Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. Multivariate Behavioral Research, 18, 115-126.

Multivariate Behavioral Research, 1983, 18, 115-126

Notes and Commentary

SOME CAUTIONS CONCERNING THE APPLICATION OF CAUSAL MODELING METHODS

NORMAN CLIFF
University of Southern California

ABSTRACT

Literal acceptance of the results of fitting "causal" models to correlational data can lead to conclusions that are of questionable value. The long-established principles of scientific inference must still be applied. In particular, the possible influence of variables that are not observed must be considered; the well-known difference between correlation and causation is still relevant, even when variables are separated in time; the distinction between measured variables and their theoretical counterparts still exists; and ex post facto analyses are not tests of models. There seems to be some danger of overlooking these principles when complex computer programs are used to analyze correlational data, even though these new methods provide great increases in the rigor with which correlational data can be analyzed.

The development of the rigorous and generalized methods for testing hypotheses concerning underlying structures in covariance matrices is perhaps the most important and influential statistical revolution to have occurred in the social sciences. Certainly it is the most influential since the adoption of analysis of variance by experimental psychology in the 1940's (Lovie, 1979). Those who heard Karl Jöreskog's paper at the meeting of the Psychometric Society in 1968 knew they were hearing something important, but certainly did not appreciate how far-reaching the application would be.

Reference here, of course, is to the various procedures that go by the names of maximum likelihood factor analysis (Jöreskog, 1969; 1971), analysis of covariance structures (Jöreskog, 1973a, 1973b; McDonald, 1978), and generalized least squares analysis (Browne, 1974; Jöreskog & Goldberger, 1972). The most familiar names are perhaps those of the programs: LISREL and ACOVS (Jöreskog & Sörbom, 1978). An excellent survey is provided by Bentler (1980).

This is a revised version of a paper presented at the special national workshop, "Research Methodology and Criminal Justice Program Evaluation," Baltimore, Maryland, March 17-19, 1980, sponsored by the National Institute for Justice. Preparation of the paper was supported in part by the National Institute for Justice, Grant 79-NI-AX-0065. The author wishes to thank Dr. Robert Cudeck for a number of valuable suggestions.

Request for reprints should be sent to Norman Cliff, Department of Psychology, University of Southern California, Los Angeles, California 90089.

Initially, these methods seemed to be a great boon to social science research, but there is some danger that they may instead become a disaster, a disaster because they seem to encourage one to suspend his normal critical faculties. Somehow the use of one of these computer procedures lends an air of unchallengeable sanctity to conclusions that would otherwise be subjected to the most intense scrutiny. These methods have greatly increased the rigor with which one can analyze his correlational data, and they solve many major statistical problems that have plagued this kind of data. However, they solve a much smaller proportion of the *interpretational*—inferential in the broader sense—problems that go with such data. These interpretational problems are particularly severe in those increasingly common cases where the investigator wishes to make causal interpretations of his analyses.

The remarks that follow are not to be taken as a blanket indictment of the practice of seriously modeling various processes, using correlational data. Indeed, many important contributions have been made in this regard, both before and after the introduction of those methods pioneered by Karl Jöreskog. The work of Bock and his collaborators contain striking early examples of the intelligent application of path analysis (Bock, 1960; Bock & Bargmann, 1966; Bock, Dicken, & Van Pelt, 1969; also, Harris, 1963). There are many current examples of equal merit. The book by Namboordi, Carter, and Blalock (1975) furnishes a good introduction to the avoidance of specific pitfalls, as well as how to apply these methods to study a variety of problems.

Principles of Scientific Inference

At an early point in our professional training, most social scientists are required to take courses in research design, and in such courses they are assumed to absorb some principles of scientific inference. The purpose of the present paper is to remind readers of these principles, and to suggest that they still apply, even though the most numerically sophisticated of available computer programs are used to analyze the data.

Of the numerous principles that underlie the scientific method, there are four which modern, computerized modeling methods seem especially likely to entice us to violate. These go by a variety of different names in different sources, but their content or meaning is fairly standard. The first principle is that the data do not

confirm a model, they only fail to disconfirm it, together with the corollary that when the data do not disconfirm a model, there are many other models that are not disconfirmed either. The second principle is that *post hoc* does not imply *propter hoc*. That is, if a and b are related, and a followed b in time, it is not necessarily true that b caused a. The third principle is that just because we name something does not mean that we understand it, or even that we have named it correctly. And the fourth principle is that *ex post facto* explanations are untrustworthy. These principles may seem so self-evident that they do not need re-stating, but observation of the behavior of researchers who use modern algorithms for the analysis of correlational data such as LISREL (Jöreskog & Sörbom, 1973) leads to the feeling that some reminders may be in order.

We learn in basic courses not to treat correlation as confirming causation, and presumably all but the most naive users of structural modeling of covariance matrices perceive this in the simple cases. If the level of complexity increases, however, there is a tendency to lose sight of the fact that nothing fundamental changes. If the correlations are among *latent* variables, each of which is represented by a few manifest variables, it seems rather easy to conclude that some causal model of relations among the *latent* variables has been verified. Surely the same principles apply to correlations among latent variables as among observed ones. Indeed, there is even more cause for concern in these cases due to the ambiguity in the definition of latent variables that is touched upon in a later section.

Models Are Not Confirmed by Data

The Unanalyzed Variable

If variables x and y correlate, this is an interesting observation. If x seems somehow more fundamental than y or precedes it in time, we may tentatively conclude that x is an explanation for, even a cause of, y. But suppose two other variables v and w are known to correlate, or are suspected of correlating, with x and y. Then the skeptic can argue that the correlation between x and y is an epi-phenomenon, and the real explanation is v and w. Then traditionally it has been the responsibility of scientists to go and see whether v and w indeed correlate with both x and y. It is also our responsibility as practitioners—administrators, politi-

cians, clinicians—to not go charging off in pursuit of x until we have been assured that reasonable alternative explanations have been ruled out.

Suppose w, but not v, is measured along with x or y, correlations among all three variables are computed, and it is found that indeed it is x which has the major explanatory role. This enhances the status of x, but does not clinch matters because there is still v to take into account, and this has not been done. In the days of simple, partial, and multiple correlational analysis, this seemed clearly recognized. It is still true, but somehow, if LISREL or ACOVS does the analysis, the user tends to forget it. If the program gives a non-significant chi-square for a model where arrows go from x to y and w, and rejects the one where they emanate from w, it is somehow felt that the model is confirmed by the data. It is not. It is just not disconfirmed. A model involving v is not disconfirmed either, and until someone gets the data and does disconfirm it, the status of this model is just as good as the one involving x. These programs are not magic. They cannot tell the user about what is not there.

Alternative Models

Even without resorting to alternative variables as explanations of data, it is well to remember that models other than the one that "fits" will fit the data equally well. Indeed, the very form of the equations underlying LISREL *guarantee* that in virtually every application there are an infinity of models that will fit the data equally well. While only a small minority of these may be legitimate alternative explanations of the data, the fact that an author's model is not discomfirmed means that these are not disconfirmed either. Where any of these alternatives have a legitimate status in the literature, of course, the researcher should recognize the ambiguity of the data.

The foregoing is largely a restatement of the widely held view of science (e.g., Popper, 1959) that asserts that data can never positively confirm a model; they can only fail to disconfirm it. The social aspect of science needs also to be recognized. Science is a group activity that relies heavily on mutual criticism to maximize the validity of conclusions. Much of what characterizes good research is the ability to anticipate, and neutralize with data, potential criticisms of conclusions. To treat models as confirmed when in fact there are plausible alternatives that cannot be ruled

out by the data is no more justifiable with a model for a covariance structure than it is anywhere else.

Post Hoc Is Not Propter Hoc

One of the first things we learn about the scientific method is the uncertainty that surrounds attempts to ascertain the causal relations between variables. This uncertainty is particularly great in those cases where the data are correlational, i.e., where for some reason or another it is not possible for the investigator to actively manipulate the values of the independent (causal) variable. Indeed, as Simon (1977, p. 76) put it in the context of criticism of structural equation models,

> It is to be noted that we have here again explicitly introduced the notion of an experimenter who, by his direct control over the parameters of the equations provided by nature, can bring about independent variations in the variables that are exogeneous to AX. If this procedure is operationally meaningful, the experimenter, confronted by a self-contained structure A, can partition the structure into its complete subsets and, isolating each of these from the whole, proceed to determine its parameters.

This is to say that the most satisfactory, almost the *only* satisfactory, method for demonstrating causality is the active *control* of variables, so that the complexity of the relations among them may be simplified, at least temporarily. With correlational data, it is not possible to isolate the empirical system sufficiently so that the nature of the relations among the variables can be unambiguously ascertained.

Correlational data, where the investigator must take whatever his samples give him, lead to almost unending controversy. A salient example is the ambiguity of the role of cigarette smoking in various diseases. It is only recently that the last remaining pockets of doubt concerning this relation have begun to disappear, in spite of the decades of correlational evidence in human beings, and in spite of the experimental evidence from animal studies.

This is not to say that correlational data cannot be suggestive of causal relations—the smoking studies again are an example—indeed they can and usefully do. It is just that they do not establish these relations, and until various lines of converging evidence support the ideas of a causal relation, one should hold in abeyance the inference of causality.

The uncritical reading of a leading text on causal modeling

may start unwary students of this topic on the wrong foot. In an example, used to introduce the ideas and methods of causal modelling analysis of some correlational data, Kenny (1979, p. 25) concludes that "Father's Occupation caused Intelligence," on the basis of a path-analytic analysis of correlations, including those between these variables. It may be that it does, but somehow I doubt it. It seems unlikely that, if ever the causal variables involved in scores on modern "intelligence" tests are sorted out, one's father's occupation will ever be one of them. There may be several variables which are correlated with father's occupation and which may be influencing one's scores on such tests, but to suggest a direct causal role for it seems naive.

The temporal order of observations is not in itself an infallible guide to the identification of causal relations. If *a* comes before *b*, and they are correlated, then there is still room for the influence of an innumerable collection of other variables to operate, particularly where the separation in time is substantial. Also, where the measurements are of rather slowly changing characteristics of the individual, as they often are in the analysis of covariance structures, the time of *measurement* may be an unreliable guide to the sequence of events. Consider the possibility that we measure a child's intelligence in the fifth grade and her father's occupation when she is in the tenth.

Let us not forget that *"post hoc ergo propter hoc"* is a conclusion we reach only after ruling out the influence of all plausible alternative causes, doing this at a minimum by their inclusion in the correlation matrix, and preferably by means of holding them constant and manipulating the causal one.

The Nominalistic Fallacy

The next problem is also an old and familiar one: If we name something, this does not mean we understand it. Suppose we posit a theoretical variable, invent a manifest variable which we think would be related to it, give it the same name and then correlate that variable with some others. The resulting correlations—or absence of them—of the manifest variable cannot be treated as if they corresponded directly to relations of the theoretical one. This is one manifestation of what has been known for many years as the *nominalistic fallacy*. The fallaciousness remains, no matter how

sophisticated the computer program which takes part in the analysis.

This gap between manifest variable and theoretical variable has two aspects. One is the *invalidity* problem: the variable at least partly measures something different from what we think it does. The other is the *unreliability* problem: the variable partly doesn't measure anything at all. Both validity and reliability affect a variable's correlations, not only its direct ones, but its higher-order ones as well. Therefore, they affect LISREL results, too, because that is all it has to work from.

This means that we can only interpret our results very cautiously unless or until we have included enough indicators of a variable in our analysis, and have satisfied not only ourselves but skeptical colleagues and critics that we have done so. The beauty of these new methods is that when we have done our work in this respect, they can provide a more solid basis for our conclusions, than we have previously had.

The nominalistic fallacy issue can take a more subtle form, particularly in those more factor analytically oriented modeling ventures, those in which the correlations among a few manifest variables are interpreted as defining a latent variable. Perhaps they do define such a variable, but the truth is that the variable so defined remains latent, and not manifest. Factors never "emerge." They stay hidden. Correlations and factors derived from them do *not* specify what they are. If it is argued that this has always been true in factor analysis, the present author will agree but think of that as a poor support for the practice. Three quarters of a century of factor analyzing mental ability data has led to little insight in the nature of individual differences in this respect, and, indeed, there is precious little agreement on how to interpret the "factors" that do "emerge." The fact is that, in the statement that verbal ability is whatever certain tests have in common, the empirical meaning is only that certain tests correlate, and "verbal ability" is nothing more than a shorthand for the observation of the correlations. It does not mean that verbal ability is a variable that is measurable in any manifest sense

Even when there are several indicators of a latent variable, there still remains residual ambiguity concerning the interpretation of its correlations. Recently, discussion related to this issue has been reviewed (e.g., McDonald & Mulaik, 1979; Mulaik & McDonald, 1978; Rozeboom, 1982; Steiger, 1979). While the details of the practical conclusions are still somewhat controversial or

obscure, a principle seems to have been established. This is that the definition or interpretation of a latent variable (or both) only becomes less uncertain as (a) the number of indicators and (b) their individual validities (communalities) increase. Furthermore, it seems that the status of a latent variable with only three or four indicators, each of which correlate .7 or so with it, remains very ambiguous.

The major exception to my indictment of this approach in the mental testing field is Guilford's Structure of Intellect research—not the theory as such (Guilford, 1967), but the research that was carried out in its name (Guilford & Hoepfner, 1971). Here, there was an active attempt to *manipulate* the correlational patterns of tests by systematically varying their demands on the test-taker. Thus, these researchers attempted to escape the correlational bounds of the typical factor, just as they should. It is unfortunate that the amount of success that even this work had is questionable (Guilford, 1974; Horn & Knapp, 1973; see also Guilford, 1982). There was little reliable success in relating factor loadings to manipulations of test content. This lack of success reinforces the position that one is unlikely to infer the nature of a true underlying variable from observing the company it keeps in a correlation matrix. The fact that this was true here in an area where there were decades of fairly systematic work makes the interpretational picture even bleaker for the average investigator who ends up correlating a grab-bag of variables which he happens to have available, or which are forced on him by circumstance. Let me hasten to say that the general usefulness of correlational data is not being questioned. This is only a reminder that such data are typically only suggestive of the nature of the true underlying variables. Even in what is called "confirmatory" factor analysis, it is not the nature of the factors which is confirmed; the only thing which is confirmed is that the observed covariance matrix is not *inconsistent* with a certain pattern of parameters. It does not tell us what those parameters mean, and experience has shown that our belief that we do know what they mean is often ill-founded.

On this particular issue, it may be pointed out that the "confirmation" of a set of parameter values, by which is meant the factor loadings, uniquenesses, factor intercorrelations, and the like, does not in any way mean that this is the only set of parameters which is consistent with the data. Quite the contrary. There are typically an infinity of alternative sets of parameters which are

equally consistent with the data, many of which would lead to entirely different conclusions concerning the nature of the latent variables (Green, 1977). Parenthetically, it may be noted that one of the most troublesome technical problems in confirmatory factor analysis is the determination of just which families of alternative solutions are consistent with a given set of data.

The Unreliability of Hindsight

There is a fourth point that needs to be made, and this again is a well-worn one, but it seems to need to be made yet again. This has to do with the tendency to treat *ex post facto* analyses as if they were confirmatory.

Let us consider a situation that is probably fairly common. The investigator has a moderately well-defined model for a set of data, specifically for the variance-covariance matrix. The model may be causal or it may be only structural. He tests this model on his data, and finds a highly significant chi-square; i.e., the model is rejected by the data. It happens to all researchers, and the natural response is to look around and try to find out what made the model fail. This can be done by some combination of residual-inspecting, second-derivative examination, or plain head-scratching. Eventually the researcher is likely to hit upon some combination of changes in the model which results in a new model which does "fit," according to the statistical criterion. How is the new model to be treated?

One possibility is to write up the study as if that were the model which one started with, and treat the results as if they confirmed this model, not mentioning the original one or the fact that it was found on the basis of trial and error. That approach would exemplify an ethical problem, and the concern here is not with ethical problems, but rather with scientific method. The scientific-statistical problem is that the investigator has looked at the data and from them made some estimates of the parameters using the same data. The relevant probability distributions no longer apply, and, thus, the goodness-of-fit value is meaningless, or very nearly so.

Before this is dismissed as a quibble, consider how complex the data are if there are more than a few variables; consider how many different possible adjustments could be made, and then consider how likely it is that, by chance, at least one could be found

which would make a big difference in the goodness-of-fit. Thus, *ex post facto* models do not have the status of models to be fit by the data.

In analysis of variance, or regression, there are ways of treating *ex post facto* tests of parameters, for example, the Tukey and Scheffé corrections (Winer, 1971). There, the researcher finds that in a problem of even minor complexity, the resulting adjustments mean that he must be several times as uncertain of *post hoc* comparisons as of *a priori* comparisons.

No one knows how to treat the *post hoc* problem in a corresponding analytic way in confirmatory covariance analysis, and the problem may be, in principle, insoluble due to the fact that there are qualitatively different ways to modify these models. The main point is that once one starts adjusting a model in the light of the data, the model loses its status as a hypothesis, and that model finally chosen represents in practice a much more unstable picture of what is really going on.

There is one thing that one can do in this case, a crude but sturdy and time-honored practice: cross-validation. One can split the original sample in half, and put one half aside. Fiddle with models to one's heart's content on the first half. When one has a model that seems to fit, bring out the other half of the data, and try that model out on it. As far as those data are concerned, the model is a legitimate hypothesis—those data did not influence the nature of the model.

If the model fits, everything is satisfactory, but there are unpleasant things which can happen here, too, of which the most common is that the model gives a horrendous chi-square when it is cross-validated. That is unpleasant, but as far as the scientific enterprise is concerned, that does not make it bad. The reason, of course, is that the model's failure to cross-validate is telling us that the apparent good fit in the original sample was largely a bootstrap effect.

There are disadvantages to the cross-validation strategy, the most obvious being that it reduces the sample size by half. In my opinion, the disadvantages are greatly outweighed by the fact that it does allow one to test his model, rather than leaving the investigator, and the consumers of his research, in the position of trying to make use of results which they know are unstable to an unknown degree, or, worse yet, trying to do so when they do *not* know that the results are unstable.

Conclusion

The considerations that have been presented here are aspects of one general point. This is, that these beautiful computer programs do not really change anything fundamental. Correlational data are still correlational, and no computer program can take account of variables that are not in the analysis. Causal relations can only be established through patient, painstaking attention to all the relevant variables, and should involve active manipulation as a final confirmation. *Post hoc non est propter hoc,* and a good fit for a causal model does not confirm it. Also, let us not forget the discrepancy that typically exists between our theoretical variables and their empirical counterparts, and that latent variables are difficult to understand correctly. Finally, let us be aware of the problem of *ex post facto* analysis, and not mislead either ourselves or our readers, concerning the stability of our results.

Two final points may be made. One is that these issues are particularly important in applied research. Someone is likely to take action, perhaps far-reaching action, on the basis of applied research findings. Therefore, both producers of such research and audiences or consumers of it need to be particularly concerned that the conclusions reached are valid ones. Finally, let it be emphasized that the programs such as LISREL and its relatives provide completely unprecedented opportunities to do this kind of research well. With their aid, conclusions can be made which heretofore would have been impossible, but only provided the analysis is approached intelligently, tough-mindedly, and honestly.

References

Bentler, P. M. Multivariate analysis with latent variables: Causal modeling. In M. R. Rosenzweig & L. W. Porter (eds.), *Annual Review of Psychology,* 1980, Vol. 31. Stanford, California: Annual Reviews, Inc., 1980, 419-456.

Bock, R. D. Components of variance analysis as a structural and discriminal analysis for psychological tests. *British Journal of Mathematical and Statistical Psychology,* 1960, *13,* 151-163.

Bock, R. D., & Bargmann, R. E. Analysis of covariance structures. *Psychometrika,* 1966, *31,* 507-534.

Bock, R. D., Dicken, C., & Van Pelt, J. Methodological implications of content-acquiescence correlation in the MMPI. *Psychological Bulletin,* 1969, *71,* 127-139.

Browne, M. W. Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal,* 1974, *8,* 1-24.

Green, B. F. Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research,* 1977, *12,* 263-287.

Guilford, J. P. *The nature of human intelligence.* New York: McGraw-Hill, 1967.
Guilford, J. P. Rotation problems in factor analysis. *Psychological Bulletin,* 1974, *81,* 498-501.
Guilford, J. P. Cognitive psychology's ambiguities: Some suggested remedies. *Psychological Review,* 1982, *89,* 48-59.
Guilford, J. P., & Hoepfner, R. *The analysis of intelligence.* New York: McGraw-Hill, 1971.
Harris, C. W. (ed.). *Problems in measuring change.* Madison, Wisconsin: University of Wisconsin Press, 1963.
Horn, J. L., & Knapp, J. R. On the subjective character of the empirical base of Guilford's Structure-of-Intellect Model. *Psychological Bulletin,* 1973, *80,* 33-43.
Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika,* 1969, *34,* 183-202.
Jöreskog, K. G. Simultaneous factor analysis in several populations. *Psychometrika,* 1971, *36,* 409-426.
Jöreskog, K. G. Analysis of covariance structures. In P. R. Krishaiah (Ed.), *Multivariate analysis - III.* New York: Academic Press, 1973a.
Jöreskog, K. G. A general method for estimating a linear structural equation system. In A. S. Goldberger and O. D. Duncan (Eds.) *Structural equation models in the social sciences.* New York: Academic Press, 1973b.
Jöreskog, K. G., & Goldberger, A. S. Factor analysis by generalized least squares. *Psychometrika,* 1972, *37,* 243-260.
Jöreskog, K. G., & Sörbom, D. *LISREL IV Users Guide.* Chicago: National Educational Research, 1978.
Kenny, W. *Correlation and Causality.* New York: Wiley, 1979.
Lovie, A. D. The analysis of variance in experimental psychology: 1934-45. *British Journal of Mathematical and Statistical Psychology,* 1979, *32,* 151-178.
McDonald, R. P. A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology,* 1978, *31,* 59-72.
McDonald, R. P., & Mulaik, S. A. Determinacy of common factors: A non-technical review. *Psychological Bulletin,* 1979, *86,* 297-306.
Mulaik, S. A., & McDonald, R. P. The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika,* 1978, *43,* 177-192.
Namboordi, N. K., Carter, L. F., & Blalock, H. M., Jr. *Applied multivariate analysis and experimental designs.* New York: McGraw-Hill, 1975.
Popper, K. R. *The logic of scientific discovery.* London: Hutchinson, 1959.
Rozeboom, W. W. The determinacy of common factors in large item domains. *Psychometrika,* 1982, *47,* 281-295.
Simon, H. *Models of discovery.* Dordrecht, Holland: D. Reidel, 1977.
Steiger, J. H. Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels. *Psychometrika,* 1979, *44,* 157-167.
Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.