Psychological Bulletin
1973, Vol. 79, No. 4, 239-242

THE DEVIL RIDES AGAIN:

CORRELATION AS AN INDEX OF FIT[1]

MICHAEL H. BIRNBAUM[2]

University of California, Los Angeles

Correlations between theoretical predictions and data can be higher for incorrect than for correct models, as illustrated by analyses of two sets of hypothetical data. This fact raises questions about the conclusions of recent studies that use correlation as an index of fit. Functional measurement provides a sounder basis for model evaluation by placing scaling in the context of model fitting and by testing deviations from prediction rather than concentrating an overall goodness of fit.

Although widely recognized as an "instrument of the devil" when used to infer causation from confounded data, the correlation coefficient is still employed as an index of the fit of theoretical models. There are two serious criticisms of this usage. First, incorrect models can correlate highly with data (Anderson, 1971; Yntema & Torgerson, 1961).

Second, based upon the assumption that the better model will correlate higher, investigators have recently used the correlation coefficient to compare the fit of rival models.[3] The point of the present paper is to warn that the devil may also be at work when correlation is used in this way: Predictions from an incorrect model sometimes correlate better with the data than predictions from the correct model.

Linear Model Correlates with Multiplicative Data

Table 1 illustrates one way this can happen. The matrix of hypothetical, errorless data might represent clinical ratings of the degree of neuroticism, Y, based upon two test scores, X_1 and X_2, with eight and three levels, respectively.

The usual *linear* model can be written as follows:

$$Y = a + bX_1 + cX_2, \qquad [1]$$

where Y is the predicted value for the data in the table, X_1 and X_2 are the a priori values of the two predictor variables (in this case, test scores), and a, b, and c are the linear coefficients.

The *multiplicative* model (Einhorn, 1970) can be written[4]

$$Y = aX^b{}_1X^c{}_2, \qquad [2]$$

where Y, X_1, and X_2 are defined as above, and a, b, and c are constants. Taking loga-

[1] Computing assistance was received from Campus Computing Network, University of California, Los Angeles. The author is grateful to Allen Parducci, Norman H. Anderson, Amos Tversky, Clairice T. Veit, and Andrew L. Comrey for their helpful comments on earlier versions of this paper.

[2] Requests for reprints should be sent to Michael H. Birnbaum, Department of Psychology, University of California, San Diego, P.O. Box 109, La Jolla, California 92037.

[3] For example, Einhorn (1970, 1971) and Goldberg (1971) have compared linear with "conjunctive" (multiplicative) models of judgment by correlating the data for each subject with the predictions from each of the rival models; when one model correlated higher for the great majority of the subjects, it was assumed to be a better representation of the process of human judgment. Einhorn (1971) obtained rankings of the attractiveness of hypothetical jobs, based on such cues as income and the opportunity to use special interests. The multiplicative model yielded higher correlations than the linear model for 32 of the 37 subjects. Goldberg (1971) found that the linear model yielded higher correlations than the multiplicative for each of 29 clinicians, who attempted to differentiate neurosis from psychosis on the basis of the Minnesota Multiphasic Personality Inventory profiles.

[4] Einhorn (1970) has defined Equation 2 as the "conjunctive model." Equation 2 captures some of the intuition behind the notion of a subjective "conjunctive" strategy; however, the multiplicative model (Equation 2) is compensatory and therefore should not be confused with the traditional definition of the conjunctive model (see Coombs, 1964).

TABLE 1

HYPOTHETICAL MULTIPLICATIVE DATA

A priori value of second cue	A priori value of first cue							
	1	2	3	4	5	6	7	8
1	15	18	21	24	27	30	33	36
2	20	22	24	26	28	30	32	34
3	25	26	27	28	29	30	31	32

TABLE 2

HYPOTHETICAL ADDITIVE DATA

A priori value of second cue	A priori value of first cue				
	1	2	4	8	16
1	14	15	16	17	18
2	15	16	17	18	19
4	16	17	18	19	20
8	17	18	19	20	21
16	18	19	20	21	22

rithms of both sides of Equation 2 gives:

$$\log Y = \log a + b \log X_1 + c \log X_2, \quad [3]$$

which, like Equation 1, can be fitted by multiple linear regression with three constants.

The fit of Equation 1 to the data in Table 1 yielded a correlation of .930. Figure 1A plots these data against the best-fit predicted values in the usual way; the fit looks reasonably good. However, Figure 1B shows the same data plotted as a function of X_1, with a separate curve for each level of X_2. The nonparallelism of these curves indicates serious violation of the additive model of Equation 1. It should be clear that the data conform instead to a multiplicative model.

Einhorn's technique would fit his multiplicative model to the data by means of Equation 3, which yields a coefficient of only .899. This example shows that even when the data are perfectly multiplicative, comparison of correlation coefficients can lead to the erroneus conclusion that the linear model provides the better representation. The source of this difficulty lies in the assumption, implicit in the use of Equation 3, that the a priori values of X and Y are positive and known to a ratio scale. As shown below, the data of Table 1 fit Equation 2 perfectly when linear transformations are applied to the a priori scales.

MULTIPLICATIVE MODEL CORRELATES WITH ADDITIVE DATA

A second example shows that predictions from the multiplicative model can correlate better than those from the linear model, even when the linear model is more appropriate. The hypothetical data in Table 2 might be

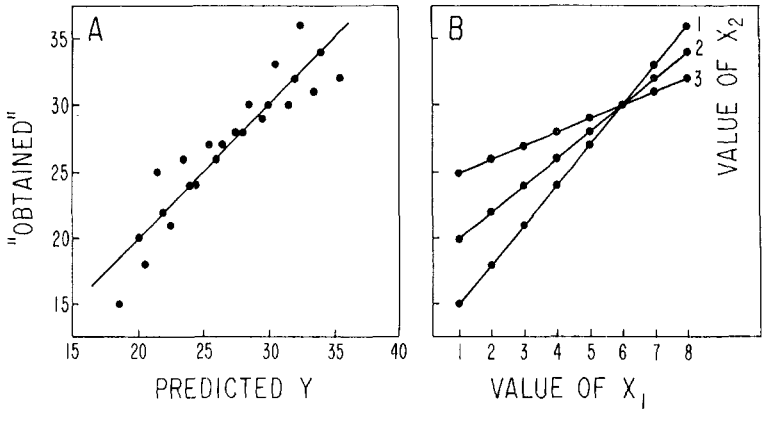

FIG. 1. A—Correlation of linear model with perfectly multiplicative data of Table 1; B—Same data plotted as a function of X_1, with a separate curve for each level of X_2. (Although the intersecting linear functions in B indicate that the data are perfectly multiplicative [with a functional zero point where the curves cross], Equation 3 provides an inferior fit because it requires the curves to intersect at the zero point of the a priori X and Y scales.)

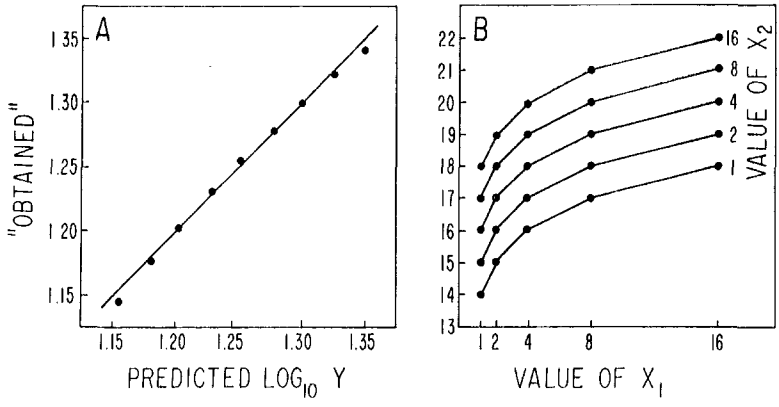

FIG. 2. A—Correlation of multiplicative model with perfectly additive data of Table 2. (Graph points represent varying numbers of data points.) B—Same data plotted as a function of X_1, with a separate curve for each level of X_2. (Although the parallelism of the curves in B indicates perfect additivity, the linear model provides an inferior fit because it requires the curves to be straight-line functions of the a priori values of X_1 and X_2.)

ratings, Y, of the attractiveness of various jobs based on two cues, X_1 and X_2.

Fitting the linear model to these data yields a correlation of only .933 compared with .998 for the multiplicative model. It can be inferred from Figure 2A, which plots the fit of Equation 3 to the data, that the use of rank-order correlation (Einhorn, 1970, 1971) would have "improved" on the already excellent fit of the wrong model. The multiplicative model is the wrong model because the data are perfectly additive, as shown by the parallelism of the curves of Figure 2B. The investigator using correlation as an index of fit might erroneously conclude that the data reveal a "conjunctive" judgmental strategy unless the data were plotted as in Figure 2B. It is only because the linear regression model requires the data to plot as linear functions of the a priori values of X_1 and X_2 that it yields a lower correlation.

Functional Measurement Provides Better Test

Functional measurement (Anderson, 1970, 1971) correctly diagnoses the same hypothetical data. The linear model is additive and therefore predicts a zero interaction between the cues. Equation 2 is multiplicative and therefore predicts an interaction which should be located entirely in the bilinear component. These interactions can be tested graphically, as in Figure 1B and 2B, and statistically, using the analysis of variance.

The intersecting linear curves of Figure 1B are indicative of a multiplicative model with a functional zero at the point where the curves intersect. When Y is replaced by $Y - 30$, X_1 is replaced by $X_1 - 6$, X_2 is replaced by $4 - X_2$, and $a = b = c = 1$, Equation 2 fits the data of Figure 1B perfectly.

The parallelism of the curves of Figure 2B indicates that there is no interaction between cues, in agreement with an additive model. In the case of additive models, the marginal means estimate the functional scale values for the cues. Equation 1 fits these data perfectly when the marginal means of Table 2 are substituted for the a priori values of X.

Comments and Conclusions

These two examples demonstrate that even with factorial designs and errorless, interval data, the correlation coefficient can be an inappropriate instrument for comparing the fit of models with equal numbers of parameters. In real-life applications, correlations may be even more misleading.

Correlations can be diabolical when factorial designs are not employed. The reader can easily select those cue combinations from

Figure 1B and 2B that would improve the fit of either model. Furthermore, with certain research designs, graphic tests like those shown in Figures 1B and 2B become impossible. Hence, "representative" or improper "contrived" designs can exaggerate the fit of a seriously defective model and also can make it impossible to evaluate.

The data need not be particularly unusual for the correlations to be so misleading. In Table 2, the functional scale values (marginal means) are linearly related to the logarithms of the a priori stimulus values. That is often the case for category ratings of psychophysical stimuli or stimuli presented in numerical form, such as incomes. Consequently, there is a very real danger that the multiplicative model would yield higher correlations than the linear in many applications, even though the data were additive.[5]

In contrast to correlational procedures, functional measurement (Anderson, 1970, 1971) requires neither a priori values for the stimuli, nor ratio scales for the responses.[6] These procedures estimate best-fit values for the stimuli directly from the data, using these estimates as the basis for statistical tests of the discrepancies. Functional measurement typically employs factorial designs, analysis of variance, and graphical inspection of the crucial features of the data. These advantages of functional measurement over the correlational approach are illustrated by the sample data in the present paper.

[5] Goldberg (1971) has proposed that some of these problems would be avoided if a greater number of models were compared in any correlational contest of fit. For example, Goldberg's logarithmic model, which replaces X with log X in Equation 1, would have correlated perfectly with the data of Table 2. However, none of his five models would have correctly diagnosed the data of Table 1. The number of possible correlation-regression models is impractically large, since each arbitrary rescaling of either the a priori stimulus values or the overt responses is treated as a different model. It should be emphasized that the usual applications of multiple regression (as in the present examples) are *not* equivalent to the analysis of variance. Equivalent techniques (e.g., Cohen, 1968) are seldom employed in this type of analysis because they would be cumbersome to apply.

[6] Functional measurement also provides for monotonic rescaling of the data under certain circumstances although no nonlinear transformation was called for in these examples. When the dependent variable is considered to be only an ordinal measure of the psychological attribute, models can be compared by examination of critical ordinal properties of the data (Krantz & Tversky, 1971). For example, the crossover interaction in Figure 1B is ordinally inconsistent with an additive representation.

REFERENCES

ANDERSON, N. H. Functional measurement and psychophysical judgment. *Psychological Review,* 1970, 77, 153–170.

ANDERSON, N. H. Integration theory and attitude change. *Psychological Review,* 1971, 78, 171–206.

COHEN, J. Multiple regression as a general data-analytic system. *Psychological Bulletin,* 1968, 70, 426–443.

COOMBS, C. H. *A theory of data.* New York: Wiley, 1964.

EINHORN, H. J. Use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin,* 1970, 73, 221–230.

EINHORN, H. J. Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance,* 1971, 6, 1–27.

GOLDBERG, L. R. Five models of clinical judgment: An empirical comparison between linear and nonlinear representations of the human inference process. *Organizational Behavior and Human Performance,* 1971, 6, 458–479.

KRANTZ, D. H., & TVERSKY, A. Conjoint measurement analysis of composition rules in psychology. *Psychological Review,* 1971, 78, 151–169.

YNTEMA, D. B., & TORGERSON, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions on Human Factors in Electronics,* 1961, HFE-2, 20–26.

(Received January 26, 1972)