Florida Journal of Educational
Research, Volume XVII, 1975

REGRESSION ON FACTOR SCORES WITHOUT FACTOR SCORES: A COMMENT AND A PROGRAM

John D. Morris
University of Florida

Regression prediction using factor scores as independent variables seems to have become increasingly popular in recent years. Part of the reason probably arises from the widespread use of electronic computers. Without their speed, the tedious calculation of the factor scores would be impractical. Three fundamental justifications for the use of orthogonal factor scores are their (1) parsimony, (2) psychological meaningfulness, and (3) lack of intercorrelation.

When moving from data-level variables to factor scores, the number of predictors is always reduced. An attempt is also made to tap higher level constructs composed of many data-level variables. The result is usually that we transform our isolated variable scores into factors of the "trait" level which are more meaningful psychologically. Finally, the troublesome problem of partitioning a dependent variable's variance among correlated independent variables is solved in the case of factor scores resulting from an unrotated or orthogonally rotated principal components analysis since they are uncorrelated. Should the researcher choose the more psychologically sound common factor analysis (Guertin and Bailey, 1970, p. 148), the problem is only partially ameliorated since even with geometrically orthogonal solutions the resulting common factor score estimates are intercorrelated (Harman, 1967, p. 347).

The typical analytic procedure seems to be to calculate the factor scores and input these into a multiple regression program. The factor score calculation is unnecessary since a regression equation can be calculated with only the unfactored data-level variable intercorrelation matrix, the correlations of the criterion with the data-level variables, the criterion mean and variance, and the factor structure. If the sample size is large, this procedure outlined could save a considerable amount of computer time, thus money.

A formula for the correlation of factor I with factor II, both linear composites of the same variables, will be derived for the special case of two variables, and then generalized to any number of variables. The factor scores for individual i on factor I and II can be represented as:

$$F_{iI} = w_{I1}z_{i1} + w_{I2}z_{i2} \quad \text{and} \quad F_{iII} = w_{II_1}z_{i1} + w_{II2}z_{i2}$$

where the w's are estimates derived from the data-level variable intercorrelation matrix and the factor structure (Harman, 1967, p. 351). Since these factor scores are deviation scores (standard scores in the case of

principal components analyses), their covariance over N subjects can be represented as:

$$\text{cov}(I,II) = \sum_{i=1}^{N} (w_{I1}w_{II1}z_{i1}^2 + w_{I1}w_{II2}z_{i1}z_{i2} + w_{I2}w_{II1}z_{i2}z_{i1} + w_{I2}w_{II2}z_{i2}^2)/N$$

Simplifying,

$$\text{cov}(I,II) = (Nw_{I1}w_{II1} + Nw_{I1}w_{II2}r_{12} + Nw_{12}w_{II1}r_{12} + Nw_{I2}w_{II2})/N$$

or,

$$\text{cov}(I,II) = w_{I1}w_{II1} + w_{I1}w_{II2}r_{12} + w_{I2}w_{II1}r_{12} + w_{I2}w_{II2}$$

or, in general, for nv variables,

$$\text{cov}(I,II) = \sum_{i=1}^{nv} \sum_{j=1}^{nv} w_{Ii}w_{IIj}r_{ij}$$

$$\text{Since } S_I^2 = \sum_{j=1}^{nv} w_{Ij}a_{jI}$$

Where S_I^2 is the variance for factor I and a_{jI} is the loading for factor I on variable j in the factor structure (Harman, 1967, p. 352),

$r_{I,II} = \text{cov}(I,II)/S_I S_{II}$.

This formula can, of course, be used to calculate the total factor score intercorrelation matrix A with any number of factors. Since A is a full rank intercorrelation matrix, it will be positive definite, thus the necessary inversion required by the multiple regression technique can be easily accomplished. The vector C of correlations of the criterion with the factor scores can be similarly derived as:

$r_{Iy} = \sum_{i=1}^{nv} w_{Ii}r_{iy}/S_I$, where r_{iy} is the correlation of data-level variable i with the criterion. At this point, the vector B of standardized regression coefficients on factor scores can be calculated as $B_s = A\text{-}1_C$, and the raw-score regression coefficients can be calculated as usual as $b_{rI} = b_{sI}(S_y/S_I)$, where b_{rI} is the raw score regression coefficient on factor I, b_{sI} is the standardized regression coefficient on factor I, and S_y is the standard deviation of the criterion. Since the factor scores all have a mean of zero, the criterion intercept is merely the criterion mean. It should be pointed out that this procedure is applicable to factor scores resulting from rotated or unrotated principal components or common factor analyses.

Program Input and Output

A computer program written in FORTRAN IV completes all these calculations. The input required is the unfactored data-level intercorrelation matrix, the factor structure, the mean and standard deviation of the criterion, and the correlations of the criterion with the data-level variables. It might be pointed out that all of the above information is obtained from a data-level regression and factor analysis, thus factor score calculation is obviated.

The program outputs the weights for estimating factors, the factor score intercorrelation matrix, the correlations of the factor scores with the criterion, the multiple correlation coefficient of the regression on factor scores, and the standardized and raw-score regression equations on factor scores.

In addition to the regression use, the program has been found useful in obtaining the intercorrelations of common factor score estimates. This routine is not included in any factor analytic package known, and produces important information which is often ignored or unavailable. Obtaining the intercorrelations is accomplished by simply leaving blank the field reserved for the criterion's standard deviation. This is interpreted as a zero and is thus not a valid criterion standard deviation. Therefore, the zero is used as a flag to cause only the factor score intercorrelation matrix to be calculated. The only input necessary in this case is the variable intercorrelation matrix and factor structure, both output from any factor analysis.

Availability

The program with complete documentation, including a sample card set-up and output, are available from the author.

REFERENCES

Guertin, W. H., and Bailey, J. P. Introduction to Modern Factor Analysis. Ann Arbor: Edwards Brothers, 1970.

Harman, H. H. Modern Factor Analysis. Chicago: University of Chicago Press, 1967.