DOCUMENT RESUME

ED 339 755 TM 017 688

AUTHOR	Witta, Lea; Kaiser, Javaid
TITLE	Four Methods of Handling Missing Data with the 1984 General Social Survey.
PUB DATE	Nov 91
NOTE	12p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (20th, Lexington, KY, November 12-15, 1991).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Evaluation Methods; Predictive Measurement; Regression (Statistics); *Research Methodology; Sampling; *Surveys
IDENTIFIERS	*General Social Survey; *Missing Data

ABSTRACT

When survey data are statistically analyzed, many times some of the data is missing. If the missing values are not correctly handled, results of the analysis may be dubious and publication may jeopardize the credibility of the organization preparing the report. This study examined four of the more commonly used methods of handling missing data. The following techniques were compared: (1) listwise deletion; (2) pairwise deletion; (3) mean substitution; and (4) regression imputation of missing data. Comparisons were made using a sample selected from the General Social Survey--1984 of the National Opinion Research Center. The sample of 829 cases was randomly divided into two sample groups: Sample 1, with 415 cases; and Sample 2, with 414 cases, which was reduced to only non-missing cases at 283. Sample 1 was used to develop regression equations after treatment by each technique. Sample 2 was used to compare the efficiency of these regression equations in predicting the criterion variable by comparing the actual criterion mean to the predicted mean using Dunnett's test for contrasts. There was a statistically significant difference between the actual mean and the mean predicted by mean substitution with the significance level at 0.01. The other methods exhibited no significant differences. Mean substitution appears inappropriate as a way of handling missing data. A seven-item list of references is included. Three data tables are provided. (SLD)

FOUR METHODS OF HANDLING MISSING DATA WITH THE 1984 GENERAL SOCIAL SURVEY

Lea Witta and Javaid Kaiser
Virginia Polytechnic Institute & State University
Department of Educational Research

Presented at the annual meeting of MSERA,
November 14, 1991, in Lexington, KY.

Correspondence should be sent to Lea Witta, Rt 3, Box 124, Abingdon, VA 24210.

Running Head: MISSING DATA

Abstract

When survey data are statistically analyzed, many times some of the data is missing. If the missing values are not correctly handled, results of the analysis may be dubious and publication may jeopardize the creditability of the organization preparing the report (Little & Smith, 1983).

This study is an examination of four of the more commonly used methods of handling missing data. The techniques of listwise deletion, pairwise deletion, mean substitution, and regression imputation of missing data were compared using a sample selected from the General Social Survey - 1984 (NORC). This sample was randomly divided into two samples. Sample 1 was used to develop regression equations after treatment by each technique. Sample 2 was used to compare the efficiency of these regression equations in predicting the criterion variable by comparing the actual criterion mean to the predicted mean using Dunnett's test for contrasts.

There was a statistically significant difference between the actual mean and the mean predicted by mean substitution with the significance level at 0.01. The other methods exhibited no significant differences.

When data is statistically analyzed, many times there are missing values. As a result of these, orthogonal designs may become correlated; unbiased regression coefficients may develop bias; or subject loss reduces the sensitivity of a statistical test to detect changes. In such cases, analysis and publication of the data may be of dubious value and may jeopardize the credibility of the organization preparing the report (Little & Smith, 1983).

Many solutions to the missing data problem have been proposed, but no one answer has been found. This study examines randomly missing data in a sample taken from the General Social Survey - 1984 (NORC), using four of the most common methods of handling missing values: listwise deletion, pairwise deletion, mean substitution, and regression. The purpose is to determine which method will produce a regression equation that is most efficient in predicting the criterion variable. It is hypothesized that there will be a difference in the effectiveness of the four methods.

Listwise Deletion

Listwise deletion, a common solution to the missing data problem, is the default option in several computer programs (ie, LISREL, SPSS, NCSS). This method discards cases with a missing value on any variable and thus is very wasteful of data. If, however, values are missing randomly, the regression coefficients are unbiased (Anderson, Basilevsky, & Hum, 1983). The loss of cases, however, results in a loss of error degrees of freedom yielding a loss of statistical power and a larger standard error (Cohen & Cohen, 1983).

Pairwise Deletion

Pairwise deletion computes co-variances between all pairs of variables having both observations eliminating only data that is missing for one of the two variables. Means and variances are computed on all available observations. This method is based on the assumption that the use of the maximum number of paired points and individual observations will yield better estimates of the relationship between the pairs and will produce estimates of the mean and variance that are more satisfactory than if they were excluded (Anderson et al, 1983).

Pairwise deletion and listwise deletion rely on the assumption of randomness of missing values. If this assumption is not met, listwise deletion may eliminate sub-populations and the covariance/correlation matrices produced by pairwise deletion may not be symmetric and may contain impossible values.

Mean Substitution

Mean substitution fills in a variable´s missing values with the mean of the observed variables. If the sample is normally distributed, the sample variable mean is the optimal estimate of that variable's most probable value. This method does not alter the sample mean, but artificially reduces the variance for the treated variable resulting in a reduction in the levels of association between the variables (Anderson et al. 1983; Gleason & Staelin, 1975).

Regression

Buck (1960) was first to use regression methods to estimate missing values. His procedure uses listwise deletion to produce an initial correlation matrix. Then each variable with missing values is treated as a dependent variable and regressed on the non-missing variables. The resulting equations are used to produce estimates for the missing values for each variable and these estimates are inserted into the incomplete data set simultaneously.

Chan and Dunn (1972) used a variation in which a variable with some missing values was regressed on all other non-missing variables. An estimate for that variable was computed, inserted into the data set, and used in the computations to replace the other missing values.

Method

Data Source

Data used in this study was obtained from the 1984 General Social Survey (NORC). There were 1473 cases in the initial sample.

Procedure

Selection of the variables to be used in this study was begun by choosing a criterion variable. The remaining variables in the initial sample were entered in stepwise regression. Seven of these variables contributed significantly to the criterion. Variables chosen are:

EDUC	Highest year of school completed (criterion variable)
SPEDUC	Spouse's highest year of school
AGEWED	Age of first marriage
PAEDUC	Father's highest year of school
SIBS	Number of siblings
MAEDUC	Mother's highest year of school
SEX	Respondent's sex
HEALTH	Condition of health

Since SPEDUC was the most significant indicator of EDUC, unmarried respondents were excluded. This left a total sample size of 829 and a criterion variable with no missing values.

This sample was split into two randomly selected groups: Sample 1 with 415 cases and Sample 2 with 414 cases. Sample 2 was reduced to only non-missing cases yielding a sample size of 283. It was assumed that if the values are missing randomly, this reduced sample is a random sample of the larger one. It was then reserved to be used in testing the regression equations developed after using the missing data methods.

The missing values in Sample 1 were treated by listwise deletion, pairwise deletion, mean substitution, and the regression variation suggested by Chan and Dunn (1972). The criterion variable was not used in this method.

After treatment by a missing value method, a regression equation was developed from the resultant matrix using the criterion variable. This produced four regression equations to be used in comparison - one for each missing value method.

Five random samples of 25 cases and five of 50 were selected from the reserved sample. The regression equation developed by each missing value method in Sample 1 was used to predict the criterion variable for each sample.

Comparison of Methods

Dunnett's test for contrasts was chosen to evaluate the efficiency of the regression equations developed by each missing data method. This test controls for the probability of false rejection using the experimentwise error rate and does not require an overall significance test prior to testing the planned comparisons (Kirk, 1983). The mean of the predicted criterion variable for each method was compared to the actual mean using this test.

Results

The only significant differences in results as shown by Dunnett's contrasts occur in the mean substitution method as is illustrated in the tables below.

TABLE I
Sample Size = 25

Method	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Listwise	.73	.00	.20	.01	.38
Pairwise	.74	-.02	.20	.00	.38
Mean Subs	-.25	-1.04*	-.87*	-1.06*	-.64
Reg Variat	.61	-.18	.04	-.18	.23
Dunnett Critical Val	.77	1.01	.69	.83	.81

TABLE II
Sample Size = 50

Method	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
Listwise	-.02	.45	.27	-.07	.36
Pairwise	-.05	.01	.26	-.09	.34
Mean Subs	-1.14*	-.63*	-.73*	-1.15*	-.72*
Reg Variat	-.24	.26	.10	-.27	.15
Dunnett Critical Val	.58	.56	.50	.53	.60

* Significant at a=.01

These results would indicate that as sample size grows and Dunnett's test becomes more powerful, mean substitution would show more deviation. What is more surprising is the number of times that this method shows significant differences with a small sample. With small samples of 25 there was a significant difference between mean substitution and the 'true' values in 3 of 5 tests. When the sample size was increased to 50, all tests showed a significant difference.

Based on this small sample, mean substitution is the most inappropriate way to handle missing values when the objective is to determine the value of the criterion variable. The other methods are similar.

Discussion

While conducting this study, it was also found that listwise deletion, pairwise deletion, and regression were not pr^ducing the same results in other areas. For example, as the regression equation for each method was developed, the following results were observed and are reported in the table below.

TABLE III

Method of Handling Missing Values	R^2	Standard Error	Sample Size
Listwise	.47830	2.11511	293
Pairwise	.49093	2.24368	300
Mean Substitution	.46098	2.30329	415
Regression Variat	.52471	2.1654	415

The standard error and variance accounted for by each method using the same variables differs. Naturally an increased sample size will increase the variance explained and reduce the standard error. Excluding mean substitution, pairwise deletion has the largest standard error, but not the smallest sample size.

In addition, after 'SPEDUC' is entered, the relative significance of the other variables (as measured by 't') is not uniform. No two equations would rank the variables in the same order. Stepwise regression does not enter the same second variable. When the variables are forced to enter in the same order, the variance accounted for by individual variables differs according to the method used.

Conclusions

These results would indicate there are differences based on the method used to handle missing values. The only question addressed in this study was the efficiency of the regression equation produced by use of each method to predict the cri^erion variable. Based on this study, if the research question is answered by predicting a criterion variable, these differences do not affect the effectiveness of the regression equations produced by listwise deletion, pairwise deletion or regression. Further research is needed to determine the influence of methods of handling missing values on predictor variab.es, variance accounted for, and standard error of these variables.

References

Anderson, A.B., Basilevsky, A. & Hum, D. P. J. (1983). Missing data: A review of the literature. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds), Handbook of survey research (pp 415-494). San Diego: Academic Press.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, 22, 302-306.

Chan, L.S. & Dunn, O.J. (1972). The treatment of missing values in discriminant analysis--1. The sampling experiment. Journal of the American Statistical Association, 67, 473-477.

Cohen, J. & Cohen, P. (1975). Missing data. In J. Cohen & P. Cohen, Applied multiple regression/correlation analysis for the behavioral sciences (pp. 265-290). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Gleason, T.C. & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40, 229-251.

Kirk, R.E. (1982). Multiple comparison tests. In Experimental design: Procedures for the behavioral sciences (2nd ed., pp. 90-127). Belmont, CA: Brooks/Cole.

Little, R.J.A., & Smith, P.J. (1983). Multivariate edit and imputation for economic data. Proceedings of the Section on Survey Research Methods American Statistical Association 1983. 518-522.