SOME ESTIMATION PROBLEMS IN MULTIPLE REGRESSION ANALYSIS

Dr. Ali Delavar
University of Allame Tabatabai

[DOR: 20.1001.1.25382640.1990.1.1.2.0]

Abstract

The Multiple Linear Regression is widely used in psychological and educational research for perdiction purposes. Misuse and misinterpretation of multiple regression techniques have prompted investigators to formulate a set of "rules of thumb" to assist researchers in their application to small and medium-sized samples ----- a condition aggravated by the use of advanced computers. Keeping this aim in mind, the problems to be subjected to empirical investigation are as follows:

1. **The effectiveness of regression weights selected by variance-reduction procedures.**
2. **The accuracy and usefulness of formulas for estimating the predictive effectiveness or population validity of a sample regression equation.**
3. **The accuracy and usefulness of formulas for estimating the population multiple correlation.**
4. **The effect of the variation of some parameters of the population distribution on the results of (1), (2) and (3).**

Introduction

A major aim of science is such precise descriptions of phenomena and their relationships that accurate forecasts of future findings or happenings become possible. In astronomy, for example, eclipses are predicted with a high degree of accuracy; similarly, in chemistry it is often possible to state properties of a compound before the substance is actually in existence. Psychologists and educators aim to understand human behavior. While it is extremely unlikely that human behavoir will ever be completely predictable, one of the most important presuppositions in the study of human behavior is the notion that people do not behave in an entirely random fashion. It is assumed that they learn ways of responding to environmental stimuli which are personally rewarding, and are more likely to respond in the future in ways that result in rewards than in some other fashion.

It is further believed that groups of individuals also display non-random patterns in their behavior and this information is valuable for predicting the actions of individuals in groups. Statistical techniques may be advantageously used in prediction behavior of both individuals and groups. The most popular technique for prediction, currently in use, is multiple linear regression.

The multiple linear regression is widely used in the psychological and educational research for prediction purposes. One of its earliest applications was in the assessment of selection procedure for assigning children to different types of schools (Overall, 1972), but it hae also been used in many fields of inquiry.

In recent years, electronic computers have made the multiple regression method readily available to psychologists, educators and other scientists, while simultaneously making it unnecessary for them to study, in full, the cumbersome computational details of the method (Darlington, 1969). Yet, the situation seems little improved from what it was when Cureton (1950, p.690) wrote: "It is doubtful that any other statistical techniques have been so generally and widely misused and misinterpreted in educational research as have those of multiple correlation." All too often the nature of the data used, or the size of the sample employed, is not standard for multiple regression purposes. There is little reason to expect improvement in this situation and, indeed, the ready availability of standard computer regression programs may make the situation worse. Therefore, empirical studies should be done to formulate a set of "rules of thumb" to aid the researcher in the application of multiple regression techniques to small and mediumsized samples.

To use a multiple regression model requires scores from a random sample of individuals from the population of

interest on a number of predictor variates, the so-called independent variates, and also scores on a criterion variate so-called dependent variate, which we wish to predict. The problem is to find a set of weights to apply to independent variates which will maximize the correlation between their combined effect on one hand and the dependent variate on the other. There are many independent variable weighting techniques which range from subjective guesses, concerning the relative importance of each predictor, to complex procedures involving factor analysis or component analysis. Each of these techniques has certain properties and advantages, and the proper one to use in a given situation depends upon the nature of the data and the purpose of study. For a discussion of many of these techniques the reader is referred to Stanley and Wang (1968).

The essential task of multiple regression analysis is to develop a prediction equation by solving a set of normal equations: "n" simultaneous linear equations derived from the intercorrelation matrix of the dependent variable and the "n" independent variables. Thus, in order to carry out a multiple regression analysis it is necessary to have measures of both independent and dependent variables. The solution of a set of "n" normal equations is termed "a least" squares solution because when the weights derived from a sample are applied to that sample, the sum of squares of deviations of the actual values from the predicted values of the dependent variable are a minimum:

$$\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \min. \qquad (1)$$

where

$$\tilde{y}_i = w_1x_{1i} + w_2x_{2i} + \ldots + w_nx_{ni}. \qquad (2)$$

y_i = ith observation on the dependent variable y.
$\tilde{y}_i$ = the predicted value of the ith observation of the dependent variable y.
w_i = weights derived by multiple regression.
x_{ki} = ith observation on independent variable x_k (k = 1, 2, ...n).
n = number of observation (i = 1, 2...n).

All variables are assumed to be in standard score form, i.e., there is a common mean for all variables. The set of weights derived by multiple regression applied to standard scores are termed "beta weights". The weights thus derived are "least square" estimates of the true population weights; the weights that would minimize the sum of squared errors for the population. Unfortunately, weights determined for one sample of subjects will not usually satisfy the minimum sum of squared errors condition in another sample. The sample regression weights can be derived as:

$$\hat{\underset{\sim}{B}} = (x'x)^{-1}x'v. \qquad (3)$$

Because of standardization of the data x'x is the sample intercorrelation matrix of the predictors R_{xx}; x'y is the matrix of sample correlations between the independent variables and dependent variable R_{xy}. Therefore formula (3) can be written as:

$$\hat{\underset{\sim}{B}} = R^{-1}_{xx} R_{xy}$$

When multiple regression is used to determine the regression weights, a multiple correlation coefficient can be determined. This is the Pearson product-moment correlation between the actual dependent variable values (y_i) and the predicted dependent variable values ($\tilde{y}_i$). Thus the multiple regression weights have two primary properties: (1) the sum of squares of differences between the actual and predicted dependent variable values will be minimum, and (2) the correlation between the actual and predicted dependent variable values will be at maximum, where both of these properties apply to the sample from which the weights were driven.

The multiple correlation, which is defined as the degree of relationship between the predictors and criterion (dependent variable), is a biased estimate of this relationship and generally larger than the true multiple correlation for the population (Herzberg, 1969). This estimation is biased because the first property (minimizing the sum of square of errors) actually determines the second property (maximizing the correlation between the actual and predicted variable values). As a result, the estimate of the multiple correlation will be higher than the actual population multiple correlation.

One problem in the application of multiple correlation techniques is, therefore, the estimation of the true multiple correlation from the biased sample multiple correlation (Herzberg, 1969).

The Shrinkage of the Multiple Correlation Coefficient in Cross-Validation

The process of applying weights obtained from one sample to predictors of another has been called "cross-validation" (Langmuir, 1954). The cross-validation problem is concerned with the fact that, in general, the correlation between the predictors and predicted scores in the second group is less than the multiple correlation coefficient computed in the first group. This phenomenon is called shrinkage of the multiple correlation coefficient by most writers. Shrinkage is attributed to "overfitting" (Cureton, 1950) in the sense that the weights obtained are the weights which insure maximum efficiency on the variable observation. Although this problem has been recognized for some time, a search of the literature has failed to reveal any studies which indicate the frequency or amount of shrinkage. In most of the published material the problem is discussed from a theoretical point of view without obtaining empirical verification. Consideration of the problem of predicting the shrinkage seems to have been first treated in the literature by S.C. Larson (1931) in which he attributed the shrinkage formula as following to B. B. Smith.

$$\hat{R}^2 = 1 - \frac{N}{N-n}(1 - r^2). \qquad (4)$$

This formula was modified by Wherry (1931) to yield what is now the most widely used formula for estimating the squared population multiple correlation, given a

[DOR: 20.1001.1.25382640.1990.1.1.2.0]

sample multiple correlation coefficient.

$$\hat{R}^2 = 1 - \frac{N-1}{N-n}(1 - r^2) \qquad (5)$$

where

N = sample size
n = number of independent variables
r = sample multiple correlation coefficient
R = population multiple correlation coefficient.

The formula (5) is applicable to a zero constant; when the constant is not zero the estimate of R^2 is:

$$\hat{R}^2 = 1 - \frac{N-1}{N-n-1}(1-r^2). \qquad (6)$$

Formula (6) is often referred to as Wherry's formula. Larson and Wherry have attempted to obtain an unbiased estimate of the population correlation coefficient. However, this estimate which involves the ratio of two unbiased estimates is not an unbiased of R^2, nor is its square root an unbiased estimate of R (Darlington, 1968).

The main idea behind the papers of Larson and Wherry seems to have been an attempt to obtain an unbiased estimate of population multiple correlation coefficient. Since the maximum likelihood estimate for the multiple correlation coefficient from a normal sample is biased upward (Nicholson, 1960), the argument for correcting the bias to account for shrinkage might proceed as follows:

> The prediction equation used in the second sample is not the least square equation which, by definition, is the sample multiple correlation coefficient. Since we expect the sample multiple correlation coefficient from samples of the same size drawn at random from the same population to be equal, apart from sampling errors, we expect to get smaller multiple correlation coefficients from samples in which we use linear predictors other than otpimum ones.

Lord (1950) and Nicholson (1960) published an unbiased estimate of R^2:

$$\hat{R}^2 = (1-(\frac{N-1}{N-n-1}))(\frac{N+n+1}{N})(1-r^2). \qquad (7)$$

This formula is applicable to the regression model with a constant term. This formula was modified by Darlington (1968) for the correlation model with a constant term (Herzberg, 1969). His formula is

$$\hat{R}^2 = 1 - \frac{N-1}{N-n-1}\frac{N-2}{N-n-2}\frac{N+1}{N}(1-r^2). \qquad (8)$$

Okin and Pratt (1958) published an unbiased estimate of R^2:

$$\hat{R}^2 = 1 - \frac{N-2}{N-n}(1-r^2)\,F(1,1;\frac{N-n+2}{2};1-r^2). \qquad (9)$$

where F(a, b, c, x) is the hypergeometric function. Pratt (1958) provided an approximation to (9), based on the first two terms of the hypergeometric series and a partial correction for the omitted later terms:

$$\hat{R}^2 = 1 - \frac{(N-3)(1-r^2)}{N-n-1}[1 + \frac{2(1-r^2)}{N-n-2.3}]. \qquad (10)$$

Herzberg (1969) has also given an approximation to (9):

$$\hat{R}^2 = 1 - \frac{(N-3)(1-r^2)}{N-n-1}[1 + \frac{2(1-r^2)}{N-n+1}]. \qquad (11)$$

Formulas (10) and (11) differ only in the value of divisor in the rightmost term.

Claudy (1970) published an unbiased estimate of $\hat{R}$:

$$\hat{R} = [1 - \frac{(N-4)(1-r^2)}{N-n+1}(1 + \frac{2(1-r^2)}{N-n+1})]^{1/2}. \qquad (12)$$

Examination of (6), (9) and (10) indicates that as the size of the sample (N) increases and/or the number of independent variables or predictors (n) decreases, the amount of shrinkage of the sample multiple correlation coefficient decreases. This is directly in accordance with Fisher's (1924) original estimate of the expected value of the square of the sample multiple correlation coefficient (Herzberg, 1969):

$$\hat{r}^2 = 1 - \frac{N-n}{N-1}(1 - R^2). \qquad (13)$$

which shows that the degree of overestimation by the sample multiple correlation coefficient is directly proportional to the number of predictors and inversely proportional to the size of the sample.

Estimation of the regression coefficient is an important aspect of regression analysis. While we do not know what the true parameter values are, we try to construct as good estimates as possible. In particular, the estimates should be unbiased so that the mean of the sampling distribution for a regression coefficient estimate equals the value of the parameter. Another desirable property is to have a small variance in the sampling distribution.

The problem arises from the fact that the sample regression coefficients differ from the population regression coefficient . Therefore, if we use the sample regression coefficient in the population, the resulting aggregate correlation will be lower than the population multiple correlation. By using the sample regression coefficients in the population, we will get population validity coefficient. The method in common use for estimating this aggregate correlation is called cross -validation: Applying weights obtained from one sample to the predictors of another from the same population. The Pearson productmoment correlation between the criterion and the predictor values in the second sample is the aggregate correlation which is termed "the cross-validation coefficient" (r_c).

Mosier (1951) suggested an extension of crossvalidation

[DOR: 20.1001.1.25382640.1990.1.1.2.0]

which he termed double cross-validation. A given sample is split into two independent subsamples and beta weights are derived from both subsamples. The beta weights from each subsample are applied to the other subsample to yield two aggregate correlations of cross-validities. The average of the cross-validities is used as an estimate of the actual validity in the population.

Cureton (1962) proposed a method to determine sample regression weights whose variability will be more nearly that of the population beta weights.

This method is termed the "least deviant" procedure:

1. Divide the original sample into two equal subsamples.
2. Determine the beta weights in each subsample.
3. Arrange the beta weights from both subsamples in a single rank order from highest positive to lowest positive or highest negative, and determine the median.
4. From the pair of beta weights for each variable, select the one nearest the median and use this as the weight for that variable in the regression equation.

A second method for reducing the variability of sample beta weights, termed here the "average" procedure, is suggested. However, it will not reduce the variability as much as will the "least deviant" procedure:

1. Divide the original sample into two equal subsamples.
2. Determine the beta weights for each subsample.
3. Determine the mean of pair of beta weights for each variable and use this mean as the weight for that variable in the regression equation.

Though their relation to the double cross-validation procedure is obvious, neither of these variance-reduction procedures has any real theoretical or mathematical basis, and neither has been empirically studied.

For many years little distinction was made between estimates of the population multiple correlation, obtained by shrinkage of a sample multiple correlation coefficient, $\hat{R}$; and estimate of the aggregate correlation, $\hat{R}_c$, obtained by cross-validation (Guilford, 1965; Guion, 1965). The implication being that they were basically equivalent, differing only in how they were obtained. This is not the case. On the average, both values are smaller than the population multiple correlation, and the population validity coefficient is smaller than the population multiple correlation. When working with a single population, especially where sample size is small, the following inequality holds (Herzberg, 1969):

$$\bar{r} > \hat{R} > \hat{R} \approx \bar{r}_c \qquad (14)$$

The cross-validity coefficient, r_c, is an estimate of the validity or predictive effectiveness of regression equation derived in one sample when it is applied to a second independent sample. As with sample multiple correlation coefficient, we are not interested in the sample cross-validity coefficient. What we do want to estimate is the population validity coefficient, R_c; that is the predictive effectiveness of a sample regression equation when it is applied to the entire population from which the sample was drawn.

Since in actual practice we do not have available the entire population of interest, we can never directly calculate R_c. Instead we must find some way to estimate this value. Herzberg (1969) provides two equations for doing this. The first is based on the work of Lord (1950) and Nicholson (1948) and applied to «regression" or Fixed-X model:

$$\hat{R}^2_c = 1 - \frac{N-1}{N-n-1}\left(\frac{N+n+1}{N}\right)(1 - r^2). \qquad (15)$$

The second is due to Darlington (1968) and is for the "correlation" or Random-X model:

$$\hat{R}^2_c = 1 - \frac{N-1}{N-n-1}\left(\frac{N-2}{N-n-2}\right)\left(\frac{N-1}{N}\right)(1 - r^2). \qquad (16)$$

Burket (1964) also provides a formula by which the population validity, which he calls the "weight validity", can be estimated:

$$\hat{R}_c = \frac{Nr^2 - n}{r(N-n)} \qquad (17)$$

The formula (17) provides a direct estimate of the population validity rather than its square. But it is applicable within the context of Fixed-X regression and may not be suitable when applied to Random-X data.

In applying the multiple regression technique we are looking for two types of outcomes. We are seeking a regression equation whereby we can predict the criterion; and, we seek a measure of the predictive effectiveness (validity) of the regression equation in the population. In order for the sample regression equation to represent the population regression equation, the sample from which it is to be derived should be as large as possible. The entire sample should be used. If this is done, one does not have an independent sample for cross-validation to determine the validity of regression equation. Therefore, it is impossible to determine both the desired outcomes of a multiple regression. This problem is discussed by Horst (1966) and Moiser (1954). So far, formulas such as (14), (15) and (16), which allow the direct estimation of the population crossvalidity have been used. However, their use has not been subject to empirical test. Therefore, it is necessary to investigate empirically the accuracy of estimation formulas for population validity. This will be especially important in studies where the size of the available sample is small. Discussing this problem Horst (1966, p. 378) says:

> Ideally, there should be procedures for estimating this shrinkage without bias, but such methods have not yet been satisfactorily worked out from a theoretical point of view and no really satisfactory methods are available for computational purposes.

When multiple regression is used to compute predictor weights, the value of the multiple correlation in the sample may be calculated. But, there is no way to obtain the value of the multiple correlation in the population; most frequently this value has been estimated by the Wherry "shrunken multiple correlation" formula (6).

The Olkin-Pratt (1958) (9) and Herzberg (1969) approx-

[DOR: 20.1001.1.25382640.1990.1.1.2.0]

imations to it (10), (11) and (12) have been advanced as unbiased estimates of the squared multiple correlation in the population. They, too, have not been subjected to empirical tests. Thus, the next problem which is subject to investigation is: The accuracy of estimation formulas for the populaton correlation.

In summary the following problems are subject to empirical investigation:

1. The effectiveness of regression weights selected by variance-reduction procedures.
2. The accuracy and usefulness of formulas for estimating the predictive-effectiveness or population validity of a sample regression equation.
3. The accuracy and usefulness of formulas for estimating the population multiple correlation.
4. The effect of the variation of some parameters of the population distribution in the results of (1), (2) and (3).

REFERENCES

Burkett, G. R. A study of reduced rank models for multiple prediction. **Psychometric Monographs,** 1964, No . 12.

Claudy, J. G. An empirical investigation of small sample multiple regression and cross-validation. Ann Arbor, Mich.: University Microfilms, 1970, No. 70-7553.

Cureton, E. E. Validity in E. F. Lindquist (Ed.), **Educational measurement.** Washington, D.C.: American Council on Education, 1950.

Cureton, E. E. Approximate linear restraints and best predictor weights. **Educational and Psychological Measurement.**

Cureton, E. E. Multivariate psychological statistics . The University of Tennessee, Knoxivlle, 1962.

Darlington, R. B. Multiple regression in psychological research and practice. **Psychological Bulletin,** 1968, 69, 161-182.

Fisher, R. A. The influence of rainfall on the yield of wheat at Rothemsted. **Philosophical Transactions of Royal Society of London,** 1928, 213B, 89-142.

Grabill, F. A. **An introduction to linear statistical models, Volume I.** New York: McGraw-Hill Book Company, 1961.

Guilford, J. P **Fundamental statistics in psychology and education** (4th ed.). New York: McGraw-Hill Book Company, 1965.

Herzberg, P. A. The parameter of cross-validation. **Psychometrica,** 1969, 34.

Horst, P. **Psychological measurement and prediction.** Belmont: Wadsworth and Company, 1966.

Larson, S. C. The shrinkage of the multiple correlation coefficient. **Journal of Educational Psychology,** 1931, 22, 45-55.

Lawshe, C. H. Statistical theory and practice in applied psychology. **Personnel Psychology,** 1969, 22, 117-124.

Lawshe, C. H., & Schucker, R . E. The relative efficiency of four test weighting method in multiple prediction. **Educational and Psychological Measurement,** 1959, 19, 103-114.

Lord, R. M. Efficiency of prediction when a regression equation from one sample is used in a new sample (Research Bulletin No. 50-40). Princeton: Educational Testing Service, 1950.

Mosier, C. I. Problems and design of cross-validation. **Educational and Psychological Measurement,** 1951, 11, 1-11.

Nicholson, G. E., Jr, The application of regression equation to a new sample. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill, 1948.

Olkin, I., & Pratt, J. W. Unbiased estimation of certain correlation coefficient. **Annals of Mathematical Statistics,** 1958, 29, 201-211.

Overall, E. John. **Applied multivariate analysis.** New York: McGraw-Hill Book Company, 1972.

Stanley, J. C., & Wang, M. D. **Differential weighting- A survey of methods and empirical results.** New York: College Entrance Examination Board, 1968.

Wesman, A. G., & Bennett, C. K. Multiple regression vs. simple addition of scores in prediction of college grades. **Educational and Psychological Measurement,** 1959, 19, 243-246.

Wherry, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. **Annals of Mathematical Statistics,** 1931, 2, 446-457.

[DOR: 20.1001.1.25382640.1990.1.1.2.0]