Validity and Cross-Validity of Metric and Nonmetric Multiple Regression

Robert C. MacCallum and Edwin T. Cornelius, III
The Ohio State University

Timothy Champney
Fels Institute

Several questions are raised concerning differences between traditional metric mutiple regression, which assumes all variables to be measured on interval scales, and nonmetric multiple regression, which treats variables measured on any scale. Both models are applied to 30 derivation and cross-validation samples drawn from two sets of empirical data composed of ordinally scaled variables. Results indicate that the nonmetric model is, on the average, far superior in fitting derivation samples but that it exhibits much more shrinkage than the metric model. The metric technique fits better than the nonmetric in cross-validation samples. In addition, results produced by the nonmetric model are more unstable across repeated samples. A probable cause of these results is presented, and the need for further research is discussed.

A common problem in data analysis involves the choice of appropriate methods for analyzing ordinally scaled data. Most traditional statistical methods invoke the assumption that scales of measurement are at least interval; such methods are described as *metric,* whereas techniques which are designed for data obtained from nominal or ordinal scales are called *nonmetric.* Since much of the data collected in the social sciences fail to meet the assumption of an interval scale, researchers must be concerned with the validity of results when such data are analyzed with metric methods. This concern has led to the development of a variety of nonmetric techniques, which should, in theory, provide more valid results than corresponding metric methods when applied to noninterval data. These nonmetric methods range from simple measures of association, such as Spearman's rank-order correlation and Kendall's tau coefficient, to complex nonmetric techniques, such as factor analysis (Kruskal & Shepard, 1974) and multidimensional scaling (Kruskal, 1964a, 1964b; Shepard, 1962a, 1962b; Takane, Young, & deLeeuw, 1977).

Comparative research often indicates that metric methods are quite robust to violations of the interval scale assumption and that there is relatively little advantage in using a nonmetric method when the assumption is violated (e.g., Havlicek & Peterson, 1974, 1977; Kruskal & Shepard, 1974; Weeks & Bentler, in press). These results, however, should not produce the general conclusion that the interval scale assumption of metric methods is irrelevant. There are situations in which nonmetric models *will* outperform metric methods, e.g., when observed data represent a severely nonlinear monotonic transformation of the true underlying variable. This has been shown in the context of factor analysis (Kruskal & Shepard, 1974) and three-way multidimensional scaling (Widaman, Hahn, & MacCallum, 1979). In addition, research is yet

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 4 Fall 1979 pp. 463–468

to be done on a number of relatively new nonmetric techniques, e.g., nonmetric principal components (Young, Takane, & deLeeuw, 1978) and nonmetric multiple regression (Young, deLeeuw, & Takane, 1976). Thus, there is a clear need for further research into the relative advantages and disadvantages of metric versus nonmetric methods of data analysis.

The present study represents an attempt to provide further information in this area. Metric multiple regression, represented by the traditional least squares procedure, is compared with nonmetric multiple regression, represented by the recently developed optimal scaling technique proposed by Young, deLeeuw, and Takane (1976). Given the popularity and practical utility of multiple regression, and given the common application of the metric model to ordinal data in practice, results of this study should be quite relevant to empirical users of the multiple regression model.

Nature of Metric and Nonmetric Multiple Regression

First, consider a brief comparison of the mathematical nature of these two techniques. Given measurements for N observations on p independent variables $X_1, X_2, \ldots X_p$, and one dependent variable Y, the metric approach solves for partial regression coefficients $b_1, b_2, \ldots b_p$ and an additive constant, a, such that $\sum_{i=1}^{N}(Y-\hat{Y})^2$ is minimized, where $\hat{Y} = b_1X_1 + \ldots + b_pX_p + a$. This is the traditional least squares criterion of fit.

The nonmetric method also optimizes this criterion but in addition is designed to take into account the scale of measurement of each of the $p + 1$ variables. This is accomplished by a scheme which alternates between two different stages of analysis: an estimation of weights phase and an optimal scaling phase. During the estimation phase, regression coefficients are estimated according to the least squares criterion defined above, treating all variables as if they were measured on at least an interval scale. During the optimal scaling phase, the coefficients are held constant and all variables are *rescaled,* or transformed, to further reduce the least squares criterion. The rescaling of the variables is based on the properties of the scale of measurement of each variable, i.e., for each variable X_j a transformation t_j is found that is permissible given the scale of X_j and that minimizes the least squares criterion. For example, for nominal variables, any t_j that maintains the relationships of equality and inequality is permissible; for ordinal variables, t_j must be monotonic.

After all variables have been rescaled, they are then held constant while regression coefficients are reestimated. Then, the coefficients are held constant while variables are again rescaled, and so on. Each iteration, composed of an estimation phase and an optimal scaling phase, further reduces the least squares criterion. When improvement becomes trivial, the process stops. Results consist of a multiple correlation and a set of regression coefficients as well as the transformations of each variable to its final optimally scaled form.[1]

Purpose

Given the common occurrence of noninterval scaled data in psychology, the availability of these two regression methods raises some important questions. For a regression problem in which at least some of the variables are not interval scaled, which method will provide more accurate predictions of the dependent variable in a given sample? Clearly, the nonmetric method must provide better fit to a given sample because of its larger number of parameters; but the degree of the advantage is un-

[1]A computer program called MORALS (Multiple Optimal Regression by Alternating Least Squares), which performs this nonmetric regression analysis, is available from Forrest Young, L. L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC 27514.

known. A much more important issue arises with respect to the matter of cross-validity, i.e., the application of a regression model derived from one sample to a new sample of observations from the same population. Typically, due to capitalization on chance in the least squares fitting process, predictions will be less accurate in the new sample (the cross-validation sample) than in the original sample (the derivation sample). This "shrinkage" of the multiple correlation is a critical issue when predictive accuracy is an important research objective. An obvious question then concerns whether nonmetric solutions will tend to exhibit more or less shrinkage than metric solutions for the same set of data.

The present study is an investigation of the relative performance of metric and nonmetric multiple regression in both derivation and cross-validation samples selected from two sets of empirical data.

Method

Data

The analyses were carried out using data collected from two different test validation studies. The first study was a concurrent validity research design. Data consisted of six variables measured on a sample of 173 foremen in two chemical processing plants. Four of the variables were used as independent variables: (1) a test of foreman job knowledge; (2) a leadership opinion questionnaire assessing two leadership factors—consideration of other workers and initiation of psychological structure; (3) an empirically keyed biographical inventory; and (4) a supervisory judgment test, assessing decision-making and problem-solving abilities. The two dependent variables were (1) a salary administration ranking, i.e., supervisors' ranking of the foremen in terms of overall worth to the company and (2) a supervisors' rating of job performance, which was a summed score across several graphic rating scales.

The second study was a predictive validity research design. Data consisted of measurements on five variables for a sample of 207 noncommissioned officers of the U. S. Army participating in the Army paramedical training program for physician assistants. Four of the variables were independent variables: (1) a measure of medical and surgical knowledge as determined by panel ratings from an oral board interview; (2) the Otis test of mental ability; (3) an alternative-functions test of creativity; and (4) a measure of general knowledge of science and medicine. The dependent variable was the percentage of points earned across the various subcourses in the year-long training program.

Derivation and Cross-Validation Samples

These data sets lent themselves well to the present study in that all of the variables could be considered to be only ordinally scaled. To establish a basis for comparison of the two regression models, each of these samples was randomly divided 15 separate times into a derivation sample, composed of two-thirds of the original sample, and a cross-validation sample, composed of the remaining one-third. Thus, for the foreman data, 15 separate derivation samples, each containing 115 observations, and 15 corresponding cross-validation samples, each with 58 observations, were constructed. Likewise, the Army sample was randomly divided 15 times into a derivation sample of 139 cases and a cross-validation sample of 68. Given the relatively small numbers of variables, all of these sample sizes would be considered fairly large—certainly large enough to satisfy the usual rules-of-thumb regarding sample sizes necessary for stable results (see Wherry, 1975).

Analysis

Each of the 30 derivation samples was then analyzed using both metric and nonmetric multiple regression. Since there were two different criterion variables in the foreman data,

this produced 45 different metric solutions and 45 corresponding nonmetric solutions. For the nonmetric analyses, all variables were specified as ordinally scaled. For each solution, squared multiple correlations were obtained. Each derived solution was then applied to the corresponding cross-validation sample. To cross-validate the metric solutions, the derived regression equations were simply applied to the independent variables in the cross-validation samples and the squared correlation was obtained between the resulting predicted values of Y and the observed values of Y. To cross-validate the nonmetric solutions, the derived optimal scaling transformations were first applied to the independent and dependent variables in the cross-validation samples; then, the derived regression equations were applied to the rescaled variables and the cross-validated squared multiple correlations were computed.

Results

Table 1 presents means and standard deviations for the derived and cross-validated squared multiple correlations from both metric and nonmetric analyses. Each mean and standard deviation in Table 1 is based on 15 samples. The table also presents the means and standard deviations of the degree of shrinkage exhibited by both models.

As mentioned above, the nonmetric regression model will necessarily produce higher R^2's in derivation samples than will the metric model. Table 1 shows this advantage to be quite substantial. On the average, the derived R^2 from the nonmetric model were nearly 2.5 times as large as those from the metric model. However, it should also be noted that the standard deviations were also higher in the nonmetric solutions. Thus, in the derivation samples, the non-

Table 1

Means and Standard Deviations of Squared Multiple Correlations Across 15 Samples from Three Data Sets

Method and Statistic	Data Set					
	Foremen[a]		Foremen[b]		Army Physicians Assts.	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Metric Solutions						
Derived R^2	.195	.042	.321	.037	.142	.028
Cross-Val. R^2	.030	.063	.261	.067	.125	.056
Shrinkage	.065	.103	.059	.103	.017	.083
Nonmetric Solutions						
Derived R^2	.526	.085	.628	.049	.441	.059
Cross-Val. R^2	.061	.041	.156	.094	.033	.036
Shrinkage	.465	.115	.472	.135	.408	.085

[a] Dependent variable was Salary Administration Ranking
[b] Dependent variable was Job Performance Rating

metric model provided substantially better fit but was more unstable over repeated sampling than were the metric solutions. For instance, in the foreman sample using the salary administration ranking as the dependent variable, derived R^2's for the nonmetric model ranged from .44 to .77, and those for the metric solutions ranged from .13 to .26.

In examining cross-validity of the two models, a distinct and important contrast is seen. The mean level of shrinkage of R^2 from the metric solutions was always less than .10, but the mean for nonmetric solutions was always greater than .40. In addition, the standard deviations again show the results produced by the nonmetric technique to be more unstable. The drastic difference in mean level of shrinkage has the effect of eliminating the large advantage which the nonmetric model held in the derivation samples. A comparison of the mean R^2's for the two models in the cross-validation samples shows that the metric model provided consistently better fit in the new samples.

Discussion

Three important effects were revealed in the results. First, the nonmetric model was distinctly superior in fitting a given sample; second, shrinkage produced by nonmetric solutions was severe, resulting in the metric model providing better fit when derived solutions were fit to cross-validation samples; and third, results of the nonmetric technique were noticeably more unstable than the metric method in terms of derived R^2's and amount of shrinkage. All of these effects can probably be attributed to the presence of the optimal scaling phase in the nonmetric algorithm. Recall that this phase involves determining a transformation of each variable so as to further reduce the least squares criterion. This process represents the critical difference between the nonmetric and metric methods. In effect, this difference means that the nonmetric procedure estimates more parameters than does the metric procedure from the same amount of data. Therefore, the nonmetric technique may be more sensitive to chance fluctuations in the derivation sample, i.e., fluctuations arising from sampling error and/or error of measurement. This sensitivity amounts to increased capitalization on chance. While it would serve to improve fit in the derivation samples, the instability of the transformations of the variables would result in poor performance under cross-validation, as has been observed. As further support for this explanation, it was found in the present study that the transformations produced by the optimal scaling phase of the nonmetric model were quite unstable across derivation samples.

It is expected that the instability and great shrinkage associated with the nonmetric model would be alleviated in very large samples. That is, with extremely large sample size, transformations produced by the optimal scaling phase should be more stable, thus reducing capitalization on chance and producing less shrinkage. It seems likely that for a large enough sample, the nonmetric model might achieve better fit than the metric model in cross-validation, as well as derivation, samples. The sample size necessary to achieve this effect is unknown, but it may be extremely large. Recall that sample sizes in the present study would be considered relatively large by most standards; derivation sample N's were more than 20 times the number of variables in all cases. Monte carlo studies would be very useful in investigating the relationship of sample size to the phenomena observed in the present project.

Conclusions

The results presented above have important implications for the empirical researcher who wishes to use multiple regression on noninterval scaled variables. In the rare case in which the researcher is primarily interested in fitting a particular sample, nonmetric regression will probably provide substantially better fit than the traditional metric method. However, when

predictions for new observations are important, present results indicate that the metric method may be preferable to the nonmetric, even when all variables are ordinal. Of course, this demonstration does not prove that the metric approach is always superior. In fact, it is probably demonstrable that there are at least two situations in which the nonmetric method would perform better under cross-validation. The first, mentioned above, would be when sample size becomes very large. A second would be when there are severely nonlinear monotonic relations between the dependent variable and one or more independent variables. However, such situations seem to be relatively uncommon in practice, and the authors believe that present results are fairly representative of the real world.

References

Havlicek, L. L., & Peterson, N. L. Robustness of the *t* test: A guide for researchers on effect of violations of assumptions. *Psychological Reports,* 1974, *34,* 1095-1114.

Havlicek, L. L., & Peterson, N. L. Effect of the violation of assumptions upon the significance levels of the Pearson *r. Psychological Bulletin,* 1977, *84,* 373-377.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika,* 1964, *29,* 1-27. (a)

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika,* 1964, *29,* 115-129. (b)

Kruskal, J. B., & Shepard, R. N. A nonmetric variety of linear factor analysis. *Psychometrika,* 1974, *39,* 123-157.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika,* 1962, *27,* 125-140. (a)

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika,* 1962, *27,* 219-246. (b)

Takane, Y., Young, F. W., & deLeeuw, J. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika,* 1977, *42,* 7-67.

Weeks, D. G., & Bentler, P. M. A comparison of linear and monotone multidimensional scaling models. *Psychological Bulletin,* in press.

Wherry, R. J. Underprediction from overfitting: Forty-five years of shrinkage. *Personnel Psychology,* 1975, *28,* 1-18.

Widaman, K. F., Hahn, J., & MacCallum, R. C. *The recovery of structure in ordinal data by INDSCAL and ALSCAL.* Manuscript submitted for publication, 1979.

Young, F. W., deLeeuw, J., & Takane, Y. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika,* 1976, *41,* 505-529.

Young, F. W., Takane, Y., & deLeeuw, J. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika,* 1978, *43,* 279-281.

Author's Address

Send requests for reprints or further information to Robert C. MacCallum, Department of Psychology, Ohio State University, 404C W. 17th Avenue, Columbus, OH 43210.