Mass Imputation of Agricultural Economic Data
Missing by Design
A Simulation Study of Two Regression Based Techniques

Matt Fetter
National Agricultural Statistics Service (NASS), U.S. Department of Agriculture

Abstract

The demand for information concerning all facets of the production of agricultural commodities is constantly increasing. This demand is placing significant burden on the relatively few large and mid-sized producers that account for a disproportionally large percentage of all agricultural production. In particular, the demand for economic data from farm operations is especially intrusive of the producer in terms of time and sensitivity. One way to reduce this burden is to simply collect significantly less data and then try to regain some of the lost precision by modeling the now unobserved data using data that is observed. This paper evaluates some of the characteristics of data sets that are incomplete by design and are "completed" via imputations obtained from two regression based imputation methods. Estimates of population means and correlations are evaluated for a set of 27 economic variables that have a fixed missing rate of 60 percent. Some standard error estimates are obtained for one of the methods and these are evaluated as well. Gains in RMSE can be made for many variables and correlations can be preserved for many pairs of incompletely observed variables.

Keywords: MCMC, Multiple Imputation, Mass Imputation, Missing By Design, Regression

The Sample Design and Imputation Strategy

The strategy being considered in this research takes the direct approach to respondent burden reduction–simply collect less data and extract as much information as possible out of it. In an effort to reduce respondent burden and increase response rates on its annual economic oriented Cost and Returns Report (CRR) survey, NASS is considering the use of a rotating 3 year Panel sample supplemented by an additional sample to serve for a single year. A complete set of data would be collected for all variables of interest from the Panel sample. These records will be referred to as the "completely observed" set of records. Data for a much smaller subset of these variables would be collected from a supplemental sample.

	Variables						
	1F	2F	3F	4P	5P	6P	7P
Panel Records (Completely Observed)	1	1	1	1	1	1	1
	1	1	1	1	1	1	1
	1	1	1	?	?	?	?
Supplemental Records (Partially Observed)	1	1	1	?	?	?	?
	1	1	1	?	?	?	?

Figure 1.

These records will be referred to as the "partially observed records". (In figure 1, variables 1 through 3 are **F**ully observed, variables 4 through 7 are **P**artially observed.). The idea is to use the observed relationships among the complete set of variables from the panel sample (completely observed) to develop models that could be used as an aid to impute values for unobserved variables in the supplemental sample (partially observed). In this research, all variables subject to imputation had a fixed missing rate of 60%. Therefore, mean and correlation estimates for partially observed variables are based on data values that are 40% observed and 60% imputed.

The completed data set is expected to provide not only point estimates, but also will be subjected to detailed analysis. Though there are simpler ways to account for nonresponse such as weighting adjustments and various forms of mean imputation, what is desired in this situation is a complete data set that will provide a basis for a consistent analysis requiring the underlying data structure to be preserved to the fullest extent possible.

Introducing the Imputation Methods

The performance of two regression based imputation methods were evaluated in this research, Markov Chain Monte Carlo Multiple Imputation (MCMI), and another method that will be referred to as the Regression with Empirical Residual (RER) method. The two methods are similar in some ways, but have important differences. Both are regression based and are using the same set of predictor variables. MCMI is a multiple imputation procedure whereas RER uses only a single imputation.

In the case of RER , no parametric distribution for the error terms is ever specified. By contrast, the MCMI method assumes that the error terms follow a conditional multivariate normal distribution. The estimated mean vector and covariance matrix of this distribution is not obtained directly from the data however, but is randomly selected from a joint posterior distribution for these parameters. By randomly selecting the mean and covariance in this manner, some *additional* uncertainty concerning the correct value to impute is injected into the variability of the MCMI imputations.

These procedures were evaluated by studying the observed simulation standard errors and the observed simulation biases of the estimates of means and correlations. A 1,500 iteration simulation was performed using the MCMI multiple imputation and RER methods. Distributional characteristics of the imputed data were noted. Standard error estimates for the MCMI method were obtained using standard multiple imputation procedures and evaluated. At the time of this writing, efforts in achieving good standard error estimates for the RER method have resulted in very limited success. Therefore, no standard error estimates will be presented here for the RER method. More work on this problem will be required. Nonetheless, it is interesting to compare the characteristics of imputations obtained using RER to those obtained using MCMI even though the "complete" RER package has not yet been produced.

Some Wrinkles

In their observed form, the imputed variables are highly skewed and are often semi-continuous in the sense that it is much more likely to observe a zero value than any particular positive value.

Through various transformations, the positive values for many of these variables can be transformed so that the resulting marginal distributions are nearly normal.

Imputation Step 1- Should it be Zero or Positive ?

Imputations for both methods were obtained using a two step procedure. First, logistic regressions were performed to aid in determining whether to impute a positive value or a zero. If it was determined that a positive value should be imputed, then the MCMI and RER methods were used to produce it in the second step.

Both MCMI and RER used the same method to determine whether a zero or a positive value should be imputed. Logistic regression models were developed using the completely observed records. The estimated coefficients were then applied to the partially observed records and an estimate of the probability of a positive response for the kth unobserved variable on record j, $\hat{P}_{jk}$ was computed.

To determine whether a zero or a positive value would be imputed for the kth unobserved variable for partially observed record j, a uniform (0,1) random number, denoted q_{jk} was generated. If $q_{jk} <= \hat{P}_{jk}$, then a positive value was imputed, otherwise a zero was imputed. For MCMI, this procedure was done 5 times, once for each of the 5 multiple imputations. This permitted a zero to be imputed for a given variable on a record in one imputation, but a positive value to be imputed for the same record and variable in another.

Imputation Step 2. MCMI-Obtaining the Positive Value

MCMI – Some Light Theory:

For the data matrix, some of the data are observed and some of the data are not observed (see figure 1). Consider a typical incomplete data vector of the transformed responses Z for any given record. Let Z be partitioned into two sub-vectors, Z_{obs} the vector of observed variables, and Z_{miss} , the vector of variables not observed (missing). The distribution of Z, P(Z), will be expressed as $P(Z)=P(Z_{obs},Z_{miss})$. Further, let's assume that $P(Z_{miss}|Z_{obs})$ are iid multinormal vectors with conditional mean vector μ and conditional covariance matrix, Σ . Let $\theta = (\mu, \Sigma)$ represent the unknown parameters from the population of interest. To create imputations, it is desired to draw Z_{miss} from $P(Z_{miss} \mid Z_{obs})$. In order to create imputations that properly reflect the uncertainty concerning θ, a prior distribution for θ is assumed. The imputations are then created in such a way as to be obtained by independent draws from the conditional predictive distribution of Z_{miss} given θ, averaged over the observed data posterior of θ (Schafer, 1977, p. 105):

$$P(Z_{miss}|Z_{obs})=\int P(Z_{miss}|Z_{obs},\theta)P(\theta|Z_{obs})d\theta \quad .$$

In general, this is achieved through the use of a procedure which begins by setting an initial temporary estimate of θ, say at $\theta^{(0)}$, and then alternately applying the following two relations:

$$Z_{miss}^{(t)} \sim P(Z_{miss}|Z_{obs},\theta^{(t-1)}) \text{ and}$$

$$\theta^{(t)} \sim P(\theta|Z_{obs},Z_{miss}^{(t)}) \text{ , for iteration t= 1,2,....}$$

These iterations produce a Markov chain that for sufficiently large t converges in distribution to $P(Z_{miss}|Z_{obs})$.

The observed values of the random vector Z_{miss} at any two iterations are correlated, with the strength of the correlation declining as the number of iterations between the two observed vectors increases. To obtain "proper" imputations, the vectors used for the imputations must be *independent* random draws from $P(Z_{miss} \mid Z_{obs})$. For sufficiently large s, draws at iteration t and iteration (t + s) will be approximately independent. See Schafer, 1997 for a thorough discussion of this method.

Many of the completely observed records contained zero values for the variables in the partially observed variable set. This created problems if one needed to assume that these variables were normally distributed. Even with the transformations, often a large spike at zero was present. The idea was to assume that only the positive values were normally distributed and set the observed zero values to missing. This way, the only completely observed records available to estimate the coefficients used for imputing values for a particular partially observed variable were those with a positive value for that variable. The MCMI procedure itself was only used for the imputation of positive values, whereas zeroes were imputed via the logistic regression procedure.

The immediate consequence of setting these observed zeroes to missing before running the MCMI algorithm was to artificially increase the amount of missing information apparent in the data set. Available information was temporarily discarded for the sake of preserving the integrity of the model for the positive values. This was particularly a problem for variables with a high proportion of zero values.

Another unfortunate consequence of doing this was that the *monotonic* missingness pattern induced by the sample design was disrupted. A monotonic missingness pattern exists if the rows and columns of the data matrix can be arranged so that for every observation, variable p observed implies that variables p-1,p-2,..,1 are also observed (see figure 1). This is important because if the missingness pattern is monotonic, the MCMI algorithm would converge almost immediately. However, by setting the zero values to missing, the resulting non-monotonic missingness pattern required MCMI to go through many more iterations before convergence was achieved.

Multiple Imputation Point Estimates and Standard Error Estimates.

Rubin (1987) outlines how the m estimates obtained from each of the m imputed data sets are combined to obtain the complete data point estimates and standard error estimates. For point estimates, let q_i, where i=1,..,m equal the estimate of the population parameter, Q, obtained from imputation i. In the application of the MCMI method, m= 5 so five estimates of Q are obtained, one from each imputation.

Then $\hat{Q}=\frac{1}{m}\sum_{i=1}^{m} q_i$ is the multiple imputation point estimate of Q.

For estimates of the standard error, let v_i, i=1,..,m be the *within* imputation variance estimate obtained for imputation i, where v_i is calculated in the usual manner, once for each of the 5 imputations.

Then, $\bar{v}=\frac{1}{m}\sum_{i=1}^{m} v_i$ is the overall within imputation variance estimate. The *between* imputation variance estimate, b, is defined as: $b=\frac{1}{m-1}\sum_{i=1}^{m}(q_i-\bar{Q})^2$.

The estimate of total variance is then defined as: $\hat{V}=\bar{v}+(1+m^{-1})b$.

Imputation Step 2. RER-Obtaining the Positive Value

Unlike MCMI, the RER method is non-iterative and uses randomly selected empirical errors to create the imputations. Again, let $Z=(Z_{obs}, Z_{miss})$ represent the typical transformed incomplete data vector.

In the notation that follows, the subscript i will be associated with completely observed records, the subscript j will be associated with partially observed records. The subscript k will be associated with partially observed variable Z_k, k=1,..., 27.

Define $Z_{miss}=(Z_1, ..., Z_k)$. Define $G_j=(g_{j1}, ..., g_{jk})$ as a 1 x k vector of 1's and 0's resulting from the logistic regression procedure described earlier for partially observed record j. If $g_{jk} = 0$ for the kth unobserved variable for record j, then 0 will be the final imputed value for Z_{jk}. If $g_{jk} = 1$, then obtain a value to impute for Z_{jk} as follows:

Let C_k^{pos} represent the set of all completely observed sampled records for which the variable $Z_k > 0$. First estimate the regression coefficients in the model:

$$z_{ik}=\beta_{0k}+\beta_{1k}x_{i1}+....+\beta_{rk}x_{ir}+\varepsilon_{ik}, \forall i \in C_k^{pos} \qquad (1)$$

using all samples in C_k^{pos}. Define $\hat{z}_{ik}$ to be the resulting fitted value for z_{ik}, obtained from applying the fitted model to each of the completely observed records in C_k^{pos}. Let $e_{ik}=z_{ik}-\hat{z}_{ik}$ represent the resulting error for record $i \in C_k^{pos}$.

Now let M_k^{pos} represent the set of all partially observed records for which $g_{jk}=1$ (records for which a positive imputed value is desired for variable Z_k). For each of the records in M_k^{pos}, compute the predicted expected value $\hat{z}_{jk}$ using the estimated coefficients from (1).

$$\hat{z}_{jk}=\hat{\beta}_{0k}+\hat{\beta}_{1k}x_{j1}+__+\hat{\beta}_{rk}x_{jr}, \forall j \in M_k^{pos} \qquad (2)$$

To obtain the value to impute for Z_{jk}, simply use simple random sampling with replacement to select an e_{ik} and add it to $\hat{z}_{jk}$. Finally, the actual value to impute for variable Z_{jk} for each record $j \in M_k^{pos}$ is computed as:

$$\hat{z}_{jk\_imp}=\max((\hat{z}_{jk}+e_{ik}),0) \forall j \in M_k^{pos}, g_{jk}=1 \qquad (3)$$

$$\hat{z}_{jk\_imp}=0, g_{jk}=0. \qquad (4)$$

In (3) , adding the error term to $\hat{z}_{jk}$ can occasionally result in a negative value. Since all variables being imputed for are non-negative, the minimum permissible imputed value is zero.

The Simulation Setup

An artificial population was assembled using reported data obtained on the 1997 Cost and Return Report (CRR) administrated by NASS. Farms with estimated sales of $500,000 and up from 7 Midwestern states were pooled together and replicated to produce an artificial population of about 6500 records. For each iteration of the simulations (the RER and MCMI simulations were run independently of each other), a random sample was selected without replacement of size 400. From these 400 samples, 240 were again randomly selected to serve as the partially observed records. The remaining 160 samples would serve as the completely observed records and would form the basis for estimating the model parameters. The values for the 27 variables in the partially observed set were then deleted. Imputed values were then obtained for each variable in this set by the MCMI and RER methods. The effect of these imputed values on the quality of estimates of the population means and correlations obtained from the artificially complete data set of 400 records was then assessed.

The fully observed (predictor) variables were largely "profile" variables. Profile variables are variables that give a description of some basic aspect of the farm such as total acres operated, acres of harvested crop land, head of hogs, head of cattle, number of landlords, geographic location, etc. A few of the predictor variables were items that would be required to complete the IRS schedule F such as livestock feed expenses. A few variables were used as predictors because of their usefulness in predicting some variable that might be particularly burdensome for the respondent to report (interest paid as a predictor of debt for example).

The variables chosen to be partially observed were not necessarily chosen for any particular reason. In the early stages of this research, partially observed variables were chosen to be variables that had a relatively low proportion of zero responses. As the research progressed more variables were added to the mix without much regard to the proportion of zero responses.

Results

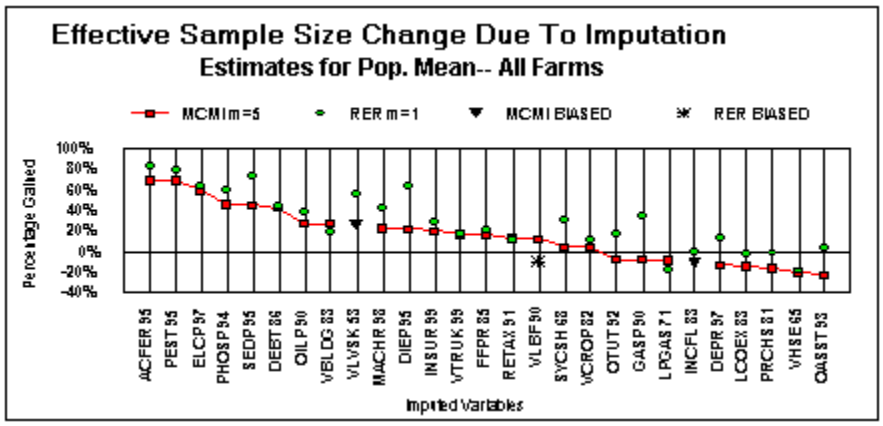

Figure 2. (The number directly above the variable name indicates its percentage of positive values in the population.)

The "Effective Sample Size Change" (expressed as a percent in *figure 2)* shows how using the imputation completed data sets (n=400) to estimate the population mean compares with using just the n=160 fully observed records (incomplete cases are deleted) to estimate the population mean for the 27 partially observed variables. This percentage (ESSC) is computed as:

$$ESSC = (100) * \{[MSE(\text{incomplete case deletion}) / MSE(\text{imputed})] - 1\}.$$

The ESSC is an ad-hoc evaluation method that treats any bias as if it were an additional part of the total variance and therefore could be reduced by increasing the sample size.

If the imputation gained nothing in terms of precision, then the ESSC is 0. A value of ESSC near 160% would indicate that the imputed values are about as good as the real values. Negative values indicate that the imputations generated only bias and noise, resulting in less precise estimates than would have been obtained if only the 160 fully observed records were used to create the estimates.

A quick glance at *figure 2* will reveal that for some variables, both methods yield substantial gains in effective sample size, but also, a substantial loss for a few variables. The RER method dominates the MCMI method for nearly every variable in terms of increase in effective sample size.

In terms of bias, neither method appears to have any serious bias problems for estimates of means for the entire population. Observed absolute biases in the simulations were less than 5% of the true population mean for all but one or two variables.

Average Correlations for all pairs of imputed variables.

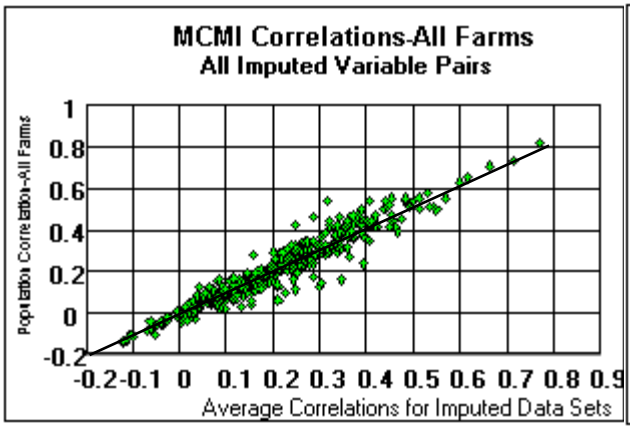

Figure 3 a.

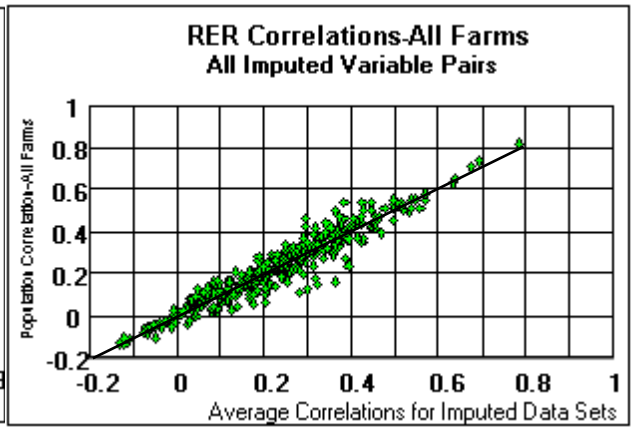

Figure 3 b.

On average, both methods perform about equally well in preserving the correlations of variables at the entire population level (see *figure 3a and figure 3b*). There is, however, clear evidence of at least some bias in correlation estimates for many of the variable pairs.

Some Comparisons of Marginal Distributions

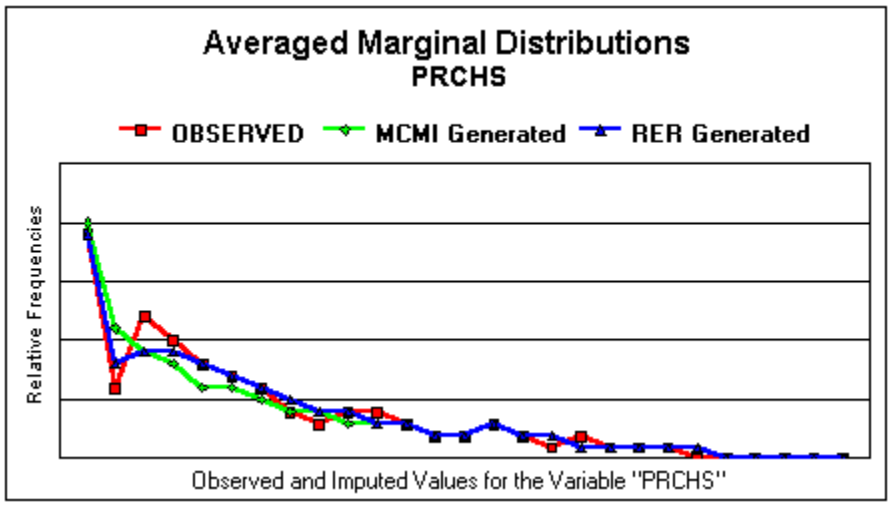

Figure 4

Figure 4 shows an example of how the pure imputed data distributions obtained by MCMI and RER compare to the real data distribution for the variable "PRCHS". Although the pure imputed data distributions do not exactly match that of the real data, the general shapes are similar. Note that neither method was particularly efficient at imputing for this variable (see figure 2).

Performance of Multiple Imputation RMSE Estimation

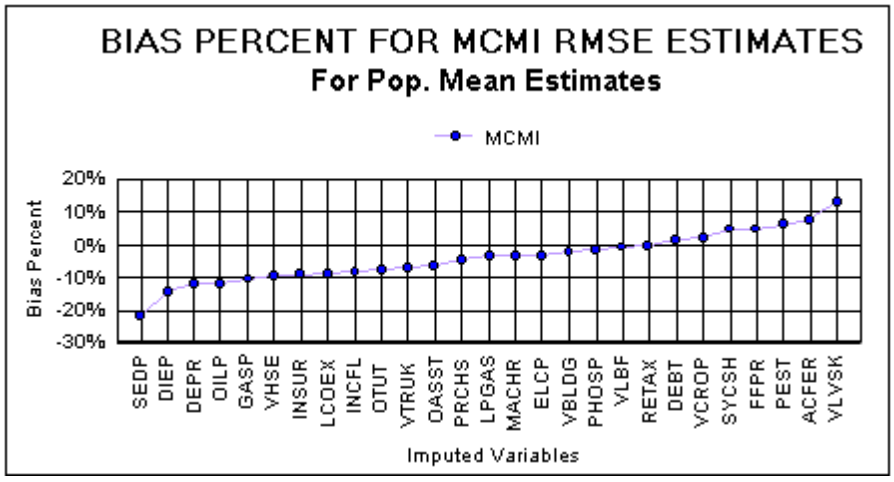

Figure 5.

Figure 5 shows how well the multiple imputation standard error estimates perform as estimators of the root mean squared error (rmse) observed in the simulation for the MCMI method. The performance varies from rather severe underestimation for a few variables to being somewhat conservative for other variables. The average underestimation across all 27 variables was about 4%.

Conclusions

Modeling does appear to make large gains in precision for estimates of means for some of the

variables under study. However, finding a model that can provide efficient imputations for a large number of variables might prove to be a difficult task. Even for some of the relatively few variables studied here, there was little gain and sometimes even a loss in the precision of estimates of population means. Additionally, neither the MCMI nor the RER method assures that the imputations are consistent within a record, and the amount and full effect of the editing that will be required after the imputations are made is not yet known.

The RER method provides better precision for estimates of population means than are obtained using the MCMI method, although the practical difference in precision might not be deemed substantial for many of the variables. The current absence of a good standard error estimator might be seen as a drawback to using the RER method, but continued efforts will be made in this area. The MCMI method does have a standard error estimation procedure that appears to work fairly well as an estimator of the rmse for most variables. Both methods appear to have very similar characteristics regarding correlation estimation.

Certainly there is no substitute for quality reported data. However, NASS is experiencing a steady decrease in CRR response rates, which arguably is attributed largely to the lengthy questionnaire. In effect, the decrease in response rates steadily increases the amount of nonresponse modeling (currently being done by weight adjustments) required. It is not whole-heartedly recommended that one should go ahead and collect significantly less data and then model what is not collected. However, everything is relative. If response rates can be increased significantly by substantially reducing the length of the questionnaire for a large portion of the sample, the overall quality of the imputed data might be considered acceptable.

References

Ford, B.L., Hocking, R.R., and Coleman, A. (1977) "Reducing Respondent Burden on an Agricultural Survey" USDA/NASS Report SF&SRB77-11.

Little, R.J.A. and Rubin, D.B.(1987) *Statistical Analysis with Missing Data*. New York: Wiley.

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

Santos, R.L., (1981) "Effects of Imputation on Complex Statistics" Income Survey Development Program-Survey Development Research Center in Nonresponse and Imputation-Report on Additional Task 2. Ann Arbor: Survey Research Center, University of Michigan.

Sarndal, C-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schafer, J.L.(1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Yuan, Y.C. (2000) "Multiple Imputation for Missing Data: Concepts and New Developments" SAS Institute Technical Report P267-25.