DOCUMENT RESUME

ED 048 365 TM 000 443

AUTHOR Remer, Rory; Burton, Nancy
TITLE Consequences of Various Procedures for Estimating Missing Data in Factor Analysis.
PUB DATE Feb 71
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Calculation, Comparative Analysis, Correlation, *Data Analysis, Data Collection, *Factor Analysis, *Goodness of Fit, Multiple Regression Analysis, *Research Methodology, Research Problems, *Research Tools
IDENTIFIERS *Principal Components Analysis

ABSTRACT

The relative precision of four methods of estimating missing data in principal components analysis was investigated. Artificial data with known characteristics, obtained from Cattell's "Plasmode: 30-10-4-2," was used with one third of the data on half of the variables being systematically eliminated. The four methods of missing data estimation were: means substitution, simple regression, stepwise regression, and multiple regression. In order to extract all possible variance, the Principal Components Analysis was employed without rotation. Factor scores from complete data and each of the estimated data solutions were obtained. Goodness-of-Fit was judged on the basis of cross-correlations of each estimated data solution with the solution derived from complete data. The study showed that all four methods of estimation compared fairly well with the criterion. The average correlations improved from the method using least concomitant information (means substitution) to that employing most (multiple regression). Indications of the study were that means-substitution may be a viable method of estimating missing data . (AE)

CONSEQUENCES OF VARIOUS PROCEDURES FOR ESTIMATING MISSING DATA IN FACTOR ANALYSIS

by

Rory Remer

and

Nancy Burton

Laboratory of Educational Research
University of Colorado

Submitted to the Annual Convention of the American Educational Research Association, New York, February 1971.

Consequences of Various Procedures for Estimating Missing Data in Factor Analysis

Introduction

Rarely in practical situations is it possible to obtain complete data on all subjects, particularly when the study is done on a large scale. These gaps can sometimes be overlooked or accommodated when certain statistics are employed. When large quantities of information are missing, problems arise concerning the best method of handling the situation. It becomes infeasible to overlook or discard the subject for which incomplete information has been obtained - such procedures can, at times, produce very misleading results.

In any factor analytic technique, missing data can do more than produce errant results. They can make it impossible for any results to be obtained. The original correlation matrix can easily be ill-conditioned and hence, not invertable, stopping any extraction procedure. The question thus becomes one of what to do about large quantities of missing data.

Little has been written concerning this problem. Guertin (1968) in an empirical study with actual data used three methods of handling missing data--means estimation, regression and omission--in producing correlation coefficients for different total N'S and for different percents of missing information. He found that it was not worth the effort to obtain multiple regression estimates for a variable with 40 percent missing scores and small samples. His results, however, were based on comparison of methods with each other, no possible outside criterion being available.

The present study represents a first attempt at finding a criterion in a factor analytic framework (principal components analysis). Complete data were located and used to specify a criterion solution. No attempt was made to be comprehensive. Accordingly those procedures judged to be simplest, most straight forward, most easily manipulated, and most easily understood have been employed. The purpose of this study was to provide an initial step toward determining the relative precision of four different methods of estimating missing data in principal components analysis. It is possible to simulate various amounts of missing data, to eliminate the data in various systematic or random ways, to use various methods in estimating the missing vlaues, to use numerous procedures for extracting and rotating, and to use various criteria to judge best fit. In the present instance the following alternatives were selected:

I. Data

Artificial data with known characteristics, obtained from Cattell's "Plasmode: 30-10-4-2" (Cattell and Jaspers, 1967), were used. One third of the data on half of the variables was systematically eliminated by excluding the last, by order, 100 (of 300) cases on the second 15 (of 30) variables.

II. Methods of Missing Data Estimation

Four least-squares methods of data estimation were selected for examination. They were: Means Substitution--estimation of each missing value by inserting the mean for that variable. This is a least-squares procedure when no concomitant information is known. Simple Regression--estimation of the missing value from the highest correlating predictor. Step-wise Regression--estimation including all independent variables contributing .01 or more to the multiple R. Multiple Regression--estimations made using all 15 possible independent predictor variables.

III. Method of Extraction

In order to extract all possible variance Principal Components Analysis was employed.

IV. Method of Rotation

No rotation method, orthogonal or oblique, was employed.

V. Criterion for Judgement of Goodness of Fit

Factor (component) scores from complete data and each of the estimated-data solutions were obtained. Goodness-of-fit was judged on the basis of cross-correlations of each estimated-data solution with the solution derived from complete data.

The BMD 03M computer program (Dixon, 1968, p. 169) was used to extract by the principal components method 30 components corresponding to the 30 variables in the Plasmode. Then the data were eliminated, 1500 pieces being considered the maximum possible amount which could be accommodated by a 30x300 matrix. All remaining cases (the first 200) for which complete data were available were used to produce estimation equations.

The BMD 02R computer program (Dixon, 1968, p. 218), a stepwise regression algorithm, was used to form the three different types of regression equations for each of the 15 variables with missing data. In the stepwise procedure variables are entered in the order of highest residual correlation with the criterion. The first step was used as the simple regression estimation equation. By specifying that all 15 predictor variables be successively entered, the last step could be employed as the full multiple-regression estimation equation. The intervening steps were examined to ascertain that step which added just more than .01 to the multiple R, thus obtaining the step-wise estimation equation. The desired means were also produced as output of the program.

The missing scores were estimated and combined with those of the 200 complete cases and four BMD 03M programs, one for each set of estimated data, were run.

When the principal components analysis is employed, an explicit criterion of best fit is possible. Component scores on the criterion, complete data, solution may be obtained explicitly by the solution of the matrix equation for the components model:

$$Z_c = F_c X_c \qquad (1)$$

where

Z_c is the nxN matrix of standardized observations of the complete data variables

F_c is the nxn factor pattern for the complete data

X_c is the nxN matrix of standardized component scores for the compiled data solution

Thus the matrix of component scores, X, can be obtained by the following equation:

$$X_c = F_c^{-1} Z_c \qquad (2)$$

provided that as many components as variables are extracted and that F is non-singular. However, the estimation-solution component scores were derived as follows:

$$X_e = F_e^{-1} Z_c \qquad (3)$$

where

X_e = the nxN matrix of estimation-solution component scores

F_e = the nxn estimation-solution factor pattern

and Z_c = the nxN standardized matrix of criterion complete data.

The cross-correlations between X_c and X_e are the cross-correlations between the respective principal components. In geometric terms, they give the cosines of the angle of separation between, for example, the first principal-axis of the criterion solution and the first principal-axis of one of the estimated-data solutions, when these areas are represented in a space determined by the original complete data.

The cross-correlations between X_c and X_e as defined in equations (2) and (3) above were then obtained.

The cross-correlation between X_c and X_e, R_{ce}, is

$$R_{ce}=\frac{X_c\, X_e^T}{N}. \qquad (4)$$

Substituting from equations (2) and (3), we obtain

$$R_{ce}=\frac{F_c^{-1}Z_c(F_e^{-1}Z_c)^T}{N}=\frac{F_c^{-1}Z_cZ_c^T(F_e^{-1})^T}{N}.$$

The middle portion of this equation equals the original intercorrelation matrix, R_{cc}:

$$R_{cc}=\frac{Z_cZ_c^T}{N}.$$

Utilizing the facts that

$$R = QD^2Q^T \qquad (5)$$

and

$$F = QD \qquad (6)$$

where Q = a matrix of latent vectors

and D = a diagonal matrix of the square roots of latent roots,

we substitute in the above equation for R_{ce} to obtain

$$R_{ce}=\frac{F_c^{-1}Z_cZ_c^T(F_e^{-1})^T}{N}$$

$$= (Q_cD_c)^{-1}Q_cD_c^2\ Q_c^T\{(Q_eD_e)^{-1}\}^T$$

$$= D_c^{-1}Q_c^{-1}Q_cD_c^2Q_c{}^T\ (Q_e^{-1})^T\ (D_e^{-1})^T.$$

But $\quad Q^{-1} = Q,^T$ since Q is orthogonal;

and $\quad (D^{-1})^T = D,^{-1}$ since D is diagonal, so

$$R_{ce} = D_c^{-1}\ D_c^2Q_c^TQ_eD_e^{-1}$$

$$= D_cQ_c^T\ Q_eD_e^{-1}.$$

By substituting from equation (6), the final result is:

$$R_{ce} = F_c{}^TF_eD_e^{-2}\ . \qquad (7)$$

Thus, taking the factor patterns and latent roots from the BMD 03M, equation (7) was employed to solve for the desired cross-correlations.

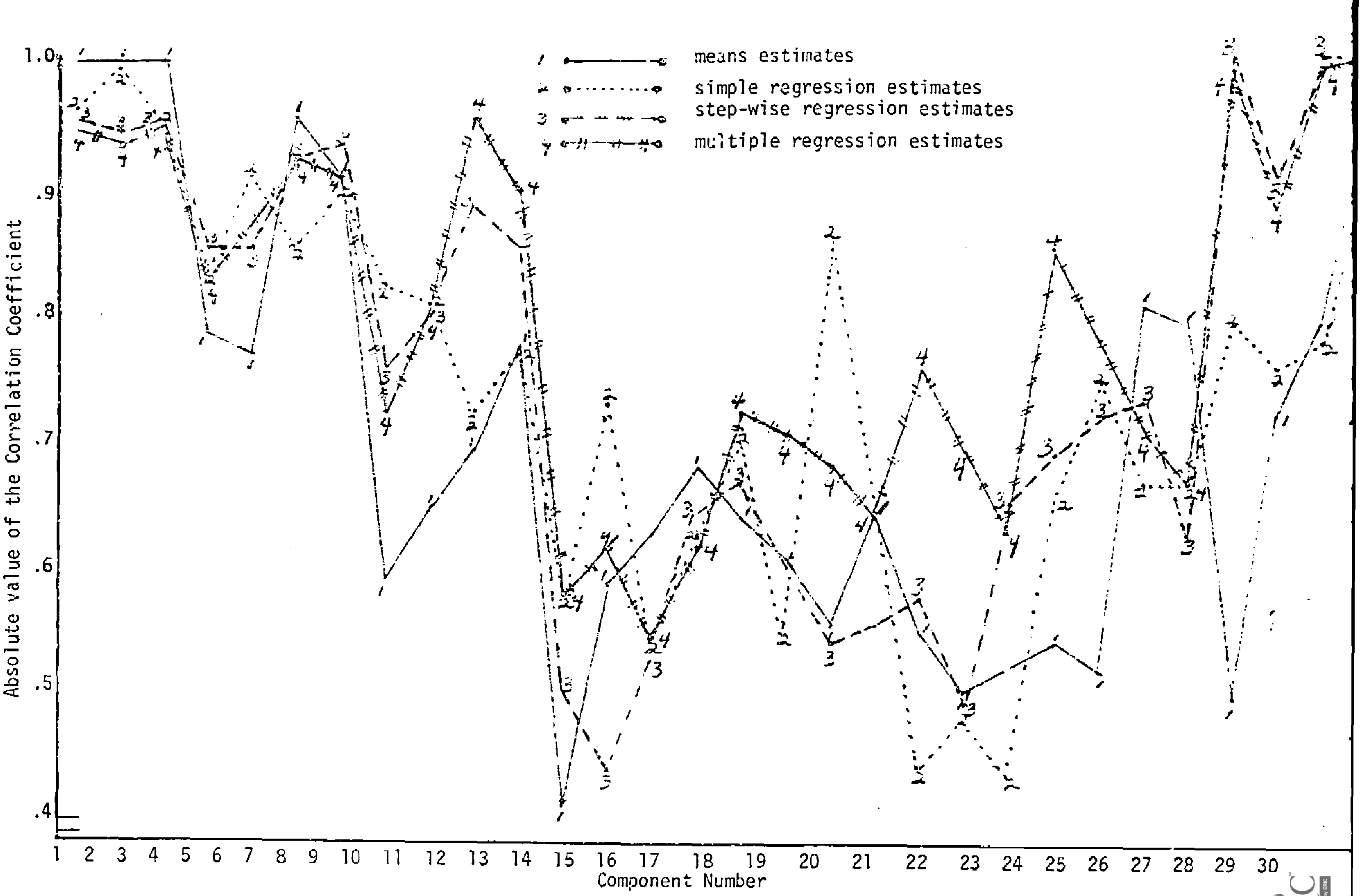
Results

If any estimation procedure were to estimate the missing data with perfect accuracy, one would expect the cross-correlation matrix, R_{ce}, to equal I, the identity matrix. All four of the cross-correlation matrices were approximately diagonal. The first eleven and the last six factors had high ($|r| \doteq .8$) correlations in the diagonal; for the middle thirteen factors the correlations were split up among several adjacent factors. Of the four estimation methods, means and simple regression resulted in fewer off-diagonal correlations greater than .4 (26 and 25, respectively); both the stepwise and the multiple regression estimates resulted in 32 off-diagonal correlations greater than .4. The first eleven components could be expected to hold up well, since Cattell's plasmode was built to contain ten first-order factors.

The rest of the variance extracted was expected to be unique variance. The authors believe that the last six factors were constituted of unique variance from the 15 variables from which no data were missing and thus would remain unchanged in the various solutions. The uniqueness of these variables would create small but stable factors.

Figure 1 is a graph of the absolute cross-correlations with the criterion component scores on each of the 30 components for each of the sets of component scores derived from the four methods of data estimation. It shows the relatively higher correlations for the first 11 and the last six components. No other general conclusion is obvious from inspection of Figure 1.

Figure 1. Absolute cross-correlations of four estimation solutions with criterion solution*



*Restriction: Isomorphism of Components.

For the four estimation methods, the average absolute correlation ($\Sigma|r|/N$) and percent of variance explained ($\Sigma r^2/N$) were computed for all 30 components and for the first 11 components (see Table I).

Table I. Average Cross-Correlation with Criterion and Squared Correlation for Four Data Estiamtion Methods.

Estimation Method	Average Absolute Correlation $\Sigma\|r\|/N$ 30 Components	Average Absolute Correlation $\Sigma\|r\|/N$ 11 Components	Average Percent of Variance Explained $\Sigma r^2/N$ 30 Components	Average Percent of Variance Explained $\Sigma r^2/N$ 11 Components
Means	.72	.84	54	74
Simple Regression	.74	.86	57	75
Step-wise Regression	.76	.88	61	77
Multiple Regression	.79	.88	65	78

All of these statistics show a trend in the anticipated direction. The *improvement in precision is more noticeable when all 30 components are* taken into account. The first 11 components, however, should account for nearly all of the non-unique variance, since the plasmode contains ten common factors. For the first 11 components, the simplest method of data estimation, means substituion, compares well with the others.

Discussion

This study is only a first step in determining the best method for estimating missing data for factor analytic studies. Research should be done with other data; with other amounts of missing data and methods of eliminating data; and using various methods of rotation. Other criteria of goodness-of-fit may be explored, but the present criterion, of the cross-correlation of component scores derived from complete-data and estimated-data solutions, deserves further exploration. Component scores from the unrotated factor matrix should be derived directly from the unrotated factor matrix by equation (2), $X=F^{-1}Z$, and cross-correlated as an extension of the present procedure.

The present study showed that all four methods of data-estimation compared fairly well with the criterion: average absolute cross-correlations ranged between .72 and .79 for all 30 components, and between .84 and .88 for the first 11 components. The average correlations improved from the method of data-estimation employing least concomitant information (means substitution) to that employing most (multiple regression). When the first 11 components were considered alone, the improvement was not great, which indicates that means-substitution may be a viable method of estimating missing data.

References

1 Cattell, R. B. and Jaspers J., "General Plasmode (No. 30-10-5-2) For Factor Analytic Exercises and Research", Multivariate Behavioral Research Monograph, Society of Multivariate Experimental Psychology, No. 67-3, 1967.

2 Dixon, W. J., BMD, Biomedical Computer Programs, University of California Press, Berkeley and Los Angeles, 1968, pages 169 (03M) and 218 (02R).

3 Guertin, W. H., "Comparison of Three Methods of Handling Missing Data Observations", Psychological Reports, 1968, 22, page 896.