Correlation and Causation: A Logical Snafu

Paul A. Games
The Pennsylvania State University

ABSTRACT. Page and various collaborators (Keith, Page, & Robertson, 1984; Page, 1981; Page & Keith, 1981, 1982) have defended the use of data from investigations (correlational studies) to make causative conclusions. These methods are not defensible logically or statistically. They can only suggest hypotheses that then should be tested by proper experiments. At worst, as in the Coleman studies, they have been used to make social policy based on unjustified conclusions. The logical flaw of the methods of path analysis or structural equation modeling is shown. A proper evaluation of the role of investigations versus experiments is cited in the work of Cochran (as described by Rubin 1983).

Page and various collaborators have promoted the use of techniques applied to correlational data that are claimed to yield “causative” conclusions (Keith, Page, & Robertson, 1984; Page, 1981; Page & Keith, 1981, 1982). In Games and Klare (1967, p. 445), we supported the conclusion that “correlations never prove causation” and extended the statement slightly to “investigations never prove causation.” The point is that experiments, studies that allow the random assignment of subjects to treatments, equate a multitude of extraneous variables so that any differences observed may be attributed properly to the treatments. With investigations, studies of intact groups or other studies lacking experimental techniques, there are always alternate explanations possible, and, thus, well-supported causative statements cannot be made. Such investigations are labeled in this article as correlational studies. These include path analysis, multiple regression, partial correlations, structural analysis, or any other technique applied to data collected without experimental control, regardless of the particular statistical technique used.

Those who read the Page articles may have concluded (falsely) that the statement that “investigations never prove causation” is old fashioned and may be ignored. Page and Keith (1982, p. 26) even wax indignant about those who question causative statements from correlational data and ask, “Indeed, where have Travers and other denigrators of causal research been lately?... For those seeking a psychological introduction we suggest a recent text by David Kenny (1979) called — you guessed it — Correlation and Causation.” Despite these assertions, correlational data alone still are not an adequate base for inferences of causation.

In logic, you are taught what are legitimate logical syllogisms, and what are not. A classic example is:

1. All normal dogs have four feet. (If A [normal dog], then B [four feet]).
2. Sparky is a normal dog. (A)
3. Therefore, Sparky has four feet. (Therefore B).

This is a legitimate argument. If statements 1 and 2 are true, then statement 3 is also true. In contrast, consider the following argument:

1. All normal dogs have four feet. (If A, then B)
2. Cedar has four feet. (B)
3. Therefore, Cedar is a dog. (Cannot conclude this, B does not imply A)
This is an illegitimate argument, even though, superficially, it looks similar to the first example. When statements 1 and 2 are true, statement 3 may still be false. Cedar happens to be a cat, and a multitude of other possibilities exists. Accepting this illegitimate argument is know as the logical error of affirming the consequent.

The correlational methods defended by Page and Kenny make this error. They build a causative model (A) that predicts certain correlational patterns (B). This is equivalent to statement 1 above (if A, then B). Then they observe some correlational data and see if they are consistent with the pattern they deduced. If the correlational data are consistent with the pattern (B), they do not prove that the model is correct. It is still illegitimate to conclude A from B. To state that the correlational data support the model is an error of affirming the consequent. There may be many other models that would also yield a correlational pattern that would be consistent with the data. The most that can be said in these correlational studies is that the data do not contradict the model.

Some people seem to think that the nonsupport of the model is in the same sense that we do not prove the null hypothesis, that we merely cannot reject the null hypothesis. This is not true. Given proper solutions for the N needed to yield a desired power in a statistical test, we can not only retain a null hypothesis, we can also conclude that the parameter difference is not as large as the important difference we specified (Cohen, 1969). If we find a 99% confidence interval for M1 - M2 is from -0.32 to +0.41 in an experiment where we defined an important difference as 15 points, we have a strong justifiable conclusion that the difference is not important. The objection to causative conclusions from a correlational study holds regardless of [begin page 241] power considerations. (In fact, the smaller the N and the greater the sources of error variance, the lower the power and the more likely we are to conclude that the data is consistent with any model, reasonable or not.) The only justifiable conclusion from a correlational study with consistent data is that the data do not contradict the model. Under no condition does it justify support for the model in a sense similar to a conclusion that —.32 < M1 - M2 < +.41 justifies support for the null hypothesis H0: M1 - M2 = 0. To claim that the correlational data do not contradict the model, therefore they support the model, is an error of affirming the consequent.

If the correlational data deviate from the pattern predicted by the causative model (not B), it is legitimate to conclude that the model or the reasoning is wrong. You can legitimately reject models by this logic. If A, then B is correct as statement 1, and not B is obtained in statement 2, then not A is a legitimate conclusion. Either the model itself is wrong, or something is wrong in the reasoning that was used to deduce B. Knowing that Sparky has only two feet (as would be the case if Sparky is a human child) may properly lead to the conclusion that Sparky is not a normal dog. Not B does indeed lead to the conclusion not A.

Lest the reader conclude that the author’s objections are idiosyncratic, let me quote from a review of Kenny’s Correlation and Causation book that appeared in the Journal of the American Statistical Association (JASA) (Ling, 1982).

Alas, the serious limitations of this book lie not in its lack of mathematical rigor, but in its faulty logic as well as its faulty presentation and interpretation of certain statistical methodology (p. 490)

In closing this review I feel obligated to register my strongest protest against the type of malpractice fostered and promoted by the title and content of this book. (p. 491)

In short, this is the most negative book review I have ever seen in JASA. Statisticians, like most professionals, do not lightly toss around the word “malpractice.” The Page articles, the Kenny book, and the above review are all unfortunate examples of extreme positions. Investigations (here correlational studies) are worthwhile as sources of hypotheses about causation. Before social policy based on those hypotheses is made, however, they should be verified by proper experimentation.
Interestingly, two journals recently devoted major space to structural equation models (SEM), and neither journal noted that the logical basis of the method was flawed. The *Journal of Educational Statistics* devoted the entire second issue of 1987 to this topic; *Child Development* devoted the first 175 pages of Volume 58 (1987) to 12 articles on this topic. Some of the articles in these two volumes had negative points about SEM, yet others illustrated the same flaws as Page's and Kenny's work.

Biddle and Marlin (1987) illustrate many fundamental errors. “Causal modeling... is a technique that is suggested for improving our ability to make causal inferences from field-study data... causal models, path diagrams, regression analyses... LISREL... represent an advance toward deriving stronger inferences from field-study data” (p. 4) and “standardized regression coefficients... can be compared directly to decide... which of two explanatory variables has a greater impact” (p. 7). That is an interesting claim. A standardized regression coefficient is the slope of a regression line in which the vertical axis is the standard score of a residualized criterion as predicted by all other predictors, $X(2)$ to $X(p)$ and the horizontal axis is the residualized predictor. The contribution each of two predictors make to a total $R^2$ can be determined by the uniqueness coefficient defined by Darlington (1968) and others, but it has no simple relation to the standardized regression coefficients. These regression coefficients are clearly affected by all $p$ predictors.

I do not wish to pretend that all the articles in these two journal issues are favorable to SEM. Freedman (1987) and Rogosa (1987) are both critical. Yet the users of SEM rarely answer criticism but continue on, undaunted in their faith.

One of the greatest educational expenditures of recent United States history was the reaction to the correlational study generally known as the Coleman report (Coleman et al., 1966). Based on mere correlational data, the Coleman report gave the following causative conclusion: “Thus, if a white pupil from a home that is strongly and effectively supportive of education is put in a school where most pupils do not come from such homes, his achievement will be little different than if composed of others like himself. But if a minority pupil from a home without much educational strength is put with schoolmates with strong educational backgrounds, his achievement is likely to increase” (Coleman, 1967, p. 509).

Federal judges, with help from the United States Office of Education, thus concluded that the discrepancy between the achievement scores of Blacks and Whites would be reduced if busing were used to integrate schools so that many schools had, say, 30% Blacks and 70% Whites. Billions of dollars of the U.S. education budget were diverted to buses, bus drivers’ salaries, gasoline, and increased smog. Years later, Coleman concluded that busing had not worked: “Insofar as one intended consequence of integration is an increase in achievement of black children, the intent is largely defeated” (Coleman, Kelly, & Moore, 1975). Newspapers and magazines had a field day pointing out the inadequacies of social scientists making government policy.

Most people want to reduce racial prejudice, and it is possible that busing has had a positive effect in this direction. Crain and Mahard (1983) give an extensive summary of studies on integration that provides far better support for busing than did the Coleman report. They even found four regions where proper randomization of volunteers permitted a comparison of Black volunteers who were bused to White suburbs versus Black volunteers who remained in central city schools. In 15 studies done in those regions at various times, the meta-analysis standardized difference measure averaged to + .235 standard deviations. Hooray for administrators who, when resources are beneath demand, are courageous and far-sighted enough to randomly assign students, thus permitting strong experimental conclusions. Although this difference of + .235 standard deviations’ superiority of the bused Black children over those who remained in the central city schools may not be enough to satisfy all critics, it is enough to render the 1975 “defeated” conclusion suspect.
It is important that when social scientists make recommendations for social policy they have an adequate factual and methodological basis for their recommendations. Here, the proper supportive evidence was collected only after improper conclusions had been implemented.

In short, we may have lucked out, no thanks to Coleman. Both the original 1966 Coleman report and the 1975 report have had the same correlational data analyzed by other investigators, who came to different conclusions. Often the defense of these different conclusions becomes rather esoteric and certainly is beyond the power of federal judges without extensive statistical training to understand. For example, the 1975 study, where Coleman gave the White flight interpretation, was based on a selected sample of only 19 cities, where the basis of the selection was not completely clear. Others have used the same data, but with more cities and different operational definitions of some variables of the multiple regression, and rejected the White flight interpretation (Farley, 1975; Pettigrew & Green, 1976; Rossell, 1975).

The point is that the causative conclusions of the 1966 report, the 1975 report, and those who provided different interpretations are insufficient grounds for making major social policy changes. Page and Keith (1981) and Coleman (Coleman, 1981; Coleman, Hoffer, & Kilgore, 1982) are another instance of different interpretations from the same correlational data. However, even if they had agreed, even if the reports had been unanimous, they still would not be an adequate base for making policy recommendations. Because they are based on correlational data only, any of these recommendations should be investigated by legitimate experiments before billions of dollars are spent by federal, state, or local governments.

How much have we learned from the busing experience? See Crain and Mahard (1983). How much more rational it would have been if the busing hypothesis had been taken as one of several possibilities, and meaningful experiments had been conducted. The feds could have said, "We will give [begin page 244] several million dollars to random cities D, A, P, L, and I for busing. We will give the same amount to cities X, R, Q, B, and Y for improving target schools in impoverished core areas. We will give the same amount to cities 0, N, H, V, and C for programs encouraging parental tutoring." Then a meaningful evaluation (with pretreatment and posttreatment measures) of achievement and attitudes in these cities and in the unfunded cities might give more interpretable results.

Ah-ha, says the critic. Here he has rejected the conclusions based on a sample of 19 cities, but he is proposing a study based on a treatment sample of only 15 cities. Yes, indeed. I would place greater trust in results from an experiment with 15 cities over a correlational study with 50 cities, let alone 19. The experiment provides control; the correlational study does not. By randomization, we have avoided at least systematic biases between the treatment groups. I am not overly comfortable about randomization producing equal pretreatment scores for the three treatment groups when there are only 5 cities per treatment. Naturally, I would prefer 20 cities per treatment, both for control and power purposes. In the correlational study, however, we have no randomization control at all. We obtain some measures on some cities and ignore many other measures and many other cities. In the experiment, the individual differences between cities become a source of error variance that we can adequately assess. In the correlational study, the individual differences between cities becomes the source of all interpretations, the source of all so-called “causative” effects. The correlational study may not even be measuring the crucial social factors, but you can be sure someone will still come out with “causative” effects, whether their methodology comes from multiple regression, first-order correlations, partial correlations, path analysis, or covariance structural analysis.

The crucial difference between experiments and investigations (correlational studies) is that in experiments the process of random assignment guarantees that in the population there is no correlation between the experimental errors and the treatment effects. For an investigation to yield proper causative conclusions, it is necessary for the same condition to hold. Any analysis of correlational data to yield causative effects makes a similar assumption; but with only correlational data, it usually is impossible to verify that assumption.
23 Of course, it would take a degree of political courage that is rarely seen in Washington for politicians to agree to spend millions of dollars on several methods that might not work, let alone to assign dollars to cities on a random basis. It should be the business of the American Educational Research Association (AERA), the American Psychological Association (APA), and the United States Department of Education to persuade the politicians that this is a rational course of action.

24 Unfortunately, there will always be “social scientists” with “causative” answers for the government, even when they have no adequate data. Research organizations such as the AERA and the APA should be shouting that these conclusions cannot be trusted.

25 Let me make it clear that this criticism is solely about the use of correlational data to come to causative conclusions. Correlational methods have many other legitimate scientific roles. When a college (or an industry) has many more applicants than it has openings, it is entirely appropriate to use correlational methods to select applicants who are most likely to succeed. The use of aptitude tests probably is the greatest practical contribution the behavioral sciences have made to our Western culture; these uses are based upon correlational studies.

26 Page and Keith even deprecate the major legitimate method for coming to causative conclusions: “The increasing difficulties of doing good experiments should drive us to other methods.…” (1982, p. 26). Nonsense. It should drive us to train graduate students adequately so that, when they become members of the U.S. Department of Education, they know enough to distinguish experimental results that are worthy of implementation from correlational results that should be taken as no more than hypotheses. It should drive us to train administrators so that school principals will recognize the need for experiments and will support them.

27 It is informative to contrast Page and Keith’s misguided view with that of an outstanding statistician who contributed at least seven papers on the analysis of observational data. Cochran states, “If nature mixes things up thoroughly, as she sometimes seems to do, statistical methods will not sort them out very well…. Experimentation… needs to be exploited as much as possible. The question: ‘Why can’t I do an experiment?’ is always worth asking, even if it sounds unrealistic” (Rubin, 1983, p. 63). Cochran also states, “Single group studies are so weak logically that they should be avoided whenever possible…. Comparison groups bring a great increase in analytical insight. The influence of external causes on both groups will be similar in many types of study and will cancel or be minimized when we compare treatment with no treatment” (Rubin, p. 57).

28 Rubin summarizes one of Cochran’s major points: “For obtaining accurate estimates of treatment effects, Cochran never viewed observational studies as a replacement for randomized experiments. He realized that data from observational studies were less expensive to obtain than data from randomized experiments and sometimes were the only data available due to practical considerations. Nevertheless, he recommended that randomized experiments should be considered whenever feasible and, further, kept in mind as an ideal objective when designing observational studies” (Rubin, 1983, p. 44). Behavioral scientists would be well advised to adopt this attitude. [begin page 246]

References


