ESTIMATION OF MISSING VALUES BY PREDICTED SCORE

Javaid Kaiser, East Carolina University
D. B. Tracy, University of Kansas

Using predicted scores as estimates of missing values is one of the most frequently used imputation technique. The method was first proposed by Buck (1960) and is based on the assumption that the values are missing at random. The quality of predicted scores as estimates of missing values depends on the number of predictors in the regression equation and their correlation with the criterion variable.

Predicted scores become good estimates of missing values when corre lation among variables is high. Theoretically, an increase in the number of predictors used improves the quality of the predicted score and makes it a better estimate of missing value. Use of too many predictors, however, overfits the regression equation causing it to produce poor estimates (Frane, 1976). Other factors that affect the quality of estimates include pattern of missing values, sample size, number of missing values in a single record, and the purpose of estimation. (Haitovsky, 1968; Frane, 1977)

The way initial correlation matrix is computed from imcomplete data to develop regression equations is controversial. Buck (1960) suggested the use of complete records only. Gleason & Staelin (1975) considered it a poor choice when data matrix has large number of variables. They suggested to impute missing values by their respective variable means, compute correlation matrix from this completed data matrix, develop regression equation, and then re-impute missing values using predicted scores. This idea is supported by the results of a simulation study (Timm, 1970) in which estimates of correlation matrix produced by complete records were found inferior to the ones produced by other methods when data matrix had more than four variables. Afifi and Elashoff (1966, 1967, 1969a, and 1969b) have compared several variations of regression method in estimating regression coefficients from data having missing values. Greenlees, Reece, and Zieschang (1982) have developed a procedure that is better than least squares methods in estimating regression coefficients from incomplete data. They claimed that unlike least squares, their procedure does not ignore the mechanism that causes missing values.

A predictor less correlated with the criterion variable is better than highly correlated variable having missing values. Huddleston and Hocking (1978) recommended that the design of the study must ensure collection of related information when 30% or more values are expected to be missing. Kalton and Kasprzyk (1983) concluded that an auxiliary variable must be added into the design of the study when a certain variable is expected to have missing values. Failure to do so causes a bias in the standard error of estimate. Kim and Kohout (1975) suggested a correction factor to adjust the predicted score if some of the predictors used have missing values. Rubin (1976) has proposed a new method to compute multiple correlation coefficient when predictor(s) have missing values.

Regression method has been compared with several other imputation techniques by various researchers. Timm (1970) found regression method superior to substitution by mean method (zero order) in estimating population correlation and variance-covariance matrices. He recommended the use of regression method for imputation purposes when the variables are at least moderately correlat- ed. Gleason et. al. (1975) found regression method better than zero order method whenever average correlation exceeded .20. Wolfe, Behrman, and Flesher (1979) reached at the same conclusion but because of univariate setting, their findings may not be applicable in multivariate situations. Duan, Marini, and Marquis (1981) and Santos (1981) found regression coefficients biased when computed form imputed data. Finkbeiner (1979), however, discovered regression method less effective than zero order method in estimating parameters of multiple factor model in samples of size 64. Champney and Bell (1982) found regression method producing biased estimates of population variance and underestimated means.

Heeringa and Lepkowski (1986) have compared the efficiency of regression method with other imputation techniques in estimating missing values in panel surveys. Marini, Olson, and Rubin (1980) concluded that regression method is inappropriate to predict missing values in follow-up waves by variables measured in earlier waves because the estimates produced underestimate variance and overestimate correlation coefficients.

METHOD

There are several variations of regression method. Besides regression with single predictor, two predictors, and all available predictors (Frane, 1976), there are several hybrids of regression method. (Champney et. al., 1982; Schieber, 1978; Walsh, 1961). The present study investigated four variations of regression method and substitution by mean method. The first variation (REG1) used a single predictor. A variable that had the highest correla tion with the variable whose value was to be imputed, was selected as predictor. The second variation (REG2) used two best predictors in the regression equation. Predictors were selected by stepwise multiple regression. The third variation (REGALL) used all predictors that had observed values. The fourth variation (REGRES) used all predictors having observed values and later modi fied it by a correction factor as suggested by Kim et. al. (1975) to adjust for predictors having missing values. The predicted scores, so produced, were used as estimates of missing values. Zero order method used variable mean as an estimate of all missing values that occurred on that variable.

A 3x3x4 factorial design was used to study factors that included sample size (n= 30, 60, & 120), the proportion of incomplete records in the sample (= 10%, 20%, & 30%), and the number of missing values on the incomplete record (m= 1, 2, 3, & 4). An 8x8 correlation matrix, given in Table 1, was used to represent population. An IMSL (1980) package of computer subroutines was used to generate nx8 data

matrices of multivariate normal deviates in standard score form.

TABLE 1

Population Correlation Matrix

V1	1.00							
V2	.318	1.00						
V3	.468	.230	1.00					
V4	.403	.317	.305	1.00				
V5	.321	.285	.247	.227	1.00			
V6	.414	.272	.263	.322	.187	1.00		
V7	.365	.292	.297	.339	.398	.388	1.00	
V8	.413	.232	.250	.380	.441	.283	.463	1.00
	V1	V2	V3	V4	V5	V6	V7	V8

The variance covariance matrix of every sample was tested against the population covariance structure for equality at .05 level of significance to ensure that the change in covariance structure after imputation is directly attributable to imputation and not to the initial deviance. A procedure described by Anderson (1958), was used for this purpose. The sample matrices whose covariance structure was found similar to the population covariance structure were retained for this study.

Missing values were artificially created in data matrices as per cell specifications of the design matrix. The pattern of missing values was random and was achieved by use of random numbers. Zero order and variations of regression method were applied one at a time to impute missing values. Means were computed before creating missing values and after imputing them. The mean discrepancy in the two values and the standard error of this discrepancy was computed for all variables and for all imputation methods to determine their efficiency in producing unbiased estimates of means.

The quality of missing value estimates was determined in terms of root-mean-square standardized residual of true and estimated values as defined by Gleason et. al. (1975) and was represented by Q. The statistic Q provided an index to compare the relative performance of imputation methods and was computed for all methods at all experimental conditions.

The variance-covariance matrix of imputed sample was computed and compared against the covariance structure of original sample. The task was accomplished by means of statistic D which represented the root-mean-square deviation of respective elements of two variance-covariance matrices as proposed by Timm (1970) and modified by Gleason et. al. (1975). The statistic D repre- sented the relative efficiency of imputation technique in retaining the covariance structure of original sample in imputed samples. Smaller the value of D, better the imputation method is. D was computed for all imputation methods at all experimental conditions.

The procedure described above, was repeated for all cells of the design matrix and each cell was replicated 500 times.

RESULTS

The results of this study indicated that all variations of regression method and zero order method produced unbiased estimates of means. All regression variations except REGRESS produced relatively better estimates of means than zero order method. However, the differences were not found large enough to suggest any ranking. There was no systematic trend establishing superiority of one method over the other across various levels of sample size, percent of incomplete records, and the number of missing values in incomplete records. REGRESS was, however, found least desirable at all experimental conditions.

The data revealed increased discrepancy between true and estimated means as the percent of incomplete records or the number of missing values per record increased. The discrepancy, however, decreased with the increase in sample size. The same trend was observed in terms of standard error for all imputation methods. Again, the observed differences were relatively small and did not seem of any practical significance.

The analysis revealed that all imputation methods studied altered the covariance structure of the original sample when missing values were imputed by their respective estimates. Regression variations were found less damaging to covariance structure than zero order method. There were relatively small differences among regression variation in retaining original covariance structure in imputed samples. REG1 was, however, found more suitable for samples of size 30 while REGALL seemed better for large samples. These findings were found consistent at all experimental conditions. The relative performance of regression variations and zero order method were plotted and is given in Figure 1.

The analysis of factors revealed that an increase in the number of missing values or the percent of incomplete records, increased damage to the variance-covariance structure irrespective of the imputation method used. The increase in sample size, however, minimized the effect when other factors were kept constant.

In terms of quality of missing value estimates, zero order method and variations of regression technique performed differently at different levels of sample size. In small samples (n=30), regression variations that used two or less predictors produced the best estimates. However, in large samples (n 30), regression variations using all predictors performed better than the remaining methods. Compared to other regression variations, REGRES produced poor estimates. Zero order method was found inferior to REG1 and REG2 for samples of size 30. In large samples, all regression variations outperformed zero order method. The relative performance of imputation methods was plotted and is given in Figure 2.

The data indicated that the quality of missing value estimates was least affected by the number of missing values present in a record. Small differences that appeared in samples of size 30, disappeared in larger samples (n 30). The percent of incomplete records, on the other hand, adversely affected the quality of estimates. Keeping other factors constant, the estimates were improved with the increase in sample size.

DISCUSSION

The data revealed that all variations of regression method and zero order method produced unbiased estimates of means. This finding is supported by Kalton et. al. (1983). When the purpose of imputation is estimation of means, zero order method may be preferred to regression variations for its simplicity and ease of computation. The method REGRES has been found least desirable because of its large discrepancy between true and estimated means and large standard error. This regression varia- tion has not been studied in the past and therefore, no data is available to compare with the results of this study. However, a logical explanation may be that the correction factor used to adjust the predicted score for predictors having missing values, over-inflated the missing value estimates causing larger discrepancy between true and estimated means. As a result, the variance of this discrepancy was also inflated. Further research on this method is however, warranted.

The results indicated that all imputation methods investigated, failed in complete restoration of original variance-covariance structure in imputed samples. This finding confirms earlier consensus that imputation increases variance and alters covariance structure (Little, 1981; Santos, 1981; Wolf et. al., 1979; Duan et. al., 1981; Champaney et. al., 1982; Kalton et. al., 1983). Although relative differences among imputation methods were small, regression variations outperformed zero order method in retaining covariance structure in imputed samples. This finding is supported by Timm (1970) and Gleason et. al. (1975) but contradics the findings of Finkbeiner (1979). The finding that increase in the percent of incomplete records and the number of missing values present in a record, individually and jointly, affected the covariance structure adversely and that this effect was minimized by increase in sample size were logical outcomes and are supported by Gleason et. al. (1975).

The variation REGRES that was not found effective in estimation of means was comparable to other regression variations in retaining original covariance structure in imputed samples. Again, no literature support is available to confirm or disconfirm this finding.

The data indicated that regression method produced better quality estimates of missing values. This may be true because regression approach used more information than zero order method in producing estimates (Frane, 1977). It was also observed that the efficiency of zero order method decreased with the increase in sample size. It makes regression a preferred approach to estimate missing values in large samples. Afifi et. al. (1967) had observed the same phenomenon in point estimation. Another interesting finding was that the quality of missing value estimates primarily depended on the percent of incomplete records and was not very sensitive to the percent of values missing on these records. Further research is needed to explain this trend.

CONCLUSION

Three out of four regression variations and zero order method pro duced unbiased estimates of means across all experimental conditions. Regression variation that adjusted the predicted score because of predictors having missing values, was found inappropriate to estimate means. All imputation methods altered the covariance structure as a result of imputation. Regression variations, however, were found relatively better than zero order method in retaining original covariance structure in imputed samples. Regression variations also outperformed zero order method in producing quality estimates of missing values. Regression method was found the best overall imputation technique to estimate missing values. Zero order method was considered appropriate for estimation of means because of its simplicity and ease of computation.

REFRENCES

Afifi, A. A., & Elashoff, R. M. (1966). Missing observations in multivariate statistics I. Journal of American Statistical Association, 61 (315) 595-605.

Afifi, A. A., & Elashoff, R. M. (1967). Missing observations in multivariate statistics II. Point estimation in simple linear regression. Journal of American Statistical Association, 62 (317) 10-29.

Afifi A. A., & Elashoff, R. M. (1969a). Missing observations in multivariate statistics III. Large sample analysis of simple linear regression. Journal of American Statistical Association, 64 (325) 337-358.

Afifi, A. A., & Elashoff, R. M. (1969b). Missing observations in multivariate statistics IV. A note on simple linear regression. Journal of American Statistical Association, 64 (325) 359-365.

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons Inc., pp. 264-267.

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, Series B, 22, 302-307.

Champney, T. F., & Bell, R. (1982). Imputation of income: A procedural comparison. Proceedings of the Survey Research Section, American Statistical Association, 431-436.

Duan, N., Marquis, K. H., & Marquis, M. S. (1981). Countering estimation bias due to response errors: A simulation example. Proceedings of the Survey Research Section, American Statistical Association, 342-345.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. Psychometrika, 44 (4), 409-420.

Frane, J. W. (1976). Some simple procedures for handling missing data in multivariate analysis. Psychometrika, 41 (3), 409-415.

Frane, J. W. (1977). Description and estimation of data. In W. J. Dixon and M. B. Brown (Eds.), BMDP Biomedical Computer Programs: P-Series. Barkeley: University of California Press, pp. 348-370. Gleason, T. C., & Staelin,

R. (1975). A proposal for handling missing data. Psychometrika, 40 (2), 229-252.

Greenlees, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. Journal of the American Statistical Association, 77, (3), 251-261.

Haitovsky, Y. (1968). Missing data in regression analysis. Journal of the Royal Statistical Society, Series B, 30, 67-82.

Huddleston, H. F., & Hocking, R. R. (1978). Imputation in agricultural surveys. Proceedings of the Survey Research Section, American Statistical Association.

International Mathematics and Statistics Library (IMSL). (1980). Library reference manual, (8th Ed.; 4 Vols.). Dallas: IMSL Inc.

Kalton, G., & Kasprzyk, D. (1983). Imputing for missing survey responses. Proceedings of the Survey Research Section, American Statistical Association, 22-31.

Kim, J. O., & Kohout, F. J. (1975). Multiple regression analysis: Subprogram regression. In N. H. Nie, C. Hadlaihull, J. G. Jenkins, K. Steinbrenner, & D. H. Bent (Eds.), Statistical Package for Social Sciences (2nd. ed.). New York: McGraw-Hill Book Company, 347-348.

Little, R. J. A. (1981). Discussion. Proceedings of the Survey Research Section, American Statistical Association, 152-153.

Marini, M. M., Olsen, A. R., & Rubin, D. B. (1980). Maximum likelihood estimation in panel studies with missing data. In Karl R. Schussler, Sociological Methodology-1980. San Francisco: Jossey-Bass Publications, 314-357.

Rubin, D. B. (1976). Comparing regressions when some predictor values are missing. Technometrics, 18 (2), 201-205.

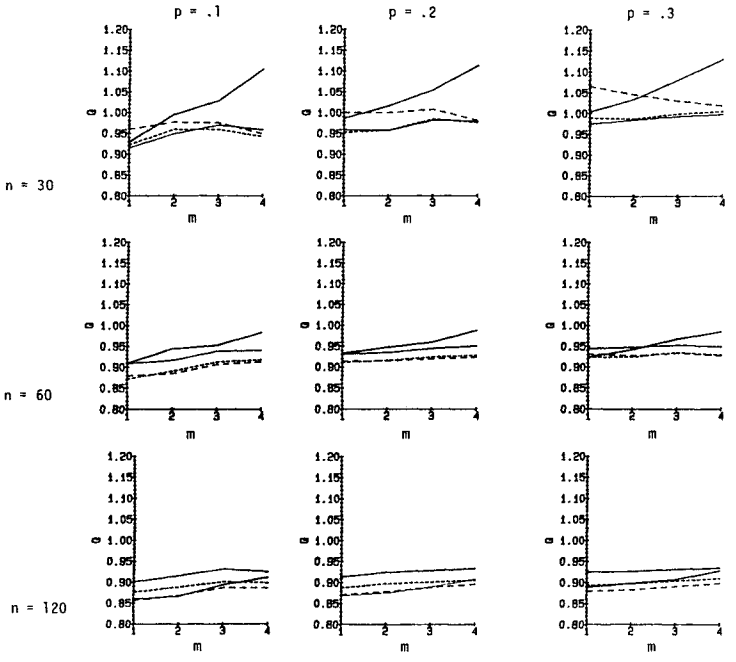
Santos, Robert. (1981). Effects of imputation on regression coefficients. Proceedings of the Survey Research Section, American Statistical Association, 140-145.

Schieber, S. J. (1978). A comparison of three alternative techniques for allocating unreported social security income on the survey of the low income aged and disabled. Proceedings of the Survey Research Section, American Statistical Association.

Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35, (4), 417-438.

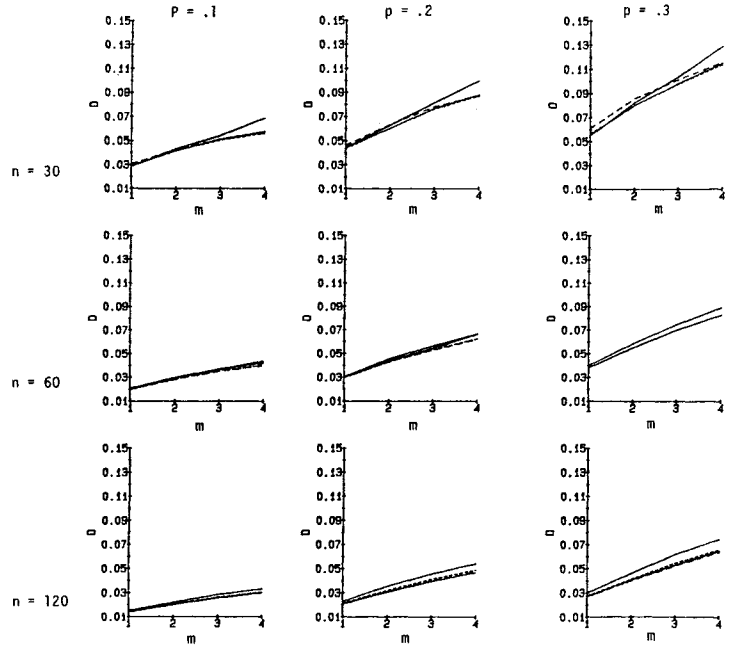
Walsh, J. E. (1961). Computer feasible method for handling incomplete data in regression analysis. Journal of the Association of Computer Machinery, 18, 647-657.

Wolfe, B. L., Behrman, J. R., & Flesher, J. (1979). A Monte Carlo study of alternative approaches for dealing with randomly missing data. Report No. DP # 587-79. Institute for Research on Poverty, University of Wisconsin Madison.



QUALITY OF MISSING VALUE ESTIMATES
AT VARIOUS LEVELS OF n, p, AND m

Figure 1



RETENTION OF COMPLETE SAMPLE COVARIANCE STRUCTURE
IN IMPUTED SAMPLES AT VARIOUS LEVELS OF n, p, AND m

Figure 2