How likely is Simpson's paradox in path models?

Ned Kock

Full reference:

Kock, N. (2015). How likely is Simpson's paradox in path models? *International Journal of e-Collaboration*, 11(1), 1-7.

Abstract

Simpson's paradox is a phenomenon arising from multivariate statistical analyses that often leads to paradoxical conclusions; in the field of e-collaboration as well as many other fields where multivariate methods are employed. We derive a general inequality for the occurrence of Simpson's paradox in path models with or without latent variables. The inequality is then used to estimate the probability that Simpson's paradox would occur at random in path models with two predictors and one criterion variable. This probability is found to be approximately 12.8 percent; slightly higher than 1 occurrence per 8 path models. This estimate suggests that Simpson's paradox is likely to occur in empirical studies, in the field of e-collaboration and other fields, frequently enough to be a source of concern.

Keywords: E-collaboration; Simpson's paradox; path analysis; numeric computation; Monte Carlo simulation.

Introduction

Simpson’s paradox, also known as the Yule–Simpson effect and the reversal paradox, is a phenomenon arising from multivariate statistical analyses (Pearl, 2009; Wagner, 1982). It is called a “paradox” because it often leads to paradoxical conclusions. Such conclusions may lead to the development of theories that incorporate causal effects disconnected from reality, based on empirical findings distorted by Simpson’s paradox. This applies to the field of e-collaboration as well as many other fields where multivariate methods are employed.

In the following sections, we provide an illustration of Simpson’s paradox in a path model, followed by a discussion of past estimates of the likelihood of Simpson’s paradox in contingency tables. We then provide a mathematical definition of Simpson’s paradox in path models, in the form of a basic inequality, which we use as a basis for the development of a more general Simpson’s paradox inequality that can be used in numeric estimations of the paradox’s probability. Finally, we illustrate the use of this general inequality to estimate the probability that Simpson’s paradox would occur at random in three-variable path models, which is found to be 12.8 percent; slightly higher than $1/8$. For simplicity, and without any impact on the generality of our discussion, we assume that all variables are scaled to have a mean of zero and a standard deviation of one.

A path model illustration of Simpson’s paradox

Let us assume that we collected data from 300 firms about two variables: degree of collaborative management (X) and firm success (Z). The variable degree of collaborative management (X) measures the degree to which managers and employees collaborate to continuously improve their firms’ productivity and the quality of their firms’ products. The variable firm success (Z) measures the profitability of each firm.

Figure 1 shows a simple path model relating these two variables. Since this path model contains only two variables, then $p_{ZX} = r_{ZX} = 0.5$; where p_{ZX} and r_{ZX} denote the path coefficient and the correlation between the two variables.

Figure 1: Two-variable path model

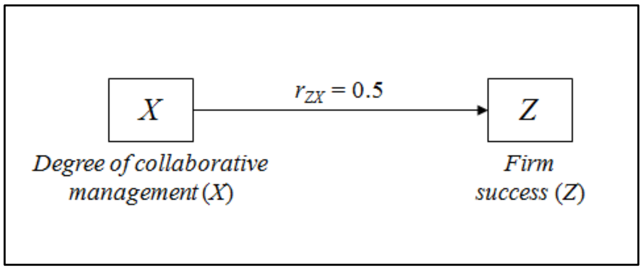
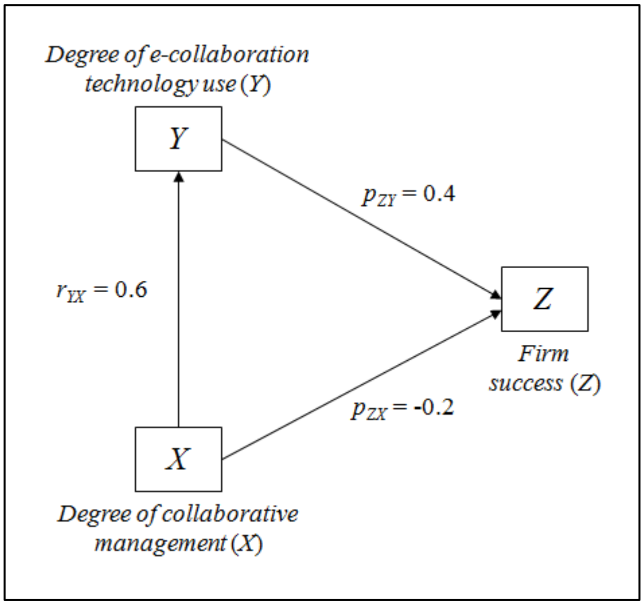


Figure 2 shows a slightly more complex path model with an additional variable pointing at Z: degree of e-collaboration technology use (Y). This new variable measures the degree to which an e-collaboration technology is used. The technology facilitates collaborative management and is available in all firms studied. Because of this, firms where the degrees of collaborative management (X) are high tend to also use the e-collaboration technology intensely, and thus present high degrees of e-collaboration technology use (Y); hence the link $X \rightarrow Y$ in the model.

Figure 2: Three-variable path model


Degree of e-collaboration technology use (Y)
Y
$p_{ZY} = 0.4$
$r_{YX} = 0.6$
Z
Firm success (Z)
$p_{ZX} = -0.2$
X
Degree of collaborative management (X)

In this example, the addition of the new variable led the path coefficient p_{ZX} for the link between the variables degree of collaborative management (*X*) and firm success (*Z*) to assume a negative value (-0.2), in contrast with the positive correlation r_{ZX} (0.5) between the same variables. This sign reversal characterizes what is known as Simpson's paradox in path models.

The likelihood of Simpson's paradox in contingency tables

Simpson's paradox is generally perceived as a problematic phenomenon, since it leads to paradoxical conclusions based on empirical research (Pearl, 2009; Wagner, 1982). In the three-variable path model illustration above, the results would lead many researchers to believe that the association between degree of collaborative management (*X*) and firm success (*Z*) is negative.

The perception that Simpson's paradox is a problematic phenomenon has motivated Pavlides & Perlman (2009), in a seminal article on Simpson's paradox, to estimate the probability that Simpson's paradox would occur at random. They focused on contingency tables, and made various assumptions, also estimating other conditional probabilities. The probability that Simpson's paradox would occur at random in contingency tables was found to be fairly low, in the neighborhood of 1 to 2 percent.

Contingency tables summarize results of multivariate analyses involving categorical variables. Often the categorical variables assume two values each, leading to 2 x 2 tables. Multivariate analyses involving such categorical variables frequently take the form of ANCOVA analyses. Analyses summarized via 2 x 2 tables could be seen as special cases of path analyses where some of the variables are dichotomous.

The probability that Simpson's paradox would occur at random obtained by Pavlides & Perlman (2009) could lead empirical researchers to reasonably conclude that Simpson's paradox is an unlikely phenomenon, which should occur only very rarely in empirical research.

However, Pavlides & Perlman's (2009) focus on multivariate analyses involving categorical variables may have led to a probability estimate that is significantly lower than that for path analyses, which typically include variables measured on ratio scales. This possibility provided the motivation for the current study.

A mathematical definition of Simpson's paradox

In this section we provide a mathematical definition of Simpson's paradox, in the form of a basic inequality. This basic inequality is necessary for the development of a more general Simpson's paradox inequality, which in turn can be used in numeric estimations of the paradox's probability.

Let X and Z be stochastic variables in a linear path model (Wright, 1934; 1960) where X is hypothesized to directly cause Z. The relationship between these two variables can be expressed as

$$Z = r_{ZX}X + \varepsilon,$$

where r_{ZX} is the correlation between X and Z, and ε is the error term that accounts for the variance in Z that is not explained by X.

Let Y be another stochastic variable that is correlated with X and that is a true direct cause of Z. The relationship among the variables X, Y and Z can be expressed as

$$Z = p_{ZX}X + p_{ZY}Y + \theta,$$

where p_{ZX} and p_{ZY} are the path coefficients for the links $X \rightarrow Z$ and $Y \rightarrow Z$, respectively, and θ is a new error term that accounts for the variance in Z that is not explained by X and Y.

Simpson's paradox in connection with the link $X \rightarrow Z$ occurs when

$$p_{ZX}r_{ZX} < 0. \qquad (1)$$

Simpson's paradox in connection with any link among two variables, such as the link $Y \rightarrow Z$, can be analogously defined (e.g., $p_{ZY}r_{ZY} < 0$).

This simple inequality refers to each link in the path model, and provides the basis for a straightforward test to identify Simpson's paradox occurrences: if the path coefficient and the correlation associated with any link in the path model assume different signs, then an instance of Simpson's paradox exists in connection with the link.

The simple inequality above, used to mathematically define Simpson's paradox, forms the foundation on which a more general Simpson's paradox inequality for numeric estimation of probabilities is developed.

A general Simpson's paradox inequality

In this section we derive a more general inequality, relying only on correlations, which can be used in numeric estimation of unconditional and conditional probabilities that Simpson's paradox would occur. The inequality is stated as a theorem, whose proof follows.

Theorem. Let r_{ZX}, r_{ZY} and r_{XY} be the correlations among the stochastic variables X, Y and Z in a linear path model, where X and Y are hypothesized to directly cause Z. Then Simpson's paradox occurs in connection with the link $X \rightarrow Z$ when

$$r_{ZX}{}^2 < r_{ZX} r_{ZY} r_{XY}. \tag{2}$$

Proof. We know from Wright (1934) that the path coefficient p_{ZX} can be expressed in terms of correlations as

$$p_{ZX} = \frac{r_{ZX} - r_{ZY} r_{XY}}{1 - r_{XY}{}^2}. \tag{3}$$

Multiplying both sides of (3) by r_{ZX} yields

$$p_{ZX} r_{ZX} = r_{ZX} \frac{r_{ZX} - r_{ZY} r_{XY}}{1 - r_{XY}{}^2}. \tag{4}$$

Comparing (1) and (4), we see that Simpson's paradox occurs when

$$r_{ZX} \frac{r_{ZX} - r_{ZY} r_{XY}}{1 - r_{XY}{}^2} < 0. \tag{5}$$

Since $1 - r_{XY}{}^2$ is always nonnegative, we can reduce (5) to

$$r_{ZX}{}^2 < r_{ZX} r_{ZY} r_{XY}.$$

We can see from this general inequality that Simpson's paradox will not occur when all of the correlations are zero. Also, given that $r_{ZX}{}^2$ is always nonnegative, Simpson's paradox will never occur when one of the correlations has the opposite sign of the other two (e.g., $r_{XY} < 0$, $r_{ZX} > 0$ and $r_{ZY} > 0$). Conversely, if all correlations are positive, then Simpson's paradox will always occur in connection with the link $X \rightarrow Z$ when $r_{ZX} < r_{ZY} r_{XY}$.

This general inequality is particularly useful in numeric estimations of the likelihood of Simpson's paradox based on Monte Carlo simulations. This is because it contains only correlations, which vary within set bounds. The earlier inequality used to derive this general inequality (i.e., $p_{ZX} r_{ZX} < 0$), on the other hand, contains a path coefficient (p_{ZX}), which can vary from $-\infty$ to $+\infty$.

How likely is Simpson's paradox in three-variable path models?

Since the correlations in (2) vary from -1 to 1, the general Simpson's paradox inequality can be used directly in Monte Carlo simulations (Delgado, 1993; Robert & Casella, 2005) to estimate unconditional and conditional probabilities of the occurrence of Simpson's paradox.

Let us consider the unconditional probability, which is the probability that Simpson's paradox would occur at random. Let C be the set of combinations of random values of correlations $\{r_{ZX}, r_{ZY}, r_{XY}\} \mid r_{ij} \in \mathbb{R} : -1 < r_{ij} < 1 : i, j \in \{X, Y, Z\}$, generated through a Monte Carlo simulation. Let S be the subset of combinations within C that satisfy (2). The probability that Simpson's paradox would occur at random is given by

$$P_S = \frac{|S|}{|C|} \text{ as } |C| \rightarrow \infty,$$

where $|S|$ and $|C|$ are the cardinalities of S and C, respectively.

We estimated P_S by generating a large number of combinations (10,000) of random values of correlations, and repeating this process a number of times (1,000) to obtain an asymptotic probability estimate. Through this process we estimated P_S to be approximately 12.8 percent. That is, of all possible three-variable path models, slightly more than 1 in 8 will contain an instance Simpson's paradox.

This suggests that Simpson's paradox is likely to occur in empirical studies frequently enough to be a source of concern. This may lead to the development of theories, based on empirical findings distorted by Simpson's paradox, which incorporate causal effects that are disconnected from reality.

Concluding remarks and future research

The probability that Simpson's paradox would occur at random in three-variable path models, which we estimated to be approximately 12.8 percent, may be an underestimation of the actual rate of occurrence of Simpson's paradox in empirical studies employing path analysis.

There are various reasons for this supposition. One of them is that empirical studies are designed to test models where correlations among variables are often expected to fall within certain ranges, where conditional Simpson's paradox probabilities may be greater than the probability that Simpson's paradox would occur at random. Also, empirical studies often employ models that are more complex than a three-variable path model.

To assess this supposition in the context of actual empirical studies employing path models, we conducted detailed reviews of articles published in the 2003-2013 period in the journal *MIS Quarterly*, a highly selective academic journal in the applied field of information systems. E-collaboration is frequently seen as a subfield of the field of information systems.

We focused on articles that employed path analysis and related methods – multiple regression analysis and structural equation modeling – where both path coefficients and correlations are normally reported. Among the articles reviewed in this period in *MIS Quarterly*, 19 contained instances of Simpson's paradox. These amounted to approximately 15.7 percent of all articles published in that journal employing path analysis and related methods.

The estimation of conditional probabilities via Monte Carlo simulations, where correlations among variables are expected to fall within certain ranges, is recommended as future research. One such range could be $0.3 < |r_{ij}| < 0.8 \mid i,j \in \{X,Y,Z\}$, which would include correlations that arguably are strong enough to yield significant path coefficients at commonly used sample sizes (e.g., $N > 150$), but not as strong as to lead to vertical or lateral collinearity (Kock & Lynn, 2012) among variables.

Acknowledgments

The author thanks Wei Ning and Pelumi Ojomu for their help in the review of published articles and identification of Simpson's paradox occurrences.

References

Delgado, M.A. (1993). Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, 17(3), 199-204.

Kock, N., & Lynn, G.S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, 13(7), 546-580.

Pavlides, M.G., & Perlman, M.D. (2009). How likely is Simpson’s Paradox? *The American Statistician*, 63(3), 226-233.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Robert, C.P., & Casella, G. (2005). *Monte Carlo statistical methods*. New York, NY: Springer.

Wagner, C.H. (1982). Simpson's paradox in real life. *The American Statistician*, 36(1), 46–48.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3), 161-215.

Wright, S. (1960). Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, 16(2), 189-202.