AN EMPIRICAL STUDY COMPARING SEVERAL METHODS OF HANDLING MISSING DATA

Fred E. Cromer

University of Alaska, Anchorage, Alaska 99504, *U.S.A.*

الخلاصة :

في هذا البحث إقترحنا إستعمال طرق مونت كارلو للتعامل مع مشكلة البيانات المفقودة . لذا ذكرنا أمثلة معينة وعرضنا دراسة تجريبية تم فيها مقارنة عدة طرق أخرى . وكانت النتيجة أننا وجدنا أن الطريقة المثلى للتعامل مع مشكلة البيانات المفقودة هو أن نحذف من التحليلات الحالات ذات القيم المفقودة لمتغير ما بالاضافة الى طرق مونت كارلو للانحدار العشوائي . .

ABSTRACT

The use of Monte Carlo procedures to handle the missing data problem is suggested. Specific examples are stated and an empirical study performed in which several replacement methods are compared. Omitting cases with missing values from analyses involving that particular variable and a Monte Carlo technique of regression–random replacement are found to be the best methods for dealing with the missing data.

0377-9211/81/010041-07$01.00

AN EMPIRICAL STUDY COMPARING SEVERAL METHODS OF HANDLING MISSING DATA

INTRODUCTION

In many projects data are likely to be missing: in a study of educational achievement, a pupil might not take the mathematics test but other data might be collected. A sample point with missing data could be discarded, but this could lead to the loss of information from good data that were collected at considerable expense. As a result, missing data are sometimes estimated using means, regression estimates, or estimates derived using the technique of principal components. Gleason and Staelin [1] reviewed several of these methods and concluded the principal components approach was superior while Frane [2] suggested several simple regression approaches for estimating missing values. The review by Gleason and Staelin [1] indicated that the regression and principal components approaches yielded similar results, with the latter method being less expensive.

Quite properly, these authors note that the various procedures for estimating missing data depend upon the assumptions that the data are missing at random and that the amount of missing data is not excessive. The regression and principal components techniques also require the variable with missing values to have at least a moderate correlation with some combination of the available variables. If any of these assumptions are seriously violated, the above procedures for handling missing data are likely to be suspect.

It should be noted that the major reason for trying to replace missing data by some estimate is to use all available information rather than throwing away or not using data that may have been expensive and time consuming to collect. In some cases replacing missing data also helps to balance a design, leading to tests that are somewhat less sensitive to minor violations of assumptions about normality and homoscedasticity. However, it should be noted that one does not gain back the degrees-of-freedom lost when missing data occur. Since these missing values have been estimated and replaced by a mathematical constant rather than a stochastic variable, the various statistical tests must take this into account.

It is frequently the case that missing values are estimated in an early, data cleaning stage of a project. In fact, on large projects this may be done months before any analyses are performed, by individuals who may or may not be involved in the analysis stage of the project. When it is time to analyze the data, the values in the data base are very likely to be treated as 'random' points, with the calculated degrees-of-freedom returned by the various statistical analysis packages being incorrect (i.e. too large).

In a study where the question of the effect of missing data arises, one is generally concerned because the data are *not* missing in a random fashion, and the above techniques will consequently be suspect. The concern is usually that there is some systematic factor related to attrition, nonresponse, etc.; which will in some way bias the results of the study.

In such instances, it may be of some interest to conduct a substudy to determine whether, on the basis of available information, the data subset containing complete information differs in any significant way from the full data set. That is, on various subsets of variables having complete information across both data sets, could these two data sets be considered random samples from the same population? If so, various of the previously discussed procedures could be used with some degree of confidence and empirical evidence would be available to support their use. An alternative would be to analyze the data twice; first, omitting sample points with missing data and secondly including all sample points with estimates included where necessary. Any substantial differences in the results of these two analyses would be reported. Thus any nonrandom or systematic factor related to the missing data might be noted.

The problem to be addressed is what method might best be used to produce the estimates used to replace missing data. Two Monte Carlo procedures are suggested and their use is compared with other commonly used replacement techniques.

MONTE CARLO REPLACEMENT PROCEDURES

One possible alternative for handling the missing data problem would be to use a Monte Carlo approach for replacing the missing values. There are many approaches to this method, the limitations being only the ingenuity of the researcher. Two alternatives are outlined below.

1. Blocked Random

Partition the data set across variables that are related to the variable with the missing values (V_m) which are to be replaced. Estimate the distribution of V_m within each block of the data set using the available values of V_m. Then for each missing value, identify the block to which the observation belongs and randomly generate a value from the estimated distribution associated with that block. This value will serve as the replacement for the missing value.

2. Regression Random

Choose some subset of variables to serve as predictors for V_m. Construct a regression model using this subset of predictors V_m, and then for each missing value, calculate its regression estimate. Use this estimate and the standard error of regression to establish a hypothetical distribution from which a random value is selected. This random value will serve as the replacement for the missing value.

The blocked random method might be viewed as a generalization of the method of replacement by the mean, but rather than using the center of the distribution as a replacement, a value is randomly selected from the distribution to serve as a replacement. Similarly, the regression random approach can be viewed as a generalization of the method of replacement by a regression estimate.

These random replacement methods have some intuitive appeal since it seems that they should do a better job of reproducing the variability in the data than a method which replaces the missing value with a fixed number. As an added bonus, if the amount of missing data is sizable, these methods allow one to 'buy back' most of the lost degrees-of-freedom (df) which result when data are missing. Since the replacements are random, they may be treated as if they are part of the 'random sample' and the only df lost are the number of parameters estimated in the construction of the hypothetical distributions. For example, in the Blocked Random case, if means and variances were estimated for each of 4 blocks, a total of a 8 degrees-of-freedom (df) are lost. If only proportions were estimated, only 4 df are lost. Thus, rather than losing 100 df when there are 100 missing values, only 8, or 4, df are lost. Similarly for the Regression Random case, it is noted that only the means, variances, and intercorrelations are needed to produce the regression model. Consequently, if k variables are used in the model to predict V_m, a total of

$$(k+1)+(k+1)+\binom{k+1}{2}=\frac{(k+1)(k+4)}{2}\ \text{df}$$

are lost.

Although these methods have an intuitive appeal, it is of interest to determine, in an actual example, how these methods compare with the more traditional methods of omission, mean replacement, and regression replacement. For this reason an empirical study was performed in which such comparisons were made.

EMPIRICAL COMPARISONS

In the interest of making the problem realistic, an example was prepared of a study where the investigators are interested in determining the impact of socioeconomic status (SES) upon mathematics achievement in a given school setting. The variables in the data set included initial achievement level (PRE), SES, and final achievement level (POST). In such a one-year program, some attrition is likely to occur and no posttest score will be available for those children. The question of interest is whether a reasonable method for handling these missing data can be found.

Three data sets were generated, each with $n=1000$. For the variables PRE, SES, and POST the first data set contained high intercorrelations (HI CORR), the second, moderate intercorrelations (MED CORR), and the third, low intercorrelations (LO CORR). These three data sets represented the 'populations' whose parameters were to serve as the basis for comparison for the various missing data strategies.

To generate uniformly distributed random numbers, a locally written generator based upon the linear congruential method [3, pp. 9–24] was used. Standard normal deviates were generated using the polar method [3, p. 104]. Normally distributed variates having specified means, variances, and intercorrelations were generated using the independent standard normal deviates produced by the random normal deviate generator and the transformations outlined by Knuth [3, p. 113].

The full data sets, each of which included 1000 records, were generated to yield approximately 10% of the cases in SES Class 1, 20% in Class 2, 35% in Class 3, 25% in Class 4, and 10% in Class 5. The pretest score was generated to be approximately normal with mean 42 and variance 196 while the posttest had mean 51 and variance 169. One of the data sets had high intercorrelations, the second had moderate in-

Table 1. Description of the Full Data Sets
(SES = 1, PRE = 2, POST = 3)

	HI CORR	MED CORR	LOW CORR
$\bar{x}_1$	3.067	3.099	3.093
$\bar{x}_2$	42.189	42.565	42.040
$\bar{x}_3$	51.273	51.001	50.888
s_1	1.3492	1.2988	1.3199
s_2	13.9558	14.5478	15.1685
s_3	13.0353	13.6236	13.5036
r_{12}	0.77215	0.37587	0.08936
r_{13}	0.65976	0.16897	0.00244
r_{23}	0.86763	0.44282	0.11104

tercorrelations, and the last had low intercorrelations among the variables. Table 1 describes these three data sets. Variable SES, is designated by subscript 1, PRE by 2, and POST by 3. The $\bar{x}$'s represent the means, the s's the standard deviations, and the r's the intercorrelations among the variables.

Two methods of generating missing data were used (for each of the above data sets). First, a uniformly distributed random variable was generated for each case; if this number was less than 0.1 the value for the POST variable was deleted so that approximately 10% of the data were missing. Second, values of POST were deleted so that a much higher percentage of low SES possttests were missing, but an overall 10% missing was again the target. Again each record was considered. If SES = 1 and a uniformly distributed random variable was less than 0.2, POST was deleted. Thus approximately 20% of the students in the lowest SES group did not have a posttest score. In a similar manner missing data were generated in approximately 15% of the cases where SES = 2, in 10% of the cases where SES = 3, and 4% of the cases where SES was 4 or 5. Thus, two data sets containing missing data were generated from each of the three full data sets; one set having data missing at random and the other having data which were selectively missing.

The following list describes the methods of handling missing data which were investigated:

1. OMIT—Cases having missing values were dropped for any calculations involving that variable.
2. MEAN—The missing value was replaced by the mean of the available values for that variable.
3. REGR—The missing value was replaced by its regression estimate, in this case: $\text{POST} = a_0 + a_1 * \text{SES} + a_2 * \text{PRE} + a_3 * \text{SES} * \text{PRE}$.
4. BR—This is the Blocked Random replacement method outlined earlier. In this case the data were blocked by SES. The mean and variance of POST were calculated for each of the five levels of SES. The values of POST were assumed to be normally distributed within each block so if a missing value was detected, it was replaced by a random value from the appropriate distribution.
5. RR—This is the Regression, Random replacement method which was also outlined earlier. Here the model
$\text{POST} = a_0 + a_1 * \text{SES} + a_2 * \text{PRE} + a_3 * \text{SES} * \text{PRE}$,
was constructed from the available data. Then each missing value was replaced by a random value from an assumed normal distribution with mean equal to the regression estimate of POST and standard deviation equal to the standard error of regression.

First an analysis was performed to see whether the group without posttest scores differed significantly on the other variables from those with posttest scores. If there are no differences, it would be reasonable to assume that the data were missing at random and analysies based on the data set with cases having missing values omitted would yield results similar to what would have been found if the full data set had been available. Differences would indicate that the data were not missing at random and may indeed influence the results of later analyses. Analyses revealed (as expected) that there were no differences between the groups, for the data sets with values deleted at random, but there were significant differences in average SES and PRE between the groups where the data was not deleted at random.

The statistics calculated for each of the above five cases were compared to their corresponding parameters in the full data set. The statistics considered were the means, variances, intercorrelations, multiple R-square, and B-coefficients for the regression model listed in case 5 above. Two basic comparisons were made. First, the percentage error for each statistic was calculated (PE = ((parameter-statistic)/parameter) * 100)), and second, a test was conducted to determine whether the statistic differed significantly from the parameter of the full data set. Tables 2–7 summarize these comparisons. The symbols $\bar{x}$, s, and r represent the mean, standard deviation, and correlation as in Table 1. R^2 represents the multiple R-square, and the B's represent the regression coefficients.

Several ways of further summarizing and condensing the information in Tables 2–7 were considered. Simple counts of the number of significant differences for various breakdowns of the tables were performed. For each statistic, the percentage errors

Table 2. Percentage Errors (HI CORR), Non-random Missing (N = 901)

Statistic	Replacement Method				
	OMIT	MEAN	REGR.	BR	RR
$\bar{x}_3$	0.9	0.9	0.2	0.1	0.1
s_3	0.4	4.7†	1.1	0.1	1.1
$r_{1,3}$	0.4	5.7†	1.2	0.2	0.1
$r_{2,3}$	0.3	4.6†	1.4	4.6†	0.4
R^2	0.5	8.6†	2.9	8.8†	0.7
B_0	1.1	37.9	1.1	25.4	4.4
B_1	15.5	466.5†	15.6	109.7	13.8
B_2	1.2	17.4†	1.2	79.7†	3.2
B_{12}	88.3	4478.3†	88.3	4125.0†	478.3

SES = 1, PRE = 2, POST = 3, SES*PRE = 12

†The calculated Statistic Differs Significantly from the Parameter ($P<0.05$). B_0 was not Tested.

Table 3. Percentage Errors (HI CORR) Random Missing (N = 893)

Statistic	Replacement Method				
	OMIT	MEAN	REGR.	BR	RR
$\bar{x}_3$	0.1	0.1	0.0	0.2	0.1
s_3	1.1	4.5	0.6	2.0	0.7
$r_{1,3}$	2.1	3.4	2.1	2.7	0.4
$r_{2,3}$	0.1	5.0†	1.3	4.5†	1.1
R^2	0.1	9.8†	2.6	8.3†	2.4
B_0	2.6	12.5	2.6	6.6	0.4
B_1	82.4	176.9	82.4	397.3†	5.9
B_2	0.3	9.5†	0.3	13.8†	0.7
B_{12}	371.7	871.7	371.7	371.7	360.0

SES = 1, PRE = 2, POST = 3, SES*PRE = 12

†The Calculated Statistic Differs Significantly from the Parameter ($P<0.05$). B_0 was not Tested.

Table 4. Percentage Errors (MED CORR), Non-random Missing (N = 911)

Statistic	Replacement Method				
	OMIT	MEAN	REGR	BR	RR
$\bar{x}_3$	0.6	1.1	0.1	0.0	0.1
s_3	0.8	5.3†	4.6†	1.1	0.6
$r_{1,3}$	2.6	2.9	6.8	1.8	3.8
$r_{2,3}$	0.9	3.2	3.8	3.1	3.2
R^2	3.0	3.9	9.0	3.8	6.7
B_0	5.2	13.5	5.2	10.1	0.3
B_1	47.0	92.1†	47.0	81.0†	8.6
B_2	16.4	36.7†	16.4	31.5†	0.5
B_{12}	46.8†	82.8†	46.8†	80.3†	8.8

SES = 1, PRE = 2, POST = 3, SES*PRE = 12

†The Calculated Statistic Differs Significantly from the Parameter ($P<0.05$). B_0 was not Tested.

Table 5. Percentage Errors (MED CORR), Random Missing (N = 891)

Statistic	Replacement Method				
	OMIT	MEAN	REGR	BR	RR
$\bar{x}_3$	0.0	0.4	0.1	0.1	0.1
s_3	0.3	5.9†	4.7†	1.0	0.4
$r_{1,3}$	2.7	3.7	6.0	0.9	3.1
$r_{2,3}$	0.0	5.6	5.1	13.0†	1.4
R^2	0.9	10.9	11.5	24.0†	1.4
B_0	4.0	4.3	4.0	1.2	6.5
B_1	42.3	12.7	42.3	60.0†	54.9
B_2	12.2	12.1	12.2	8.4	20.9
B_{12}	39.5†	9.6	39.5†	33.5†	56.4†

SES = 1, PRE = 2, POST = 3, SES*PRE = 12

†The calculated Statistic Differs Significantly from the Parameter ($P<0.05$). B_0 was not Tested.

were ranked across the various replacement methods, again for various breakdowns of the tables. Median percentage errors were determined for each method. Counts of the number of percentage errors under 1.0, under 2.0, under 5.0, and under 10.0 were performed. Sign tests were conducted to determine whether the pattern of percentage errors for various pairs of replacement method differed significantly. Rather than actually tabulate the results of these various summaries, which would produce a voluminous set of tables, all of which can be derived from Tables 2–7, the results will merely be reported.

The five replacement methods were fairly consistently ranked with OMIT and RR being clearly superior, REGR a close third, BR a distant fourth, and MEAN a much more distant fifth, especially for data sets with moderate or high correlations.

The full regression model $\text{POST} = a_0 + a_1 * \text{SES} + a_2 * \text{PRE} + a_3 * \text{SES} * \text{PRE}$, and all of its possible restricted models were also constructed for each of the six data sets. Any patterns of differences from the models constructed for the full data set were noted. Again, the results were entirely consistent with, and led to the same conclusions as, the summaries of Tables 2–7. Of a total of 40 possible significant component differences, the observed differences ranged from 1 for RR to 7 for BR.

Since the OMIT and RR were the best techniques

Table 6. Percentage Errors (LO CORR), Non-random Missing (N = 896)

Statistic	Replacement Method				
	OMIT	MEAN	REGR	BR	RR
$\bar{x}_3$	0.3	0.3	0.3	2.8	0.2
s_3	0.3	5.1†	5.0†	1.1	0.5
$r_{1,3}$	464.3	431.1	434.0	1215.2	262.3
$r_{2,3}$	1.7	3.9	7.0	25.9	21.1
R^2	4.2	6.0	15.4	29.1	34.0
B_0	0.3	0.6	0.3	4.0	2.9
B_1	1.2	0.1	1.2	47.2	33.3
B_2	12.4	30.3	12.4	125.1†	55.1
B_{12}	20.8	21.4	20.8	118.4†	29.8

SES = 1, PRE = 2, POST = 3, SES*PRE = 12

†The Calculated Statistic Differs Significantly from the Parameter ($P<0.05$). B_0 was not Tested.

Table 7. Percentage Errors (LO CORR), Random Missing (N = 898)

Statistic	Replacement Method				
	OMIT	MEAN	REGR	BR	RR
$\bar{x}_3$	0.2	0.2	0.3	1.9	0.3
s_3	0.2	5.1†	5.0†	3.7	0.2
$r_{1,3}$	424.6	408.6	545.9	103.7	237.7
$r_{2,3}$	10.5	5.7	15.6	2.7	3.7
R^2	23.8	13.1	35.3	12.1	4.4
B_0	0.4	0.4	0.4	0.7	2.6
B_1	7.3	7.5	7.3	83.5	68.3
B_2	30.7	8.8	30.7	48.7	51.0
B_{12}	18.7	10.4	18.7	89.5†	67.7

SES = 1, PRE = 2, POST = 3, SES*PRE = 12

†The Calculated Statistic Differs Significantly from the Parameter ($P<0.05$). B_0 was not Tested.

found for handling missing data and it was not clear from the available evidence which was superior, three new data sets were constructed to study further these two techniques. A similar procedure was used to generate the missing values. This time each of the HI, MED, and LO correlation data sets were examined. A missing rate of 40% for students in the SES = 1 group was used, 30% when SES = 2, 20% when SES = 3, 10% when SES = 4, and 5% when SES = 5 for a total of approximately 20% missing heavily concentrated at the low SES levels. Analyses similar to those previously discussed were carried out with the results still being inconclusive. RR was slightly better for the data sets with moderate-to-high correlations and OMIT was slightly better for the low correlation data set, but a sign test on the patterns of percentage errors was nonsignificant.

CONCLUSIONS

Based upon the results of this study, omitting missing data from analyses involving that variable (OMIT) appears to be the best technique if one were able to choose only one method. It is certainly least expensive in terms of computing time and appears to do the best job in data sets with low correlations or with data missing at random.

However, if the data are not missing at random and moderate-to-high intercorrelations exist, the regression random (RR) approach appears to be slightly better. This is especially true if a sizable percentage of the data points contain missing values. Then, this method has the added bonus of regenerating some of the degrees-of-freedom lost by the missing values.

The regression replacement approach (REGR) was also a good technique, although not as good as OMIT or RR. The blocked random (BR) technique is useful if the blocks are chosen such that the blocking variables are highly related to the variable with missing values. In general the technique of replacing missing values by means (MEAN) gave much poorer results.

This empirical study only addresses the problem of missing data in the criterion variable. The techniques discussed can be used just as easily for imputation of missing values for one (or more) predictor variables; however, the question of how well the data sets are reproduced in those cases has not been addressed. This remains the subject of future inquiry along with such questions as those relating to the effects of greater percentages of missing data, the effects of different patterns of missing data among predictors, and the effects of data missing for various reasons (e.g. refusal to answer, ambiguity, lack of required knowledge, absence at testing, etc.).

In summary, it appears that the use of Monte Carlo procedures for imputation or for estimating the effects of missing data is not only conceptually appealing, but also backed by empirical evidence. Obviously, this approach is limited only by the ingenuity and resources of the investigator. To perform many such reanalyses would become expensive, but if the alternative is to be able to draw no conclusions at all from a study, these reanalyses may be called for. The

ultimate concern would, of course, be whether such extra work would be cost-effective in terms of time, money, and potential impact of the study.

REFERENCES

[1] T. C. Gleason and R. Staelin, 'A Proposal for Handling Missing Data', *Psychometrica*, **40** (1975), pp. 229–252.

[2] J. W. Frane, 'Some Simple Procedures for Handling Missing Data in Multivariate Analysis', *Psychometrica*, **41** (1976), pp. 409–415.

[3] E. E. Knuth, *The Art of Computer Programming, V. 2/Seminumerical Algorithms*, Addison-Wesley, 1969.

Paper Received 10 July 1979; Revised 26 January 1980.