One Ecosystem 7: e72780
doi: 10.3897/oneeco.7.e72780

OPEN ACCESS

Review Article

General guidance for custom-built structural equation models

James Benjamin Grace ‡

‡ U.S. Geological Survey, Lafayette, United States of America

Corresponding author: James Benjamin Grace (jimgrace001@gmail.com)

Academic editor: Alessandro Gimona

Received: 09 Aug 2021 | Accepted: 27 Jan 2022 | Published: 01 Feb 2022

Citation: Grace JB (2022) General guidance for custom-built structural equation models. One Ecosystem 7: e72780. https://doi.org/10.3897/oneeco.7.e72780

Abstract

Structural Equation Modelling (SEM) represents a quantitative methodology for specifying and evaluating causal network hypotheses. The application of SEM typically involves the use of specialised software packages that implement estimation procedures and automate model checking and the output of summary results. There are times when the specification details an investigator wishes to implement to represent their data relationships are not supported by available SEM packages. In such cases, it may be desirable to develop and evaluate SE models “by hand”, using specialised regression tools. In this paper, I demonstrate a general approach to custom-built applications of SEM. The approach illustrated can be used for a wide array of specialised applications of non-linear, multi-level and other custom specifications in SE models.

Keywords

structural equation modelling, custom specifications, Acadia National Park

Introduction

Structural equation modelling has grown in popularity as a method for quantitative analysis for natural systems in the past two decades (Shipley 2000, Grace 2006, Shipley 2016).

During this same time period, scientists have adopted increasingly sophisticated statistical models to represent their data. The development of SE models is most commonly conducted using covariance procedures, such as those implemented in the R package 'lavaan' (Rosseel 2012) or piecewise solution procedures, such as those found in the R package 'piecewiseSEM' (Lefcheck et al. 2018). Other approaches, such as Bayesian implementation (Lee 2007) are sometimes used in SEM and the general approach demonstrated here can be adapted to those cases as well.

Covariance methods, such as lavaan, provide tremendous flexibility with regard to the kinds of models that can be estimated. These include those with latent variables, error correlations, reciprocal interactions and causal loops. The greatest limitations associated with modelling covariances in lavaan include: (a) response types are limited to Gaussian and binary, (b) linkages must be linear equations and (c) incorporation of random effects is limited. The piecewiseSEM package permits a wide variety of response types and also the use of mixed models. It too runs into limitations, however, when dealing with non-linear linkages and more complex responses, such as multi-parameter zero-inflated response variables. In cases where lavaan and piecewiseSEM do not support particular specifications of interest, it is possible to develop SE models as collections of submodels built from a series of regressions. In this paper I illustrate basic procedures for such custom implementations.

Necessary Steps Required for Customised Structural Equation Modelling

Grace et al. 2012 recommend ten main tasks to consider as part of the SEM workflow process, spanning from consideration of the goals of the analysis to the reporting of methods, findings and interpretations. Here, I focus on only a subset of the total workflow steps:

1. Establish Causal Assumptions and Testable Implications
2. Consider Data Characteristics
3. Deciding on and Incorporating Custom Specifications
4. Checking for Omitted Links that Should be Added
5. Model Pruning, Model Comparisons and Model Selection
6. Development of Summary Quantities

The Ecological Example

For this illustration, I utilise data from a study of wetland biotic integrity, conducted at the Acadia National Park (NP) in Maine, USA (data from Grace et al. 2016). Acadia NP is located on a 24,000 ha granite bedrock island that includes the highest mountain on the Atlantic coast of the United States (upper panel of Fig. 1) Due to its mountainous topography, wetlands on the island are in relatively small catchments. Wetland watersheds range from relatively undisturbed (panel A of Fig. 1) to highly impacted (panel B). A wide

range of disturbance levels were included in the sample, as represented by HDI (human disturbance index) values in the map occupying panel C of Fig. 1. In the referenced study, 37 non-forested wetlands were examined as part of an assessment of the relationships between degree of human development, hydrology and biological integrity. Details related to the measurements taken are summarised in Table 2 of Grace et al. (2012). Only four variables are utilised in the illustrations below. These four variables are: (a) intensity of land use in the watershed, ranging from low to high (0 to 3), (b) count of the number of hydrologic alterations, from 0 to 6, (c) proportion of the year the soil surface was flooded and (d) average number of native plant species per study plot. Modelling choices for these variables are discussed below.

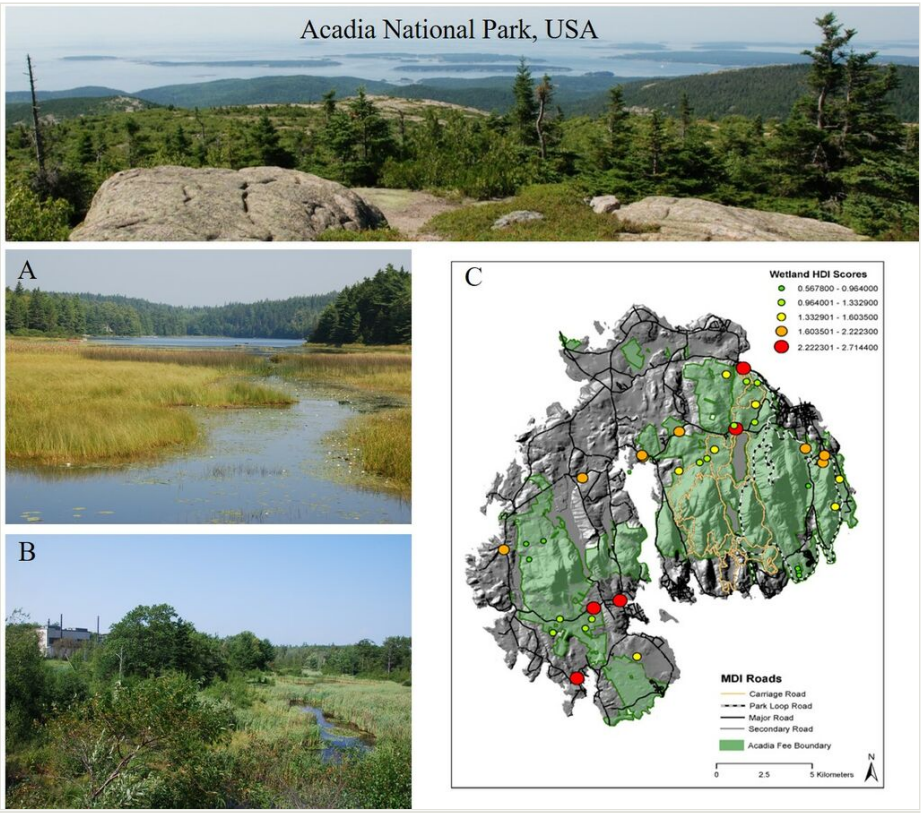

Figure 1.

The study area for the ecological example – Acadia National Park, Maine, USA. Upper panel is the view towards the Atlantic Ocean from the top of Cadillac Mountain, the highest point in the Park. Panel A is from an area where human disturbance of the watershed is minor, while Panel B is from an area with intense human disturbance (sewage treatment facility). Panel C is a map of the Park showing variation in human disturbance intensity (HDI) for the wetlands sampled. Photographs by the author. Map produced by Kathryn Miller, US National Park Service.

The ambition of the original study is represented by the metamodel shown in Fig. 2. Metamodels describe the hypothesis under examination at a very general, conceptual level. Here, it can be seen that the overall objective was to understand how human activities might influence key elements of biotic integrity (Karr 1991) through alterations in hydrology and nutrient loading. Fig. 3 shows the submodel of four variables considered here (in coloured boxes), as well as the full SE model that was evaluated in the original study (which includes the grey boxes and arrows). In words, the hypothesis considered here is that intensity of land use (**Use**) drives the number of hydrologic alterations (**Hyd**), which in turn influences the proportion of time a wetland remains flooded (**Flood**), which influences the number of native wetland species at a site (**Rich**).

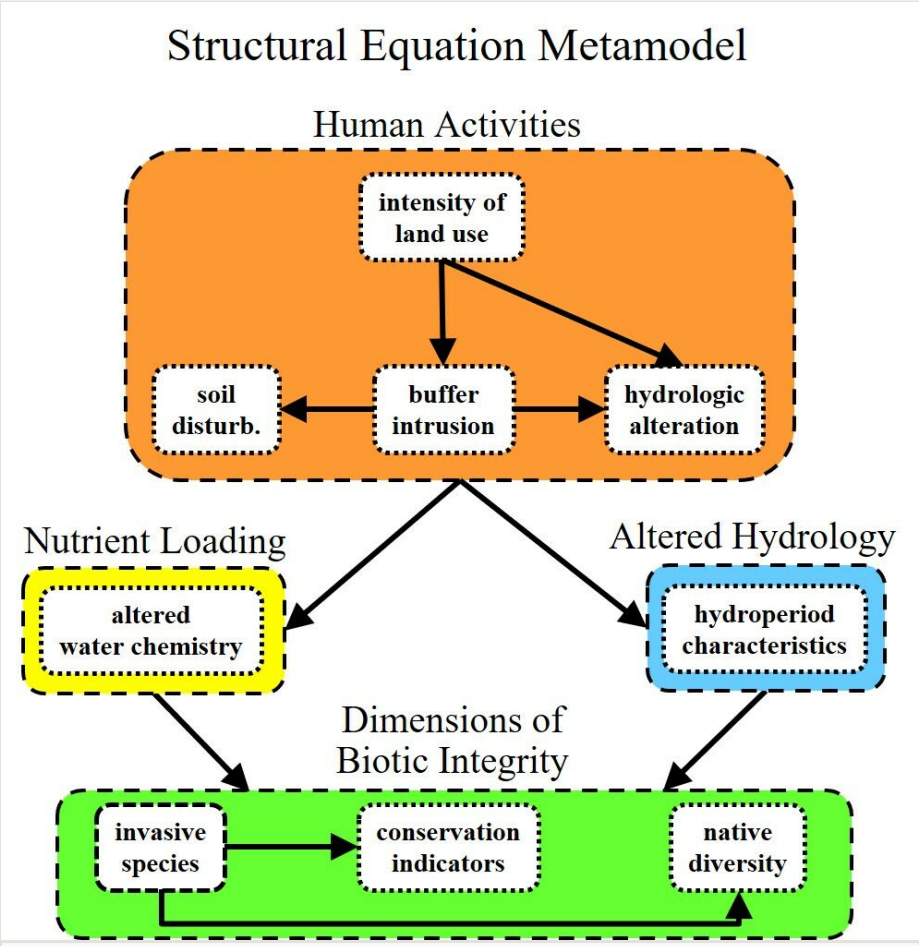

Figure 2.

Metamodel representing the general hypothesis considered in the source publication (Grace et al. 2012).

Protocol for Specialised SEM Applications

1. Establish Causal Assumptions, Testable Implications and Alternative Structural Models

In SEM, we evaluate data expectations that follow from hypothesised model architectures using the principles of causal analysis. This can and should be addressed prior to considering specification details, as it directs the investigator's focus to important causal assumptions that are separate from any statistical assumptions. Grace and Irvine (2020) provide a succinct description of these principles. Fig. 4 presents an initial causal diagram for our illustration. There are important distinctions between Structural Models and Causal Diagrams. Causal diagrams constitute a formal graphical mathematical language suitable for causal analysis. Causal diagrams ignore statistical details and make no assumptions about the forms of either linkages (linear or non-linear) nor about the distributions of the response variables and error distributions. Knowing only the architecture of the hypothesis, we can state the causal assumptions that are encoded and conditional independences that are implied.

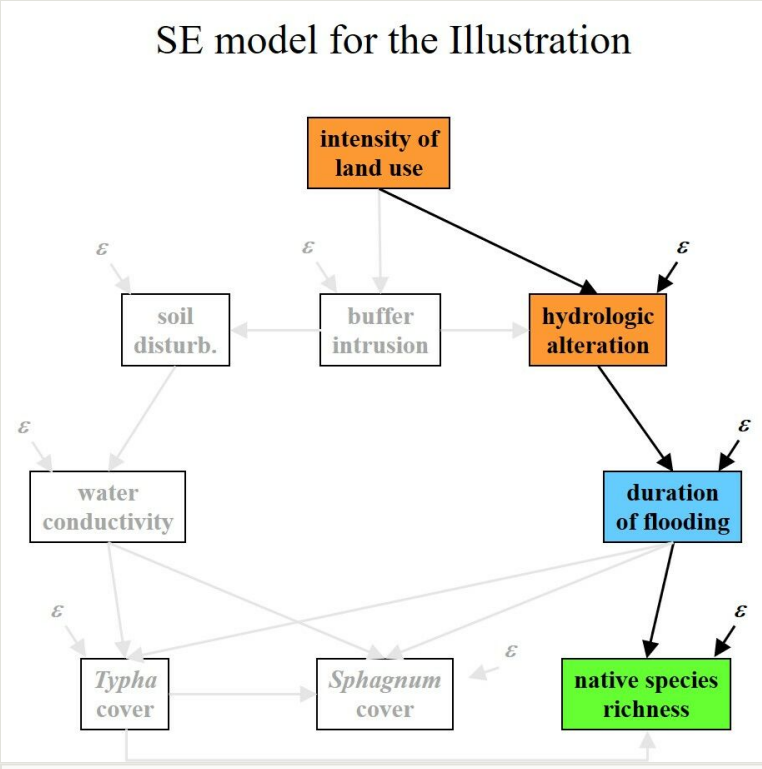

Figure 3.

SE model made up of four variables extracted for use in the illustration contained in this paper. The complete model, including the portion greyed out, was the structural equation model used in Grace et al. (2012) to represent the general hypothesis shown in Fig. 2.

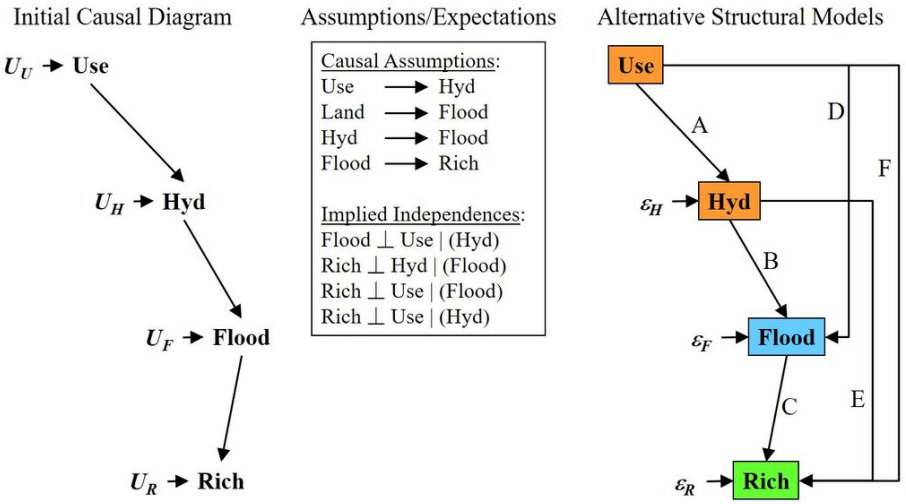

Figure 4.

Initial causal diagram (on left), relevant causal assumptions and potential alternative structural models. U variables in the causal diagram refer to unspecified causes of variations in models. In structural models, which are informed by data, U variables are replaced by prediction error variables (epsilons) where appropriate.

In this case, the omission of arrows from Use to Flood, Use to Rich and Hyd to Rich, suggest formal tests for our hypothesis that apply regardless of statistical details. In words, what is hypothesised is that the effect of land use (human activity) on native richness (biological integrity) can be explained by associated changes in hydrology and water level stability. Once data are brought to bear, the conditional independence claims can be tested (which ultimately leads us to learn that more processes were at work than were implied by our initial hypothesis).

Also of importance in our causal diagram (Fig. 4) is the absence of arrows connecting the U variables. The U variables represent the unspecified additional causes creating variations in our variables. If we hypothesised an omitted confounding factor as part of the data generating process, we would include double-headed arrows connecting the U variables influenced by the omitted factor. Grace et al. (2012) considered several possible common-cause confounders (distance from shore, salinity), but found no evidence of their importance. Further discussion of omitted confounders and their remedies can be found in Grace 2021.

When developing SE models by hand, it is important for the investigator to contemplate the alternative models that could be discovered through the analyses. When using specialised SEM programmes and packages, this is not essential (though can be valuable) because automated procedures will compare estimated models to hypothetical saturated models. Fig. 4 shows the nested set of possible structural models in the right panel. Here, arrows

indicated by the letters D-F represent linkages that may be added to the initial model, while letters A-C indicate linkages that might be dropped due to lack of empirical support.

2. Consider Data Characteristics

None of the variables included in this analysis is a classical Gaussian variable (Fig. 5). All variables have a minimum value of zero and three appear to possess zero-inflated distributions (Use, Hyd and Flood). Use is an ordered categorical variable with 4 integer values (0-3), but since this variable is exogenous in the model, its distribution will not influence model specifications (though its distribution is relevant to coefficient interpretations). Hyd is count data, with six integer values. Since there was a maximum of six types of hydrologic alteration, the variable Hyd represents the proportion of possible alterations found at each site. Flood is integer numeric, constrained by the number of days of the year. The Flood variable represents the proportion of the days of the year when a site had water at or above the surface (which is of physiological importance to plant tolerances). Rich, in contrast, is a count variable that is constrained to be non-negative.

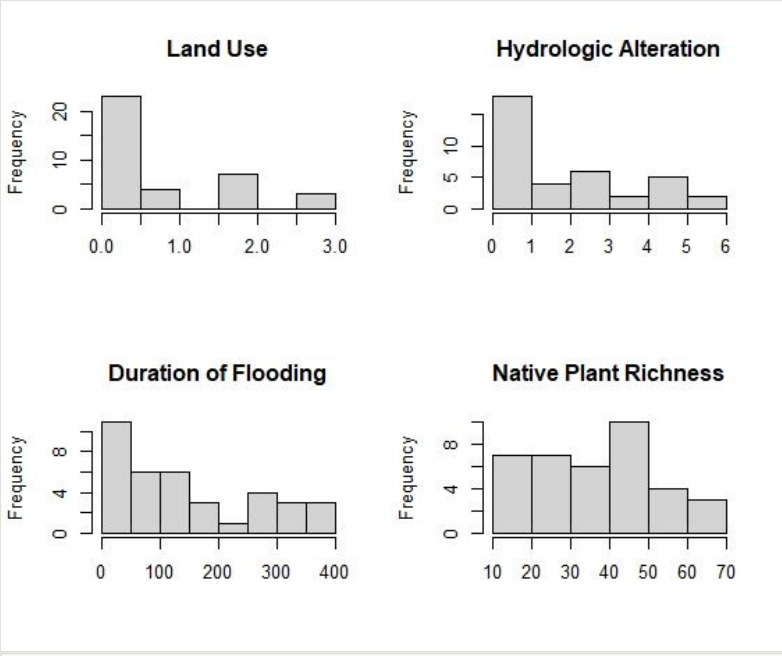

Figure 5.

Histograms for the four variables used in the illustration.

3. Deciding and Incorporating Custom Specifications

Here, we develop the overall model as a series of submodels, one submodel for each endogenous variable. The specification of submodels involves both an initial consideration of variable characteristics and an evaluation of model diagnostics. As a result, it is possible for final specifications to be of a different form from those initially considered. The R-code

for proceeding from initial evaluation to final submodel selection is given in Suppl. material 1.

Hyd – Proportion of Possible Hydrologic Alterations found at a Site: This variable is bounded by zero and the maximum number possible and can be construed to represent proportional counts. For this submodel, I utilised a GLM for Proportional Data as suggested by Zuur et al. 2009, pp 254-257. The implementation of a proportional count model in R using the 'glm' function involves handing off a matrix of two variables, the raw number of counts and the difference from the maximum score, rather than a single response variable. This can be thought of as the number of alterations present and the number of alterations absent for each wetland. The fact that the response variable is actually a matrix of two variables instead of a vector poses challenges for extraction of parameter estimates by piecewiseSEM and, thus far, this option has not been implemented in any SEM software as far as I know (I note that it is possible to approximate this model by converting the count matrix to a vector of proportions, though an error message will be returned). A binomial distribution with logit link was initially specified for the estimation process. For diagnostic evaluation, a second GLM, using a quasi-binomial distribution, was run to obtain an estimate of overdispersion. Results returned a value greater than 1.5, which suggests the quasi-binomial model is more appropriate for these data. A QQ plot comparing observed to expected residuals for the quasi-binomial model is shown in Panel A of Fig. 6.

Flood – Proportion of the Year that each Site was Flooded: This variable is of the same general form as Hyd, though with many more values. There is, however, an important difference. The Flood variable is based on a truncation of information from the underlying variable "water table level". The actual water table level is not part of the dataset analysed here and thus constitutes a "latent cause". As discussed in Grace et al. 2018, binomial data, representing truncated information, can be modelled either as binomial observations or as manifestations of latent linear propensities. In this case, I felt both viewpoints are plausible and, therefore, examined both types of models. As shown in Suppl. 1, diagnostics from using a quasi-binomial specification indicated considerable overdispersion. However, QQ plots for residuals from a quasi-binomial model (Fig. 6, Panel C) were not improved compared to the binomial model. Transforming responses to proportions and analysing with a linear model produced residual plots that appear to be improved at the extremes (Fig. 6, Panel D). The interpretational advantages of coefficients from latent linear models over those from non-linear GLMs has been extensively discussed in literature (Long 1997, Greene 2012, Fox 2016, Grace et al. 2018) and, in this case, I have selected the latent linear model results for final interpretations. Some discussion of the consequences of this choice is given in the Discussion section.

Rich – Number of Native Species found at a Site: The Poisson distribution is often idealised as the appropriate expectation for species richness values. In many real-world situations, the variable can be modelled using log-linear models, though in this case, I chose to examine both Poisson and quasi-Poisson GLMs. Diagnostics from the quasi-Poisson indicated modest levels of overdispersion (dispersion parameter value < 1.5). Additionally, a QQ plot of the residuals showed reasonably good fit (Fig. 6. Panel B). Therefore, I chose the Poisson model for interpretations.

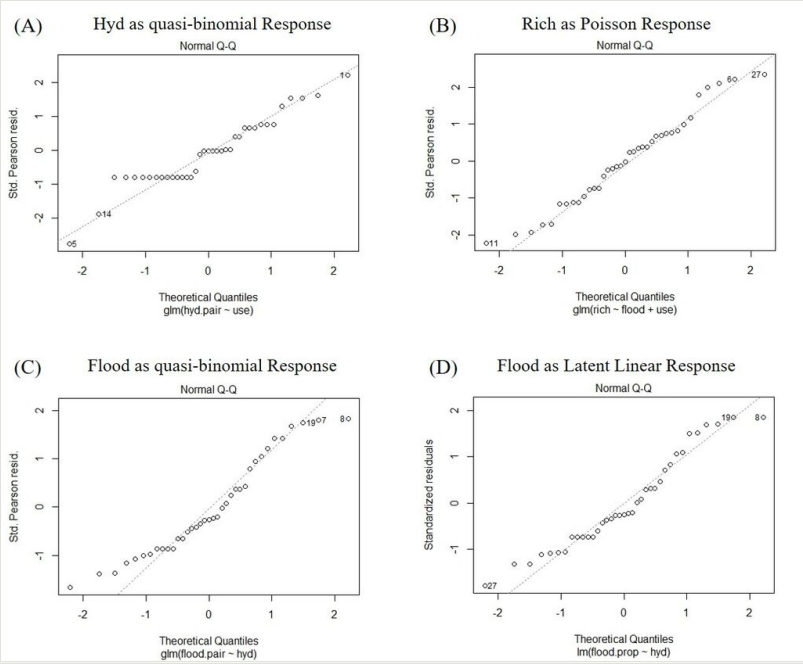

Figure 6.

QQ plots comparing residuals to model expected values.

4. Checking for Omitted Links That Should be Added

Our approach with SEM is nearly always one of: (1) first check for omitted links that should be included and (2) only after additional links are included do we consider whether any links should be removed. It is easy, in this simple example case, to recognise fairly quickly the alternative models that could be considered (see Alternative Models in Fig. 4). Things will not always be this simple and, in other situations, I would adopt a more sophisticated approach to model checking (refer to Grace 2020 for a more complex example). Here, I simply evaluated all possible predictors for the submodels.

5. Model Pruning, Model Comparisons and Model Selection

The p-values reported from the submodel analyses provided very clear support for or against alternative submodels. Direct model comparisons were conducted where such comparisons seemed necessary (Suppl. 1). The selected submodels are given in Fig. 7.

6. Development of Summary Quantities

The first step in summarising results is to assemble the submodels into the whole model (Fig. 7). In this illustration of methodology, I simply present the raw parameter estimates and standardised quantities, including both standardised parameter estimates and approximate R-squares. Standardised parameter estimates are quantities typically

returned for the investigator by SEM software. In this case, however, these need to be computed by hand.

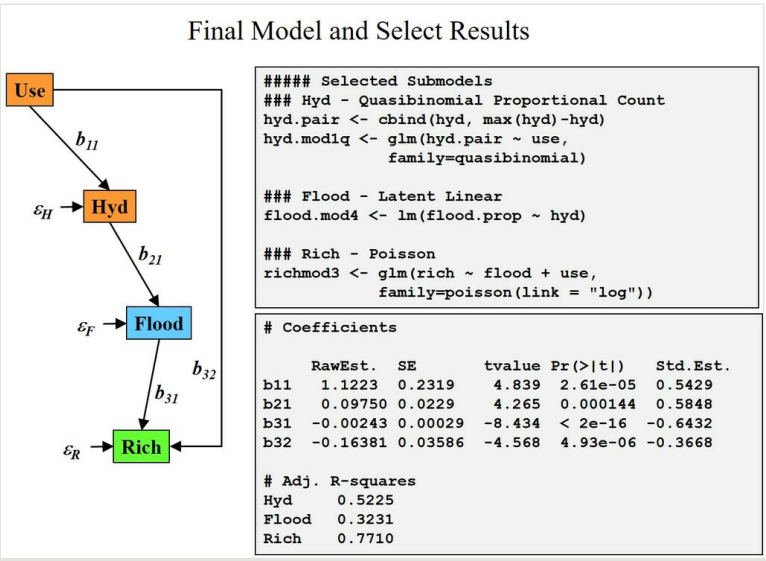

Figure 7.

Final model, R specifications used for the three subcomponent submodels, along with raw parameter estimates, standardised coefficients and adjusted R-squares.

The R code and details for how standardised coefficients were computed are presented in Suppl. 1. For Hyd, which relied on a quasi-binomial specification, I used the latent-linear standardisation method, described in Grace et al. (2018). For Flood, I used a linear model and computed standardised coefficients by hand. For Rich, I created a log-linear approximation to the Poisson model and, to illustrate a slightly simpler approach than used for Flood, extracted standardised coefficients using the QuantPsyc R package (Fletcher 2010). The “rsq” package (Zhang 2021) was used to extract R-square values for these submodels (Suppl. 1). Summary results are presented in Fig. 7.

Discussion of Results

In this ecological example, the primary hypothesis of a causal chain, whereby intensity of land use influences changes in hydrology that ultimately impact native plant richness, is supported by the results (Fig. 7). This causal chain provides an explanation for the previous observation that measures of biotic integrity are lower in wetlands whose watersheds have been substantially altered. Standardisd coefficient values indicate all of the direct effects in the model are moderately strong contributors to explaining the observed variation in the sample of wetlands studied. Support was also found for an unanticipated effect of land use on richness independent of variations in hydrology. At

present, it is not clear what this effect might represent, leaving an important mechanism for further investigation.

It is worth considering the merits and demerits associated with detailed model specifications. Generalised linear models (GLMs), as employed here for count data, rely on link functions that attempt to match the distributions of the observations. This has two proposed benefits. First, assumptions about error distributions tend to be more closely adhered to compared to linear models. This is especially true at the extremes of the distributions. Second, predicted values will fall within the observed limits of measured variables, which will be helpful when the equations are used for forecasting. What is often undiscussed in standard statistical presentations is the fact that GLMs are non-linear functions and, as a result, the coefficients returned can be quite difficult to interpret scientifically. For Poisson models, interpretational issues are minimal because the link function is simply the log of the counts (thus, coefficients represent log-linear relationships). Binomial models, however, are typically based on logit link functions, which can be challenging to interpret. Scientists are accustomed to interpreting coefficients as consistent slopes of response. However, binomial models produce coefficients that represent log odds ratios and the actual relationships between observations and predictors are non-linear outside of the middle range of values. This issue is sufficiently problematic that it is not uncommon for experienced investigators to use linear models when analysing binomial data (Mood 2010 ,Greene 2012). In structural equation modelling, we solve this problem by relying on standardised coefficients, which represent the fit between predicted and observed values regardless of any non-linearities (Grace et al. 2018). In this example, diagnostics showed better fit of the flood frequency data to a linear model than to a binomial model, presumably because the flood frequency data are driven by an underlying continuous variable, water level table. Thus, I chose the model with better model-data fit, which was the linear model. Comparisons between the linear and binomial model (at the end of Suppl. 1) showed nearly identical R-squares. However, standardised coefficients, extracted from the linear model, are more consistent with the R-square, further supporting the linear model as a suitable approximation in this case. The take-home message from these findings is that linear models are not necessarily inferior to GLMs in all situations. Further discussion of this issue can be found in Mood 2010).

Conclusion

Ecologists are increasingly adopting powerful and sophisticated regression techniques for their analyses. By implementing SEM as a network of regressions, our modelling options are greatly expanded. What is often not returned by existing regression packages are the summary quantities we might need to interpret networks of relationships. For example, aside from generating comparable standardised coefficients, which are vital for interpreting SE models, we might also wish to compute indirect and total effects by multiplying path coefficients, which is easily accomplished for custom applications. This paper provides a general demonstration for how to develop custom-built SE models.

It is important to note that, while custom-build modelling allows for a greater variety of statistical specifications, the classical approach of modelling covariance relationships (e.g. using 'lavaan') has its own unique strengths. Covariance modelling permits the inclusion of latent variables, error covariances and reciprocal interactions, all of which have special uses in SEM. Thus, custom-built SE models should be seen as an additional tool, but not a replacement for other existing methods.

Acknowledgements

I thank Darren Johnson of the US Geological Survey, Alessandro Gimona of The James Hutton Institute, Frank Pennekamp of the University of Zurich, Jackie Potts of The James Hutton Institute and an anonymous reviewer for helpful comments and suggestions. This work was supported by the USGS Ecosystems and Land Change Science Climate Research and Development Programs. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Hosting institution

US Geological Survey, Wetland and Aquatic Research Center, 700 Cajundome Blvd, Lafayette, Louisiana, USA

Conflicts of interest

The author declares no conflicts of interest related to this publication.

References

- Fletcher TD (2010) Package QuantPsyc. R Package URL: https://cran.r-project.org/web/packages/QuantPsyc/QuantPsyc.pdf/
- Fox J (2016) Applied regression analysis and generalized linear models. Third. Sage Publications, Los Angeles.
- Grace JB (2006) Structural equation modeling and natural systems. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511617799
- Grace JB, Schoolmaster DR, Guntenspergen GR, Little AM, Mitchell BR, Miller KM, Schweiger EW (2012) Guidelines for a graphtheoretic implementation of structural equation modeling. Ecosphere 3: 1-44. https://doi.org/10.1890/ES12-00048.1
- Grace JB, Schoolmaster DR, Guntenspergen GR, Little AM, Mitchell BR, Miller KM, Schweiger EW (2016) Data from Guidelines for a graph-theoretic implementation of structural equation modeling,. URL: https://doi.org/10.6084/m9.figshare.c.3308274.v1
- Grace JB, Johnson DJ, Lefcheck JS, Byrnes JE (2018) Quantifying relative importance: computing standardized effects in models with binary outcomes. Ecosphere e02283 https://doi.org/10.1002/ecs2.2283

- Grace JB (2020) A 'Weight of Evidence' approach to evaluating structural equation models. One Ecosystem 5 URL: https://doi.org/10.3897/oneeco.5.e50452
- Grace JB, Irvine KM (2020) Scientists guide to developing explanatory statistical models using causal analysis principles. Ecology e02962 URL: https://doi.org/10.1002/ecy.2962
- Grace JB (2021) Instrumental variable methods in structural equation models. Methods in Ecology and Evolution 11: 1148-1157. https://doi.org/10.1111/2041-210X.13600
- Greene WH (2012) Econometric analysis. Pearson Education, New York,.
- Karr JR (1991) Biological integrity: a long-neglected aspect of water resource management. Ecological Applications 1: 66-84. https://doi.org/10.2307/1941848
- Lee SY (2007) Structural equation modeling: a Bayesian approach. John Wiley & Sons, New York, NY, USA. https://doi.org/10.1002/9780470024737
- Lefcheck JS, Byrnes JE, Grace JB (2018) piecewiseSEM: Piecewise structural equation modeling. An R package version 2.1 URL: https://cran.r-project.org/web/packages/piecewiseSEM/piecewiseSEM.pdf
- Long JS (1997) Regression models for categorical and limited dependent variables. Sage Publications, Los Angeles.
- Mood C (2010) Logistic regression: why we cannot do what we think we can do, and what we can do about it. European Sociological Review 26: 67-82. https://doi.org/10.1093/esr/jcp006
- Rosseel Y (2012) Lavaan: An R package for structural equation modeling and more. Version 0.512 (BETA). Journal of Statistical Software 48: 1-36.
- Shipley B (2000) Cause and correlation in biology. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511605949
- Shipley B (2016) Cause and correlation in biology. Second Edition. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139979573
- Zhang D (2021) rsq: R-Squared and Related Measures. R package version 2.2. URL: https://CRAN.R-project.org/package=rsq
- Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009) Mixed Effects Models and Extensions in Ecology with R. Springer Publishing, New York. https://doi.org/10.1007/978-0-387-87458-6

Supplementary material

Suppl. material 1: General guidance for custom-built structural equation models doi

Authors: Grace, J.B.
Data type: R code
Brief description: This file contains the R code used to develop the demonstrations included in Grace JB (2021) General guidance for incorporating custom specifications in structural equation models. One Ecosystem.
Download file (8.99 kb)