Bulletin of the Psychonomic Society
1977, Vol. 10 (2), 134-136

Correlation as a deceiving measure of fit

JAMES SHANTEAU
Kansas State University, Manhattan, Kansas 66506

There has been a continuing controversy over the use of correlations for evaluating fit of a model. Lost in this controversy, however, has been any evidence that correlations have in fact proved to be deceptive. This note discusses three instances in the human judgment/decision making literature where such deception has actually taken place. In each instance, high correlations provided initial support for the models involved. Reanalyses and later evidence, however, have shown that these correlations were deceptive. Thus, correlations not only *can* be deceptive, as has been argued earlier; these instances illustrate that correlations *have* been deceptive.

There have been a number of recent articles (Anderson, 1969; Birnbaum, 1973; Zeleny, 1976) that have warned about the pitfalls of using correlation coefficients for evaluating the fit of mathematical models. Such articles have produced a considerable controversy (Alf & Abrahams, 1974; Rorer, 1974) in which a key point has been neglected: Have these pitfalls actually occurred in the published literature?

Neither Anderson, Birnbaum, nor Zeleny adequately address this question. Anderson cited instances in which incorrect models provided extremely high correlations. However, these correlations were computed after the fact and played no role in the model evaluation. Birnbaum showed that an incorrect model can lead to higher correlations than a correct model. However, he used artificial data that are unlikely to occur in practice. Zeleny demonstrated that the generalizability of models derived using correlation/regression techniques is questionable. However, there were no instances cited to show that this is a real problem. It is an open question, therefore, whether the alleged pitfalls of correlation should be a matter of practical concern. This note examines three instances from the published literature on human judgment and decision making where correlations were employed.[1]

AN ADDITIVE MODEL FOR DUPLEX BETS

In an often-cited study, Slovic and Lichtenstein (1968) used an additive regression model for two-part bets. The model produced average correlations of .86 with both bidding and rating responses (p. 8). This high correlation value has been widely interpreted to mean that the parameters of a bet combine by adding (Payne, 1973).

The additive model contrasts sharply with the multiplicative model assumed by most theories of risky decision making, such as utility theory. In a direct comparison, Tversky (1967) found clear statistical support for a multiplicative model over an additive model. Anderson and Shanteau (1970) reported even more striking graphical results which visually demonstrate the superiority of the multiplicative model.

While these results raise questions about the additive model, they do not bear directly on Slovic and Lichtenstein's data. Accordingly, the data for their preliminary subjects were reanalyzed using functional measurement techniques introduced by Anderson and Shanteau (1970).[2] The results for three sample subjects in Figure 1 show the diverging fan of straight lines

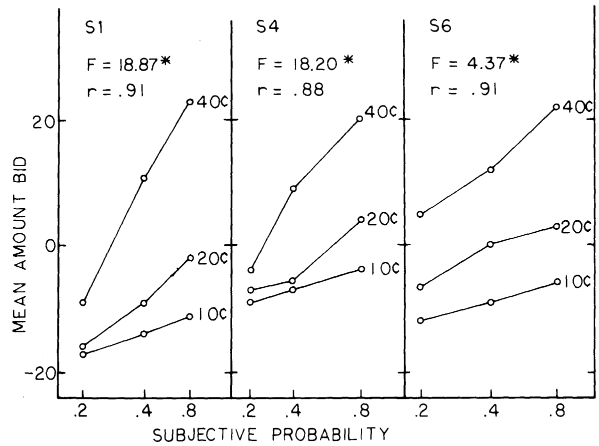

Figure 1. Mean amount bid for three preliminary subjects in Slovic and Lichtenstein (1968). Left panel shows a highly multiplicative subject, middle panel shows a median subject, and right panel shows one of the least multiplicative subjects. All subjects have high correlations (r) with an additive model, but all subjects also reveal significant Probability by Payoff interactions (F). These interactions are consistent with a multiplicative model, but represent deviations from an additive model (from Shanteau, 1975).

Preparation of this paper was supported in part by Public Health Service Grant MH 22610 from the National Institute of Mental Health. I wish to express my gratitude to Paul Slovic for kindly providing the preliminary data from Slovic and Lichtenstein (1968), and to Donald Campbell for kindly providing the data from Hicks and Campbell (1965). I also want to thank Norman Anderson and Michael Birnbaum for their helpful comments on the manuscript. Requests for reprints should be sent to James Shanteau, Department of Psychology, Kansas State University, Manhattan, Kansas 66506.

predicted by multiplying. There is little evidence of the parallelism predicted by adding.

Statistically, all three subjects have high correlations with an additive model, but all subjects also reveal highly significant Probability by Payoff interactions. This interaction violates the additive model, but is predicted by a multiplicative model. Therefore, despite the high correlations of the additive model, the multiplicative model provides a better account of these data.

A BAYESIAN MODEL OF INFERENCE

In a typical Bayesian inference study, subjects' judgments are compared with optimal values derived from Bayes' theorem. While the judgments are almost always conservative or less extreme than optimal, they do correlate highly with Bayesian values. For example, Peterson, Schneider, and Miller (1965) report a correlation of .93 (p. 527).

Such results led many investigators to conclude that inference judgments are basically Bayesian except for the conservatism effect. For instance, Beach and Wise (1969) state that "Bayes' theorem is a generally appropriate model" (p. 564). This conclusion produced a large body of research aimed at finding the psychological locus of conservatism.

It is now known, however, that inference judgments are nonoptimal in more fundamental ways than conservatism. For example, order effects are nonoptimal but have been found repeatedly (e.g., Shanteau, 1970). Also, Slovic and Lichtenstein (1971) have catalogued a whole set of nonoptimal behaviors ranging from misuse of prior information to insensitivity to the informativeness (diagnosticity) of information. Thus, it is clear that the Bayesian model provides a far from adequate description of human inference behavior. Any early successes of the Bayesian model are now viewed as "purely a matter of coincidence" (Slovic & Lichtenstein, 1971).

AN ADDITIVE UTILITY MODEL FOR COMMODITY BUNDLES

The value or utility of a commodity bundle of goods has been frequently claimed to equal the sum of the values of the individual goods. For example, Hicks and Campbell (1965) reported an average correlation of .95 between predicted and observed values (p. 802). Results like this provided such seemingly strong support that Edwards and Tversky (1967) observed that the additive utility model "so completely dominates the literature ... that it has no competitors" (p. 255).

It was with some surprise, therefore, that Anderson and Shanteau (1970) found marked deviations from additivity. This was substantiated by Shanteau (1974), who observed a subadditivity effect: The value of the combinations was consistently less than the sum of the parts. Similar results have since been found with a broad variety of stimuli, procedures, and analyses (Shanteau, 1975).

This subadditivity finding prompted reanalyses of previous studies reporting additivity. As an illustration, the data of Hicks and Campbell (1965, p. 806) have been plotted in Figure 2 using techniques described in Shanteau (1974). As can be seen, the additive utility model, which corresponds to the 45-deg diagonal, is called into serious question. Instead of additivity, the points fall well below the diagonal; the observed (combination) responses are consistently less than the prediction (sum of the parts) responses. This subadditivity effect apparently reflects a more general law of diminishing returns (Shanteau, 1975).[3]

WHY HAVE CORRELATIONS DECEIVED?

There are several reasons why correlations have proved so deceptive. These reasons have been discussed elsewhere (e.g., Anderson & Shanteau, in press) and will only be summarized here. In the case of duplex bets, an additive regression model was used. Such models will almost always produce high correlations (Dawes & Corrigan, 1974; Wainer, 1976), even when the underlying process is nonadditive. The same point was made by Lichtenstein, Earle, and Slovic (1975), who noted that correlations derived for regression models "were not useful in uncovering serious discrepancies from the model" (p. 85). One of the causes of this is that correlations ignore the number of parameters estimated in the model; given enough parameters, any model can provide a perfect correlation (Nimh, 1976).

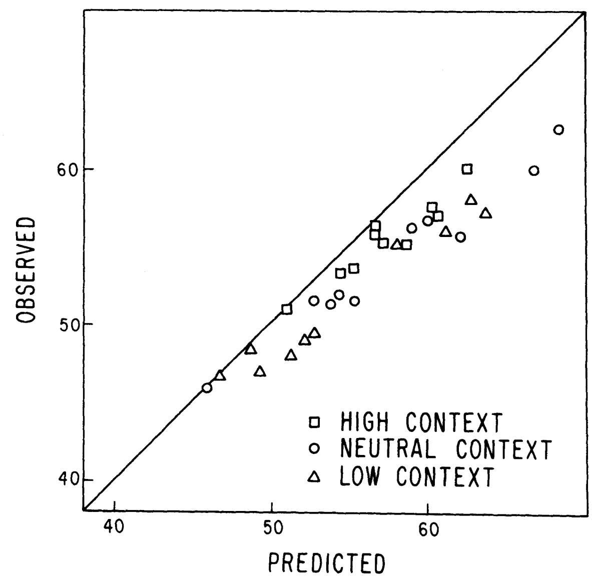

Figure 2. Predicted vs observed ratings of commodity bundles of gifts in reanalysis of Hicks and Campbell (1965, p. 806). Data plotted from high-, neutral-, and low-context conditions for (observed) combination of gifts as a function of (predicted) sum of single gifts. Perfect additivity corresponds to the diagonal line; subadditivity appears as a downward divergence from the diagonal (from Shanteau, 1974).

For the Bayesian model, high correlations were almost guaranteed by the task. That is, the experimental procedure itself defines for subjects the appropriate response direction, and this alone is enough to produce high correlations. Indeed, Slovic and Lichtenstein (1971) concluded that a subject "seems capable of little more than revising his response in the right (Bayesian) direction" (p. 714). Since it is not difficult to design a study to produce high correlations, such values can hardly be taken as evidence for a model (Anderson, 1969).

The commodity bundles case illustrates that, while a correlation can specify the degree to which two variables are related, it is incapable of saying why or in what way they are related. In Figure 2 the points fall roughly around a straight line, which leads, of course, to a high correlation. However, the additive utility model requires in addition that the points fall along the diagonal, and this is not evaluated by a correlation.

Another common problem is that correlation measures often seem to discourage plots of the data (other than scatter plots). As revealed in both Figures 1 and 2, rather obvious deviations can be seen by simply looking at the data. Indeed, Birnbaum (1973) noted that many of the pitfalls of correlations can be avoided by use of appropriate plots along with adequate goodness-of-fit tests.

CONCLUSIONS

In each of the examples, high correlations provided initially compelling support for the models involved. This led to further, often extensive, research on the presumed psychological processes behind these models. Reanalyses and later evidence, however, have now shown that these initial correlations were deceptive. Indeed, correlations often seem to have done more to cover up than to reveal the underlying processes.

The statistician John Tukey (1969) offered the following advice to psychologists: "Sweeping things under the rug is the enemy of good data analysis [and] using the correlation coefficient is 'sweeping under the rug' with a vengeance" (p. 89). Tukey's negative views on correlations are reinforced by the present examples in which systematic model deviations were swept under a rug of large correlation coefficients. Thus, correlations not only *can* be deceptive, as has been argued earlier; these instances illustrate that correlations actually *have* been deceptive.

REFERENCES

ALF, E. F., & ABRAHAMS, N. M. Let's give the devil his due: A response to Birnbaum. *Psychological Bulletin*, 1974, **81**, 72-73.

ANDERSON, N. H. Comment on "An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment." *Psychological Bulletin*, 1969, **72**, 63-65.

ANDERSON, N. H., & SHANTEAU, J. C. Information integration in risky decision making. *Journal of Experimental Psychology*, 1970, **84**, 441-451.

ANDERSON, N. H., & SHANTEAU, J. Weak inference with linear models. *Psychological Bulletin*, in press.

BEACH, L. R., & WISE, J. A. Subjective probability revision and subsequent decisions. *Journal of Experimental Psychology*, 1969, **81**, 561-565.

BIRNBAUM, M. H. The devil rides again: Correlation as an index of fit. *Psychological Bulletin*, 1973, **79**, 239-242.

DAWES, R. M., & CORRIGAN, B. Linear models in decision making. *Psychological Bulletin*, 1974, **81**, 95-106.

EDWARDS, W., & TVERSKY, A. (Eds.). *Decision making*. Baltimore: Penguin Books, 1967.

HICKS, J. M., & CAMPBELL, D. T. Zero-point scaling as affected by social object, scaling method, and context. *Journal of Personality and Social Psychology*, 1965, **2**, 793-808.

LICHTENSTEIN, S., EARLE, T. C., & SLOVIC, P. Cue utilization in a numerical prediction task. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, **1**, 77-85.

NIMH, S. D. Polynomial law of sensation. *American Psychologist*, 1976, **31**, 808-809.

PAYNE, J. W. Alternative approaches to decision making under risk: Moments versus risk dimensions. *Psychological Bulletin*, 1973, **80**, 439-453.

PETERSON, C. R., SCHNEIDER, R. J., & MILLER, A. J. Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, 1965, **69**, 522-527.

RORER, L. G. "What, can the devil speak true?" *Psychological Bulletin*, 1974, **81**, 355-357.

SHANTEAU, J. C. An additive model for sequential decision making. *Journal of Experimental Psychology*, 1970, **85**, 181-191.

SHANTEAU, J. Component processes in risky decision making. *Journal of Experimental Psychology*, 1974, **103**, 680-691.

SHANTEAU, J. Information integration analysis of risky decision making. In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes*. New York: Academic Press, 1975.

SLOVIC, P., & LICHTENSTEIN, S. Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology*, 1968, **78**(3, Part 2).

SLOVIC, P., & LICHTENSTEIN, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, **6**, 649-744.

TUKEY, J. W. Analyzing data: Sanctification or detective work? *American Psychologist*, 1969, **24**, 83-91.

TVERSKY, A. Utility theory and additivity analysis of risky choices. *Journal of Experimental Psychology*, 1967, **75**, 27-36.

WAINER, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 1976, **83**, 213-217.

ZELENY, M. On the inadequacy of the regression paradigm used in the study of human judgment. *Theory and Decision*, 1976, **7**, 57-65.

NOTES

1. The inclusion of instances here is not meant as criticism of the authors of these studies. For the most part, it was not their intention that correlations be taken as a measure of fit. Nevertheless, high correlations have been taken by others as evidence of a good fit.

2. Unfortunately, the subjects in the main study of Slovic and Lichtenstein (1968) cannot be reanalyzed so simply since an incomplete stimulus design was used. However, the procedures and results in the preliminary and main studies appear to be substantially the same. Moreover, the authors base their support of the additive model on the preliminary data (pp. 7-8).

3. This is a good example of where a complete linear regression, rather than a correlation, would have been appropriate. The slope of the regression line, which is clearly less than one, would have provided direct evidence of subadditivity. Thus, the present negative comments should not be taken as criticism of correlation techniques in general. Rather, the comments are restricted to the use of correlations as an index of fit for a model.

(Received for publication March 31, 1977.)