be systematic reasons why data are missing. Hence, in general it is difficult to observe MCAR in actual survey data.

The study by Ahern et al. (2011) provides a comparison of the USDA's conditional mean imputation method with sequential regression multivariate imputation (SRMI), using the ARMS

dataset. SRMI is based on imputing missing observations for each variable separately, based on the observed distribution of each variable (Ahearn et al., 2011) [1]. Missing data for this study was also created by random removal of observations from the ARMS dataset. Their results showed that the method used by the USDA has mimicked the distribution of some variables more closely to the full dataset than the SRMI method, while the opposite is true for some other variables. Hence, no definite general conclusion was made on which imputation method is preferable. Although the studies by Robbins and White (2011) and Ahearn et al (2011) analyzed the impact of missing data imputation on certain variables, neither of these studies analyzed the impact of missing data imputation on regression coefficient estimates. Lastly, the study by Moss and Mishra (2011) applies Gibbs Sampling (a MCMC method), which is different than the multiple imputation method developed by Rubin (1987), to estimate a Leontief production function using the ARMS data, using also synthetically generated missing data. They find that results using imputed data and results using only complete observations did not differ significantly for regression coefficient estimates. However, multiple imputation resulted in higher standard errors than using only complete observations, which is unexpected as multiple imputation should decrease the standard errors by using more observations. Authors conclude that this unexpected result was due to the collinearity problem caused by the arbitrarily created missing data pattern. One of the major limitations of the studies reviewed was generating missing data randomly from a complete dataset and not accounting for survey sampling design. The objective of this paper is to evaluate the impact of survey sampling design on multiple imputation.

Multiple Imputation

Multiple imputation methods, both multivariate and univariate, are based on simulation from a Bayesian posterior predictive distribution of missing data (Rubin, 1987: Schafer, 1997).

[1] See Little and Rubin (2002) for a detailed review of each imputation method.

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 univariate logistic model is used to represent the adoption of soil testing. Lastly, during the pooling step, the results obtained from M completed-data analyses are combined into a single multiple imputation based estimation results. Below we provide the detailed description for each step of multiple imputation.

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 and ordinal (discrete variables), and continuous. Although multivariate normal imputation is originally developed for imputing continuous variables, studies show that a multivariate normal model can be used for discrete variables, given that imputed observations 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 others are converted to 1.

Multivariate Normal Multiple Regression

The basic multivariate normal regression model for imputing missing variables can be represented as follows. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ be a random sample from a *p*-variate normal distribution, where *p* represents the number of variables with missing values for observation *i* = 1,…,*N*. The multivariate normal regression can be represented as:

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

where $\mathbf{x}_i = (\mathrm{x}_{i1}, \mathrm{x}_{i2}, \ldots, \mathrm{x}_{ip})$ is the vector of values of the variables to be imputed for observation *i*, $\boldsymbol{z}_i$ is a *q* x 1 vector of values of the complete (independent) variables for observation *i*, $\boldsymbol{\Theta}$ 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 (positive definite) matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ are referred as the model parameters. Next we provide the information on data augmentation.

Data Augmentation

Multivariate normal multiple regression model uses Data Augmentation, which is an iterative MCMC method, to impute missing values (Rubin, 1987). 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 $\mathbf{x}_i = (\boldsymbol{x}_{i(0)}, \boldsymbol{x}_{i(m)})$ 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 $\boldsymbol{\Theta}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$ 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 $\boldsymbol{\Theta}^{(t+1)}$ and $\boldsymbol{\Sigma}^{(t+1)}$ 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: $\mathbf{x}_{i(m)}^{(t+1)} \sim P\left(\boldsymbol{x}_{i(m)} \middle| \boldsymbol{z}_i, \boldsymbol{x}_{i(0)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Sigma}^{(t)}\right), i = 1, \ldots, N$

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

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

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

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

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 Bayes rule $P(\boldsymbol{\Theta}, \boldsymbol{\Sigma} | \boldsymbol{X}) \propto L(\boldsymbol{\Theta}, \boldsymbol{\Sigma} | \boldsymbol{X}) f(\boldsymbol{\Theta}, \boldsymbol{\Sigma})$, where $L(\boldsymbol{\Theta}, \boldsymbol{\Sigma} | \boldsymbol{X})$ is the standard multivariate normal likelihood function, the I and P steps become (Schafer, 1997):

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

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

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

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 $\boldsymbol{\Sigma}_{mm.o}^{(t)}$ are the mean and variance of the conditional distribution of $\boldsymbol{x}_{i(m)}$given $\boldsymbol{x}_{i(0)}$ based on $\mathbf{x}_i \sim N_p(\boldsymbol{\Theta}^{(t)'}\boldsymbol{z}_i, \boldsymbol{\Sigma}^{(t)})$. The matrix $\widehat{\boldsymbol{\Theta}}^{(t+1)} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}^{(t+1)}$ 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 $\Lambda_*^{(t+1)}$ and the posterior degrees of freedom for the inverted Wishart distribution λ_* are defined as (Johnson and Wichern, 2002):

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

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

Values for the degrees of freedom and the scale parameter are determined based on the requested prior distribution for $\boldsymbol{\Theta}$. For the uniform prior distribution for $\boldsymbol{\Theta}$, the values are λ = -(*p*+1) and $\Lambda^{-1} = \mathbf{0}_{pp}$, where $\mathbf{0}_{pp}$ is a zero matrix (Johnson and Wichern, 2002). In the current study, to reflect uncertainty about model parameters, noninformative uniform prior distribution is used.

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):

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

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):

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

The E step and M steps are processed 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 $\boldsymbol{\Sigma}$ correspond 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 $\mathrm{E}(\sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{x}'_i)$ and $\mathrm{E}(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}'_i)$ are computed with respect to the conditional distribution $P\left(\mathbf{X}_m \middle| \boldsymbol{\Theta}^{(\mathrm{t})}, \boldsymbol{\Sigma}^{(\mathrm{t})}, \boldsymbol{X}_0\right)$ (Little and Rubin, 2002):

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

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 $\boldsymbol{\Theta}'_{i(\mathrm{m})}\boldsymbol{z}_i + \boldsymbol{\Sigma}_{i(\mathrm{mo})}\boldsymbol{\Sigma}^{-1}_{i(\mathrm{oo})}(\mathbf{x}_{i(o)} - \boldsymbol{\Theta}'_{(\mathrm{s})}\mathbf{z}_i)$, and c_{ij} is the element of the matrix $\boldsymbol{\Sigma}_{i(\mathrm{mm})} - \boldsymbol{\Sigma}_{i(\mathrm{mo})}\boldsymbol{\Sigma}^{-1}_{i(\mathrm{oo})}\,\boldsymbol{\Sigma}'_{i(\mathrm{oo})}$ (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:

$$\boldsymbol{\Theta}^{(t+1)} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\mathrm{E}(\sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{x}'_i)$$

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

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).

Completed-data Analysis Step

Promoting adoption of new technologies is an important policy issue in agricultural economics. In the current study we analyze adoption of soil testing. The adoption decision of farmers can be represented using a logistic regression as (Greene, 2008):

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

where $y_i = 1$ if the farmer adopts soil testing and $y_i = 0$ if the farmer does not adopt soil testing. $\boldsymbol{q}$ is the vector of coefficients in interest, in the completed-data analysis, to be estimated and $\boldsymbol{x}_i$ is the vector of independent variables. This model is performed separately on each set of imputed data (completed data) $m = 1,\ldots,M$ (Gelman et al, 2004).

Pooling Step

The results obtained from *M* completed-data analyses are combined into a single multiple-imputation based estimation results (Enders, 2010). Let $\{(\widehat{\boldsymbol{q}_i}, \widehat{\mathbf{U}_i}): i = 1,2,\ldots,M\}$ be the completed-data estimates of **q** and the respective variance-covariance estimates **U** from *M* imputed datasets (Enders, 2010). The multiple imputation estimate of **q** is $\overline{\boldsymbol{q}}_M = \frac{1}{M}\sum_{i=1}^{M} \widehat{\boldsymbol{q}_i}$. 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 $\overline{\mathbf{U}} = \frac{1}{M}\sum_{i=1}^{M} \widehat{\boldsymbol{U}_i}/M$ is the within-imputation variance-covariance matrix and $\mathbf{B} = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{q}_i - \overline{\boldsymbol{q}}_M)(\boldsymbol{q}_i - \overline{\boldsymbol{q}}_M)'/(M-1)$ is the between-imputation variance-covariance matrix (Enders, 2010).

Sampling Based Estimation

In sample surveys, observations are selected through a random process, but different observations may have different probabilities of selection. Weights are equal to (or proportional to) the inverse of the probability of being sampled. Various postsampling adjustments to the weights are sometimes made, as well. A weight of *wj* for the *j*th observation means, roughly speaking, that the *j*th observation represents wj elements in the population from which the sample was drawn. Omitting weights from the analysis results in estimates that may be biased, sometimes seriously so.

Results

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 discover if the farmers had adopted new technologies and how the farmers' 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 who had stopped farming, farmers who had returned the survey due to not being the farm operator, and undeliverable surveys to farmers (due to an address change) 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 the comparison of logistic regression results between the no imputation case and MVN multiple-imputation with *M* 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 at 10 percent significance level in the multiple-imputation case (e.g., age and owned land). It is important to see that almost all of the variable estimates have *lower* standard error in the MVN imputed regression than in the no-imputation regression. Hence, MVN imputation significantly increased the efficiency of the estimates.

Table 2 provides the comparison of logistic regression results between the MVN imputation (non-sampling based) and MVN imputation (sampling based). The results show that MVN imputation (sampling based) leads to larger standard errors than MVN imputation (non-sampling based) for the variables other than farm sales. This is expected, as stratification was done based on farm sales, which leads to lower standard errors. Hence, overall sampling based MVN imputation leads to more realistic standard errors.

Conclusion

Multiple imputation results have important policy implications. Policy makers can end up enforcing different policies based on whether they used no-imputation regression or the multiple imputation regressions. This is because multiple imputation based regression and the regression using only complete observations can have differences in the sign, magnitude, and statistical significance of coefficient estimates. Since multiple-imputation methods provides unbiased estimates and increase the efficiency of the regression estimates, it is recommended that policy recommendation should be made using multiple imputation methods rather than using a regression with missing observations.

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., and D. B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science*,7: 457–472.

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

Graham, J. W. 2009. "Missing data analysis: Making it Work in the Real World." *Annual Review of Psychology*, 60: 549–576.

Graham, J.W., S.M. Hofer, and A.M. Piccinin. 1994. Analysis with Missing Data in Drug Prevention Research. *Advances in Data Analysis for Prevention Intervention Research*, National Institute on Drug Abuse.

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

Horton, N. J., and K. P. Kleinman. 2007. "Much Ado about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *American Statistician*, 61: 79–90.

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.

Moss, C., and A. K. Mishra. 2011. "Imputing Missing Information in the Estimation of Production Functions and Systems." *American Journal of Agricultural Economics*, 93(2): 619-626.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger. 2001. “A Multivariate Technique for Multiply Imputing Missing Values Using a sequence of regression Models.” *Survey Methodology*, 27: 85–95.

Reiter, J. P., and T. E. Raghunathan. 2007. “The Multiple adaptations of Multiple Imputation.” *Journal of the American Statistical Association*, 102: 1462–1471.

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.

Schafer, J. L., and J. W. Graham. 2002. “Missing data: Our View of the State of the Art.” *Psychological Methods*, 7:147–177.

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. Regression Results for Adoption of Soil Testing

Variables	No-Imputation			Multivariate Normal Imputation				
	Coeff.	Std. Err.	p-Value	Coeff.	Std. Err.	p-Value	DOF	Inc. S.E. (%)
Age	1.001	0.015	0.948	0.024	0.010	0.021	12358	1.38
Owned Land	1.001	0.001	0.174	0.001	0.001	0.056	639	6.52
Land Rented Out	0.999	0.002	0.819	-0.001	0.001	0.302	4247	2.38
Land Rented In	1.003	0.001	0.005	0.002	0.001	0.032	243	11.29
Missouri (Base=Iowa)	0.319	0.118	0.002	-1.037	0.261	0.000	4210	2.4
Non-Family Labor	1.301	0.494	0.488	-0.260	0.294	0.378	1749	3.79
Environmental Perceptions								
Water Quality	0.746	0.133	0.100	-0.250	0.129	0.053	429	8.13
Managing Manure	1.130	0.209	0.510	0.226	0.151	0.135	140	15.77
Global Warming	0.868	0.111	0.271	-0.135	0.091	0.138	24002	0.98
Farm Sales								
$50,000-$99,999	3.586	1.630	0.005	0.955	0.329	0.004	4274	2.38
$100,000-$249,999	7.554	4.030	0.000	1.368	0.374	0.000	3419	2.67
$250,000-$499,999	16.078	12.982	0.001	2.169	0.569	0.000	653	6.44
$500,000 or more	9.137	11.341	0.075	2.263	0.964	0.019	576	6.91

Table 2. Regression Results for Adoption of Soil Testing

Variables	Multivariate Multiple Imputation					Multivariate Multiple Imputation (Sampling Based)				
	Coeff.	Std. Err.	p-Valu	DOF	Inc. S.E. (%)	Coeff.	Std. Err.	p-Value	DOF	Inc. S.E.(%)
Age	0.024	0.010	0.021	12358	1.38	0.014	0.013	0.276	374	2.63
Owned Land	0.001	0.001	0.056	639	6.52	0.002	0.001	0.094	88	17.32
Land Rented Out	-0.001	0.001	0.302	4247	2.38	-0.004	0.002	0.050	155	10.54
Land Rented In	0.002	0.001	0.032	243	11.29	0.002	0.001	0.073	56	24.95
Missouri (Base=Iowa)	-1.037	0.261	0.000	4210	2.4	-1.192	0.351	0.001	410	1.51
Non-Family Labor	-0.260	0.294	0.378	1749	3.79	-0.019	0.400	0.962	325	3.95
Environmental Perceptions										
Water Quality	-0.250	0.129	0.053	429	8.13	-0.219	0.152	0.149	365	2.86
Managing Manure	0.226	0.151	0.135	140	15.77	0.159	0.183	0.385	95	16.21
Global Warming	-0.135	0.091	0.138	24002	0.98	-0.102	0.128	0.428	399	1.89
Farm Sales										
$50,000-$99,999	0.955	0.329	0.004	4274	2.38	1.156	0.485	0.018	431	0.64
$100,000-$249,999	1.368	0.374	0.000	3419	2.67	1.720	0.547	0.002	401	1.81
$250,000-$499,999	2.169	0.569	0.000	653	6.44	2.541	0.929	0.007	291	4.93
$500,000 or more	2.263	0.964	0.019	576	6.91	1.921	1.468	0.192	336	3.67