The Negative Aspects of the Eta Coefficient as an Index of Curvilinearity

Keith A. McNeil
Southern Illinois University

The objective of this paper is to discuss the eta coefficient and to point out some limitations and misconceptions about the coefficient. Specifically, we will discuss the fact that; 1) the eta coefficient is a global measure of curvilinearity; 2) the eta coefficient has limited interpretability; 3) there are a number of other curvilinear relationships that might be of more significance and of more interpretability; 4) these other curvilinear relationships do not suggest nor encourage grouping of data as does the eta coefficient; and 5) these other curvilinear relationships may tend to be more amenable to replication than is the eta coefficient.

The eta coefficient

The eta coefficient indicates the general or global relationship between two variables. Essentially, the "line" of the best fit is drawn through the means of each separate predictor (X) score. The mean Y value of all of the Y values for those subjects who received a value of, say, 3 on the X variable is calculated and used as the predicted Y score for those subjects. The squares of the deviations of the actual Y values from the predicted Y value gives the error sum of squares remaining in the data. This error sum of squares can be compared to the error sum of squares due to the linear prediction to indicate the extent of global curvilinearity existing in the data, over and above the linear relationship in the data. The implication of significant curvilinearity is that the Pearson Product Moment Correlation (the measure of linear relationship) is underestimating the degree of relationship between the two variables. Note that this conclusion does not say that the Pearson Product Moment is inappropriate, as many users and even some statisticians would have us believe, but that we are underestimating the degree of relationship in our data when we report the Pearson Product Moment coefficient when, in fact, curvilinearity exists.

The eta coefficient as a global measure of curvilinearity

We are saying that the eta coefficient is a general measure of the degree of curvilinearity because it includes a number of different kinds of curves. The eta coefficient includes the linear fit of the data, the second degree fit, the third degree fit, the fourth degree fit, and so on, up to the $(K-1)^{th}$ degree of fit, where K is the number of different values along the X axis. Once we have a significant eta coefficient, we are not aware of what kind of a curve we really have which explains that data, just that some kinds of curvilinearity exist in the data.

The interpretation of an eta coefficient

A statistical index has little value unless one can interpret it and we are proposing that the eta coefficient has extremely limited interpretability. Because the single coefficient includes all of the various degrees of curvilinearity, one is not told the exact nature of the curve. All that one is told is that a non-linear model fits the data better than a linear model. But the specific non-linear model is not divulged.

The eta coefficient has often been used in a negative sense, in that it is used to disqualify the application of the Pearson Product Moment Correlation. Very seldom does a researcher get excited when he finds a significant eta coefficient. Too many researchers have been brought up under the guise that all X's and Y's are inherently rectilinearly related! We are suggesting that the eta coefficient should be discarded in favor of other models wherein more specific curvilinear relationships are investigated -- models that will more likely yield interpretable results and have some empirical or theoretical import (McNeil & Spaner, 1970).

Other curvilinear models

The simplest curvilinear model is one that involves a second degree polynomial. That kind of model allows for either a continually accelerating curve, or a continually decelerating curve (Figure 1). Curves that have either maximums or minimums require a third degree component, as well as possibly a second degree component and a linear component (Figure 2). We would prefer to have our data fit by these models because they indicate a predictable pattern more clearly than does the eta coefficient. A more objective reason for prefering the above models rests upon the ability of these models to predict Y scores for X values which might not have been observed in the original sample. This kind of flexibility is not available with the eta coefficient.

Furthermore, the significance of the second degree model might be greater than the significance of the eta coefficient. The second degree model might even be significant when the eta coefficient is not significant. This could occur when the number of X values is relatively large and when the data tend to fit a second degree curve. The fictitious data in Figure 3 is an extreme case of the above two conditions.

The above notions can probably best be depicted in multiple linear regression terminology (See Kelly, Beggs, McNeil, Eichelberger, & Lyon, 1969 or Bottenberg & Ward, 1963). The model which simulates an eta coefficient by allowing each predictor variable to have its own mean value is:

(Model 1) $Y_1 = a_oU + b_oX_o + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8$

Where: Y_1 = the criterion to be predicted
U = a "1" for all subjects
X_o = 1 if the X value is 0; 0 otherwise
X_1 = 1 if the X value is 1; 0 otherwise
X_2 = 1 if the X value is 2; 0 otherwise
X_3 = 1 if the X value is 3; 0 otherwise
X_4 = 1 if the X value is 4; 0 otherwise
X_5 = 1 if the X value is 5; 0 otherwise
X_6 = 1 if the X value is 6; 0 otherwise
X_7 = 1 if the X value is 7; 0 otherwise
X_8 = 1 if the X value is 8; 0 otherwise

a_o, b_o,...b_8 are weighting coefficients selected so as to minimize the sum of the squared components in E_1.

E_1 is the difference between the predicted Y_1 value and the actual Y_1 value.

The linear model which allows only a linear fit to the data is:

(Model 2) $Y_1 = a_oU + a_1X_9 + E_2$

Where: Y_1 = the criterion to be predicted
U = 1 for all subjects
X_9 = the X score for all subjects

a_o and a_1, are weighting coefficients selected so as to minimize the sum of the squared components in E_2.

E_2 is the difference between the predicted Y_1 value and the actual Y_1 value.

Model 1 can be statistically compared to Model 2 via the general F test for regression models: $F = \frac{(R^2_F - R^2_R)/(m_1 - m_2)}{(1 - R^2_F)/(N - m_1)}$,

Where R^2_F and R^2_R indicates the proportion of criterion variance accounted for by the full and restricted models, respectively, and m_1 and m_2 indicate the number of linearly independently vectors in the full and restricted models, respectively. The full model in this case is Model 1 and the restricted model is Model 2.

The degrees of freedom for the numerator of the F ratio is $(m_1 - m_2)$ and the degrees of freedom for the denominator is $(N-m_1)$. In comparing Model 1 against Model 2, we determine that the degrees of freedom are (9-2) or 7 and (18-9) or 9.

The reader should note that an equally good full regression model would be of the following form:

(Model 3) $Y_1 = a_oU + a_1X_9 + a_2X_9{}^2 + a_3X_9{}^3 + a_4X_9{}^4 + a_5X_9{}^5 + a_6X_9{}^6 + a_7X_9{}^7 + a_8X_9{}^8 + E_3$

Where: Y_1, U, and X_9 are defined as in Model 2.

$X_9{}^2$ is the squared value of the corresponding value in X_9

$X_9{}^3$ is the cubed value of the corresponding value in X_9

etc.

a_o, a_1,... a_8 are weighting coefficients selected so as to minimize the sum of the squared components in E_3.

E_3 is the difference between the predicted Y_1 value and the actual Y_1 value.

The restriction in going from Model 3 to Model 2 may be more obvious than in going from model 1 to model 2. If the higher order polynomials are not needed to fit the data, then $a_2 = 0$; $a_3 = 0$; $a_4 = 0$; $a_5 = 0$; $a_6 = 0$; $a_7 = 0$; and $a_8 = 0$. These 7 restrictions are reflected by the 7 degrees of freedom in the numerator of the F ratio. The F ratio resulting from testing either Model 3 against Model 2 or Model 1 against Model 2 is 1.76. With 7 and 9 degree of freedom, this is not significant at the .05 level of significance. We would thus conclude that there is no significant eta coefficient in this data. That we cannot conclude that there is no significant curvilinear relationship should be obvious from the data and from the test of significance of the following model:

(Model 4) $Y_1 = a_o U + a_1 X_9 + a_2 X_9^2 + E_4$

Where: All symbols are defined as in Model 3, except E_4, which will be the error in prediction using the particular set of predictor vectors and their associated weights.

Model 4 has 3 linearly independent vectors and allows for a second degree curve. Another look at the data in Figure 3 should satisfy the reader that the line of best fit produced by model 4 is a relatively good fit. As can be seen in Table 1, the R^2 produced by Model 4 is as high as the R^2 produced by the eta coefficient model (Model 1 or Model 3). Comparison of Model 4 to Model 2 will in this case result in a lower probability value than will the comparison of either Model 1 or Model 3 to Model 2. The comparison of Model 4 to Model 2 is significant at the .05 level of significance ($F = 20.53$, probability $< .001$).

The grouping of data in the eta coefficient model

The eta coefficient does not demand the scores along the predictor axis be grouped, but most examples in statistical books and most applications do involve grouping. Whenever data are grouped, some error is probably going to be introduced, particularly if there is some kind of continuous function in the data. Grouping of data is introduced in order to minimize the computational problems, but with the computer available, computation no longer is a problem.

There is one advantage to grouping of data that needs to be discussed. The advantage lies in the fewer number of predictor values, and consequently fewer degrees of freedom in the numerator of F and more in the denominator of F. Thus, with a constant R^2_F and R^2_R, we will more likely find a significant eta coefficient. But the grouping along the predictor axis will most likely (not necessarily) reduce the R^2_F and R^2_R. The R^2_F will be reduced drastically by grouping if there is a systematic function, and thus the apparent advantage turns out to be a distinct disadvantage.

Replicability of curvilinear models

We have already indicated that the eta coefficient cannot predict Y values for X scores that are not observed in the original sample. Indeed, the eta coefficient produces the maximum overfit to the data. Because of this, the eta coefficient is less likely to yield a high coefficient of replicability, unless of course, the phenomenon under investigation is relatively stable. If the phenomenon is stable, the curvilinear models which fit the data as well as does the eta coefficient model, but yet require fewer predictor variables, are preferable because they are more parsimonious. That is, these models require fewer pieces of information about the data. As scientists, we expect some orderly fashion in our data, but the eta coefficient does not encourage an orderly investigation of the data.

Discussion and implications

We have tried to demonstrate the inapplicability of the eta coefficient for most behavioral science research. We have tried to demonstrate that the eta coefficient does not help the researcher, but actually hinders him by reducing the liklihood of finding significance, by discouraging the orderly explanation of the data, by encouraging the researcher to arbitrarily group his data, and by producing a model which is extremely difficult to interpret and extremely difficult to replicate.

The original impetus for this paper was an article by Hawk (1970). Hawk systematically investigated a large number of General Aptitude Test Battery (GATB) validation studies to determine the frequency of non-linear relationships. He computed the eta coefficient and tested the eta against the linear relationship. He found that the number of significant non-linear relationships fell very close to the chance level. Hawk coarsely grouped the GATB scores into five intervals and excluded scores which deviated more than 2 1/2 standard deviations from the mean. We don't approve of grouping data nor of eliminating Ss from the face of the Earth, but that is his preference.

What we would like to criticize is the conclusion that Hawk arrives at, mainly: "The author's inclination is to assume that, especially in GATB validation studies, the relationships are linear unless there is some theoretical or empirical reason to believe otherwise" Hawk (1970, p. 251).

At best, the author demonstrated that the eta coefficient model was not applicable to GATB. But as the data in Figure 3 indicates, there might well be some extremely significant, specific, non-linear relationship between GATB and the particular validation criterion. As McNeil and Kelly (1970) have pointed out, we can never see most of the variables in behavioral science research, and as a consequence, it is quite inappropriate to think that all the variables are rectilinearly related. As a consequence, researchers should investigate specific non-linear relationships, while realizing the limitations of the eta coefficient.

References

Bottenberg, R. A. & Ward, J. H., Jr. Applied Multiple Linear Regression PRL-TDR - 63-6, AD-413 128. Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division, March 1963.

Hawk, J. A. "Linearity of Criterion -- GATB Aptitude Relationships" Measurement and Evaluation in Guidance 1970, 2, 249-251.

Kelly, F. J., Beggs, D. L., McNeil, K. A., Eichelberger, T., & Lyon, J. Research Design in the Behavioral Sciences: Multiple Regression Approach. Carbondale, Illinois: Southern Illinois University Press, 1969.

McNeil, K. A., & Kelly, F. J. "Express functional relationships among data rather than assume intervalness." Journal of Experimental Education, In Press.

McNeil, K. A., & Spaner, S. D. "A Defense for Including Highly Correlated Predictor Variables in Multiple Regression Analysis." Paper presented at American Educational Research Association, 1970.

Table 1

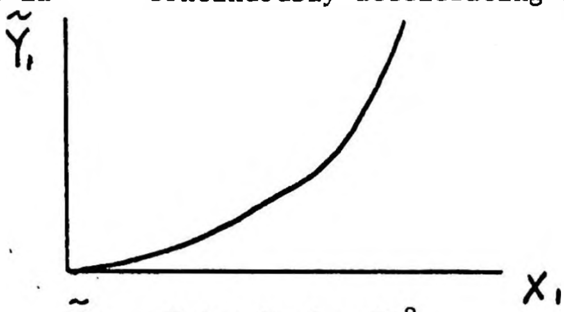
Information Relevant to the Various Models

Model #	Model Name	R^2	linearly independent vector
Model 1	"Grouped X values"	.9485	9
Model 2	"Linear Model"	.8781	2
Model 3	"Polynomial Model"	.9485	9
Model 4	"Second degree Model"	.9485	3
Model 5	"Unit vector model"	.0000	1

F Ratios

Full	Restricted	F	df_1	df_2	P
Model 1	Model 2	1.7596	7	9	$>$.2111
Model 3	Model 2	1.7596	7	9	$>$.2111
Model 4	Model 2	20.5333	1	15	$<$.0004
Model 1	Model 5	20.7400	8	9	$<$.0001
Model 2	Model 5	115.2720	1	16	$<$.0001

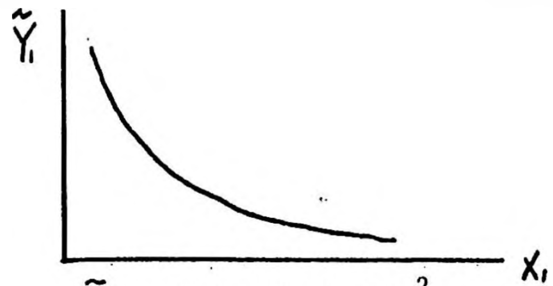
Figure 1a Continuously accelerating curve



$\tilde{Y}_1 = a_oU + a_1X_1 + a_2X_1^2$

a_2 will be a positive value

Figure 1b Continuously decellerating curve



$\tilde{Y}_1 = a_oU + a_1X_1 + a_2X_1^2$

a_2 will be a negative value

Figure 2 One possible situation depicted by a third degree curve

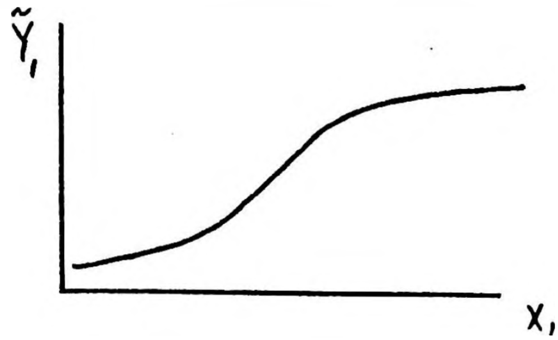


Figure 3 Ficticious data depicting the situation wherein the eta coefficient is not significant and the second degree coefficient is significant.