DEPARTMENT OF ECONOMICS AND FINANCE WORKING PAPER SERIES • *July 2005*

Financial Well-Being in an Urban Setting: An Application of Multiple Imputation

David A. Penn *
Middle Tennessee State University, Murfreesboro, TN

Abstract

Many studies delete incomplete data prior to model estimation, resulting in less efficient and potentially biased parameter estimates. Multiple imputation provides a model-based method of simultaneously estimating missing values for several variables, conditioned on the observed values. The technique is applied to financial well-being data collected by survey from householders in Oklahoma County, Oklahoma. Ordered logistic models are estimated for both complete cases and multiply imputed data. Estimates from the complete case model are somewhat biased and less efficient compared with the multiple imputation model.

Key words: Missing data, multiple imputation

JEL category: C15

*David A. Penn, Associate Professor Department of Economics and Finance, Middle Tennessee State University, Murfreesboro, TN 37132, phone: 615-904-8571, fax: 615-898-5041, email: dpenn@mtsu.edu.

Financial Well-Being in a Urban Setting: An Application of Multiple Imputation

I. Introduction

Recent studies of subjective well-being (SWB) rely heavily on survey data, but none adequately addresses an important issue: how best to deal with missing survey information? Missing information of concern in this paper consists of item non-response, defined as a refusal to respond or simply a lack of response to a particular survey question. Some householders are very reluctant to respond to survey questions, particularly regarding age and income. The proportion of householders who decline to answer questions on income can be significant, with item non-response ranging from 15 percent to more than 20 percent of the completed sample (Gronhaug 1988, Bell 1983). The manner in which researchers choose to deal with missing information can significantly affect parameter estimates and standard errors (Schafer 1997). Simply deleting incomplete observations is acceptable in some, but not all, circumstances. Estimating, or imputing, the missing information may be a more methodologically sound approach. This study applies an approach for estimating missing values that has become relatively well developed in the statistics and public health literature[1] but has received little attention from economists.

Research on the determinants of happiness and subjective well-being is becoming more prevalent in the economics literature.[2] For example, an entire recent issue of the

[1] Raghunathan (2004) is a recent example in the public health literature.
[2] Easterlin (2001) offers an introduction to this literature.

Journal of Economic Behavior and Organization[3] is devoted to the topic, and a 1997 issue of the *Economic Journal* contains three relevant articles. In a more recent contribution, Bukenya, Gebremedhim, and Shaeffer (2003) model self-reported quality of life in West Virginia using data collected from households by mailed questionnaire. However, the researchers offer no discussion of the prevalence of non-response for items such as age or income or how missing information is dealt with. Van Praag, Frijters, and Ferrer-I-Carbonell (2001) use data from the German Socio-Economic Panel to estimate various domains for subjective well-being. A large number of observations (19,000) are used in the study, but the authors do not discuss the prevalence of missing information or what is done with observations that have missing information. McBride (2001) analyzes subjective well-being as function of relative income and demographic variables using data collected by the U.S. General Social Survey. Starting with more than 2,000 observations, he excludes households with incomes greater than $75,000 and eliminates dozens of observations because of missing education, health status, marital status, well-being, or parent's well-being. After the exclusions and deletions, his model estimates are based on 324 observations. The author offers no discussion regarding possible bias and inefficiency created by deleting incomplete observations.

This study models the determinants of household financial well-being for a city in the Southwest U.S. using survey data adjusted for item non-response. Specifically, the study applies the technique of multiple imputation to a survey of financial well-being in Oklahoma County, Oklahoma. Multiple imputation is well established in public health

[3] July 2001.

and psychology but not widespread in the economics literature.[4] This paper proceeds as follows. First, a brief overview of the issue of how to deal with missing data is offered, followed with an introduction to multiple imputation (MI). The next section applies multiple imputation to data collected in Oklahoma County, Oklahoma. Two models of financial well-being are estimated, one using just the complete cases and the other using multiply imputed data. Two ordered logit models are estimated, one with complete cases only and the other model with complete cases and multiply imputed cases combined. Results of the two models are compared and final conclusions offered.

II. What to do about missing data?

Deleting records with missing values is a common practice for dealing with item non-response. This practice produces a reduced-size dataset of complete cases in which all the variables are fully observed. Reducing the dataset to complete cases has its advantages: it offers simplicity, since standard statistical packages can now be easily applied, and comparability, as all calculations proceed from a common base (Little and Rubin, 2002). List-wise deletion is simple and may be perfectly appropriate in numerous situations, particularly if the number of deleted incomplete cases is relatively small or if the deleted cases are very similar to the complete cases.

In many situations, however, discarding incomplete cases creates disadvantages. First, estimates based on complete cases are biased if the deleted cases differ from the complete cases. Second, the precision of model estimates will be lower due to the smaller sample size. It is possible that the extent of the bias and loss of precision will be

[4] Schafer and Graham (2002) provide an excellent discussion of the principles and applications of multiple imputation. Recent applications of multiple imputation include Davey (2000) and Raghunathan (2004).

small; rules of thumb are difficult to formulate, however, since the degree of bias depends not only on the proportion of incomplete cases but also on the differences between complete and incomplete cases and the pattern of missing data (Little and Rubin, 2002).

The statistical literature uses the term *missingness* to refer to the manner in which missing data are distributed. A classification scheme developed by Rubin helps to sort out general relationships between the missingness pattern and values of the data. Suppose a dataset consists of both complete cases and other cases that have missing items; together these are the observed data. The missing data mechanism is said to be *missing at random* (MAR) if the probability of missingness depends on the complete data but not the missing data. If the data are MAR, then the probability that an item is missing can be related to the values of observed items alone, not the missing items.

A special case of MAR is *missing completely at random* (MCAR). If the missingness pattern is MCAR, then the probability of missingness cannot be related to the observed data; missing data occur as if randomly distributed throughout the dataset. If MCAR is true, then analyzing the complete cases only is appropriate: standard errors are higher because of the smaller number of observations but estimators will not be biased.

Finally, missing data are termed *missing not at random* (MNAR) if the pattern of missingness is related to *the missing values.* If MAR is not true, then the missing data are MNAR. For example, if older householders tend to refuse to respond to the age question more than other households, then the missingness is related to the missing values and the mechanism is MNAR.

Survey data are tested later in this study for MCAR. Failing MCAR, the data are

assumed to be MAR. Testing the MAR assumption typically is not possible except by follow-up surveys with non-respondents. However, multiple imputation methods are robust to deviations from the MAR assumption, as violations may only have minor effects on estimates and standard errors (Schafer, 1997).

III. Multiple imputation

Instead of generating just one estimate for a missing value as is the case in single imputation, MI produces several estimates for *each* missing value; the variation of the estimates measures the uncertainty of imputation, an improvement compared with single imputation. Other techniques for estimating missing data have serious shortcomings. Replacing the missing values by the mean of the variable biases the variance and covariance toward zero, for example. Using regression models to estimate missing data also is problematic, since the correlation of the estimates and explanatory variables are biased away from zero. Weighting is proper to adjust for missing cases when all values of an observation are missing. Using weights to adjust for item non-response is equivalent to assuming that the missing data are similar to the complete data, an assumption that may not be true.

Multiple imputation offers a comprehensive method of simultaneously estimating missing values. Marginal associations among variables are preserved, and missing value uncertainty is estimated. MI produces stochastic estimates for missing values, drawing from the predictive distribution of missing values given the observed data (S-Plus Manual). In multiple imputation, the analyst specifies an imputation model; the model incorporates information about the relationships among the observed data, using this

information to estimate missing values.

Typically five to ten sets of estimated values are generated and combined with complete cases, with each combined dataset containing a set of estimates for the missing values and a replicate of the complete cases. The estimated values will typically differ from dataset to dataset. Each of the datasets is analyzed separately, estimating parameters (means, standard errors, and regression coefficients, for example) from each dataset; the parameter estimates are then combined to obtain overall estimates.

Little and Rubin (2002) provide rules for combining the parameter estimates. The multiple imputation point estimate for a mean is simply the sum over the imputed datasets divided by the number of datasets. For example, for variable *x*, the point estimate for the mean is

$$\bar{x} = \sum x_i / k ,$$

where k is the number of imputed datasets, each consisting of complete cases and imputations of missing values. Let v_j denote variance for variable *x* within the $_j$th dataset. Using Rubin's rules (Little and Rubin, 2002), average within-imputation variance is defined as

$$\bar{v} = \sum v_i / k ,$$

and between imputations variance is

$$B = \sum (v_j - \bar{v})^2 / k$$

and total variance is $T = \bar{v} + B/(1+k)$. If B = 0 there is no missing information and estimated variance is $\bar{v}$. Dividing the term on the right-hand side by $\bar{v}$ results in the ratio $r = B(1+k)^{-1}/\bar{v}$. Rubin calls *r* the relative increase in variance due to nonresponse; it

estimates how much the variance of an estimator has increased due to imputed values.

3.1. Estimating the missing data model

The missing data model is an analytical tool that predicts missing values based on relationships among observed values. The missing data model is not a model as economists understand the term, as it does not specify dependent or independent variables or operate in a causal framework. The model uses a multivariate framework to estimate missing values and adds random noise to preserve an appropriate degree of variability in the imputed data (Schafer and Graham, 2002). Two important algorithms used for the missing data model in this study are the expectation maximization algorithm and data augmentation.[5] The expectation maximization algorithm produces maximum likelihood estimates of parameters such as means, variances, and cell probabilities in the presence of missing data. A Markov chain Monte Carlo technique, the data augmentation algorithm generates posterior distributions for parameters and sequences of imputations for missing data. Both are discussed next.

3.1.1. Expectation Maximization

The EM algorithm is a powerful iterative technique for estimating missing values. For categorical data, EM allocates incomplete cases, adding them to a table of complete cases. An example will help illustrate.[6] Suppose a dataset of *n* observations consists of two categorical variables, Y_1 and Y_2. Y_1 is completely observed, but Y_2 is missing for some cases of Y_1. The complete cases of Y_1 and Y_2 can be summarized by a matrix of

[5] Gill (2002) provides a readable introduction to the expectation maximization algorithm and the data augmentation algorithm.

[6] Refer to Schafer (1997) for details regarding the use of the EM algorithm for continuous variables and mixed continuous and categorical variables.

cell counts with dimension $j \times k = n$. Similarly, the incomplete cases of Y_1 comprise a column vector with dimension j x 1 (Figure 1). Adapting an example from Little and Rubin (2002), let C_{jk} denote the completely observed cell count for row j and column k and let C_{j+} denote the sum of the cell counts across the columns of Y_2, where the position of the '+' subscript indicates summation across columns. Similarly, C_{+k} is the sum of the cell counts across rows of Y_1. Let m_j denote the counts of Y_1 with missing Y_2, and C_{++} the sum of the complete case cell counts. Consequently, the missing counts of Y_2 are equal to the total number of observations less the complete case cell counts ($m_j = n - C_{++}$).

The Expectation step of the EM algorithm adds counts of m_j to the complete case cell counts C_{jk}. The M step, or Maximization, uses the new cell counts to update the cell probabilities. The updated probabilities are then used to reallocate the incomplete cases to the complete case cell counts (E step), followed by a new M step. The process proceeds iteratively until the change in the likelihood becomes very small. Typically, just a few iterations are needed to achieve convergence.

As Little and Rubin (2002) demonstrate, EM factors the updated cell probability into two terms, the marginal distribution of Y_1 (labeled π_{j+}), and the conditional distribution of Y_2 on Y_1 (labeled π_{kj}). The marginal distribution of Y_1 is the probability that Y_1 will fall in a particular row, whether or not Y_2 is missing, defined as

$$\pi_{j+} = (c_{j+} + m_j) / n$$

The conditional distribution of Y_2 on Y_1 is the probability of Y_2 for a given a row of Y_1, calculated from the complete cases:

$$\pi_{kj} = c_{jk} / c_{j+}.$$

The updated cell probability is the product of the two factors,

$$\pi_{jk} = \pi_{j+}\pi_{kj}.$$

Substituting terms for π_{j+} and π_{kj} results in

$$\pi_{jk} = (c_{j+} + m_j) / n * c_{jk} / c_{j+}.$$

and rearranging produces

$$\pi_{jk} = (c_{jk} + (c_{jk} / c_{j+})m_j) / n,$$

which is the maximum likelihood estimate for a cell probability that incorporates missing data. Incomplete cases are assigned to complete case cell counts according to the conditional probability computed from the complete cases. For example, values of Y_2 are added to the cell count for cell (1,1) with probability c_{11} / c_{1+} and to cell (1,2) with probability c_{12} / c_{1+}, and so on. Note that if no missing data exist, then $m_j=0$ and the expression reduces to the well-known equation for a cell probability, $\pi_{jk} = c_{jk} / n.$

3.1.2. Data Augmentation

First presented by Tanner and Wong (1987), data augmentation (DA) is a Markov chain Monte Carlo algorithm for estimating missing values. Similar to the Expectation-Maximization algorithm, data augmentation proceeds in an iterative fashion, simultaneously estimating parameters and missing values. The critical difference is that EM converges to a single set of parameters and imputations, while DA converges to a distribution of multiple sets of parameters and imputed values.

The DA algorithm consists of two steps, the Imputation or I step and the Posterior or P step. The I step draws from the missing values of Y_2, adding them to the complete case counts with probability π_{kj} / π_{j+} (Little and Rubin, 2002). For example, π_{11} / π_{1+}

percent of the missing values in row one are added to the cell count for row one of column one, π_{12} / π_{1+} percent are assigned to column two of row one, and so on for the remaining columns and rows. In this manner the complete case cell counts are *augmented* with new data.

Next, based on the augmented cell counts the P step draws new values for the cell probabilities from the complete data posterior distribution,[7] updating the parameters π_{kj} and π_{j+}. The I step is then repeated, drawing values for Y_2 with for each row of Y_1, followed by a new P step.

The DA algorithm produces stochastic sequences for both missing values and cell probabilities. Assessing convergence is less straightforward than for EM, as convergence occurs in *distribution*. Convergence can be judged by examining the sequences of the cell parameters, with nonstationarity suggesting nonconvergence. Convergence is also indicated by independence of successive iterations. Convergence by iteration *t* means that the cell parameters for iteration *t* and iteration *t+s* are independent. Autocorrelation functions can be used to assess the independence of parameter iterations; a small autocorrelation by iteration *t* suggests independent iterations.

Once DA has converged, a small number (*z)* of the imputed sequences are set aside and combined with replicates of the complete cases to form *z* separate datasets. Models are then estimated for each of the *z* datasets and the results combined.

IV. Description of the Data

[7] In practice, the draws of the parameters are from a Dirichlet distribution, typically used to simulate draws from an unknown target distribution when the data are distributed multinomial, as in the case with categorical data.

Data for this study were collected by the Center for Economic and Management Research, The University of Oklahoma, under contract with Community Council of Oklahoma County, a not-for-profit social services agency responsible for coordinating research needs for local social service agencies. Community Council wished to develop annual indicators of quality of life for Oklahoma County with special attention to the quality of life perceived by sub-sets of the population such as the elderly, households with children, and households with health problems. Data were collected by telephone interview with randomly selected households during the spring of 2002. The phone sample was generated by random digit dialing, stratified by age and gender. A response rate of 45 percent was achieved during interviewing, resulting in 1,265 interviews.

Six items from the survey are used in this study: financial well-being, householder's age, the presence of minor children, household income, home ownership, whether the householder has health insurance, and gender. Table 1 presents descriptions and summary statistics for each of the variables. Financial well-being (FWB) is measured as the response to the question, "How would you say you feel about the overall financial security of your household? Would you say you feel very secure, somewhat secure, somewhat insecure or very insecure?". The number of householders who respond 'Somewhat insecure' to this question is very small, making the missing data model difficult to estimate. Consequently, FWB is re-coded from four values to three, where Insecure is combined from Somewhat insecure and Very insecure.

Figure 2 shows the distribution of missingness in the data. Fourteen distinct patterns of missingness exist, with each pattern characterized by a unique combination of missing variables. In Pattern 1 no variables are missing; this pattern characterizes 80.6

percent of the records. Pattern 2 is missing income, Pattern 3 is missing age, Pattern 6 is both missing income and age, and so forth. Clearly, the pattern of missingness is complex. Table 2 shows the percent of missing data for each variable, sorted by increasing missingness. Overall, 19.4 percent of the observations have at least one missing value, largely attributable to missing income and missing age.

Before proceeding with the imputation algorithm, it is worthwhile to determine whether the data are missing completely at random (MCAR). If MCAR is not true, model estimates based on complete cases only may be biased. Testing for MCAR is similar to a Chi-square test for independence. The categorical variables HEALTH, HOME OWNER, CHILDREN, FWB, and INCOME have 2, 2, 2, 3, and 3 levels, respectively, describing a contingency table with 72 cells. The dataset consists of both complete cases and incomplete cases, together called the *observed* data. Testing for MCAR proceeds in three steps. First, calculate cell probabilities for two tables, with the first table using complete cases only (complete case table) and the second table using both complete and incomplete cases (observed data table). Next, calculate cell counts by multiplying the cell probabilities in each table by the total sample size (1,265), and third, compare the cell counts in the two tables using Pearson's Chi-square statistic. This test determines whether the complete case table is reasonably representative of the observed data table. If the cell counts from the complete data table differ substantially from those in the observed data table, then the missing data are not MCAR. Calculating the complete case table is relatively straightforward: cell probabilities from the complete cases are figured and multiplied by the total sample size of 1,265.

The EM algorithm is used to calculate cell probabilities for the second table. Each

cell of this table consists of counts of the complete cases plus a proportional share of the incomplete cases. Schafer (1997) shows that the EM algorithm generates maximum likelihood (ML) cell probabilities that incorporate both the complete and incomplete cases.

If the complete case cell counts are not statistically different from the maximum likelihood cell counts, then missing values are MCAR and model estimates using just the complete cases will not show bias due to missing information. To test the MCAR hypothesis, we use two goodness-of-fit tests: the likelihood ratio statistic (G^2) and Pearson's chi-squared statistic (X^2). The likelihood ratio statistic is

$$G^2 = 2\sum (Complete)\log(Complete / ML)$$

and Pearson's chi-square statistic is

$$x^2 = \sum (Complete - ML)^2 / ML$$

with summation over all the cells in the contingency table. 'Complete' indicates complete case cell counts and 'ML' (maximum likelihood) counts are from the expectation maximization algorithm. Both statistics are distributed Chi-square with degrees of freedom equal to the number of cells minus the number of parameters estimated. The contingency table formed from the five categorical variables has 72 cells estimated with 72-1 parameters, so degrees of freedom are unity. If the complete case table is accurate, then the null hypothesis will not be rejected.

The relatively high values for X^2 and G^2 in Table 3 and low p-values suggest that the null hypothesis should be rejected; the data are not missing completely at random.

Therefore, limiting our analysis to just the complete cases may result in biased parameter estimates.

V. Estimating the Missing Data Model

A missing data model can be specified for categorical variables, for continuous variables, or a combination of both categorical and continuous variables as is the case in this study; Schafer (1997) calls the latter model *conditional Gaussian*. The missing data model estimates missing values for the categorical variables and the continuous variables separately, combining the results. For the categorical variables, the missing data model estimates cell probabilities for the contingency table; in our case, the model estimates the probability that an observation falls into each of the 72 possible cells.

For the continuous variable, each observation of AGE falls into one of the cells of the contingency table; parameters for AGE are the mean for each cell and overall variance; mean AGE can vary from cell to cell, but variance is assumed common for all cells.

In general, the total number of parameters estimated for the conditional Gaussian model is

$$(D-1)+Dq+q(q+1)/2$$

where D is the number of cells and q the number of continuous variables (Schafer, 1997). The last term on the right is the number of parameters needed to estimate the covariance. In our case, the total number of parameters is (72-1) + 72 + 1(1+1/2) = 144. If these cell probabilities, cell means, and the variance are estimated with an equal number of

estimators, we have a *saturated* model. Unfortunately, coverage of the data over the 144 parameters is somewhat sparse; a significant number of cells have no observations. Consequently, a saturated model cannot be estimated with this data; the missing data model must be simplified by either eliminating variables or reducing the number of parameters needed for estimation.

Since all the variables in the missing data model are important for predicting financial well-being, we choose not to eliminate any of them. Instead, following examples in Schafer (1997), we restrict the number of parameters needed to estimate the missing data model. The number of parameters is reduced in two ways: by placing log linear constraints on the cell probabilities, and by estimating the cell means of AGE with a simple linear model with the categorical variables as regressors.

Loglinear models are widely used by biological scientists and social scientists for the analysis of complex contingency table data. In this type of analysis, log cell probabilities are estimated using additive terms for various levels of interaction among the categorical variables. For example, consider a table with three binary variables forming a 2x2x2 table with 8 cells. A *saturated* log linear model predicts cell probabilities using terms for each of the possible variable combinations: three main effects, three two-way interaction terms, and one three-way interaction term for a total of seven terms. A loglinear model for this example can be expressed by the equation

$$\log m_{ijk} = u + u_A + u_B + u_C + u_{AB} + u_{BC} + u_{AC} + u_{ABC}$$

where m_{ijk} is the cell count for cell *ijk*. On the right side of the equation, u is the grand mean, defined as $u = \frac{1}{8}\sum_{i=1}\sum_{j=1}\sum_{k=1} \log m_{ijk}$, u_A is the main effect of variable A expressed as

deviation from the grand mean, $u_A = \frac{1}{4}\sum_{j=1}\sum_{k=1} \log m_{ijk} - u$, and u_{AB} is the interaction term for A and B expressed as deviation from the grand mean, $u_{AB} = \frac{1}{2}\sum_{k=1} \log m_{ijk} - u$, and so on with the remaining interaction terms (Fienberg, 1979).

The loglinear model can be simplified by setting one or more of the interaction terms to zero. Suppose we set the three-way interaction term to zero; the restricted model now consists of 8-1-1 terms to estimate the 8 cell probabilities. Setting the three-way interaction term to zero is equivalent to assuming that any two variables taken together are independent of the third variable; this is termed the *no second order interaction model* (Fienberg, 1979).

A number of other restrictions are possible; for example, we could restrict the three-way interaction term and one or more of the two-way interaction terms. However, too many restrictions on the interaction terms may cause a poor fit. For example, the simplest restricted model consists of the main effect terms only; this model assumes that no interaction exists at all among the three variables. In this case, all the interaction terms are set to zero and the number of parameters is reduced to just three (8-1-4). Simple to estimate, this model is very unlikely to perform well, as it assumes away all the relationships of interest and will likely produce poor estimates of the actual cell probabilities.

Returning to the financial well-being dataset, a simplified loglinear model is specified for the categorical variables with some of the interaction terms to zero. After much experimentation, it was determined that the data augmentation algorithm will run if the three-way and four-way interaction terms and the five-way interaction term are set to

zero, retaining the main effects and the two-way interaction terms. The two-way interaction model requires just 40 terms, a considerable saving compared with the 72 terms needed for the saturated model.

As for the continuous variable, we use a simple linear model to predict AGE given the values of the categorical variables. The simplified model for AGE consists of eight parameters, one each for the three binary variables CHILD, HEALTH, and HOME OWNER, two for INCOME, two for FWB, one for HEALTH, and one for the variance. Compared with the saturated model, the restricted AGE model requires far fewer terms, reducing the number of parameters from 72 to just eight.

To summarize this section, two missing data models are specified: a restricted log linear model for the categorical variables and a simple linear model to predict cell means for AGE conditioned on the categorical variables. Restrictions reduce the number of parameters required to estimate our model from 144 to 48, a substantial saving of degrees of freedom.

Next, the expectation maximization algorithm is run incorporating the two missing data models, converging to single set of parameters and imputed values. Following the recommendation of Schafer (1997), the EM parameter estimates are used as starting values for the data augmentation algorithm. We ran 7,000 iterations of the data augmentation algorithm, discarding the first 100 iterations as a burn-in period.

Convergence for data augmentation can be assessed by examining the time series plots of the parameter iterates and examining autocorrelation plots for each of the parameters. The rate of convergence of the parameter iterates depends on the fraction of missing information; the higher the fraction of missing data, the greater the number of

iterations required.

Convergence is indicated as the absence of trend in the parameter iterates and by a rapidly declining autocorrelation function; the autocorrelation function (ACF) plots should die out after a finite number of iterates if the algorithm converges. Our missing data model estimates several dozen parameters; thus, examining iterate plots and ACFs for each individual parameter is not feasible. As an alternative, Schafer (1997) recommends monitoring a particular linear function of the parameters that converges slowly compared with other linear functions. This *worst linear function* (WLF) offers evidence of global convergence. The ACF plot of the worst linear function is shown in Figure 3. The ACF dies out quickly after just 25 iterations suggesting quick convergence.

Schafer also recommends monitoring the distribution of the likelihood ratio statistic as a measure of global convergence. The likelihood ratio statistic is the difference between the likelihood of the parameter estimates from the EM algorithm and the likelihood of the parameter estimates for iteration *i* from the DA algorithm:

$$dl(\theta_i) = 2[l(\eta \mid Yobs) - l(\phi_i \mid Yobs)]$$

where η is the maximum likelihood estimate of the parameters from the EM algorithm and ϕ_i are the estimated parameters from iteration *i* of the DA algorithm. The distribution of this function is approximately Chi-square with degrees of freedom equal to the number of parameters estimated under the null hypothesis of equivalence (Schafer, 1997). A good fit of the distribution of the likelihood ratio statistic to Chi-square indicates global convergence.

Panel A of Figure 4 shows a plot of 500 iterations of the likelihood ratio statistic

with the Chi-squared distribution superimposed. The degrees of freedom for the Chi-squared distribution is 36, which is the total number of parameters minus the number of restrictions. The fit of the likelihood ratio statistic to Chi-squared is poor in Panel A, indicating the absence of convergence. Increasing the number of iterations to 1,000 offers improvement (Panel B), but the fit still is not good. Panel C with 3,500 iterations shows a better fit, but the best fit occurs with 7,000 iterations (Panel D), indicating global convergence has been achieved for the data augmentation algorithm.

The next step involves making draws from the data augmentation iterates, each draw consisting of both complete data and imputations of the missing data. In order to assure independent draws, the selection interval is set high at every 250^{th} iteration, with the limit set at ten draws. Logit models are then estimated for each of the ten datasets, and results combined as described in the next section.[8]

VI. A Model of Financial Well-Being

Self-reported financial well-being is modeled as depending on age, household income, home ownership, health insurance, and the presence of children under 18 years of age. Following van Praag, AGE is modeled as a second-order polynomial. Some of the variables can be expected to have strong predictive power for FWB. Income will almost certainly be a very strong predictor, and homeowners are likely to enjoy higher FWB than households who do not own homes. Householders with health insurance will feel more financially secure than those who are not insured. Supporting a family creates stress on household finances, so we expect the presence of children to be negatively related to FWB. The expected influence of AGE on FWB is less clear, although both

[8]The specific S-Plus and SAS code used for this study is provided in the Appendix.

Van Praag (2003) and Bukena (2001) found a quadratic relationship between age and quality of life, suggesting that householders towards either end of the age distribution are more likely to report higher FWB than are middle-age householders, holding other variables constant.

6.1. Estimating the Ordered Logit Model

Ten ordered logit models are estimated with FWB as the dependent variable, one estimate for each of the ten imputed datasets. The model includes AGE*AGE as an explanatory variable in order to capture the U-shaped relationship between FWB and AGE. Parameter estimates from the ten models are combined using Rubin's rules (discussed above).

Incorporating variance due to imputation is an important aspect of multiple imputation. Variance for model estimators depends not only on variation within the datasets, but also variation between the datasets. Table 4 shows estimator variance due to multiple imputation. *Within variance* is average variance of the eight imputations, and *between variance* is the additional variance attributable to the imputed values. *Total variance* is the sum of within and between variance. The table also shows the relative increase in variance caused by the imputed values. Variables with relatively small amounts of missing information show only modest increases in variance, while those with large amounts of missing information show large increases. Not surprisingly, the income variables exhibit the largest increases in variance, with LOWINC showing a 14 percent increase and MEDINC a 21 percent increase. Similarly, variance for HOME OWNER and HEALTH are 8 percent higher and AGE and CHILD show increases in the 4 percent to 5 percent range.

Coefficient estimates and standard errors for the multiple imputation model and the complete case model are presented in Table 5. The variable MALE is added to the models to capture possible gender differences. Comparing these estimates two things are apparent. First, the MI coefficients are more efficient: on average, the standard errors for the MI model are 9 percent smaller than for the CC model. The reason for the smaller standard errors is the larger sample size: the MI model utilizes all 1,265 observations, while the CC model is limited to a much smaller 1,020 observations. Thus, even though the multiple imputations generate additional variance, the variance increase is more than offset by a larger number of observations.

Secondly, the coefficients show both similarities and differences. The coefficients for two income variables show little difference; both LOWINC and MEDINC have about the same value and very high t-values in both models. Most of the other variables, however, show differences. HOME OWNER is 19 percent larger in the MI model, and CHILD 39 percent larger in absolute value, while HEALTH is somewhat smaller. AGE is 24 percent smaller in the MI model and AGE*AGE is 27 percent smaller. Finally, the intercept terms are smaller in the MI model.

The extent of bias in the CC model is apparent in Figure 5, which shows the predicted probability of 'Very Secure' by age for householders with no health insurance. The CC model underestimates predicted financial well-being for householders 45 years and younger by 10 percent to 20 percent; predicted probabilities converge as age approaches 70 years.

If we follow the usual approach and delete observations with missing values, we are left with the CC model. As shown in the table, the CC model under-estimates the effects

of home ownership and the presence of children on financial well-being and over-estimate the effects of health insurance, income, and age.

VII. Summary and Conclusion

This study applies multiple imputation to a model of financial well-being, using data collected from Oklahoma County, Oklahoma, as a case study. The model shows that self-reported financial well-being depends on household income, age, the presence of children, home ownership, and whether the householder has health insurance.

Missing values are estimated using multiple imputation. A missing data model is estimated, with interactions limited to main effects and two-way interactions. Ordered logit models are estimated for each of the ten imputed datasets and the results combined. A model is also estimated for just the complete cases. Results from the complete case model are less efficient and tend to be biased compared with the multiple imputation model. The presence of missing values may bias the parameter estimates if the analysis is limited to just the complete cases.

The study shows that economists should give more attention to the problem of missing data in surveys. If observations with missing data are simply deleted, model estimates will be less efficient and may be biased. Multiple imputation offers a comprehensive method of estimating missing values and estimating the uncertainty resulting from missing value estimates.

References

Analyzing Data with Missing Values in S-Plus, Seattle: Insightful Corporation, 2001.

Bell, R. (1984).'Item nonresponse in telephone surveys: an analysis of who fails to report income', *Social Science Quarterly*, Vol. 65, pp. 207-215.

Bukenya, J. O., Gebremedhin, T. G., and P. V. Shaeffer. (2003), 'Analysis of Quality of Life and Rural Development: Evidence from West Virginia Data', *Growth and Change*, Vol. 34, pp. 202-218.

Davey, A., Shanahan, M. J., and Schafer, J. L. (2001). 'Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth Using Multiple Imputation', *The Journal of Human Resources*, Vol. 36, pp. 500-519.

Easterlin, R. A. (2001), 'Subjective well-being and economic analysis: a brief introduction', *Journal of Economic Behavior & Organization*, Vol. 45, pp. 225-226.

Fienberg, S. E. (1979), *The Analysis of Cross-Classified Categorical Data*, Cambridge, Mass.: The MIT Press.

Gill, J. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach*, London: Chapman & Hall/CRC.

Gronhaug, K., Gilly, M. C., and Enis, B. M. (1988). 'Exploring Income Non-response: A Logit Model Analysis', *International Journal of Market Research*, Vol. 30, pp. 371-378.

McBride, M. (2001). 'Relative-income effects on subjective well-being in the cross-section', *Journal of Economic Behavior & Organization*, Vol. 45, pp. 251-278.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, Hoboken, N.J.: John Wiley & Sons, Inc.

Raghunathan, T. E. (2004). 'What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data', *Annual Review of Public Health*, Vol. 25, pp. 99-117.

Schafer, J. L. and Graham, J. W. (2002) 'Missing Data: Our View of the State of the Art', *Psychological Methods*, Vol. 7, pp. 147-177.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, New York: Chapman & Hall/CRC.

Tanner, M.A. and Wong, W.H. (1987). 'The Calculation of Posterior Distributions by Data Augmentation', *Journal of the American Statistical Society*, Vol. 82, pp. 528-550.

Van Praag, B.M.S., Frijters, P., and Ferrer-i-Carbonell, A. (2003). 'The Anatomy of Subjective Well-Being," *Journal of Economic Behavior & Organization*, Vol. 51, pp. 29-49.

Appendix: S-Plus code and SAS code used to generate imputations

S-Plus is used to estimate missing values and SAS is used to estimate the logit models and combine the results.

S-Plus code
(more details are provided in the S-Plus missing data manual)

1. Restrict the margins for the loglinear model to two-way interactions and main effects:
margins.em<-~FWB+INCOME+HOME+HEALTH+CHILD+FWB:INCOME+
FWB:HOME+FWB:HEALTH+FWB:CHILD+INCOME:HOME+INCOME:HEALTH+
INCOME:CHILD+HOME:HEALTH+HOME:CHILD+HEALTH:CHILD

2. Estimate the linear model for AGE:
design.form<-~FWB+HEALTH+CHILD+HOME+INCOME

3. A preliminary tabulation of the raw data
qol.s<-preCgm(Data)

4. Run the EM algorithm using the loglinear model and the linear model for AGE:
qol.em<-emCgm(qol.s,margins=margins.em,design=design.form,prior=1.02)

5. Run the data augmentation using the results of EM as the starting point. Run DA for 7,000 iterations, throwing out the first 100 as a burn-in period:
qol.da<-daCgm(qol.em,control=list(niter=7100,save=100:7100))

6. Create ten sets of imputations, selecting every 250^{th} iteration from DA:
qol.imp<-impCgm(qol.da,nimpute=10,control=list(niter=250))

7. Save the ten imputed datasets:
M1<-miSubscript(qol.imp,1)
M2<-miSubscript(qol.imp,2)
M3<-miSubscript(qol.imp,3)
M4<-miSubscript(qol.imp,4)
M5<-miSubscript(qol.imp,5)
M6<-miSubscript(qol.imp,6)
M7<-miSubscript(qol.imp,7)
M8<-miSubscript(qol.imp,8)
M9<-miSubscript(qol.imp,9)
M10<-miSubscript(qol.imp,10)

8. The ten datasets are then appended to form one dataset with 12,650 records, then imported into SAS and re-labeled CCDAT]

SAS code

```
DATA NEXT;
  SET CCDAT;

PROC SORT;
BY _IMPUTATION_;
```

9. Ten sets of logit model estimates are generated, one for each imputed dataset:

```
PROC LOGISTIC data=NEXT outest=logout covout;
CLASS FWB;
MODEL FWB = INC1-INC3 AGE AGE*AGE HOME CHILD SEX HEALTH;
BY _IMPUTATION_;
```

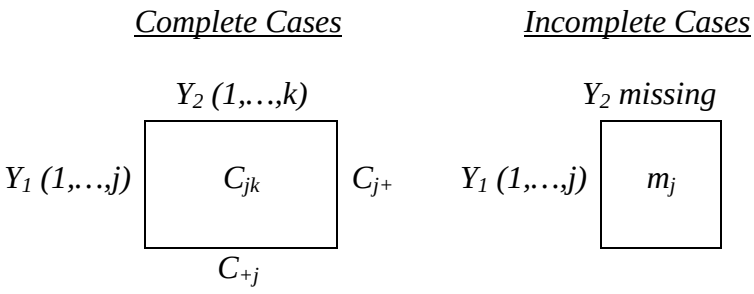
10. The ten estimates are consolidated according to Rubin's rules and results reported using the new SAS MIANALYZE procedure:

```
PROC MIANALYZE data=logout;
VAR Intercept INC1-INC3 AGE AGE*AGE HOME CHILD SEX HEALTH EDU;
TITLE 'MI MODEL';


RUN;
```

TABLES AND FIGURES

FIGURE 1
Example showing allocation of missing values for categorical variables



Note: adapted from Little and Rubin, 2002.

TABLE 1

Variable Descriptions

Variable name	*Description*	*Mean*	*Standard Deviation*	*Min*	*Max*
FWB	Financial well-being (1=Very secure, 2=Somewhat secure, 3=Insecure)	1.831	0.711	1	3
LOWINC	Income less than $25,000	0.275	0.447	0	1
MEDINC	Income from $25,000 to $50,000	0.318	0.466	0	1
HIGHINC	Income more than $50,000	0.407	0.491	0	1
AGE	Age of respondent	47.370	17.132	18	99
HOME OWNER	Home ownership (1=Home owner, 0=Not home owner)	0.718	0.450	0	1
CHILDREN	Children under 18 present (1=Present, 0=Not present)	0.377	0.485	0	1
MALE	1=Male, 0=Female	0.469	0.499	0	1
HEALTH	Health insurance (1=Have, 0=Don’t have)	0.803	0.398	0	1

TABLE 2

Information Missing by Variable

Variable	*Number Missing*	*Percent Missing*
Health	4	0.3%
Home	5	0.4%
Children	8	0.6%
Financial well-being	16	1.3%
Age	51	4.0%
Income	221	17.5%
At least one variable missing	245	19.4%

FIGURE 2

Patterns of missingness (shading indicates missing data)

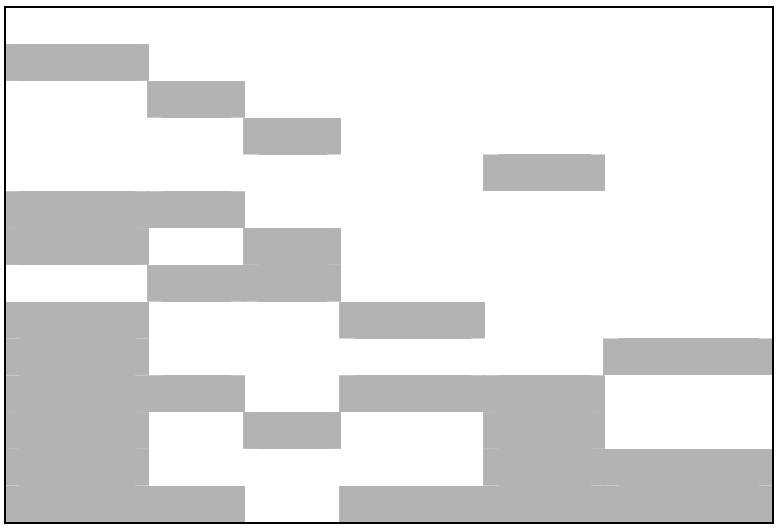


TABLE 3

Testing for MCAR
(null hypothesis: the data are MCAR)

X^2	9.01	(0.0027)
G^2	5.00	(0.0253)

Note:
(p-values in parentheses, df=1)

FIGURE 3
Autocorrelation Plot for the Worst Linear Function

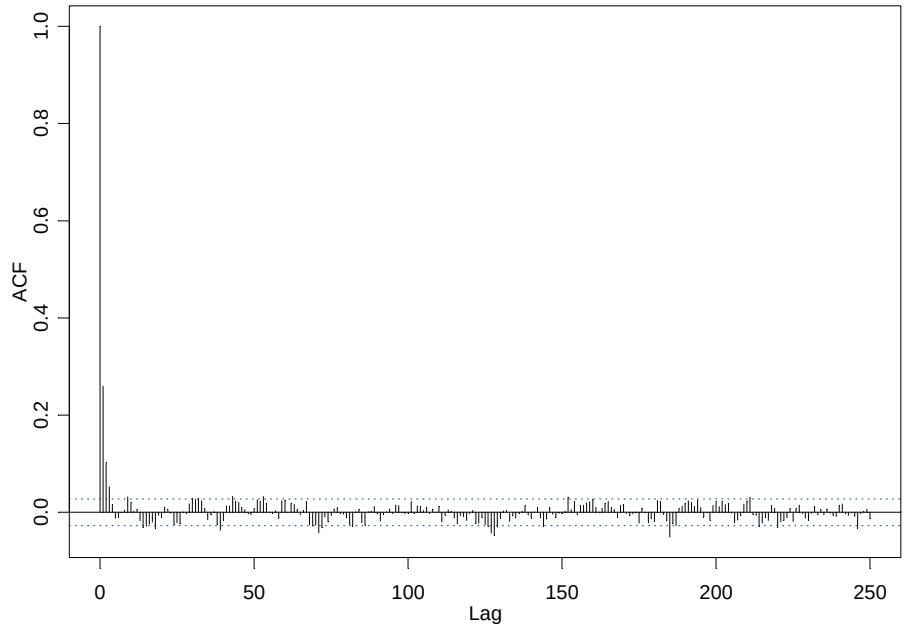


FIGURE 4

Convergence of data augmentation iterations

(columns show the distribution of the likelihood ratios; the solid line is the Chi-square distribution with 36 degrees of freedom. A good fit is indicated when the Chi-square distribution and the distribution of the likelihood ratios converge.)

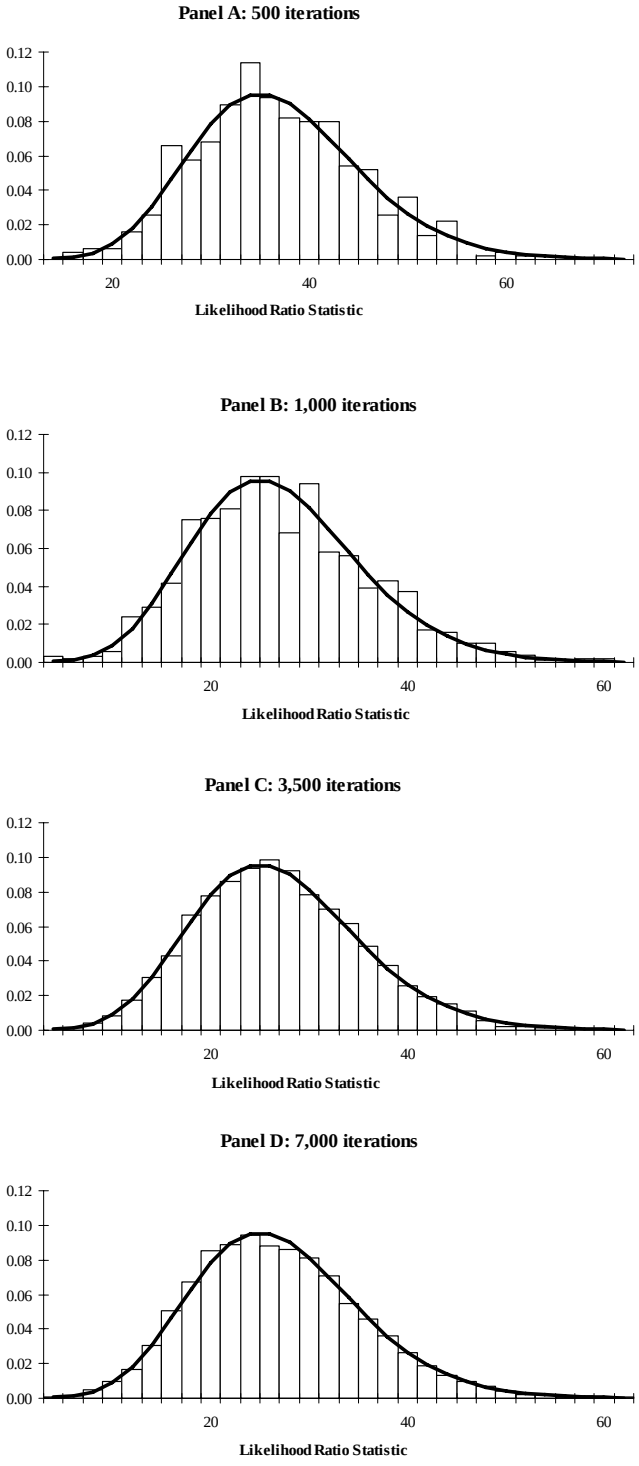


TABLE 4

Variance for Multiple Imputation Estimated Parameters

Variable	*Between imputation variance*	*Average within imputation variance*	*Total variance*	*Percent increase in variance due to multiple imputations*
LOWINC	0.003663	0.028032	0.032061	14.4%
MEDINC	0.003621	0.018623	0.022606	21.4%
HOME OWNER	0.001762	0.021727	0.023665	8.9%
HEALTH	0.001830	0.023617	0.025629	8.5%
CHILD	0.000607	0.016280	0.016948	4.1%
AGE (x 1,000)	0.016059	0.354000	0.370059	5.0%
AGE*AGE (x 1,000,000)	0.001642	0.034266	0.035908	5.3%
Intercept 1	0.01387	0.226249	0.241507	6.7%
Intercept 2	0.016112	0.238036	0.255759	7.4%

TABLE 5

Ordered Logit Estimates for the Multiple Imputation Model and Complete Case Model

	Multiple Imputation			*Complete Case*		
Variable	*Coefficient*	*Standard error*	*t-statistic*	*Coefficient*	*Standard error*	*t-statistic*
LOWINC	-1.632	0.180	-9.05	-1.767	0.194	-9.13
MEDINC	-0.941	0.151	-6.23	-1.014	0.154	-6.57
HOME OWNER	0.452	0.154	2.94	0.380	0.164	2.31
HEALTH	0.748	0.160	4.67	0.899	0.175	5.15
CHILD	-0.530	0.130	-4.06	-0.381	0.141	-2.70
AGE	-0.091	0.019	-4.72	-0.119	0.022	-5.38
AGE*AGE (x 1,000)	0.874	0.190	4.60	1.210	0.222	5.45
Intercept 1	1.363	0.495	2.76	1.786	0.567	3.15
Intercept 2	3.862	0.509	7.59	4.329	0.582	7.44

FIGURE 5

Predicted Probability of 'Very Secure' for Householders with No Health Insurance, Multiple Imputation Model and Complete Case Model