Encyclopedia of Measurement and Statistics

Spurious Correlation

Contributors: Brian Haig
Editors: Neil J. Salkind
Book Title: Encyclopedia of Measurement and Statistics
Chapter Title: "Spurious Correlation"
Pub. Date: 2007
Access Date: December 10, 2014
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412916110
Online ISBN: 9781412952644
DOI: http://dx.doi.org/10.4135/9781412952644.n427
Print pages: 938-941

http://dx.doi.org/10.4135/9781412952644.n427
In social science research, the idea of spurious correlation is taken to mean roughly that when two variables correlate, it is not because one is a direct cause of the other but rather because they are brought about by a third variable. This situation presents a major interpretative challenge to social science researchers, a challenge that is heightened by the difficulty of disentangling the various concepts associated with the idea of spurious correlation.

Correlation and Causation

Drawing appropriate causal inferences from correlational data is difficult and fraught with pitfalls. One basic lesson social scientists learn in their undergraduate statistics education is that correlation does not imply causation. This adage is generally taken to mean that correlation *alone* does not imply causation. A correlation between two variables *X* and *Y* is not sufficient for inferring the particular causal relationship "*X* causes *Y*" because a number of alternative causal interpretations must first be ruled out. For example, *Y* may be the cause of *X*, or *X* and *Y* may be produced by a third variable, *Z*, or perhaps *X* and a third variable, *Z*, jointly produce *Y*, and so on.

The statistical practice in the social sciences that is designed to facilitate causal inferences is governed by a popular theory of causation known as the *regularity theory*. This theory maintains that a causal relation is a regularity between different events. More specifically, a relationship between two variables *X* and *Y* can properly count as causal only when three conditions obtain: (a) *X* precedes *Y* in time; (b) *X* and *Y* covary; and (c) no additional factors enter into, and confound, the *X*-*Y* relationship.

The third condition requires a check for what social scientists have come to call *nonspuriousness*. A relationship between *X* and *Y* is said to be nonspurious when *X* is a direct cause of *Y* (or *Y* is a direct cause of *X*). A relationship between *X* and *Y* is judged nonspurious when we have grounds for thinking that no third variable, *Z*, enters into and confounds the *X*-*Y* relationship. In this regard, researchers typically seek to establish that there is neither a common cause of *X* and *Y* nor a cause intervening between *X* and *Y*.

Senses of Spurious Correlation

The term *spurious correlation* is ambiguous in the methodological literature. It was introduced by Karl Pearson at the end of the 19th century to describe the situation in which a correlation is found to exist between two ratios or indices even though the original values are random observations on uncorrelated variables. Although this initial sense of a spurious correlation remains a live issue for some social science researchers, it has given way to a quite different sense of spurious correlation. In the 1950s, Herbert Simon redeployed the term to refer to a situation where, in a system of three variables, the existence of a misleading correlation between two variables is produced through the operation of the third causal variable. H. M. Blalock's extension of Simon's idea into a testing procedure for more-complex multivariate models has seen this sense of a spurious correlation come to dominate in the social sciences. As a consequence, the social sciences have taken the problem of spuriousness to be equivalent to checking for the existence of third variables.

[p. 938 ↓]

A Typology of Correlations

In order to understand that this problem of the third variable is not really a matter of spuriousness, it is important to be able to identify different types of correlations in terms of their presumed causes. The following is a typology of such correlations identified in terms of different kinds of presumed causes. These various correlations are sometimes confused when considering the problem of spuriousness.

At the most general level, the typology identifies two kinds of correlation: accidental and genuine. Accidental correlations are those that cannot be given a proper causal interpretation. There are two types of accidental correlation: nonsense and spurious. By contrast, genuine correlations are amenable to a proper causal interpretation. There are two types of genuine correlation: direct and indirect.

Nonsense correlations are those accidental correlations for which no sensible, natural causal interpretations can be provided. Statisticians delight in recounting the more amusing of these cosmic coincidences, such as the high positive correlation between birth rate and number of storks for a period in Britain or the negative correlation between birth rate and road fatalities in Europe over a number of years. In the statistics literature, these are sometimes called *illusory correlations*. These correlations exist, of course, but they cannot be given a plausible causal interpretation.

As characterized here, spurious correlations are accidental correlations that are not brought about by their claimed natural causes. To be true to their name, spurious correlations cannot be genuine correlations because they are false. They are artifacts of method and arise from factors such as sample selection bias; use of an inappropriate correlation coefficient; large sample size; or errors of sampling, measurement, and computation. Karl Pearson's original sense of spurious correlation mentioned above belongs here because the misleading value of ratio correlations depends, not on the relationship between the variables in question, but on their sharing of highly correlated components.

Direct correlations are genuine correlations for which one of the correlates is said to be a direct cause of the other. For example, heavy trucks are a direct cause of road damage, and frequent and intense sun spots directly cause radio transmission noise. The social sciences are replete with empirical studies that are concerned with establishing direct causal relations. For example, manifest independent variables are examined in outcome studies on the assumption that they impact measured dependent variables in a causally direct way. Just as indirect correlations are often misleadingly called spurious correlations, so direct correlations are sometimes misleadingly called nonspurious correlations.

Indirect correlations are the genuine correlations that are produced by common or intervening causes and that we misleadingly call spurious correlations. However, there is nothing spurious about them at all. So-called spurious correlations are really genuine correlations, so their existence can hardly be denied by claiming that they are brought about by some underlying third variable. For example, if general intelligence is the common cause of correlated IQ performance on the verbal and numerical subtests of an intelligence test, then those subtest performances are indirectly and genuinely

correlated. Clearly, this correlated IQ performance is not spurious because general intelligence explains why the correlation obtains; it does not render the correlation nonexistent or give us grounds for thinking that this is so.

From this analysis, it is clear that spurious correlations, properly named, must be regarded as a class of accidental correlations; otherwise we cannot sensibly deny the causal relations they are mistakenly thought to express.

Generative Causation

Tests for so-called spurious correlations are generally carried out to determine whether causal relations are empirical regularities. For this task, the regularity theory of causation is adequate, but only up to a point. Its requirements of temporal priority and covariation capture the idea of regular succession, but in order to properly understand the so-called problem of spuriousness, it is necessary to go beyond the restrictions of the regularity theory. A theory that allows us to do this is the generative theory of causation. The generative **[p. 939 ↓]** theory depicts causation as a relation where, under appropriate conditions, a causal mechanism *produces* its effect. For this to happen, the causal mechanism must connect to its effect and have the power to generate that effect, usually when stimulated by the appropriate causal condition. It should be noted that it is the productivity of a generative mechanism that makes it a causal mechanism, and for this to be possible, there must be a naturally necessary connection that allows for a transmission of power from a cause to its effect. This causal power exists irrespective of whether it is currently being exercised. As such, it is properly viewed as a *tendency*, that is, an existing state of an object which, if unimpeded, will produce its effect. When it is unimpeded, we are able to diagnose the presence of the causal mechanism on the basis of the triggering condition(s) or its presumed effect(s) or both.

Unlike the regularity theory of causation, the generative theory is able to accommodate explanatory theories that are concerned with illuminating unobserved causal mechanisms. We need a theory of causation that affords us the conceptual space to do this because many of the world's causal mechanisms are not open to direct inspection. This is certainly true of the social sciences, where many of the postulated causal

mechanisms are internal states. Intellectual abilities, personality traits, and emotional states are obvious cases in point.

Causation and Spuriousness

Simon's influential analysis of spurious correlation reveals a commitment to something like the regularity theory of causation. He notes that in order to distinguish true from spurious correlation, the term *cause* must be defined in an empiricist manner, with no reference to necessary connections between events, as the generative theory of causation makes.

Simon believes that a commitment to empiricist thinking about causality enables him to distinguish true from spurious correlations as he understands them. Ironically, however, this commitment actually prevents him from drawing his distinction properly. Recall that, for Simon, correlations are spurious if they are brought about by common or intervening causes. Now, given that many of these causes will be latent or unobserved, it follows from a commitment to the regularity theory of causation that for a methodologically acceptable treatment of these variables to be possible, Simon and fellow empiricists must focus on altogether different variables at the manifest level. But this cavalier ontological attitude threatens to wreck our efforts to obtain genuine causal knowledge because the manifest replacement variables cannot act as surrogates for their latent variables, which are common and intervening causes. They are ontologically distinct from such causes and, although as causal conditions they may trigger their latent counterparts, they do not function as major causal mechanisms that determine so-called spurious correlations. Clearly, a coherent perspective on third variables that are latent variables requires a generative theory of causation.

Conclusion

When addressing the third variable problem, methodologists and researchers employ the misleading term *spurious correlation* to speak about genuine, indirect correlations. This practice only muddies the waters. Drawing causal inferences from correlational information is as difficult as it is important, and being clear about our key concepts can

only facilitate such an undertaking. Not only is the terminology confusing, and thereby an impediment to understanding, but it also encourages a misleading view about the relation between causation and spuriousness that has the potential to misguide our causal modeling practices.

BrianHaig

http://dx.doi.org/10.4135/9781412952644.n427
Further Reading

Aldrich, J. Correlations genuine and spurious in Pearson and Yule . Statistical Science 10 364–376 (1995).

Baumrind, D. Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar . Journal of Personality and Social Psychology 45 1289–1298 (1983). http://dx.doi.org/10.1037/0022-3514.45.6.1289

Blalock, H. M. (1964). Causal inference in nonexperimental research . Chapel Hill: University of North Carolina Press.

Haig, B. D. What is a spurious correlation? Understanding Statistics 2 125–132 (2003). http://dx.doi.org/10.1207/S15328031US0202_03

Harré, R., & Madden, E. H. (1975). Causal powers . Oxford, UK: Blackwell.

Pearson, K. Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs . Proceedings of the Royal Society of London 60 489–498 (1897). http://dx.doi.org/10.1098/rspl.1896.0076

Prather, J. E. (1988). Spurious correlation. Encyclopedia of statistical science (Vol. 8, pp. 613–614). New York: Wiley.

Simon, H. (1985). Spurious correlation: A causal interpretation . In H. M. Blalock (Ed.), Causal models in the social sciences (2nd ed., pp. 7–21). New York: Aldine.