DOCUMENT RESUME

ED 389 727 TM 024 266

AUTHOR Witta, E. Lea
TITLE Are Values Missing Randomly in Survey Research?
PUB DATE Nov 94
NOTE 25p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Nashville, TN, November 9-11, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Chi Square; *Longitudinal Studies; *National Surveys; *Regression (Statistics); *Research Methodology; Simulation
IDENTIFIERS *Missing Data; National Education Longitudinal Study 1988; *Random Variables

ABSTRACT

Many missing data studies have simulated data, randomly deleted values, and investigated the method of handling the missing values that would most closely approximate the original data. Regression procedures have emerged as the most recommended methods. If the values are missing randomly, these procedures are effective. If, however, the values are not missing randomly, the use of regression procedures to impute values for missing data is questionable. The purpose of this study was to determine if values were missing randomly in samples selected from the National Education Longitudinal Study of 1988. Four samples were selected: 2 samples of 8 variables, average inter-correlation of 0.2 and 0.4 respectively; and 2 samples of 4 variables, average inter-correlation of 0.2 and 0.4 respectively. All cases containing one or more missing values were selected, and the pattern of missing values for each was determined. Chi square analysis indicated that the missing values are not missing randomly ($p<.001$). Implications of the use of regression procedures to handle non-randomly missing values are discussed. Appendix A contains four tables of descriptive statistics. (Contains 12 references and 2 tables.) (Author/SLD)

Are Values Missing Randomly in Survey Research?

E. Lea Witta

East Tennessee State University

Department of Educational Leadership

Paper presented to the annual conference of Mid-South Educational Research Association in Nashville, TN, November, 1994.
Correspondence should be sent to Lea Witta, Statistical Analysis & Evaluation, 19556 Stone Mountain Road, Abingdon, VA 24210 - (703) 628-9760.

Running Head: Randomness

Abstract

Many missing data studies have simulated data, randomly deleted values, and investigated which method of handling the missing values would most closely approximate the original data. Regression procedures have emerged as the most recommended methods. If the values are missing randomly, these procedures are effective. If, however, the values are not missing randomly, the use of regression procedures to impute values for missing data is questionable.

The purpose of this study was to determine if values were missing randomly in samples selected from the National Education Longitudinal Study of 1988. Four samples were selected: two samples of eight variables, average inter-correlation of .2 and .4 respectively; and two samples of four variables, average inter-correlation of .2 and .4 respectively. All cases containing one or more missing values were selected.

The pattern of missing values for each selected case was determined. Chi square analysis indicates the missing values are not missing randomly ($p \leq .001$). Implications of the use of regression procedures to handle non-randomly missing values are discussed.

When data are analyzed in survey research, often there are missing values. Ignoring this problem may lead to analysis of data that is of dubious value. Publication of the results of this analysis without correctly handling the missing values may "jeopardize the credibility of the organization conducting the survey and preparing the analysis and report:..." (Little & Smith, 1983, p. 518). Unfortunately, there is no established correct method for handling missing values when the mechanism causing them is unknown.

Methods that are currently being used to handle missing values include deletion methods (listwise and pairwise deletion) and imputation methods (mean substitution and various regression procedures). The most common solution to the missing data problem is probably listwise deletion. This procedure is the default option in several computer programs (LISREL, SPSS, NCSS). This method discards cases with a missing value on any variable and thus is very wasteful of data. If the data are assumed to be missing completely at random, this procedure is acceptable. Nevertheless, the loss of cases 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, attributed to Glasser (1964), computes covariances 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. The assumption made is that the use of the maximum number of paired points and individual observations (all available data) will yield more valid estimates of the relationship between the pairs.

Missing values are estimated and imputed to avoid non-representation if cases are dropped from a sample; to avoid power loss; to capitalize on inherent information in the missing/nonmissing pattern; and to utilize information present in the other variables (Cohen & Cohen, 1983). Mean substitution, attributed to Wilks' (1932) fills in a variable's missing values by the mean of that variable's sample values.

The regression methods rely on information contained in non-missing values of variables to provide estimates of missing values. Theoretically, the more variables considered, the better the estimate. Bucks's (1960) procedure regresses the missing variable on all non-missing variables. Mean regression begins by replacing missing values by that variable's mean and then regressing that variable on all other variables. Iterative regression adds the aspect of repetitive estimation until further estimates do not change. This process can be very slow.

These different methods of handling missing values may produce different results. When Jackson (1968) entered data on all the available variables in a discriminant analysis, the significance of the regression coefficients, as well as the interpretation of the importance, of individual variables changed with the missing value method used. Witta and Kaiser (1991) reported that the regression coefficients and total variance accounted for by the variables changed depending on the method used to handle missing values. After reanalyzing three studies of private/public school achievement, Ward Jr. and Clark III (1991) concluded that the method used to handle missing data influenced the outcome of these studies. They further add that the iterative regression procedures are considered the most effective.

Prior research suggests that the relative effectiveness of various methods of handling missing data is based on five characteristics of the data: average intercorrelation, number of variables, sample size, proportion of missing values, and pattern of missing data (Gleason & Staelin, 1975; Kaiser & Tracy, 1988). Of these, the pattern of missing values (randomness) has been least investigated.

Anderson, Basilevski and Hum (1983) add that the literature is not consistent even in the case of randomly missing values. They distinguish four types of randomness. Type 1 refers to data

missing due to legitimate skips. This would refer to directions that instruct the respondent 'if never married skip to'. The mechanism causing the missing values is known. Type 2 refers to data missing due to the magnitude of the answer. These would be situations in which the respondent refuses to report salary or age due to its value. In this instance, the result would be a censored sample containing only middle salaries or ages. Type 3 refers to the occurrence of missing elements on a variable in association with the occurrence of missing data on another variable. Type 4 refers to the occurrence of missing values on the independent variable in association with the occurrence of a missing values on dependent variables. Cohen and Cohen (1975) add that in survey research the absence of data on one variable may well be related to another independent variable or the dependent variable.

The purpose of this study was to determine if simultaneously missing values were missing randomly from four samples selected from the National Education Longitudinal Study of 1988 (NELS-88). Since the variables for this study were not designated as independent or dependent this test would include Type 3 and Type 4 non-randomness.

Method

Variable Selection

Variables for this study were selected from the student and parent supplements of the National Education Longitudinal Study of 1988 (NELS-88). NELS-88 was designed as a nationally representative two-stage stratified probability sample with schools selected in stage one and students at the second stage (Ingels et al., 1990).

The number of variables for this study, 4 or 8, was chosen arbitrarily. However, the selection of specific variables depended on how they intercorrelated with each other. Two levels of average intercorrelation (0.2, range = 0.103 to 0.270 and 0.4, range = 0.265 to 0.722), were computed by Kaiser's Gamma (1962).

Selection of samples from the NELS-88 data base was accomplished using SPSS-X. Analysis of this data was conducted using SPSS/PC+.

Analysis

After selection of the variables, all cases containing one or more missing values was selected. This resulted in four samples. Samples 1 and 2 contained four and eight variables respectively with low average intercorrelation. Samples 3 and 4 contained four and eight variables respectively with high average intercorrelation. Descriptive information and correlation matrices of the selected variables are included in Appendix A.

The variables for each case were recoded to indicate missing or non-missing status. Each sample was then split into sub-samples of cases containing one missing value, two missing values, etc. The assumption was made that if values are missing randomly, the occurrence of any group of variables missing simultaneously from a case and the occurrence of any other same size group of variables missing simultaneously would be identical. For example, there would be no difference in the number of cases containing missing values for variables 1 and 2 and the number of cases containing missing values for variables 1 and 3. Each sub-sample was then tested by the χ^2 goodness of fit statistic. The χ^2 statistic was used repeatedly in this study. To control for Type I error, the level of significance was set at α=0.001.

Since this study was investigating the relationship of values missing on a case simultaneously, cases containing only one missing value are not reported. Cases in which all of the variables were missing values were also excluded. These cases would represent a non-response in survey research.

Results

Statistical significance ($p \leq .001$) was detected in all four samples in this study. In Sample 1, significance was detected with two variables missing simultaneous and with three variables

missing simultaneously (see Table 1[1]). In Sample 2, significance was detected with two variables missing simultaneously, with three variables missing simultaneously, and with four variables missing simultaneously.

Insert Table 1 about Here

With Sample 3, significance was detected with two variables missing concurrently and with three variables missing concurrently (see Table 2[2]). In Sample 4, significance was detected with two variables missing concurrently, with three variables missing concurrently, and with four variables missing concurrently.

Insert Table 2 about Here

Discussion

For regression to predict a value effectively, at least one variable must be highly correlated with the dependent variable.

[1]The variable combination contributing positively to statistical significance is shown in Table 1.

[2]The combination of variables contributing positively to statistical significance is listed in Table 2.

In Sample 1, 70%[3] of the cases contained missing values on variables X3 and X4 jointly. These variables had the highest zero order correlation in the matrix (see Table A-1). In Sample 3, 78% of the cases contained missing values on variables X3 and X4 jointly. The highest zero order correlation in this correlation matrix was between these two variables (see Table A-2).

The estimates produced by Buck's regression procedure are generated by regression of the variable with a missing value (X3 or X4) on X1 and X2 only. The estimate thus constructed is based on only two variables rather than three. Neither of the two variables on which the estimate is based is the one most highly correlated with the variable containing the missing value.

When using the mean or iterative regression methods, the initial estimate for the missing value is the mean. Regression equations for estimating missing values are developed using this estimate. When the individual estimate for a missing value in a case is produced, the previously used estimates are used to predict the new estimate. When the variable (X3) most highly correlated with the variable (X4) for which an estimate was being calculated also contained a missing value on this case, the largest contribution to the individual estimate (X4) was made by

[3]Determined by dividing cases in this condition (650) by the total cases (925).

an estimated value (the mean of X3). As this procedure is iterated this became progressive. Since most of the incomplete cases in these two samples contained jointly missing values from the most highly correlated variables, the regression estimates are produced by estimated values.

In Sample 2, 46% of the cases contained missing values on variables X6 and X7 jointly. While these variables did not have the highest zero order correlation in the matrix (see Table A-3), missing value estimates in Buck's procedure are produced by regression of the variable with a missing value (X6 or X7) on X1, X2, X3, X4, X5, and X8. The resulting estimate is, therefore, based on six variables instead of seven.

When X6 and X7 were missing simultaneously and the mean regression and iterative regression methods are used, estimates for each missing value are based on one estimated value that contributed to the variable, and six actual values (the first four variables).

In Sample 4, 76% of the cases contained missing values on variables X7 and X8 jointly and 73% contained missing values on variables X5, X6, X7, and X8. The highest zero order correlation in this correlation matrix was between these variables (see Table A-4). Missing value estimates in Buck's procedure are produced by regression of the variable with a missing value (X5, X6, X7, or

X8) on X1, X2, X3, and X4. The resulting estimate is, therefore, based on four variables instead of seven. Moreover, the variables used are not highly correlated with the variable being estimated.

When X5, X6, X7, and X8 were missing simultaneously and the mean regression and iterative regression methods are used, estimates for each missing value are based on one estimated value that contributed most to the variable, two other estimated values, and four actual values (the first four variables).

Conclusion

When faced with missing data, researchers have two choices: deletion or imputation. Of the two deletion methods, pairwise deletion has the advantage of retaining all known data. If the jointly missing values were to be used as the dependent variable in a study (such as the standardized test scores), the researcher may prefer to eliminate cases that do not contain these values. In this instance, the choice would be listwise deletion. Whenever imputation of missing values is desired, an imputation method would be used.

Prior to the decision to use any of the regression methods, the researcher must determine the pattern of missing values. If the concurrently missing values are missing on the most highly correlated variables, the use of regression methods to impute estimates is highly questionable.

Since the incomplete cases used in this study were selected from those existing in the NELS-88 data base, this suggests that the occurrence of jointly missing values can be expected when using NELS-88. Further research is needed to determine the effects of the pattern of missing values on the effectiveness of various missing-data-handling methods.

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 Inc.

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.

Cohen, J. & Cohen, P. (1983). Missing data. In J. Cohen & P. Cohen, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (pp. 275-300). Hillsdale, N.J.: Lawrence Erlbaum Associates, Publishers.

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

Kaiser, H.F. & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika, 27, 179-182.

Kaiser, J. & Tracy, D.B. (1988). Estimation of missing values by predicted score. American Statistical Association 1988 Proceedings of the Section on Survey Research Methods, 631-635.

Ingels, S.J., Abraham, S.Y., Karr, R., Spencer, B.D., Frankel, M.L., & Owings, J.A. (1990). National Education Longitudinal Study of 1988 Base Year: Student Component Data File User's Manual. U.S. Department of Education, Office of Educational Research and Improvement, NCES 90-464, Washington DC: U.S. Government Printing Office.

Jackson, E.C. (1968). Missing values in linear multiple discriminant analysis. Biometrics, 24, 835-844.

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.

Ward, Jr., T.J. & Clark III, H.T. (1991). A reexamination of public-versus private-school achievement: the case for missing data. Journal of Educational Research, 84, 153-163.

Wilk's, S.S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. Annals of Mathematical Statistics, 3, 163-195.

Witta L. & Kaiser, J. (1991, November). Four methods of handling missing data with GSS-84. Paper presented at the meeting of the Mid-South Educational Research Association, Lexington, KY

Table 1

Randomness of Simultaneously Missing values for Samples with Low Average Intercorrelation

Missing Condition[a]	Total Cases[b]	χ^2	Significant Combination[c]	Number of Cases[d]	Standardized Residual[e]
			Sample 1 - Four Variables		
2	769	2582**	X3, X4	650	46.09
3	156	40.9**	X1, X2, X3	72	5.28
			Sample 2 - Eight Variables		
2	835	6376**	X6, X7	452	77.31
3	354	2873**	X2, X6, X7	156	50.83
4	178	1217**	X1, X2, X6, X7	106	33.49
All Other	173				

Note. [a]Number of variables missing simultaneously. [b]Total number of cases in this condition. [c]Variables with simultaneous missing values contributing to significance. [d]Number of cases in this condition. [e]Standardized residual for this condition.

Table 2

Randomness of Simultaneously Missing values for Samples with High Average Intercorrelation

Missing Condition[a]	Total Cases[b]	χ^2	Significant Combination[c]	Number of Cases[d]	Standardized Residual[e]
			Sample 3 - Four Variables		
2	848	2494**	X3, X4	678	45.14
3	91	57.7**	X2, X3, X4	51	5.92
			Sample 4 - Eight Variables		
2	132	268**	X7, X8	48	12.56
			X3, X4	34	8.00
3	41	46.6**	X5, X7, X8	10	7.95
			X2, X3, X4	10	7.95
			X1, X3, X4	12	9.97
4	999	2020**	X5, X6, X7, X8	857	38.42
All Other	28				

Note. [a]Number of variables missing simultaneously. [b]Total number of cases in this condition. [c]Variables with simultaneous missing values contributing to significance. [d]Number of cases in this condition. [e]Standardized residual for this condition.

Appendix A

Tables of Descriptive Statistics

Tables in this appendix contain the correlation matrices, variable names, variable labels. These are as follows:

Table A-1: Sample 1 - Low Intercorrelation (.2), Four Variables

Table A-2: Sample 3 - High Intercorrelation (.4), Four Variables

Table A-3: Sample 2 - Low Intercorrelation (.2), Eight Variables

Table A-4: Sample 4 - High Intercorrelation (.4), Eight Variables

Table A-1

Sample 1 - Low Intercorrelation (.2), Four Variables

	X1	X2	X3	X4
X1	1.0000			
X2	.1808	1.0000		
X3	.1253	.1537	1.0000	
X4	.1140	.1804	.2427	1.0000

Note. X1=BYP57C Contacted about H.S. Course Selection.

X2=BYP57F Contacted about School Fund Raising.

X3=BYP58A Contacted School About Academic Performance.

X4=BYP58E Contacted School About Info for School Records.

Table A-2

Sample 3 - High Intercorrelation (.4), Four Variables

	X1	X2	X3	X4
X1	1.0000			
X2	.5353	1.0000		
X3	.4505	.2823	1.0000	
X4	.3033	.4298	.5953	1.0000

Note. X1=BYP57A Contacted about Academic Performance.

X2=BYP57B Contacted about Academic Program.

X3=BYP58A Contacted School about Academic Performance.

X4=BYP58B Contacted School about Academic Program.

Table A-3

Sample 2 - Low Intercorrelation (.2), Eight Variables

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.00							
X2	-.199	1.00						
X3	.170	-.103	1.00					
X4	.158	-.156	.154	1.00				
X5	.165	-.131	.266	.170	1.00			
X6	.149	-.157	.185	.152	.150	1.00		
X7	.270	-.109	.151	.115	.146	.142	1.00	
X8	.228	-.230	.224	.291	.152	.243	.153	1.00

Note. X1=BYS56C Students in Class see R as a Good Student.

X2=BYS60A R's Ability Group for Mathematics.

X3=BYS36C Discuss Things Studied in Class with Parents.

X4=BYS48B How Far in School R's Mother Wants R to go.

X5=BYS50A Talk to Father about Planning H.S. Program.

X6=BYS79A Time Spent on Math Homework Each Week.

X7=BYS78C How Often Come to Class Without Homework.

X8=BYTXRSTD Hist/Cit/Geog Standardized Score.

Table A-4

Sample 4 - High Intercorrelation (.4), Eight Variables

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.00							
X2	.383	1.00						
X3	.464	.397	1.00					
X4	.476	.364	.539	1.00				
X5	.320	.265	.394	.433	1.00			
X6	.371	.276	.399	.429	.722	1.00		
X7	.367	.399	.422	.414	.691	.710	1.00	
X8	.296	.275	.393	.373	.712	.701	.717	1.00

Note. X1=BYS81A English Grades from Grade 6 until now.

X2=BYS81B Math Grades from Grade 6 until now.

X3=BYS81C Science Grades from Grade 6 until now.

X4=BYS81D Social Studies Grades from Grade 6 until now.

X5=BYTXHSTD History/Cit/Geog Standardized Score.

X6=BYTXRSTD Reading Standardized Score.

X7=BYTXMSTD Mathematics Standardized Score.

X8=BYTXSSTD Science Standardized Score.