THE EFFECTS OF PROCEDURES WHICH IMPUTE FOR MISSING ITEMS: A SIMULATION STUDY USING AN AGRICULTURAL SURVEY

Barry L. Ford, Douglas G. Kleweno, and Robert D. Tortora
U.S. Department of Agriculture

ABSTRACT

This simulation study compares the effects of six procedures which impute values for missing items. Using data from an agricultural survey, this experiment covers a range of conditions which account for the method of designating which values are missing and the rate at which they are missing. An analysis of the mean square errors, the effects on the correlations, and the costs show that two versions of a ratio procedure give the best results for sample sizes which are very large.

1. INTRODUCTION

The problem of incomplete data, i.e. missing values, is one of the most common problems of survey work. Incomplete data is of two types -- missing units and missing items. Missing units are the result of nonresponse for a sample unit and, thus, consist of refusals and inaccessibles. Missing items refer to those units which have missing values but also have some reported values. For example, the respondent answers some questions but not others, or he answers some questions incorrectly. The problem of missing units is the subject of previous studies at the U.S. Department of Agriculture [2]. The purpose of this study is to compare six procedures which impute for missing items.

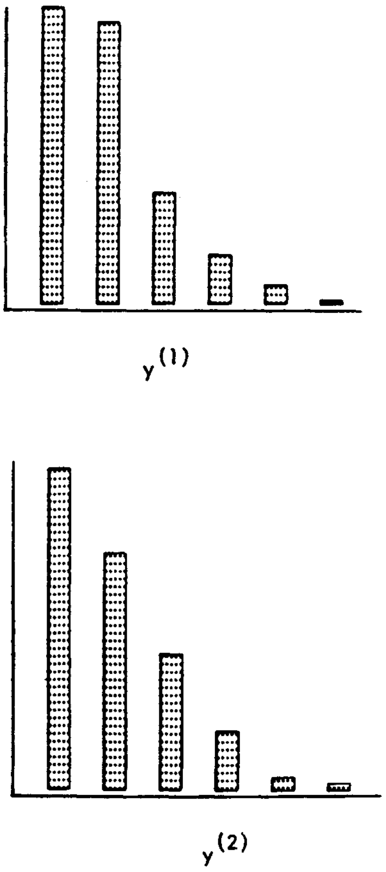
The basic research tool of this study is simulation. Using a complete data set (no missing values) from a current survey, the authors simulate which values are missing. Six missing item procedures are then applied, and the imputed values are compared to each other and to the original values. Although simulation experiments are in a sense artificial, they do allow analysis over a wide range of conditions and a comparison against "true" values.

The simulations in this study are over various levels of two effects: 1) the randomization mechanism used to designate which values are missing and 2) the rate at which values are missing. For each level of these two effects, there are several incomplete data sets simulated from the original data set. The original data set is divided into three replicates, and this replicate structure is carried over into each simulated data set. Missing item procedures are applied to each replicate independently in order to obtain unbiased estimates of standard errors.

The original data set used in this study is from one stratum of a hog survey conducted by the U.S. Department of Agriculture. There are 201 complete sample units which are divided into three replicates of 67 units each. Each unit has 15 quantitative variables -- 14 survey variables and 1 control variable which has been used to stratify the population. For the purposes of this study the authors confine the simulation of missing values to two major survey variables $y^{(1)}$ and $y^{(2)}$, of the 14 survey variables. Values for either or both of these variables can be designated as missing. All imputations must obey the edit check $y^{(1)} + y^{(2)} \leq w$, where w is another of the 14 survey variables.

The distributions of $y^{(1)}$ and $y^{(2)}$ are both highly skewed. Figure 1.1 gives two bar graphs to show the general shape of the distributions. The mean of $y^{(1)}$ is 22.11 and the variance is 509.32; the mean of $y^{(2)}$ is 21.45 and the variance is 502.22. Thus $y^{(1)}$ and $y^{(2)}$ are similar in distribution. The correlation between w and $y^{(1)}$ is 0.82 and between w and $y^{(2)}$ is 0.81. Both $y^{(1)}$ $y^{(2)}$ have integer values greater than or equal to zero.

Figure 1.1: Bar graphs to show the general shape of the distributions of $y^{(1)}$ and $y^{(2)}$.



2. THE PROCEDURES

This study compares the effects of six procedures which impute for missing items. This section gives a description of each procedure, a description which includes the estimation techniques and assumptions used by the procedures. The descriptions are written in general terms of how the procedures would impute for a data set which has both complete and incomplete units.

2.1 The Ratio Procedure (Variations 1 and 2)

The ratio procedure examined in this study imputes a value for each missing value by using the equation:

$$y_{\text{ratio}} = \hat{R}\, x^*$$

where:

$\hat{R}$ is the estimated ratio between the variables x and y

x^* is the value of an x variable for a sample unit which has a missing y value

y_{ratio} is the value imputed for the missing y value.

An estimate of R is based on the sample units which are complete. If x' and y' are totals for the complete units in the sample, then $\hat{R} = \frac{y'}{x'}$. Thus, this estimator of R assumes that the ratio for the complete units is a good estimate of the ratio for the incomplete units.

Although called an auxiliary variable, the x variable may be a survey variable or the control variable. When a y value is missing, the ratio procedure uses as the x variable that variable which is most highly correlated with the y variable. If the value of the most highly correlated variable is missing from the unit, the procedure uses the next most highly correlated variable. If that value is also missing, then the procedure continues in the same fashion until a reported value is found. Correlations are estimated by using only the complete sample units of a data set which has missing values.

This study uses two variations of the ratio procedure. These two variations arise because of the linear restriction imposed on the two variables -- $y^{(1)} + y^{(2)} \leq w$. The first variation simply imputes independently for $y^{(1)}$ and/or $y^{(2)}$ and then checks to see whether $y^{(1)} + y^{(2)} \leq w$. When $y^{(1)} + y^{(2)} > w$, then the procedure adjusts any imputed values so that $y^{(1)} + y^{(2)} = w$. The second variation uses the constructed variable $z = y^{(1)} + y^{(2)}$ as though it is a survey variable. If z is missing (either $y^{(1)}$ or $y^{(2)}$ is missing), the procedure: 1) finds an x variable by using correlations with z where the correlations are estimated from the complete units, 2) imputes a value for z, 3) makes z = w if z > w, and 4) imputes for missing values of $y^{(1)}$ and/or $y^{(2)}$ so that $y^{(1)} + y^{(2)} = z$. If both $y^{(1)}$ and $y^{(2)}$ are missing, z is split into $y^{(1)}$ and $y^{(2)}$ proportionally by using relationships from the complete units in the data set.

2.2 The Array Procedure

The array procedure is not a procedure in general use but a procedure designed within the U.S. Department of Agriculture in 1971 [1] and proposed as a method of imputing for missing values on the Department's hog survey. Although not designed by the authors, the array procedure is included among the test procedures because its effects have never been assessed.

The array procedure uses a two-way table to impute for missing values. Two survey variables, a_1 and a_2, are chosen to define the table. If these two variables have c_1 and c_2 classes respectively, then the array procedure would form a table for $y^{(1)}$, as an illustration, of the form:

		Variable a_2			
		1	2	. . .	c_2
Variable a_1	1				
	2				
	:				
	c_1				

Cell values must be initialized with an estimate of the ratio $y^{(1)}/a_3$ where a_3 is another survey in a variable. As the procedure processes the units a sample, each unit is classified into a cell of the table by the values of a_1 and a_2 for that unit. If $y^{(1)}$ is reported, the ratio $r = y^{(1)}/a_3$ from the unit is added into a cell by using the weighted formula:

$$\frac{2(\text{previous value for the cell}) + r}{3}$$

The purpose of this weighted formula is to prevent the imputation of extremely large values, i.e. outliers. If $y^{(1)}$ is missing, the value of a_3 from the unit is multiplied by the ratio from the appropriate cell and imputed for the value of $y^{(1)}$. Obviously, the ordering of the data has some importance for the estimates from the array procedure. Although data from surveys by the U.S. Department of Agriculture are often in a roughly geographic order, the data of this study were in a random order except that complete units were processed before incomplete units.

The array procedure is similar to the ratio procedure because the array procedure also uses a type of ratio to impute values. However, the array procedure is a more complex method of obtaining the ratio and a more rigid process. For example, the array procedure can use the information from three auxiliary variables -- a_1, a_2 and a_3. However, these three variables must be chosen before applying the procedure to the data set and are not allowed to have missing values. Another difference is that once the array procedure processes all complete units in a data set, then the procedure can also use incomplete units to change the ratio values in the cells as long as a_1, a_2, a_3 and $y^{(1)}$ are not missing. The ratio procedure, as used in this study can only use estimates of ratios and correlations from the complete units in a data set.

In this study the a_1 and a_3 variables are the same variable, w. The variable a_2 is another survey variable which is highly correlated with $y^{(1)}$ and $y^{(2)}$.

2.3 The ESTMAT Procedure

The ESTMAT procedure is an iterative solution to the problem of finding the maximum likelihood estimates for a multivariate data set in which some values are missing [4]. The ESTMAT procedure imputes by using multivariate regressions as defined by the reported values. As long as the same regression relationships apply to both reported and missing values, the ESTMAT procedure should be able to impute accurately even if the reported and missing values have different means.

The ESTMAT procedure represents an extension of the double sampling regression estimator to a multivariate setting. However, the ESTMAT procedure can take into account many different patterns of missing data in the data set. For example, once the data is collected for two variables, there are four possible patterns of missing data -- both variables are reported, only the first variable is reported, only the second variable is reported, or both variables are missing. With k variables there are 2^k possible patterns if one also counts as a pattern the set of complete units.

The estimation formulas which the ESTMAT procedure uses are complex and are not given in this paper. However, they can be found in the references. Convergence of the iteration process used by ESTMAT is not assured in general, but in practical applications the convergence has usually taken less then ten iterations.

The two major assumptions of the ESTMAT procedures are: 1) values follow a multivariate normal distribution, and 2) the values are missing at random. The first assumption is necessary, of course, for the derivation of the maximum likelihood estimators used in the ESTMAT procedure. One example to show robustness to the normality assumption has been given [5], but no one has made a thorough study. The second assumption is unlikely to hold when the data are missing because of refusals, inaccessibles, editing, etc. The second assumption emphasizes the fact that the ESTMAT procedure seems more appropriate for survey situations in which the missing values are planned -- double sampling schemes, triple sampling schemes, etc. [3]. However, if the procedure is robust to the randomness assumption, then applying multivariate regressions seems as reasonable as applying the ratio of a ratio procedure. The data set in this study does not obey either of the two assumptions for the ESTMAT procedure.

The ESTMAT procedure was not initially designed to impute individual values but to estimate directly the mean vector of the population. However, the procedure also estimates the variance-covariance matrix, and this estimate allows the computation of multivariate regression equations which can be used to impute individual values. These imputed values lack what Pregiborn [6] calls "commutativity" with the estimated mean vector. In other words, if one averages the reported and imputed values in a data set, this average does not equal the mean estimated directly by the ESTMAT procedure. Thus, the reader must be aware that the results of the ESTMAT procedure in this study are affected by an imputation process which may not be a part of other ESTMAT applications.

2.4 The Zero Spike Procedure

The zero spike procedure takes its name from the fact that zeros often dominate the response space of many surveys -- thus resulting in a "spike" of zeros when one draws a histogram of the distribution. The data set of this study has this characteristic. The first bar in each of the graphs of Figure 1.1 represents the zeros in the data set. For $y^{(1)}$ 33 percent of the 201 original values are zeros, and for $y^{(2)}$ 38 percent of the original values are zeros.

The zero spike procedure forms an indicator vector for each unit in the sample. For each variable, there is an element in the vector. The value of this element is "0" if the value of the variable is zero, "1" if the value is positive, and "2" if the value is missing. If unit A has a "2" in its indicator vector, then the "2" is changed to a "0" or "1" using probabilities based on S -- that subset of the sample units which: 1) is complete, and 2) matches the indicator vector of unit A for those variables reported on A. For example, if there are two variables, then the complete units can form four groups -- (0,0), (0,1), (1,0) and (1,1). If a unit has the form (0,2) then the "2" is changed to a "0" with probability $\frac{n_{(0,0)}}{n_{(0,0)} + n_{(0,1)}}$ or changed to a "1" with probability $\frac{n_{(0,1)}}{n_{(0,0)} + n_{)0,1)}}$ where $n_{(i,j)}$ is the number of complete units in the (i,j) group; i, j = 0,1. If a "2" is changed to a "0", the missing value becomes zero. If a "2" is changed to a "1", the missing value becomes a positive number of the form $\hat{R}x$, where x is the most highly correlated variable which also has a "1" in the indicator vector of A. R is the ratio which relates x to y and is estimated from the units in S.

Pregiborn actually recommends the use of any, even subjective, information to estimate the probabilities for assignments of "0" and "1" and not just the use of units in the sample. Thus, his recommendations allow a Bayesian approach to the imputation through the estimation of the probabilities. Also, Pregiborn notes that there are many possible methods -- hot decks, regression, averages, etc. -- to decide what positive value to impute for a missing value. This study uses a ratio method because the first three procedures described also use a ratio or regression method in some way. Thus, in the comparisons of estimates from the procedures, any differences for the zero spike procedure are not mainly a result of the method used to determine positive values but mainly a result of the "zero-positive" structure employed.

2.5 The Princomp Procedure

This procedure uses the first principal component when imputing for missing values. The first principal component is applied as a distance measure to select the complete unit which is most like a unit with a missing value. The reported value for this complete unit is then substituted for the corresponding missing value. The first principal component is a linear combination of all reported variables and has the maximum variance of all possible linear combinations of these variables. It is the line of closest fit in the sense that it minimizes the sum of squares of distances from data points to the line (note that a regression line minimizes the sum of squares in particular directions).

For this study the princomp procedure: 1) constructs four subsets of the data -- S1 contains the complete units, S2 contains those units with the variable $y^{(1)}$ missing, S3 contains those units with the variable $y^{(2)}$ missing, and S4 contains those units with both variables $y^{(1)}$ and $y^{(2)}$ missing; 2) computes the first principal component for S2 by using all 15 variables except $y^{(1)}$ and then computes the value of the first principal component for each unit in S1 and S2; 3) for each unit in S2, finds the S1 unit which has a principal component value closest (minimum absolute deviation) to the unit in S2 and substitutes the corresponding values of $y^{(1)}$ from the S1 unit into the missing values of $y^{(1)}$ in the S2 unit; 4) repeats steps 2 and 3 to substitute reported values from S1 for missing values in S3 and S4 by using the principal component that corresponds to each subset.

The princomp procedure is essentially a hot deck procedure (a hot deck procedure is defined as a procedure which substitutes reported values for missing values) which substitutes by the minimization of a distance function rather than substituting randomly. There are many distance measures which could have been tested, but the authors felt that only one procedure of this type could be added to the experiment due to time and cost constraints and that the princomp procedure is a distribution-free method which has the potential for accurate imputation.

3. ANALYSIS

The goal of this analysis is to identify the "best" procedure of the six described procedures which impute for missing items. There are five criteria for selection of the "best" procedure: 1) the accuracy of estimated means, 2) the standard errors, 3) the accuracy of imputations on a unit level, 4) the effect on correlations between variables, and 5) costs.

3.1 Experimental Design

Three methods designate units which have missing items: 1) a random designation, 2) a 15 percent designation of incomplete units below the median and 85 percent above, and 3) an 85 percent designation of incomplete units below the median and 15 percent above. (The median of $z = y^{(1)} + y^{(2)}$ is used in these designations.) For each of these three methods, there are two rates to designate how many units have missing items -- 10 percent and 30 percent. The combined effect of the type of designation and the rate of designation results in six different situations in which means are estimated for the entire population.

Five data sets are simulated for each level of bias. Thus, a total of 30 data sets are generated from the original data set. Each data set consists of three replicates, and each procedure is run independently on each replicate to provide unbiased estimates of the standard errors. Within each data set the group of units with missing values contains 40 percent of the units with $y^{(1)}$ missing, 40 percent with $y^{(2)}$ missing, and 20 percent with both $y^{(1)}$ and $y^{(2)}$ missing.

The structure of the simulations corresponds to an analysis of variance model. If an analysis of variance shows a significant difference due to an effect, then Duncan's multiple range test is used to identify which levels of the effect caused the differences. All tests are at a five percent level of significance.

3.2 Results

An analysis of variance shows significant differences among the six missing item procedures when the dependent variable is the average difference between the imputed values and the "true" values. Table 3.2.1 gives the results of Duncan's multiple comparison test and the patterns that are characteristic of each procedure. The ratio 1 and ratio 2 procedures are usually significantly different from the other procedures but not from each other. The princomp and zero spike procedures also tend to be different from the other procedures but are not significantly different from each other. The array procedure does not show consistent trends but tends to group with the princomp and zero spike procedures. The ESTMAT procedure tends to be by itself. Apparently the ESTMAT procedure is not robust to its normality and random error assumptions because the estimated means from this procedure are not very accurate under the random designation of missing values. All procedures tend to underestimate the mean -- even when values are randomly missing. This underestimation may not only be a result of biases inherent in the procedures but also a result of the skewness in the underlying data.

Table 3.2.1: Results of Duncan's multiple range test* when the dependent variable is the average difference between the imputed value and the corresponding original value.

Variable	Designation Method					
	Random		15% Below Median/85% Above		85% Below Median/15% Above	
	Average Difference	Procedure	Average Difference	Procedure	Average Difference	Procedure
$y^{(1)}$	-0.133	Ratio 2	3.781	ESTMAT	5.781	ESTMAT
	-1.281	Array	-5.719	Ratio 2	1.315	Ratio 2
	-2.359	Ratio 1	-5.922	Ratio 1	0.041	Ratio 1
	-5.285	ESTMAT	-10.715	Array	-4.104	Zero Spike
	-5.933	Zero Spike	-14.170	Princomp	-4.337	Princomp
	-6.756	Princomp	-15.870	Zero Spike	-5.759	Array
$y^{(2)}$	-0.852	Ratio 2	-7.567	Ratio 2	3.463	Ratio 2
	-1.500	Array	-8.711	Ratio 1	1.204	Ratio 1
	-2.104	Ratio 1	-14.378	Array	-2.641	Princomp
	-5.156	ESTMAT	-17.219	Princomp	-3.552	Array
	-5.815	Zero Spike	-18.330	Zero Spike	-4.285	Zero Spike
	-6.026	Princomp	-24.748	ESTMAT	-12.189	ESTMAT

*Any two means connected by the same bracket are not significantly different at $\alpha = 0.05$.

The interaction between the designation methods and the procedures is a significant effect. However, as Table 3.2.1 shows, this significance is a result of the fluctuation of ESTMAT procedure in relation to the other procedures. The remaining tables in this paper give overall results across designation methods and rates. These overall results do not imply that the interactions are insignificant, but, as in Table 3.2.1, they are not important enough in this study to warrant the complexity of presenting the results in each cell. Table 3.2.2, for example, is much simpler and clearer than Table 3.2.1 and does not lose much information.

Table 3.2.2 gives overall results for Duncan's multiple comparison test in terms of average difference and relative bias. In this table the relative bias is the average difference in imputed and original values divided by the "true" mean of the sample. Across both variables the ratio 1 and ratio 2 procedures give the best results. It is disturbing that the ESTMAT procedure can give the best results for $y^{(1)}$ and the worst for $y^{(2)}$. This result may be an effect of the imputation part of the ESTMAT procedure since direct estimates from ESTMAT showed a relative bias of -1.3 percent and -0.2 percent for $y^{(1)}$ and $y^{(2)}$ when estimating the mean of the entire population -- a result which seems more reasonable. Thus, imputations using the ESTMAT procedure appear to be unreliable.

Table 3.2.2: Overall results of Duncan's multiple comparison test*.

Variable	Procedure	Average Difference in Imputed Values and Original Values	Effect on Mean Estimates of Entire Population (Relative Bias)
$y^{(1)}$	ESTMAT	1.426	+0.3%
	Ratio 2	-1.512	-0.3%
	Ratio 1	-2.747	-0.6%
	Array	-5.918	-1.2%
	Princomp	-8.421	-1.7%
	Zero Spike	-8.636	-1.7%
$y^{(2)}$	Ratio 2	-1.652	-0.4%
	Ratio 1	-3.204	-0.8%
	Array	-6.477	-1.3%
	Princomp	-8.629	-1.8%
	Zero Spike	-9.477	-2.0%
	ESTMAT	-14.031	-2.9%

*Any two means connected by the same bracket are not significantly different at $\alpha = 0.05$.

To judge the accuracy of the imputations at a unit level, Table 3.2.3 gives the total of the absolute differences between the imputed values and "true" values. The optimum procedure should minimize this total. Table 3.2.3 confirms the superiority of the ratio 1 and ratio 2 procedures and explains the contradictory results in Table 3.2.2 between $y^{(1)}$ and $y^{(2)}$ for the ESTMAT procedure. The ESTMAT procedure gives the lowest difference for $y^{(1)}$ in Table 3.2.2 because of offsetting extremes in positive and negative directions. Thus, when absolute differences are calculated, the ESTMAT procedure gives the largest totals for both variable in Table 3.2.3.

Table 3.2.3: Total of the absolute differences between each imputed value and the corresponding "true" value.

Variable	Procedure	Absolute Difference
$y^{(1)}$	Ratio 2	6,683
	Ratio 1	6,711
	Array	9,648
	Zero Spike	9,909
	Princomp	10,129
	ESTMAT	14,643
$y^{(2)}$	Ratio 2	7,132
	Ratio 1	7,439
	Array	9,862
	Princomp	10,873
	Zero Spike	10,922
	ESTMAT	14,137

Table 3.2.4 gives the coefficient of variation of the estimated mean for the entire population. The coefficient of variation is the standard error (an unbiased estimate calculated using replicates) for a procedure divided by the "true" mean of the sample. The coefficients of variation in Table 3.2.4 are similar in size except that the ESTMAT procedure is larger for $y^{(1)}$.

Table 3.2.4: Coefficients of variation for the estimated mean of the entire population.

Variable	Procedure	Coefficient of Variation
$y^{(1)}$	Zero Spike	0.062
	Princomp	0.063
	Array	0.065
	Ratio 1	0.065
	Ratio 2	0.070
	ESTMAT	0.100
$y^{(2)}$	Princomp	0.065
	Ratio 1	0.065
	Zero Spike	0.068
	Ratio 2	0.068
	ESTMAT	0.070
	Array	0.070

An overall measure of the quality of the procedures is the root mean square error. This measure is defined as:

$$\sqrt{MSE'} = [(\text{Relative Bias})^2 + (\text{Coefficient of Variation})^2]^{\frac{1}{2}}$$

The $\sqrt{MSE'}$ is sensitive to the sample since the sample size affects the magnitude of the coefficient of variation and sometimes the magnitude of the relative bias. Assuming, however, the relative bias is not affected by the sample size, Table 3.2.5 displays $\sqrt{MSE'}$ for several sample sizes by using the relative biases in Table 3.2.2 and the coefficients of variation in Table 3.2.4. Only for sample sizes larger than 1000 does the relative bias component dominate the root mean square error rather than the component due to the coefficient of variation. Thus, for very large sample sizes, such as those often used in government surveys, the two ratio procedures give the best results. For smaller sample sizes, however, there is little difference in the procedures except that the ESTMAT procedure is substantially larger for $y^{(1)}$.

Table 3.2.5: Root mean square error relative to the "true" sample mean.

Variable	Procedure	Sample Size				
		50 (%)	100 (%)	1,000 (%)	10,000 (%)	∞ (%)
$y^{(1)}$	Ratio 1	13.0	9.2	3.0	1.0	0.6
	Ratio 2	14.0	10.0	3.1	1.0	0.3
	Array	13.1	9.3	3.2	1.5	1.2
	Princomp	12.7	9.1	3.3	1.9	1.7
	Zero Spike	12.5	8.9	3.3	1.9	1.7
	ESTMAT	20.1	14.2	4.5	1.4	0.3
$y^{(2)}$	Ratio 1	13.0	9.2	3.0	1.2	0.8
	Ratio 2	13.6	9.6	3.1	1.0	0.4
	Array	14.1	10.0	3.4	1.6	1.3
	Princomp	13.2	9.4	3.4	2.0	1.8
	Zero Spike	13.8	9.8	3.6	2.2	2.0
	ESTMAT	14.3	10.3	4.3	3.1	2.9

When a data set contains imputed values, estimates of standard errors are often calculated by ignoring the imputation process and treating the imputed data set as though all the values are reported. This method may lead to biases in the estimates of standard errors. Table 3.2.6 gives the ratio of the variance calculated by using the conventional formula and the variance calculated by using replicates. Table 3.2.6 shows there can be large biases in either direction.

Table 3.2.6: Ratio of estimated variances of estimated means -- variance estimate assuming imputed values are reported values divided by unbiased variance estimate using replication.

Variable	Designation Method	Imputation Procedure					
		Ratio 1	Ratio 2	Array	Zero Spike	Princomp	ESTMAT
$y^{(1)}$	Random	0.922	1.049	0.967	0.867	0.806	1.057
	15% Below Median/ 85% Above	1.084	1.172	0.970	0.902	0.746	1.316
	85% Below Median/ 15% Above	1.283	1.242	1.172	1.217	1.103	1.387
	Overall	1.096	1.154	1.036	0.995	0.885	1.253
$y^{(2)}$	Random	0.889	0.961	1.226	0.933	0.819	1.009
	15% Below Median/ 85% Above	0.869	0.933	0.980	1.063	0.838	1.136
	85% Below Median/ 15% Above	1.244	1.325	1.233	1.330	1.262	1.144
	Overall	1.000	1.073	1.146	1.109	0.973	1.096

Another important aspect of imputation is the effect on the correlation structure of the data set. Although correlations are not important for estimates of univariate statistics such as means and standard errors, correlations are important when the data set is used to explore and assess relationships among variables through regression analysis, principal components, or other multivariate techniques. Table 3.2.7 gives an example of the effects of the missing item procedures on the correlation structure. This table shows the correlations between w and $y^{(1)}$ and between w and $y^{(2)}$. Most of the procedures tend to lower the correlations, but the ratio 1 and ratio 2 procedures tend to inflate the correlations.

Table 3.2.7: Correlations between w and $y^{(1)}$ and between w and $y^{(2)}$ for six missing item procedures.

Procedure	Variable					
	$y^{(1)}$			$y^{(2)}$		
	Random	15% Below Median/ 85% Above	85% Below Median/ 15% Above	Random	15% Below Median/ 85% Above	85% Below Median/ 15% Above
(Actual)	*.82*	*.79*	*.84*	*.81*	*.72*	*.77*
Ratio 1	.97	.94	.97	.89	.66	.93
Ratio 2	.88	.86	.94	.89	.72	.94
Array	.57	.73	.83	.54	.62	.76
Zero Spike	.80	.53	.74	.67	.42	.62
Princomp	.72	.60	.77	.68	.28	.74
ESTMAT	.79	.59	.81	.72	.33	.20

The cost of each procedure for imputing data is shown in Table 3.2.8. This cost is based on imputation for all 30 data sets for the two variables $y^{(1)}$ and $y^{(2)}$. The system resource units (SRU's) -- a measure of computer usage -- required by each procedure are reasonably close except for the ESTMAT procedure. ESTMAT requires more SRU's than the other five procedures combined. This requirement is because of the complexity of the procedure. Thus, cost alone imposes a severe restriction on the use of the ESTMAT procedure. The other five imputation techniques are very similar in cost with the ratio 1 and ratio 2 procedures costing the least.

Table 3.2.8: Processing costs of six missing item procedures.

Procedure	SRU's [1/]	Cost [2/]
Ratio 1	937	$ 145
Ratio 2	938	$ 150
Array	1278	$ 205
Zero Spike	1215	$ 194
Princomp	1103	$ 177
ESTMAT	9604	$1537

[1/] SRU: System resource unit

[2/] Cost projected at 16¢ per SRU

4. SUMMARY

This study serves as an example of using a simulation experiment to make a preliminary assessment of the impact of imputation procedures on a specific survey. The scarcity of theory and guidelines about imputation procedures in the statistical literature causes the need for simulation experiments which assess the effects of these procedures. The preliminary nature of the simulation experiment in this study deserves emphasis. For large scale, repetitive surveys the objective of a simulation study is to winnow the many possible procedures and their variations to a few procedures. Further research under actual survey conditions, efficient computer programming, and other more costly requirements should be used to narrow these few procedures to one operational procedure.

This study compares the effects of six procedures which impute for missing items -- two versions of the ratio procedure, the array procedure, the ESTMAT procedure, the zero spike procedure, and the princomp procedure. The comparison of these procedures is from an experiment in which a complete data set has values deleted to simulate an incomplete data set.

The two versions of the ratio procedure perform the best for very large sample sizes (at least as large as 1000). For smaller sample sizes all of the procedures except the ESTMAT procedure have approximately the same mean square errors. The main disadvantage of the ratio procedure is an inflation of the correlations between variables in the data set.

The ESTMAT procedure emerges as the least attractive procedure because it does not impute very accurately and it has an extremely high cost relative to the other procedures. This result only applies to the ESTMAT procedure as in imputation process and not as a missing data procedure in general. For example, the ESTMAT procedure is probably a suitable method for sample designs in which missing data is planned -- in other words, a survey design in which one plans to collect only partial information on some designated units.

Finally, the reader is cautioned that the results of the study are based on one data set. The variables on this data set have skewed distributions dominated by zero values. These distributions are characteristic of much survey data but not all. Generalizations must wait until other survey organizations supply and compare procedures on their own data so that an empirical body of knowledge about imputation procedures can be created.