PREDICTING A SINGLE CRITERION FROM MULTIPLE PREDICTORS: MATHEMATICAL WEIGHTING[1]

Jacob G. Beard
Florida Institute for Continuing University Studies

The preceding presentation gave a procedure for logically combining predictors for the estimation of a single criterion. The purpose of this paper is to discuss a mathematical method of combining predictors for the same purpose, MULTIPLE REGRESSION. Multiple regression is often used for predicting academic success from two or more test scores and this discussion will center around that kind of application.

In the multiple regression technique, weights are computed for each predictor which will produce the maximum correlation between the combination of predictors and the criterion. It is a mathematical "least squares" solution in that the sum of squares of the differences between the predicted scores and the obtained scores will be a minimum. This technique, when properly used, will give the most correct predictions for the greatest number of subjects in a group situation.

We can use this technique in our school prediction problems, provided that the following requirements are met in the data.

1. The variables in the original correlation matrix must be linearly related. (Techniques exist for non-linear regressions, but they will not be discussed here.) A simple scattergram will reveal whether this is true or not.

2. The number of cases should be <u>large</u>, at least one-hundred. The validity coefficient of a battery of predictors will usually shrink when it is cross-checked on another sample. This cross-check should always be made and the shrinkage in validity will be greater for small samples than for large ones. If a cross-check is not possible, then a formula for a mathematical correction in the validity coefficient can be found in most statistics texts.

In addition to the above requirements, multiple regression will be unprofitable unless the predictors have a relatively low correlation with each other.

[1] Paper presented to the Eighth Annual FERA Testing Conferences, Tallahassee, Florida, January 25, 1964.

As an example let us consider the prediction of twelfth grade test scores from scores of tests given in the ninth grade. The correlation matrix for these variables is shown in Table 1.

Table 1

Intercorrelations Among the Scores of Three Tests

Variable	1	2	3
1	-	.80	.66
2		---	.57
3			---

1 = Total score from Florida twelfth grade tests.
2 = SCAT V from Florida ninth grade tests.
3 = SCAT Q from Florida ninth grade tests.

The SCAT Verbal and Quantitative scores are the predictors and the Florida twelfth grade total score is the criterion measure.

Our object is to determine the optimum weights to assign to each of the two predictor variables (2 and 3) in order to maximize their multiple correlation with the criterion (1). The correlation matrix (Table 1) contains the information needed to compute the two beta weights. Formula 1 shows the calculation of the beta weight for variable two, called β_2, and Formula 2 gives the calculations for β_3.

The beta weights of .62 and .29 can then be used in computing the multiple correlation coefficient between our two predictors and the criterion. The coefficient obtained is .83 (Formula 3).

Formula 1.

$$\beta_2 = \frac{r_{12} - r_{13}\, r_{23}}{1 - r^2_{23}}$$

$$= \frac{.80 - (.66)\ \ (.57)}{1 - (.57)^2} \quad = \quad .62$$

Formula 2.

$$\beta_3 = \frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}$$

$$= \frac{.66 - (.80)\,(.57)}{1 - (.57)^2} = .29$$

Formula 3.

$$R = \sqrt{\beta_2 r_{12} + \beta_3 r_{13}}$$

$$= \sqrt{.62(.80) + .29(.66)} = .83$$

The addition of the quantitative score increases the coefficient of correlation slightly over that of our best single predictor, which was the Verbal score. However, the correlation coefficient of .57 between (2) and (3) indicates that the two predictors are not unrelated measures. If the correlation coefficient between the scores of the two predictors had been decreased by one-tenth (.10) to .47, other values remaining the same, then the multiple R would have been .86 instead of .83.

Remember that we should not accept .83 as the validity coefficient ofthe battery until it has been cross-checked on another sample. The shrinkage should be small in this case because of the large number of subjects in the sample, 2735.

The matrix of correlation coefficients in Table 2 more nearly meets the criteria of relatively high correlation with the criterion and low correlation between the predictors. The low correlation of .09 between the predictors indicates that each measures a different aspect of the criterion we are attempting to predict. By considering the scores on X_3 together with X_2 the validity of our prediction would be .84, a subtantial increase over the validity of either of the predictors when used alone.

Table 2

Correlation Matrix Showing Relatively High Correlations Between the Criterion (1) and the Two Predictors (2 and 3), and Low Correlation Between the Two Predictors, (2) and (3)

Variable	1	2	3
1	-	.68	.55
2		---	.09
3			---

$\beta_2 = .64$, $\beta_3 = .50$, and $R = .84$

In actual practice, it is quite difficult to locate psychological variables which can be combined to a significant advantage. The most productive combinations are likely to be intellectual with personality variables. It is seldom worthwhile to combine more than three or four variables, since each additional predictor, after the first, contributes much less to prediction than those previously added.

We are now ready to construct the regression equation. The beta weights are used in conjunction with the correlations to predict the criterion scores. The multiple regression equation in standard score form is shown in Formula 4, where Z is the predicted standard score and Z_2 and Z_3 are the standard scores of the two predictors.

Formula 4.

$$Z' = \beta_2 Z_2 + \beta_3 Z_3$$

In order to predict raw scores we merely substitute the appropriate term $\frac{X - \bar{X}}{S.D.}$ for each Z in the above equation and this has been done in Formula 5. X' is the predicted raw score and the solution for that value is not complicated.

Formula 5.

$$\frac{X' - \bar{X}_1}{S.D._1} = \beta_2 \frac{X_2 - \bar{X}_2}{S.D._2} + \beta_3 \frac{X_3 - \bar{X}_3}{S.D._3}$$

The correlation between a group of obtained scores and the predicted values of those scores from this equation is the multiple correlation coefficient.

Multiple Regression can be used profitably in a variety of school prediction settings, if the investigator will select predictor variables which are not highly related to each other--such as intellectual, personality, and psychomotor variables. It is unwise to combine measures of the same ability in the hope of improving prediction. The examples given here involve only two predictor variables for simplicity of illustration. The computational labor and difficulty of interpretation increase greatly as the number of predictors increase; however, computer programs are available now which can analyze up to 30 variables, pick a preselected number of the most efficient predictors, and compute the regression equations in a few minutes. (Personnel from the State Universities would be helpful if you have an application of this technique which requires a large scale computer.) The availability of electronic computers can make multiple correlation and regression a highly useful tool in educational research.

References

1. Ferguson, George A. Statistical Analysis in Psychology and Education. McGraw-Hill Book Company, Inc., New York, 1959.

2. Guilford, J. P. Fundamental Statistics in Psychology and Education. McGraw-Hill Book Company, Inc., New York, 1956.

3. Lindquist, E. F. (ed.) Educational Measurement. American Council on Education, Washington, D. C., 1951.