REGRESSION EFFECTS WHEN THE ASSUMPTION OF RECTILINEARITY IS NOT TENABLE[1]

JOHN T. POHLMAN, Department of Guidance and Educational Psychology, Southern Illinois University, Cardonbale, IL 62901
ISADORE NEWMAN, Department of Educational Foundations, University of Akron, Akron, OH 44325

Abstract. When analyzing data which deals with repeated testing, one may find that extreme scores are regressing away from the mean, contrary to what one would expect based on the regression effect. This paper discusses the regression effect and presents the argument that when these contrary results occur, they are indicative of the violation of the underlying assumption of rectilinearity. One should be required to look for non-linear relationships when interpreting such data. In addition, three methods for determining whether or not non-linear relationships exist in data are suggested and briefly discussed.

OHIO J. SCI. 78(2): 96, 1978

Introductory texts in measurement and statistics typically present the topic of regression effect by examples of bivariate distributions in which extreme scorers on the independent variable (X) tend to score closer to the mean on the dependent variable (Y). In the limiting case, where there is no systematic relationship between X and Y, the predicted Z_y ($\hat{Z}_y$) values for all values of Z_y equal the mean of Z_y. On the other hand, when the correlation between X and Y is perfect, the regression effect is, by definition, absent. This implies that knowledge of X allows perfect prediction of Y. When both X and Y have been transformed to Z scores, $\hat{Z}_y = Z_x$ for all values of Z_x. Predicted Z scores, then, are assumed to take on absolute values that are either equal to or less than the X value from which they were derived (fig. 1). In this discussion, raw scores have been transformed to standard scores because differences in means and/or variances between X and Y can camouflage the regression phenomenon.

This conception of the regression phenomenon has implications in many areas where correlation is employed in analyzing and inferring from data. In covariance analyses, control is exerted on a covariate while comparing the performance of treatment groups against a criterion. The scores and groups means are adjusted in accordance with regression notions. When extreme groups in pre-test/post-test designs are studied, the use of a control group is recommended so that change scores are not falsely attributed to intervening treatments, when the regression phenomenon is the real source of change. In repeated testing paradigms where change scores are the object of interest, expectations about changes by extreme scorers are often

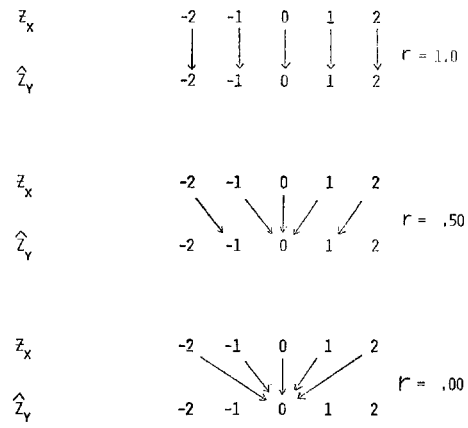

FIGURE 1. The projection of Z_x to Z_y under various degrees of association. Relationship between Z and predicted Z ($\hat{Z}$) as a function of R values.

[1]Manuscript received February 17, 1976 and in revised form July 6, 1977 (#76-18).

based on the regression phenomenon. A study reported by Vernon (1954) provides an example of expectations based on the regression phenomenon failing to hold in a repeated testing paradigm. In that study, changes in IQ were being studied as a function of repeated testing, and the expectations were that high scorers would score lower and low scorers would improve. Similar expectations have been suggested by Campbell and Stanley (1969), Kerlinger (1965), Borg and Gall (1971) and others. Vernon found, however, that high scorers scored even higher upon re-testing and gained more than low scorers.

When r_{xy} does not equal one, there is error about the least squares regression line relating X to Y. This error has 2 possible sources: pure error and lack of fit error. Pure error is error that has a mean of zero and is normally distributed in the population. Lack of fit error is characterized by having mean values that are significantly different from zero over certain ranges of X. That is, within certain ranges of the independent variable X, $\hat{Z}_y$ will be below the actual Z_y values and for other ranges of X, $\hat{Z}_y$ will be above the corresponding Z_y values. It should be stressed here that these departures must be significant at some specified alpha level before the presence of lack of fit error can be entertained.

The fitted regression line has been calculated using a model that assumes a linear relationship between X and Y. If the assumption of linearity is tenable, the error about the regression line is pure error. If, on the other hand, the as-

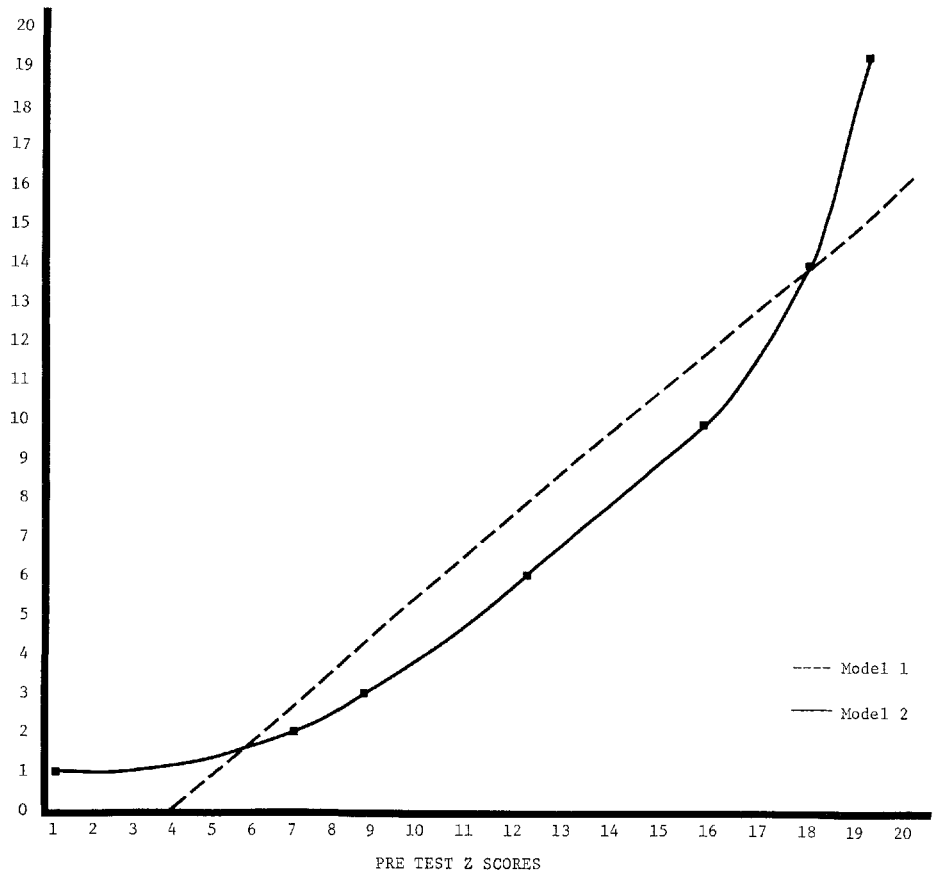

FIGURE 2. Comparison of linear (model 1) and quadratic (model 2) regression lines fitted to data reflecting a curved relationship between X and Y.

sumption of linearity is not correct, error about the regression line is composed of both pure error and lack of fit error.

If a straight line is the best fitting regression line, and if r_{xy} is not unity, then regression to the mean of Y must occur. On the other hand, if the relationship between X and Y can best be represented by a curved line of any sort, the standard phenomenon of regression effects toward the mean will not result. For example, in figure 2, data points have been drawn such that the relationship between Z_x and Z_y is clearly non-linear. Regression line 1 has been drawn through the scatter plot using the least squares criterion, and a certain amount of error has been realized. The model that reflects line 1 in figure 2 is as follows:

Model 1: $Y = a_0U + a_1X_1 + E_1$
where:
Y is the dependent variable
U is a unit vector which when multiplied by a_0 yields the regression constant
a_1 is the regression weight
E_1 is the error vector
X_1 is the independent variable

Admittedly, the magnitude of error in this model is small, but note what has occurred in our data. The subjects who had a Z score of 8 on X have an average Z score of 2.5 on Y. The Z_y value of 2.5 is greater than the Z_x value of 8. In other words, these subjects have "regressed" away from the mean of Y instead of regressing toward the mean of Y, as would be expected according to the regression phenomenon. The reason for this, of course, is that the relationship between X and Y is best represented by a regression model which allows for curvature. Model 2 is just such a model.

Model 2: $Y = a_0U + a_1X_1 + a_2X_1^2 + E_2$
where:
Y is the criterion score
U is the unit vector
X_1 is the X value
X_1^2 is the X values squared
a_1, a_2, and a_3 are the regression weights assigned to U, X, and X^2, respectively
E_2 is the vector of residuals

Since the relationship between X and Y that is depicted in figure 2 is curved, the assumption of rectilinearity is not correct. Consequently, expectations based upon the regression phenomenon would be in error and might mislead the researcher in interpreting his results. This example is particularly noteworthy because the linear model reflected in Model 1 did an excellent job of accounting for variance in Y, but Model 2 did a much better job.

The fact that the scores did not distribute themselves in the bivariate distribution as the regression phenomenon predicted should have served as a cue for the researcher to attempt a non-linear fit to the data. By determining the kind of curvilinear relationship that existed, inferences might be made as to the underlying causes for that relationship.

METHODS FOR DETECTING CURVATURE

Various procedures exist for screening paired observations for possible non-linear trends. Oftentimes, inspection of the scatter plot is sufficient to note the presence of curvature in the relationship. Inspection methods lack certainty when there are few data points or if the correlation is rather small. The following methods are offered as suggestions to assist in determining if a non-linear trend is present in a set of data.

Method I. If there are only a few observations (fewer than 25) the X variable can be divided into quintiles and the mean Y values can be calculated for each quintile group on X. The plot of these means could then be inspected for curvature. The researcher should recognize that for a set of data this small, only distinctly non-linear trends will approach significance. If the investigator suspects, as a result of this inspection, that a non-linear trend is present he may want to test the significance of that trend (Edwards 1960).

Method II. Calculation of the correlation ratio (Eta^2) for a set of data. The Eta^2 could then be compared to r^2 to test, if a significant non-linear trend exists in the data. An F-ratio could be generated as follows:

$$F = \frac{(Eta^2_{xy} - r^2_{xy})/df_1}{1.0 - Eta_2)/df_2}$$

where: Eta^2_{xy} is the Eta^2 for our data

r^2_{xy} is the square of the Pearson r for the same data

df_1 is the number of linearly independent variables or scores on X used to calculate Eta^2, minus 2 and,

df_2 is the number of paired observations minus the number of linearly independent variables or scores used to calculate Eta^2.

If this F-ratio is found to be significant, one may entertain the assumption that a non-linear trend is present in the data due to the global nature of Eta^2, one cannot, from these analyses, describe the type of relationship that exists. Eta^2 could be significantly greater than r^2 because of any departure from linearity.

Method III. The most direct method of ascertaining if the relationship between X and Y is significantly nonlinear, and describing the nature of the relationship is a curve fitting approach employing multiple regression methods. Various functional relationships can be tested and the partial regression weights may be used to describe the nature of the relationship. The use of a multiple regression approach which allows the investigator to try various data transformations and possible linear combinations is particularly useful (McNeil *et al* 1975).

CONCLUSION

One of the underlying assumptions necessary for the regression phenomenon to be accurate is that there is a linear relationship between X and Y. If the relationship is linear, all error will result in extreme scores on X, scoring on the average closer to the mean of Y. If the relationship between X and Y is nonlinear, the regression phenomenon will not represent the true state of affairs in the data. Indeed, if researchers note that expectations based on the regression phenomenon are not supported by their data, it would be advantageous for them to seek out the nature of the curved relationship that exists. In fact, the nature of the curvature may serve as a clue in determining why the regression phenomenon was insufficient for explaining the data.

LITERATURE CITED

Borg, W. R. and M. D. Gall 1971 Educational Research: An Introduction. David McKay Co., New York. 533 p.

Campbell, D. T. and J. C. Stanley 1969 Experimental and Quasi-Experimental Design for Research. Rand McNally, Chicago. 84 p.

Edwards, A. L. 1960 Statistical Methods for Behavioral Sciences. Rinehart, New York. 542 p.

Kerlinger, F. N. 1965 Foundations of Behavioral Research. Holt, Rinehart & Winston, New York. 739 p.

McNeil, K., F. J. Kelly and J. McNeil 1975 Testing Research Hypotheses Using Multiple Linear Regression. Southern Illinois Univ. Press, Carbondale, IL. 587 p.

Vernon, P. E. 1954 Practice and coaching effects in intelligence testing. *In:* The Educational Forum 18: 269–280.