Ann. Zool. Fennici 32: 439–440 ISSN 0003-455X
Helsinki 8 December 1995

Commentary

Statistical inference of island occupancy: a reply

Hannu Rita & Esa Ranta

Rita, H., Department of Forest Resource Management, P.O. Box 24 (Unioninkatu 40B), FIN-00014 University of Helsinki, Finland
Ranta, E., Integrative Ecology Unit, Department of Ecology and Systematics, Division of Population Biology, P.O. Box 17 (P. Rautatiekatu 13), FIN-00014 University of Helsinki, Finland

Received and accepted 7 July 1994

Lomolino, Brown and Davis (1995; from now on LBD) have reacted to our criticism concerning the use of ordinary regression technique when analysing species' incidence data (Rita & Ranta 1993). We advocated the use of logistic regression models when analysing response variables with two possible values, 0 or 1, like absence and presence of species in distributional data. We shall not repeat here our criticism of using ordinary regression with such data. Instead, we shall respond to the arguments raised by LBD to further explain why to use logistic regression technique with binary response variables.

To begin with, however, we shall admit that we were not aware of Adler & Wilson (1985). For anybody, who takes the trouble to read this paper, it should become evident that our non-awareness is only of poor coverage of literature. Almost in all detail we agree with the arguments of Adler and Wilson supporting the use of logistic models. This goes to the extent that to some degree we even repeat them in our critique. It seems to us, that this is in contrast to what Lomolino et al. (1986, 1995) propose. Strange enough, Adler & Wilson (1985) used logistic regression — with good reasons — already four years prior to appearance of Lomolino et al. (1989).

LBD argue against the use of logistic models by saying that the sample sizes are not large enough. Doing this, they tend to forget that even behind ordinary regression models there are many assumptions (see, however Lomolino et al. [1989] who wrote "... each statistical approach has its assumptions"). For example, r^2-values are used, and behind them is a specific way to measure the amount of variation, assuming *inter alia* homogeneity of variances, which is not the case with binary data. In Lomolino et al. (1989) these arguments were never met. We can also ask: were the sample sizes used by Lomolino et al. (1989) large enough to fade away the discrete nature of the binary response and make it behave like observations from normal distribution and thus to justify the use of ordinary regression models?

LBD also comment "... the fit of their model seems poor (based on high p-values in their Table 1)". As this is a point generally misunderstood, we shall dwell a bit to clarify it. The usual test of goodness-of-fit is based on likelihood ratio. The asymptotic distribution of this tests statistic is χ^2 with degrees of freedom, ν, determined by the number of model parameters. Thus, the expected value of the test statistic is equal to ν. Statistically significant results of this test indicate lack-of-fit. It seems to us, that LBD, and many others, do not know how to interpretate the results of model goodness-of-fit tests. If the test gives a statistically significant result, it means that the amount of residual variation of the model is larger than is expected in the model under test. Thus, large *p* values are an argument for a good model fit, not for a poor fit, all this conditional to sample size, of course. The Hosmer-Lemeshow goodness-of-fit tests were used, because the asymptotic results concerning the distribution of the test statistic are in doubt due to the binary nature of the data. On this topic, we refer to Hosmer & Lemeshow (1989), frequently also quoted by LBD. As a cautionary note, significant results of the *F*-test in ordinary regression do not indicate good model fit, but only that the current model fits statistically significantly better than the constant-only model.

LBD argue that we should pay "... more attention to the biological significance of the results". Our paper was primarily not intended to biological results, it was aiming to criticise the use of ordinary regression technique when analysing a binary response variable. However, in the tool we are favoring — logistic regression technique — is included a parameter, odds ratio, through which also "biologically significant" interpretations concerning, e.g., the comparison of effects of different variables, becomes possible. In passing, we did not encounter any difficulties when calculating lines of separation (see LBD, fig. 1) based on logistic regression analysis.

Did we really suggest use of an implausible modeling technique? For example, presence and absence as a response variable scales either to 1 or 0. In Rita & Ranta (1993) we give examples how ordinary regression technique can yield presence larger than 1, and absence smaller than 0. These are biologically impossible, and are thus *never* encountered when using logistic regression. That is to agree with LBD who say that the analysis "... must be guided by the biological characteristics of the problem under study". We shall add that emphasizing biological characteristics should not, however, override use of proper statistical tools.

References

Adler, G. H. & Wilson, M. L. 1985: Small mammals on Massachusetts islands: the use of probability functions in clarifying biogeographic relationships. — Oecologia 66: 178–186.

Hosmer, D. W. & Lemeshow, S. 1989: Applied logistic regression. Wiley Interscience, New York.

Lomolino, M. V., Brown, J. H. & Davis, R. 1989: Island biogeography of montane forest mammals in the American Southwest. — Ecology 70: 180–94.

Lomolino, M. V., Brown, J. H. & Davis, R. 1995: Analyzing insular distribution patterns: statistical approaches and biological inferences. — Ann. Zool. Fennici 32: 435–437.

Rita, H. & Ranta, E. 1993: On analysing species incidence. — Ann. Zool. Fennici 30: 173–176.