Biodiversity and Conservation **11:** 1397–1401, 2002.

Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables

RALPH MAC NALLY
Section of Ecology, School of Biological Sciences, Monash University, Victoria 3800, Australia (e-mail: dacelo@silas.cc.monash.edu.au; fax: +61-3-99055613)

Received 15 May 2001; accepted in revised form 6 August 2001

Key words: Correlated variables, General linear model, Prediction, Variable subsets

Abstract. Ecologists and conservation biologists frequently use multiple regression (MR) to try to identify factors influencing response variables such as species richness or occurrence. Many frequently used regression methods may generate spurious results due to multicollinearity. Mac Nally (2000, Biodiversity and Conservation 9: 655–671) argued that there are actually two kinds of MR modelling: (1) seeking the best predictive model; and (2) isolating amounts of variance attributable to each predictor variable. The former has attracted most attention with a plethora of criteria (measures of model fit penalized for model complexity – number of parameters) and Bayes-factor-based methods having been proposed, while the latter has been little considered, although hierarchical methods seem promising (e.g. hierarchical partitioning). If the two approaches agree on which predictor variables to retain, then it is more likely that meaningful predictor variables (of those considered) have been found. There has been a problem in that, while hierarchical partitioning allowed the ranking of predictor variables by amounts of independent explanatory power, there was no (statistical) way to decide which variables to retain. A solution using randomization of the data matrix coupled with hierarchical partitioning is presented, as is an ecological example.

Introduction

Ecologists and conservation biologists rely heavily on multiple regression (MR) to develop inferences about the determinants of patterns affecting species distributions or numbers (Mac Nally 2000). MR typically is used in conservation ecology to model the occurrence or density of a species (response variables) as a function of landscape and habitat-patch-specific variables (predictor variables) (e.g. Loyn 1987). Experimentation usually is inappropriate, unethical or intractable, so MR (including multiple logistic and multiple Poisson regression) is used to derive inferences about which predictors are important. Unfortunately, these predictor variables are frequently significantly intercorrelated (multicollinearity) so that identifying the likely causal variables is problematic.

Mac Nally (2000) reviewed some of the main methods used by ecologists for sifting through data sets with many potential predictor variables, highlighting weaknesses of many commonly used methods (e.g. all stepwise-selection techniques). He emphasized the distinction between finding the best model to describe the data and drawing inferences about the likely causality of variables, which is often of more interest in ecological management. The former involves the explicit development of predictive functions while the latter need not produce a computed function. Criteria (functions of model fit penalized for model complexity) and Bayes-factor-based methods are widely advocated for finding a single best predictive function [e.g. AIC, AIC_c, QAIC, BIC; see Burnham and Anderson (1998) for an extensive review], with methods explicitly addressing the role of model misspecification in inference also offering promise (e.g. Buckland et al. 1997). Nevertheless, these approaches still focus on finding the single best model. However, in many ecological and conservation applications, managers would have little confidence in the values of regression coefficients associated with terms in the selected model. They are likely to be happier with information on the directions and relative magnitudes of change in the response variables associated with management scenarios based on manipulations of the most likely causal variables.

Hierarchical partitioning

Seeking a single model is not the most effective way of identifying those variables most likely to influence variation in the response variable. Multicollinearity is difficult to deal with in one-model approaches. Predictor U may be included in the best model because it contributes to the overall best fit, but U may have no influence on the response variable. In the specific data set U may happen to be correlated with the true causal variables, V and W say, in such a way as to pick up the explanatory power of both V and W, which then are left out of the model.

A possible solution to this dilemma, hierarchical partitioning (HP; Chevan and Sutherland 1991), considers all models in an MR setting jointly to identify the *most likely* causal factors. In the illustration above, the increase in model fit generated by U is estimated by averaging its influence over all models in which U appears (i.e. U, UV, UW, UVW). This averaging is likely to alleviate multicollinearity problems that are effectively ignored by using *any* single-model-seeking technique. The hierarchical organization in exhaustive regression-model building arises because of the relationship of simpler models to more complex ones, of which the former are subsets.

HP employs goodness-of-fit measures for each of the 2^K possible models for K predictor variables (e.g. χ^2 in log-linear models, R^2 in MR, etc.). These measures are partitioned so that the total *independent* contribution of a given predictor variable is estimated. HP allows identification of variables whose independent, as distinct from partial, correlation with a response variable may be important from variables that have little independent effect. Thus, HP involves calculation of incremental improvement (i.e. increased goodness of fit) in models by the addition of a given variable (U, say), and averages these over all combinations in which U occurs to provide a measure of the effects of predictors (Christensen 1992). The independent impact of variable U is estimated by comparing goodnesses of fit for all possible models involving U. What is the average effect of including U in all first-order (i.e. just V, W, ...), second-order (VW, VZ, ...) and higher-order (VWZ, VWT, ..., $VWZT$, ...) models? For each predictor variable, explanatory power is ultimately segregated into independent effects, I, and effects that cannot be unambiguously associated with that single variable but are due to joint effects with other variables, J. Note that HP does not lead to the development of a predictive function *per se* because this is not its purpose.

Hierarchical partitioning and statistical significance

How does one use results of HP to determine which variables to retain? The output of an HP analysis is just a list of predictor variables and their independent (I) and joint (J) influences on the response variable Y. In many ecological and in most conservation problems, one wishes to identify those predictor variables that have the most independent impact on Y. The outcome of an HP analysis is a list of variables that can be ranked by their independent contributions but where there is little guidance as to which variables to retain for the purpose of making management decisions (Mac Nally 1996).

Randomization approaches have become popular as ecologists deal with data that are not easily treated by other means (Clarke 1993; Manly 1997; Anderson 2001). The same approach seems to be useable in the current problem: randomize the data matrix many times (e.g. 1000 from a possible $[N!]^K$ combinations) and compute the distribution of Is for each predictor variable. If the observed value I_{obs} is extreme (>95 percentile) relative to the generated distribution, then that predictor variable is worth retaining as a potentially important one for management purposes. Results of HP analyses for each variable can be expressed as Z-scores ([observed – mean {randomizations}]/SD{randomizations}), and the statistical significance based on upper 95% confidence limit ($Z \geq 1.65$).

An example

We collected historical and contemporary data on the characteristics of fragments of forest in central Victoria, Australia [see Mac Nally et al. (2000) for descriptions]. These predictor variables include ln(area 1996), absolute area change from 1963 to 1996, relative area change 1963–1996, land tenure (public/private), grazing pressure, shrub regeneration, distance to closest extensive existing forest blocks, distance to closest other fragment, area of nearest fragment, connectivity, and three compound variables describing habitat structure within each fragment (MDS1, MDS2, MDS3).

The response variable was species richness – deviation from expected, which represents the species richness of a fragment of a given area (i.e. 10–80 ha) minus the mean richness of a reference area of the same size but set within continuous forest of the same kind [see Bolger et al. (1991) for a rationale]. Thus, if the average richness in 10 ha reference areas was $N = 20.4$ species, then the value for a 10 ha fragment with 16 species was -4.4.

Two separate analyses were conducted, one with absolute area change and the other with relative area change, since these cannot both be incorporated into the one analysis. By using the $Z \geq 1.65$ criterion, current area is the most significant factor controlling species richness in these fragments (Figure 1). Z-scores for the independent contributions for this variable were >5.6 ($P < 10^{-4}$) whether absolute or relative area changes are used. The independent influence of only one other variable bordered on significance, the second compound variable MDS2, but only when absolute area change was a variable ($Z \approx 1.67$; $Z \approx 1.0$ for relative area). This sensitivity to which version of area change is included, and the marginal significance, suggest that MDS2 should not be regarded as a probable causal factor in controlling species richness in this system.

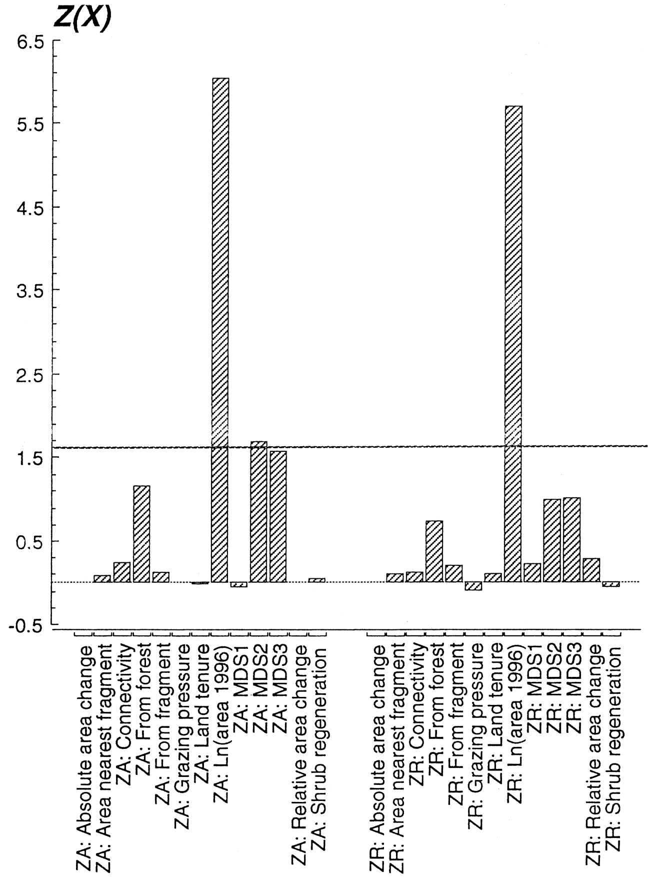

Figure 1. Plot of Z-scores for independent contributions, I, from randomizations of data matrices for potential explanatory predictor variables for two sets of analyses: one including absolute area changes (ZA) and the other relative area changes (ZR). The dashed horizontal line represents upper 95% confidence value for the Z-scores.

Conclusions

Given the ongoing reliance on MR analyses in much of conservation and ecology, use of robust methods for inference and model construction clearly is important. While criterion-like approaches are useful for locating the best single functional model, hierarchical partitioning offers the great advantage of considering the whole web of relationships between predictor variables as an ensemble. The previous drawback of having few grounds for discerning which of the predictor variables to retain in HP seems to be relatively easily solved by using the randomization approach related here.

Acknowledgements

The author acknowledges support from the Australian Research Council (grant no. F19804210).

References

Anderson M.J. 2001. A new method for non-parametric multivariate analysis of variance. Australian Ecology 26: 32–46.

Bolger D.T., Alberts A.C. and Soulé M.E. 1991. Occurrence patterns of bird species in habitat fragments: sampling, extinction, and nested species subsets. American Naturalist 137: 155–166.

Buckland S.T., Burnham K.P. and Augustin N.H. 1997. Model selection: an integral part of inference. Biometrics 53: 603–618.

Burnham K.P. and Anderson D.R. 1998. Model Selection and Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York.

Chevan A. and Sutherland M. 1991. Hierarchical partitioning. The American Statistician 45: 90–96.

Christensen R. 1992. Comment on Chevan and Sutherland. The American Statistician 46: 74.

Clarke K.R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 18: 117–143.

Loyn R.H. 1987. Effects of patch area and habitat on bird abundances, species numbers and tree health in fragmented Victorian forests. In: Saunders D.A., Arnold G.W., Burbidge A.A. and Hopkins A.J.M. (eds), Nature Conservation: the Role of Remnants of Native Vegetation. Surrey Beatty and Sons, Chipping Norton, Australia, pp. 65–77.

Mac Nally R. 1996. Hierarchical partitioning as an interpretative tool in multivariate inference. Australian Journal of Ecology 21: 224–228.

Mac Nally R. 2000. Regression and model-building in conservation biology, biogeography and ecology: the distinction between – and reconciliation of – 'predictive' and 'explanatory' models. Biodiversity and Conservation 9: 655–671.

Mac Nally R., Bennett A.F. and Horrocks G. 2000. Forecasting the impacts of habitat fragmentation. Evaluation of species-specific predictions of the impact of habitat fragmentation on birds in the box-ironbark forests of central Victoria, Australia. Biological Conservation 95: 7–29.

Manly B.F.J. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall, London.