Causal Analysis in Theory and Practice

Comments on an article by Grice, Shlimgen and Barrett (GSB): "Regarding Causation and Judea Pearl's Mediation Formula"

Filed under: Discussion, Mediated Effects,Opinion — moderator @ 3:00 pm, 09/04/2011

Opening remarks

Stan Mulaik called my attention to a recent article by Grice, Shlimgen and Barrett (GSB) (linked here http://psychology.okstate.edu/faculty/jgrice/personalitylab/OOMMedForm_2011A.pdf) which is highly critical of structural equation modeling (SEM) in general, and of the philosophy and tools that I presented in "The Causal Foundation of SEM" (Pearl 2011) (http://ftp.cs.ucla.edu/pub/stat_ser/r370.pdf.) In particular, GSB disagree with the conclusions of the Mediation Formula -- a tool for assessing what portion of a given effect is mediated through a specific pathway.

I responded with a detailed account of the disagreements between us (copied below), which can be summarized as follows:

Summary

1. The "OOM" analysis used by GSB is based strictly on frequency tables (or "multi-grams") and, as such, cannot assess cause-effect relations without committing to some causal assumptions. Those assumptions are missing from GSB account, possibly due to their rejection of SEM.

2. I define precisely what is meant by "the extent to which the effect of X on Y is mediated by a third variable, say Z," and demonstrate both, why such questions are important in decision making and model building and why they cannot be captured by observation-oriented methods such as OOM.

3. Using the same data and a slightly different design, I challenge GSB to answer a simple cause-effect question with their method (OOM), or with any method that dismisses SEM or causal algebra as unnecessary.

4. I further challenge GSB to present us with ONE RESEARCH QUESTION that they can answer and that is not answered swiftly, formally and transparently by the SEM methodology presented in Pearl (2011). (starting of course with the same assumptions and same data.)

5. I explain what gives me the assurance that no such research question will ever be found, and why even the late David Friedman, whom GSB lionize for his staunch critics of SEM, has converted to SEM thinking at the end of his life.

6. I alert GSB to two systematic omissions from their writings and posted arguments, without which no comparison can be made to other methodologies:

(a) A clear statement of the research question that the investigator attempts to answer, and
(b) A clear statement of the assumptions that the investigator is willing to make about reality.

Below is my full response to GSB (slightly edited) which is also posted on SEMNET (by Stan Mulaik), August 31.

Judea

Full Response to GSB:

I begin with the mediation example presented in (Pearl 2011, p. 28-30) which carries the promise of illuminating the source of many other disagreements on SEM and its role in causal analysis.

The problem

The example entails three dichotomous variables labeled X, Z, and Y. X represents a drug treatment (drug/no drug), Z stands for the presence of a certain enzyme in the blood stream (enzyme/no enzyme), and Y represents physical recovery from an ailment (cured/not cured). Drug treatment is the initial cause, the enzyme is the mediator, and recovery is the outcome. The three variables are linked in a standard mediation model format showing both direct and indirect connections between X and Y (as in Figure 8(a) of (Pearl 2011) and Figure 1 of GSB). The model further shows no correlations between the error terms, thus asserting that X is assumed to be randomized, and Z and Y unconfounded. Based on these assumptions, we need to define and assess what is usually called "the effect of mediation", namely "what portion of the effect of X on Y is mediated through Z"

To facilitate the discussion, I reprint the available data in the following two tables.

Drug	Enzyme	Percentage cured
YES	YES	80%
YES	NO	40%
NO	YES	30%
NO	NO	20%

Drug	Percentage of Subjects with Enzyme present
NO	40%
YES	75%

The solution

Mediation analysis asks two questions:

1. DE - The extent to which mediation is NECESSARY, i.e., what improvement in cure rate would still be realized had the drug acted alone, without enhancing enzyme secretion.

2. IE - The extent to which mediation is SUFFICIENT, i.e., what improvement in cure rate would be realized by mediation alone, namely, by the drug acting solely on enzyme secretion, barring its direct effect on recovery.

GSB pose a different question: "What are the pragmatic conclusions that can be drawn about the importance of administering the drug with or without the presence of the enzyme?" Driven by this question, GSB then calculate the effect of the drug with the enzyme present (obtaining 50% improved cure rate), subtract from it the effect of the drug when the enzyme is absent (showing only 20% improvement) and conclude: "it appears the drug augments the enzyme to produce a 30% overall increase (50% - 20%) in recovery rate among the 1000 persons in the study."

These calculations fall under the rubric of "controlled direct effect" (CDE), treated extensively in (Pearl, 2009, page 126-130) where we hold the mediator constant (say by external means) at two different levels (absent and present) and estimate the effect of the drug separately under the two conditions. The controlled direct effect may a be legitimate quantity to estimate, but it does not capture the research question of interest in mediation analysis, which falls under the rubric of "NATURAL direct and indirect effects". The natural Direct Effect (labeled DE in Pearl 2011) is not concerned with the contrast between the presence and absence of the enzyme (a contrast that would be impractical to realize at the population level), but between the presence and absence of the capacity of the drug to cure the ailment on its own, unassisted by enhanced enzyme production.

Pragmatically, such questions arise when we fear a possible reduction in recovery rate if, for some reason, the drug were to lose its current ability to enhance enzyme production. For example, suppose someone proposes the development of a much cheaper drug, equal in all respects to the one under study, but less effective in enhancing enzyme secretion. To evaluate such proposals, we need to assess the effect of the drug on a hypothetical population in which the enzyme level (for each individual) remains constant, at the same level it had just before administering the drug.

The Mediation Formula tells us (Pearl, 2009, page 132) that, under conditions of no-confounding, the effect of the drug on this hypothetical population (so called Natural Direct Effect, DE) is given by a weighted average of the two controlled direct effects computed by GSB, namely,

$$DE = 0.40 \times 0.50 + 0.60 \times 0.20 = 0.32$$

Thus, the conclusions drawn from the Mediation Formula do not conflict with those drawn from the OOP analyses; they simply answer different policy questions that, in the more general case of non-randomized trials (or in the presence of mediator-outcome confounding) cannot be answered by the multigram analysis advocated by GSB or by any analysis based solely on frequency tables.

The same applies to the analysis of sufficiency (for which there is no "controlled" version). which is concerned with the natural indirect effect (IE), or the effect transmitted solely through the mediating pathway. Here, again, we wish to compare, not the presence and absence of the enzyme, as is done by GSB, but the presence and absence of the drug's capacity to enhance enzyme secretion while suppressing all other effects the drug may have on the disease. In other words, we compare the cure rates for two hypothetical populations; one with the enzyme distribution as it was just before administering the drug (showing 40% presence), and one with the enzyme distribution as it was after administering the drug (with 75% of the subjects carrying the enzyme). The cure rates of these two populations are now compared under no-drug condition, in order to suppress any effect the drug may have on the outcome, save for its capacity to stimulate the mediator.

For this hypothetical question, the Mediation Formula instructs us to take the difference in enzyme counts (0.75 - 0.40) and multiply it by the increase in cure rate observed under no-drug condition (0.30 - 0.20). The resulting difference, IE = (0.75 - 0.40) x (0.30 - 0.20) = 0.035 amounts to about 7% of the 0.46 total improvement caused by the drug. The intuition is simple: Under no drug condition, a subject carrying the enzyme has a (0.30 - 0.20) greater chance of recovering than one without the enzyme. The drug increases the proportion of the former subjects by (0.75 - 0.40). Multiplying the two, we get IE = 0.035, which is approximately 7% of the observed total effect (TE) of the drug. This ratio, IE/TE, represents what I verbally described as the fraction of "recoveries that would be sustained by enzyme stimulation alone", namely, recoveries attributed solely to the drug effect on enzyme stimulation, discounting all other effects the drug may have on the outcome. (The word "alone" is meant to exclude those other effects, not the drug itself)

GSB complain that this quantity does not "mesh" with the quantity they estimate, namely, the proportion of people (12%) who were cured among those who did not take the drug and still produced the enzyme. I totally agree. Whereas the proportions calculated by GSB may be of interest in some statistical surveys, they do not answer the causal question at hand: the effect of the drug on the cure rate, mediated solely by its capacity to enhance enzyme production. The proportion computed by GSB does not take this enhanced production into account.

General comments on GSB arguments and methodology

The need for causal analysis

In this simple example of three dichotomous variables operating under the assumptions of partial mediation (with no confounding), causal relations such as total, direct and indirect effect can be expressed in terms of simple proportions that are easily discernible from the 2x2x2 frequency table. This simplicity may tempt one to conclude, as did GSB, that there is no utility to causal algebra or counterfactual vocabulary, and that the analysis of causation as a whole can be replaced by traditional analysis of frequency histograms, and contingency tables, as is done in OOP.

Such conclusion would be wrong and misleading; there is no such thing as a "model-free approach to detecting cause from data like these" (quoted from Barrett' message). Even the obvious claims listed by GSB, e.g., that "the drug augments the enzyme to produce a 30% overall increase (50% - 20%) in recovery rate" cannot be deduced from the data alone, without the assumptions encoded in the mediation model, and without the help of causal analysis that makes these assumptions explicit and formally derives the stated claim.

To appreciate the need for distinct causal analysis, consider a slight modification of our model to one that represents total mediation, i.e., X-->Z--->Y. Further assume that the data shown above were obtained in a non-randomized observational study, in which an unobserved set of confounders, U, was present, affecting both subjects' choice of treatment (X) and the rate of subjects recovery (Y), but not directly affecting the mediator Z. Given that 15% of the subjects chose to take the drug in this study, we wish to estimate the total effect TE of the drug, that is, the overall increase in cure rate that the drug would produce in a randomized study.

Surprisingly, mathematics permits us to conclude that, regardless of the dimensionality of U, and regardless of its distribution, the total effect of X on Y in our example is equal to TE = 0.05075, thus predicting an increase of only about 5% in cure rate. I challenge GSB to derive this result with OOP or with any other methodology that dismisses causal algebra as unnecessary. I also hope the challenge of solving this toy problem would entice curious SEM students to further exploit the marvels of causal analysis and to become part of the modern age of causation.

Model-based versus model-free analyses

GSB complain about my refusal to commit to either a "full mediation model" (i.e., no direct effect of X on Y), or to a "full moderator model" (no effect of X on Z). They view this non-commitment as "conceptual confusion" and a "conflation of two entirely different models" I disagree. The standard mediation model that I used subsumes both the "full mediation model" and the "full moderator model" as special cases, thus allowing for mediation to co-exist with moderation, and enabling us to disentangle the two by estimating both the degree to which any variable (say Z) acts as a mediator and the degree to which it acts or a moderator (for the effect of interest.) I see no reason to commit in advance to restricted models that are manifestly incompatible with the data at hand or with available scientific knowledge, and that do not permit us to estimate the quantity of interest: the effect of mediation.

This difference in methodology persists throughout the comment posted by GSB. Whereas, in SEM, a model reflects the investigator's perception of reality, in GSB analysis a "model" serves as a classification tool that is applied to frequency tables in an attempt to label certain proportions as "classified correctly" or "not correctly classified" or "not consistent with the expectation", or "ambiguously classified". Presumably, GSB deem these labels to be more meaningful than the research questions that SEM undertakes to answer, for example, effects, counterfactuals, model equivalence, statistical implications, retrospection, policy evaluation, effect decomposition, attribution, external validity, and more. (Some of these questions are discussed in algorithmic details in (Pearl 2011), others are analyzed in my book (2009) and working papers http://bayes.cs.ucla.edu/csl_papers.html

GSB further complain that the Mediation Formula does not go deep enough into the biochemical subprocesses that account for mediation, for it does not distinguish, for example, between a model in which the enzyme and the drug work together, simultaneously, to neutralize the virus, and one in which it is the drug that neutralizes the virus with the enzyme acting merely as a catalyst. Likewise, it does not tell us "why did 60 people recover without consuming the drug or secreting the enzyme?"

Indeed, the Mediation Formula was derived to answer specific mediation related problems (explicated above), at a given level of granularity, regardless of the deep microscopic structure behind the mediation process; it was not recruited to deal with the more ambitious task of uncovering the anatomy of unobserved sub-processes.

At the same time, the deep-structure problems that GSB aspires to solve are not excluded from SEM analysis, as implied by GSB. Nonparametric SEM analysis is in fact perfectly equipped to tell investigators whether a certain nuance in process structure can be corroborated by the data at hand, given the assumptions we are willing to make about the model. Moreover, none of the methodologies so highly revered by GSB, including those of Aristotle, Aquinas, Freedman and Manicas, can address these issues of model building and data corroboration.

I therefore challenge GSB to present us with ONE mechanism-related question that they can answer and that is not answered (swiftly and formally) by the non-parametric SEM methodology presented in Pearl (2011). I dare to pose this challenge because mathematics assures me that it cannot be met. In other words, modern causal analysis comes with mathematical guarantees that, if a scientific question is not answered by SEM methodology then it is not answerable by any analytical means whatsoever (given the same data and same assumptions about nature). Of course, we would all prefer to acquire a fully blown integrated model like the Krebs Cycle shown in GSB's Figure 4. But to acquire such detailed model would take more than analytical methodology; it would take painstaking microscopic experimentation and microscopic observations, guided by a methodology such as SEM (not Aristotle) that can tell us which portion of the model deserves our attention at any stage of the model building process.

Speaking of the Krebs Cycle model that GSB's present as the apex of scientific achievement, observe the arrows going from one variable to another. What do these arrows represent? Do they not represent the same "structural equations" that GSB criticize? They do, indeed. Even the late David Freedman, who could not stand the sight of structural equations throughout his life, changed his mind in the mid 2000s and came to realize that causality should be understood in terms of deterministic functions that map causes into their effects (Freedman, 2004, Evaluation Review 28:267-293) He could not of course admit conversion to SEM thinking, so he anointed what we know as a "structural equation" with a newly coined name: "response schedule" -- "the schedule by which a response is generated." This now gives bystanders the lisence to admire the scientific substance of SEM equations when they are clad in the innocence of Freedman's "response schedule", and ridicule them as "sea of assumptions" (GSB, page 10) when the time comes to derive their logical implications in concrete examples.

Summary

I wish I could offer a more comprehensive comparison between SEM and the methodology advanced by GSB but, unfortunately, two ingredients are glaringly missing from GSB's account without which any comparison would not be constructive:

(1) A clear statement of the research question that the investigator attempts to answer, and

(2) A clear statement of the assumptions that the investigator is willing to make about reality, and how these assumptions are to be encoded formally.

I will be happy to continue this discussion once we have items (1) and (2) explicated, hopefully in the framework of a specific example, preferably in response to the two challenging exercises I proposed.