Bill Shipley, Cause and Correlation in Biology. A User's Guide to Path Analysis, Structural Equations and Causal Inference with R. Second edition, 2016, Cambridge University Press, ISBN: 978-1-107-44259-7, 314 pp., £39.99 (paperback)

Anyone who has taken a statistics class has heard the old adage “correlation does not mean causation”. Distinguishing correlation from causation is not obvious. Correlation belongs to the language of statistics, while causation belongs to the language of logics. Good use of statistics allows a researcher to investigate whether correlations might reveal potential causal relationships. Although correlation and causation must not be conflated, in most cases correlation is an indicator of causation. This book examines how correlation analyses can be legitimately used to disclose and test causation.

There are many possible ways to statistically test causal hypotheses via correlation approaches. This book is dedicated to the application of structural equation modelling (which includes path analysis as a special case). Structural equations are a statistical tool particularly well suited to deal with biological data. Cause-effect relationships in biology are typically multivariate, which means that a given effect may be the result of multiple factors that act directly or indirectly; structural equations allow the researcher to help make sense of complex relationships. The use of structural equation modelling in biology and environmental sciences is still relatively limited, however. Structural equation modelling is seldom explained in general textbooks, although it is not conceptually more complicated than other methods. Shipley's book is recommended reading because it offers one of the most comprehensive presentations of structural equation modelling for biologists. In fact, it may be useful to any scientist who operates in circumstances where causation cannot be investigated by randomised experiments (a situation that is typical of, but not unique to, biology).

Use of structural equations has been subject to many critiques (e.g. Tomarken & Waller, 2005, Annual Review of Clinical Psychology, 1: 31-65; VanderWeele, 2012, American Journal of Epidemiology, 176: 608-612; Lindquist & Sobel, 2013, NeuroImage, 76, 446-449) and I had few occasions to read papers using structural equations in ecology, although their number is growing. Apart from philosophical considerations (Hall, 2007, Philosophical Studies, 132: 109–136; Hitchcock, 2009, Philosophical Studies, 144: 391–401), structural equations, like all statistical procedures, have their pros and cons, and cannot be viewed as a panacea, but a good tool kit includes them. Thus, a comprehensive book such as this is really welcome.

The book is organised into eight chapters that starts with basic concepts and proceeds to explicate progressively more complicated and detailed aspects of models of analysis and testing. The first chapter presents a very interesting discussion about the meaning of correlation and causation and the distinction between controlled and randomised experiments. The second chapter explains how causal relationships can be translated into statistical models, and therefore expressed in the language of probability. This chapter illustrates the logic of causal inference, how to construct graphs describing the causal relationships between variables, how to deduce probabilistic independence relationships using the d-separation technique, the underlying probability assumptions and a series on counter-intuitive consequences and limitations of d-separation. Chapter 3 presents the basis of path analysis and the use of d-sep tests--a set of methods which are alternative to structural equations and which were developed by the author of this book to deal with for small sample sizes, non-normally distributed data or non-linear functional relationships.

Chapter 4 illustrates how to test path models using maximum likelihood, how to disentangle the effects in path diagrams, and how to use multiple regressions in path analysis. Chapters 5 and 6 provide details about measurement error and latent variables; in these three chapters, the use of the R package “lavaan” (latent variable analysis) to construct and test path models is also described. Chapter 7 deals with multigroup and multilevel models and correction for the non-independence of observations, with a section specifically dedicated to phylogenetic non-independence. The last chapter presents intriguing clues about how to generate causal hypotheses in the light of structural equation modelling with some empirical examples that can be used as analogies for analysing readers’ cases.

The book also includes the prefaces to both editions, an appendix that summarises the usage of two R libraries (“ggm” and “lavaan”), a reference section and an index. Although the book uses a minimum of statistical jargon, it is quite demanding reading. Conceptually very dense, the book is written in a didactic style that does not allow the reader to skip sections. Although the book includes detailed instructions throughout for running analyses in R, there are no comprehensive examples presented in a schematic way. These would have greatly aided the reader in getting some practice.

The book has a hybrid nature. It is a guide to structural equation modelling (and probably one of the best guides a biologist can find) and it contains technical sections that may be annoying a reader more interested to practical applications. Thus, it is an excellent text for those seeking to understand how they can better use statistics to identify causation, but it is not a cookbook.

Simone Fattorini

Department of Life, Health and Environmental Sciences, University of L’Aquila, Via Vetoio, Coppito, 67100, L’Aquila, Italy and CE3C – Centre for Ecology, Evolution and Environmental Changes / Azorean Biodiversity Group, Universidade dos Açores, 9700-042, Angra do Heroísmo, Açores, Portugal

E-mail address: simone.fattorini@uivaq.it