vol. 173, no. 1 the american naturalist january 2009

Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution

Roger Mundry[1,*] **and Charles L. Nunn**[1,2,3,†]

1. Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany;
2. Department of Integrative Biology, University of California, Berkeley, California 94720;
3. Department of Anthropology, Harvard University, Cambridge, Massachusetts 02138

Submitted March 4, 2008; Accepted July 9, 2008;
Electronically published December 2, 2008

abstract: Statistical inference based on stepwise model selection is applied regularly in ecological, evolutionary, and behavioral research. In addition to fundamental shortcomings with regard to finding the "best" model, stepwise procedures are known to suffer from a multiple-testing problem, yet the method is still widely used. As an illustration of this problem, we present results of a simulation study of artificial data sets of uncorrelated variables, with two to 10 predictor variables and one dependent variable. We then compared results from stepwise regression with a regression model in which all predictor variables were entered simultaneously. These analyses clearly demonstrate that significance tests based on stepwise procedures lead to greatly inflated Type I error rates (i.e., the probability of erroneously rejecting a true null hypothesis). By using a simple simulation design, our study amplifies previous warnings about using stepwise procedures, and we follow others in recommending that biologists refrain from applying these methods.

Keywords: stepwise modeling, statistical inference, multiple testing, error rate, multiple regression, GLM.

Introduction

Ecological, evolutionary, and behavioral research commonly involve multivariate tests in which the investigator examines which of several predictor variables influence a single response variable. Most commonly, such analyses are conducted using a generalized linear model (GLM). GLMs can be used for the analysis of data sets encompassing any combination of continuous and categorical predictor variables and for continuous and discrete response variables, provided that the distribution of the residuals fulfills certain assumptions. Well-known examples of GLMs are multiple (linear, logistic, or Poisson) regressions, multiway ANOVA, and ANCOVA (see, e.g., Dobson 2002). A frequently applied extension of the GLM is the generalized linear mixed model (GLMM), which allows users to control for random effects factors, such as individual subjects (Faraway 2006).

Recently, the question of how to draw statistical inference from such models has caused considerable debate. Information-criterion-based multimodel inference (Burnham and Anderson 2002) has been advocated strongly in ecological and evolutionary research (e.g., Johnson and Omland 2004; Lukacs et al. 2007; Stephens et al. 2007). On the other hand, the classical statistical approach of null hypothesis statistical testing (NHST) is still commonly used in many research fields (including ecology and evolution) and presumably will remain so for a considerable time in the future (e.g., Stephens et al. 2005; Steidl 2006; Sleep et al. 2007). Our goal here is to point to a special problem that arises when stepwise procedures are applied in combination with NHST.

When using GLMs, two fundamentally different approaches are available to investigate the effect of predictor variables on the response variable: variables can be entered simultaneously into the model, or they can be entered sequentially. When predictor variables are entered simultaneously (also referred to as the "forced entry" or "all-variables-together" method), all predictor variables are entered at the same time into the (full) model. Their joint contribution in explaining the response variable is subsequently determined and summarized in a single global significance test of the full model. When predictor variables are investigated sequentially (also referred to as "stepwise"), variables are sequentially entered into and/or removed from the model. When variables are sequentially entered into the model ("forward selection"), the initial model comprises only a constant, and at each subsequent step the variable that leads to the greatest (and significant)

* E-mail: roger_mundry@eva.mpg.de.

† E-mail: cnunn@oeb.harvard.edu.

Am. Nat. 2009. Vol. 173, pp. 000–000.
0003-0147/2009/17301-50288$15.00.
DOI: 10.1086/593303

improvement in fit is added to the statistical model. In "backward deletion," the initial model is the full model including all variables, and at each step a variable is excluded when its exclusion leads to the smallest (nonsignificant) decrease in model fit. A "combination" approach is also possible, which begins with forward selection, but after the inclusion of the second variable it tests at each step whether a variable already included can be dropped from the model without a significant decrease in model fit. The final model of each of these stepwise procedures is supposed to comprise that (sub-)set of the predictor variables that have an effect on the response variable and that best explain the response (Sokal and Rohlf 1995; Zar 1999; Tabachnick and Fidell 2001; Quinn and Keough 2002; Field 2005; note that different terms have been used to denote the stepwise procedures by different authors, and selection criteria other than $P \leq .05$ have been suggested).

The application of stepwise procedures has been criticized on multiple grounds (for a review, see Whittingham et al. 2006). In fact, stepwise methods frequently fail to include all variables that have an actual influence on the dependent variable, while frequently also including other variables that do not influence the dependent variable (Derksen and Keselman 1992). Consequently, the final model is not generally the best model (Miller 1984). In addition, stepwise procedures tend to be unstable, meaning that only slight changes in the data can lead to different results as to which variables are included in the final model and the sequence in which they are entered (James and McCulloch 1990). As a consequence, stepwise methods also fail to provide a valid means for ranking the relative importance of the predictor variables.

Here we focus on an additional serious drawback of stepwise methods that occurs when they are used in conjunction with significance testing. Specifically, stepwise procedures can produce vastly elevated Type I error rates, that is, the inference of a significant result when in fact none exists (false positives). Indeed, a considerable number of articles and statistical textbooks clearly state that stepwise procedures represent a case of multiple testing without error-level adjustment, thus making the approach invalid (i.e., too liberal) in the context of statistical null hypothesis testing (e.g., Pope and Webster 1972; Wilkinson 1979; Cohen and Cohen 1983; Lovell 1983; Tabachnick and Fidell 2001; Quinn and Keough 2002; Whittingham et al. 2006). For instance, a stepwise forward selection conducted on a data set with 10 predictor variables conducts 10 significance tests in the first step, nine significance tests in the second step, and so on, and each time includes a variable in the model when it reaches a specified criterion (conventionally, the significance level, which is set at 5%, but see below). Conducting a number of significance tests without an error-level adjustment, however, considerably increases the probability of rejecting at least one of them by chance, that is, even in the complete absence of any influence of the predictor variables on the response (a Type I error). Hence, statistical inference in the classical sense—in which the user attempts to reject the null hypothesis about a set of predictor variables at a prespecified error level—is not possible when using stepwise procedures. Several methods have been proposed to overcome this issue (e.g., Pope and Webster 1972; Wilkinson 1979). However, none of these has been applied regularly in ecological or evolutionary research.

Despite the warnings about stepwise procedures, statistical inference based on them is commonly used, obviously because many authors are not aware of the serious drawbacks of doing so. A recent study of three top behavioral and ecological journals published since 2004, for example, found that 57% of the publications in which a multiple regression was feasible used some form of stepwise regression (Whittingham et al. 2006). And, in quick a survey of the issues of the *American Naturalist* from 2007 (vols. 169 and 170), we identified 10–12 articles in which at least one significance test was based on a stepwise procedure.

To bring more attention to this overlooked but serious issue, we systematically investigated the Type I error rates resulting from different stepwise methods, using multiple linear regression as an example. We did this by applying a simulation approach and comparing results from stepwise regression with a regression model in which all predictor variables were entered simultaneously. We systematically varied the number of predictor variables from two to 10, but our simulation did not include any effects of predictors on the response variable (for details of the simulation, see appendix). Thus, we tested data sets for which the null hypothesis is, by definition, true.

Results

When applying stepwise multiple regression, the proportion of erroneously significant results was above chance expectation for all stepwise procedures and for each number of predictor variables (fig. 1). This contrasts markedly with the forced entry multiple regression, in which the number of significant models never exceeded chance level (fig. 1). The probability of getting a significant result when using a stepwise procedure clearly increased with the number of predictor variables included. Remarkably, the error rate reached almost 40% when the data comprised 10 predictor variables. Thus, in case of the null hypothesis being true, an investigator would have an approximately 40% chance of incorrectly identifying the set of predictor variables as having a statistically significant effect when using 10 predictor variables. Even with just two predictors, the

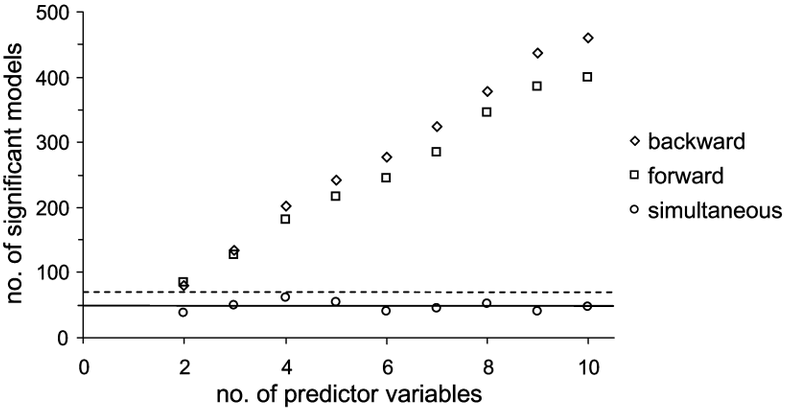

Figure 1: Numbers of significant multiple regressions (out of 1,000) based on random data for which the null hypothesis is, by definition, true. The plot shows results for two different methods of stepwise regression (backward deletion and forward selection, respectively) and for simultaneous entry of different numbers of predictor variables (*simultaneous*). Results for the combination approach matched those for forward selection. Symbols above the dashed horizontal line represent proportions significantly in excess of chance expectation (50, *solid line*; binomial test). Note that for both stepwise procedures, the probability of getting a significant result was above the desired 5% for all numbers of predictor variables.

probability of incorrectly rejecting the null hypothesis was significantly above chance level. Differences between forward selection and backward deletion were small; when using backward deletion, however, the probability of getting an erroneously significant finding was slightly higher (fig. 1; Wilcoxon test: $T^+ = 43.5$, $N = 9$, $P = .012$).

Discussion

Our results clearly demonstrate that using stepwise procedures rather than simultaneous entry of predictor variables greatly inflates the probability of incorrectly rejecting the null hypothesis of no effect (i.e., Type I error rate). Specifically, the probability of making a Type I error was almost doubled when using two predictor variables rather than one and dramatically increased with increases in the number of predictor variables. This was the case for both forward selection and backward deletion, although in the latter procedure the effect was slightly more pronounced. As a result, significance tests based on stepwise procedures are invalid statistically, and the degree of their invalidity increases with increasing number of predictor variables. Nevertheless, statistical tests based on stepwise regression are commonly seen in the ecological, behavioral, and evolutionary literature (see "Introduction" and Whittingham et al. 2006). On the basis of our findings, it seems likely that some of the published findings based on stepwise methods represent Type I errors.

Why do stepwise procedures produce elevated Type I error rates, while forced entry regression methods produce expected error rates? From our findings, it seems that the inflated Type I error rate is largely due to multiple testing. In fact, the Type I error rates we found very closely followed what is theoretically expected based on multiple testing, where the probability of at least one Type I error in a number of tests of true null hypotheses equals $1 - (1 - \alpha)^n$ (with α being the error probability, i.e., 0.05, and n being the number of tests). This finding also clearly implies that what we found is not specific to multiple regression. Instead, we are convinced that it applies to stepwise methods in general, used in conjunction with any GLM, GLMM, or discriminant function analysis.

Based on our findings, we recommend that stepwise procedures should not be used in the context of testing null hypotheses about a set of predictor variables. In fact, the only valid options for combining stepwise procedures with statistical inference based on significance testing would be to adjust error levels for the number of variables considered at each step or to use adjusted sampling distributions of test statistics (e.g., Pope and Webster 1972; Wilkinson 1979). However, using error-level adjustment would come with its well-known cost, which is greatly reduced power, that is, the probability of correctly rejecting a false null hypothesis (e.g., Moran 2003; Nakagawa 2004), and neither of the two options is implemented in statistical standard software. Hence, in the context of NHST, one should use an overall and simultaneous test of the statistical significance of all the predictor variables together (the full model). Selecting a subset of predictor variables explaining the response variable most parsimoniously could and should be done only after the initial full model revealed significance. It has been frequently pointed out,

however, that stepwise procedures also have serious drawbacks in this context. For instance, they do not necessarily find the best model, they may be unstable in the sense that only slight changes in the data lead to great changes in the variables included in the final model, and they do not necessarily allow for a valid ranking of variables by their importance (see "Introduction"). Hence, we recommend the use of information-theory-based model selection procedures for this purpose (e.g., Burnham and Anderson 2002; Whittingham et al. 2006).

It is worth noting that we do not present anything new here but just confirm what has already been stated in statistical articles and texts for more than two decades (e.g., Wilkinson 1979; Cohen and Cohen 1983; Lovell 1983; Derksen and Keselman 1992; Tabachnick and Fidell 2001; Quinn and Keough 2002; Whittingham et al. 2006). In fact, even in the 1970s, Pope and Webster (1972, p. 328) complained about "the widespread use of stepwise procedures and the lack of understanding (by nonstatisticians) of their weaknesses." Despite these many warnings, stepwise procedures remain in widespread use, perhaps because suitable examples have not been provided to convince ecologists and evolutionary biologists of the dangers of using stepwise methods. Our simulations of variables under a null hypothesis provide solid evidence for the highly elevated Type I errors associated with stepwise methods, and hopefully will convince others that these statistical procedures are statistically flawed for null hypothesis testing purposes.

Acknowledgments

We thank an anonymous reviewer and P. A. Stephens for very helpful comments on an earlier version of this article. This research was supported by the Max Planck Society.

APPENDIX

Methods

In all simulated data sets, the predictor variables and the response variable comprised pseudorandom numbers that were drawn from a uniform distribution with the range $0 \leq x < 10$. The number of predictor variables ranged from 2 to 10 (increment 1). For each number of predictor variables we generated one set of predictor variables and 1,000 response variables. Multiple regression requires a large number of cases compared to the number of predictor variables (Tabachnick and Fidell 2001; Quinn and Keough 2002; Field 2005). In addition, the validity and stability of the result of a multiple regression depends on the relation between the number of predictor variables (k) and the number of cases (i.e., data points, N). Hence, Field (2005) recommends that when the significance of the overall model should be tested, the minimum sample size should be $N = (50 + 8 \times k)$. Accordingly, to achieve comparable power in analyses of data sets with different numbers of predictor variables, we set the number of data points to be a function of the number of predictor variables, with $N = 3 \times (50 + 8 \times k)$. Pseudorandom numbers were generated using the function "rnd()" implemented in Visual Basic for Applications (Excel, ver. 2002, SP3).

Each of the simulated data sets was analyzed using the four different multiple regression methods described above: the simultaneous approach using forced entry and three stepwise approaches involving forward selection, backward deletion, and the combination approach. However, because forward selection and the combination approach always produced identical results with regard to significance of the final model and the number of predictor variables included, we present results only for the forward selection and backward deletion procedures. All statistics were calculated using SPSS 13.0.1 for Windows. In stepwise regression analyses, we used the default criteria for entering and removing variables (i.e., the P value associated with an F-test, with $P_{\text{entry}} = .05$ and $P_{\text{removal}} = .1$). For each number of predictor variables, we determined the number of data sets for which the final model was significant ($P \leq .05$). Given that, for all generated data sets, the null hypothesis was, by definition, true, we expected the proportion of significant findings to approximately equal 5%. We tested whether the number of significant results corresponded to chance expectation using a one-tailed binomial test. We used exact nonparametric tests when small samples required their use (Mundry and Fischer 1998).

Literature Cited

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference. 2nd ed. Springer, Berlin.

Cohen, J., and P. Cohen. 1983. Applied multiple regression/correlation analysis for the behavioral sciences. 2nd ed. Erlbaum, Hillsdale, NJ.

Derksen, S., and H. J. Keselman. 1992. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology 45:265–282.

Dobson, A. J. 2002. An introduction to generalized linear models. Chapman & Hall/CRC, Boca Raton, FL.

Faraway, J. J. 2006. Extending linear models with R. Chapman & Hall/CRC, Boca Raton, FL.

Field, A. 2005. Discovering statistics using SPSS. Sage, London.

James, F. C., and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annual Review of Ecology, Evolution, and Systematics 21:129–166.

Johnson, J. B., and K. S. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology & Evolution 19:101–108.

Lovell, M. C. 1983. Data mining. Review of Economics and Statistics 65:1–12.

Lukacs, P. M., W. L. Thompson, W. L. Kendall, W. R. Gould, P. F. Doherty Jr., K. P. Burnham, and D. R. Anderson. 2007. Concerns regarding a call for pluralism of information theory and hypothesis testing. Journal of Applied Ecology 44:456–460.

Miller, A. J. 1984. Selection of subsets of regression variables. Journal of the Royal Statistical Society A 147:389–425.

Moran, M. D. 2003. Arguments for rejecting the sequential Bonferroni in ecological studies. Oikos 100:403–405.

Mundry, R., and J. Fischer. 1998. Use of statistical programs for nonparametric tests of small samples often leads to incorrect *P*-values: examples from animal behaviour. Animal Behaviour 56: 256–259.

Nakagawa, S. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behavioral Ecology 15:1044–1045.

Pope, P. T., and J. T. Webster. 1972. The use of an *F*-statistic in stepwise regression procedures. Technometrics 14:327–340.

Quinn, G. P., and M. J. Keough. 2002. Experimental designs and data analysis for biologists. Cambridge University Press, Cambridge.

Sleep, D. J. H., M. C. Drever, and T. D. Nudds. 2007. Statistical versus biological hypothesis testing: response to Steidl. Journal of Wildlife Management 71:2120–2121.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. 3rd ed. Freeman, New York.

Steidl, R. J. 2006. Model selection, hypothesis testing, and risks of condemning analytical tools. Journal of Wildlife Management 70: 1497–1498.

Stephens, P. A., S. W. Buskirk, G. D. Hayward, and C. M. del Rio. 2005. Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology 42:4–12.

Stephens, P. A., S. W. Buskirk, and C. M. del Rio. 2007. Inference in ecology and evolution. Trends in Ecology & Evolution 22:192–197.

Tabachnick, B. G., and L. S. Fidell. 2001. Using multivariate statistics. 4th ed. Allyn & Bacon, Boston.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology 75:1182–1189.

Wilkinson, L. 1979. Tests of significance in stepwise regression. Psychological Bulletin 86:168–174.

Zar, J. H. 1999. Biostatistical analysis. 4th ed. Prentice Hall, Upper Saddle River, NJ.

Associate Editor: Emília P. Martins
Editor: Michael C. Whitlock