Ann. Zool. Fennici 30:173–176
Helsinki 30 June 1993

ISSN 0003-455X

Commentary

On analysing species incidence

Hannu Rita & Esa Ranta[1]

Rita, H., Department of Forest Resource Management, P.O. Box 24, FIN-00014 University of Helsinki, Finland
Ranta, E., Integrative Ecology Unit, Department of Zoology, Division of Ecology, P.O. Box 17, FIN-00014 University of Helsinki, Finland

Received 23 February 1993, accepted 13 April 1993

Diamond (1975) introduced the incidence functions to describe the effect of area on the occurrence of species. Such functions describe the likelihood of a species' presence on an island, or in an island-like habitat patch, of known characteristics. For example, a given species could be absent from islands smaller than a minimum area, then gradually increase its occurrence with increasing area and finally be present on all islands larger than a certain threshold area. Obviously, characters other than area and entities other than islands are suitable for incidence analyses.

Taylor (1991) examined presence/absence data from a different angle: she associated probability (with reliability estimates) for a species' presence on an island of given area. Our approach presented here is supplemental to Taylor's (1991): we shall stress the proper use of statistical tools in conjuction with presence/absence data. We feel that this is important because incidence data will be — in the wake of the biodiversity boom (e.g., Pimm 1991) associated with metapopulation studies (Gilpin & Hanski 1991) — increasingly used in the near future to evaluate probabilities of maintaining species on reserves with a given quality and quantity of habitat. Such analyses also aim at understanding occurrence patterns of different species (ranking the importance of abiotic and biotic factors).

A simple statistical tool to analyse the behaviour of probabilities is the logistic regression model: the probability of occurrence π is given as a function of the vector of explaining variables $\mathbf{x}$,

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta\,\mathbf{x})}{1 + \exp(\alpha + \beta\,\mathbf{x})}. \qquad (1)$$

The familiar linear nature of this model becomes apparent after the logit transformation, giving the form

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \alpha + \beta\,\mathbf{x} \qquad (2)$$

(here, ln denotes the natural logarithm). Fitting a logistic regression model to incidence data is a straightforward task and algorithms are available in several statistical program packages. We shall maintain that the use of the ordinary regression model is improper and may lead to unsound conclusions in analysing species' incidence.

There are only two possible values for the response variable in incidence data: 0 for absence and 1 for presence. In a model that attempts to explain the variation in such a variable, the residuals cannot be normally distributed (Fig.

[1] Responsible Editor: Kai Lindström

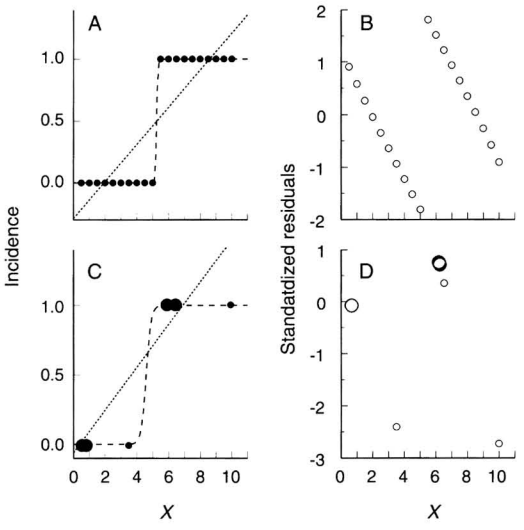

Fig. 1. Two hypothetical examples (A, C) to demonstrate the superiority of the logistic regression method (LR, dashed line) over the ordinary regression method (OR, dotted line). Incidence data for A are: absence when $0.5 \leq X \leq 5$ and presence when $5.5 \leq X \leq 10$, both with a step of 0.5 ($n = 20$), and for C: absences with x_i: 0.501, 0.501, 0.504, 0.505, 0.507, 0.507, 0.508, 3.500 and presences with x_i: 9.000, 6.014, 6.022, 6.024, 6.051, 6.059, 6.078, 6.085, 10.000). The model parameters for A are: LR $\alpha = -34.6$, $\beta = 25.6$; OR $a = -0.289$, $b = 0.150$, and for B they are: LR $\alpha = -29.2$, $\beta = 6.4$; OR $a = -0.06$, $b = 0.153$. B and D give residuals for the OR regression models in A and C, respectively (large dots mark closely located data points).

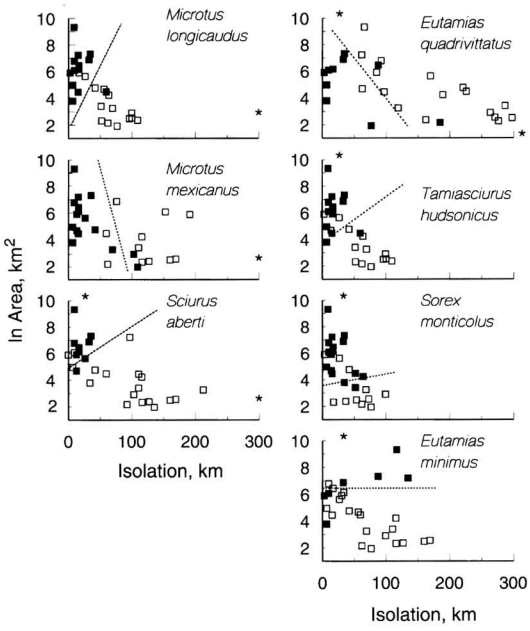

Fig. 2. Incidence (■ = presence, = absence) of seven mammal species on island-like mountain tops (data from Lomolino et al. 1989) evaluated against isolation and area. Continuous line indicates $\pi(X) = 0.5$ occurrence as indicated with logistic regression models incorporating both isolation and area as independent variables. This is done regardless of whether the variable is needed (* at the end of the corresponding axis) or not in statistical terms (for the parsimonious models, see Table 1). The model parameters for *M. longicaudus* are: $\alpha = -1.1$, $\beta_I = -0.06$, $\beta_A = 0.67$; *M. mexicanus:* $\alpha = 5.7$, $\beta_I = -0.06$, $\beta_A = -0.29$; *S. abeberti:* $\alpha = -6.4$, $\beta_I = -0.04$, $\beta_A = 1.32$, *E. quadrivittatus:* $\alpha = 6.3$, $\beta_I = -0.04$, $\beta_A = -0.62$; *T. hudsonicus:* $\alpha = -3.5$, $\beta_I = -0.03$, $\beta_A = 0.93$; *S. monticola:* $\alpha = -3.8$, $\beta_I = -0.07$, $\beta_A = 0.99$, and for *E. minimus:* $\alpha = -6.8$, $\beta_I = -0.001$, $\beta_A = 1.08$.

1B, D), as they should be in ordinary regression. Thus, the statistical theory developed for such models is not applicable to incidence data. In particular, the least squares method, generally used to estimate ordinary regression model parameters, is inefficient when applied to incidence data (Hosmer & Lemeshow 1989). In logistic regression, the binary nature of the response variable variation is the basis of parameter estimation.

Furthermore, ordinary regression models may easily lead to estimates without biological — or even mathematical — realism when applied to incidence data (Fig. 1A, C). Logistic regression models never predict inappropriate values ($\pi(X) > 1$ or $\pi(X) < 0$) for the probability of species' presence. Even the confidence intervals for the probability, in contrast to ordinary regression models, never include negative values or values greater than unity (Hosmer & Lemeshow 1989).

The coefficients in the logistic regression model have a natural interpretation, which is based on the odds ratio (see, e.g., Agresti 1984). This interpretation is consistent with the proportional nature of the probability as a measure of presence intensity. A more technical presentation of logistic regression models is given by McCullagh & Nelder (1989: chapter 4). In the particular case, when there is only one explaining variable, the parameters α and β of the lo-

gistic regression model have a useful property: the value of the explaining variable that indicates species' presence with probability 0.5 is $-\alpha/\beta$. This can easily be seen from Eq. (1) with replacement $X = -\alpha/\beta$. In fact, after estimation of parameters α and β, Eq. (1) can be solved for any value p of the probability $\pi(X)$. Confidence intervals for the obtained value of the explaining variable can be constructed.

When there are several explaining variables in the model, the solution of the equation $p = \pi(\mathbf{x})$ is a straight line (see Fig. 2) or a plane in the space of the explaining variables.

As the logistic regression model is not well known among population ecologists, we shall provide an example of its use with data on the occurrence of mammal species on mountain tops in the American Southwest (Lomolino et al. 1989). They applied ordinary regression models to examine the significance of area and isolation (predictor variables, X_i) on occurrence (response variable, Y) of the mammal species studied. Based on this analysis, they solved the equation "$Y = 0.5$" to link together the effect of area and isolation, giving the "line of separation".

Similar approaches have subsequently been used (e.g., Hanski 1991, Peltonen & Hanski 1991, among many others). Therefore, we feel we should point out the unsuitability of this method for any analyses of incidence data and strongly suggest avoiding ordinary regression models when examining factors describing species' presence/absence on islands or habitat fragments. Logistic regression analysis is far superior for these purposes. We illustrate our point by comparing the results of logistic regression models with those of ordinary linear regression. For this purpose, we use data from Lomolino et al. (1989: table 1), which refer to the occurrence of seven mammal species on 27 isolated mountain tops for which information on area and a measure of species-specific isolation are available. Occurrence of each of the seven species, in turn, was the response variable, and mountain-top area and isolation were taken as the explaining variables. In accordance with Lomolino et al. (1989), we used natural logarithm based transformation of area and left isolation non-transformed. Estimation of the model parameters was based on maximum likelihood. The decrease in model deviance was used as the means to select variables in the models. Goodness of fit of the final model was studied using the Hosmer-Lemeshow test statistic (Hosmer & Lemeshow 1989).

Distribution of two species *(Microtus longicaudus, M. mexicanus)* was influenced only by isolation, two species *(Sciurus aberti, Eutamias quadrivittatus)* were apparently affected by both area and isolation, and three species *(Tamiasciurus hudsonicus, Sorex monticola, Eutamias minimus)* were affected only by area (Table 1). This list differs from thhat of Lomolino et al. (1989). The discrepancies are greatest with *Microtus longicaudus, Eutamias quadrivittatus*

Table 1. Results of presence/absence analysis of the relationship between area (km^2) and isolation (km) for species inhabiting 7–16 of the total of 27 montane forest islands in the American Southwest. Logistic regression parameters α, β_A (for ln-transformed area) and β_I (for isolation) are tabulated. N denotes that the corresponding variable is not needed in the logistic model; y (variable needed), n (variable not needed) refer to the ordinary regression analyses of Lomolino et al. (1989). Goodness of fit of the model (Hosmer-Lemeshow χ^2, always with 8 degrees of freedom) and corresponding P-values are also listed. An asterisk before the species name indicates a discrepancy between the results of the ordinary regression analysis by Lomolino et al. (1989) and those of the logistic analysis.

Species	Islands	α	β_A		β_I		χ^2	P
**Microtus longicaudus*	13	3.28	N	y	–0.09	y	2.90	0.94
Microtus mexicanus	16	3.83	N	n	–0.05	y	6.40	0.60
Sciurus aberti	9	–6.44	1.32	y	–0.04	y	2.32	0.97
**Eutamias quadrivittatus*	10	6.21	–0.62	n	–0.04	y	4.54	0.81
**Tamiasciurus hudsonicus*	13	–6.26	1.30	y	N	y	5.73	0.68
Sorex monticolus	16	–4.39	1.08	y	N	n	12.31	0.14
Eutamias minimus	7	–6.91	1.08	y	N	n	5.89	0.66

and *Tamiasciurus hudsonicus*. The tentative explanations by Lomolino et al. (1989: 188) for distributional patterns and processes are thus liable for changes.

In Fig. 2, the incidence of the seven mountain-top mammal species is evaluated against the line of occurrence for $\pi(X) = 0.5$ in a two-dimensional space of isolation and area: on one side of the line, the probability of a species occurrence is above 0.5, and on the other side, it is below. This line can be taken to serve as the separation between good and poor islands or habitat fragments. Similar lines can be constructed for any other value of the probability. For our purpose, we used models in which both variables were forced, regardless of whether they should be included or excluded in accordance with the parsimony principle of empirical modelling. A comparison should be made between Fig. 2 herein and figure 6 of Lomolino et al (1989).

Lomolino et al. (1989: 186) also refer to discriminant analysis when they touch upon the problems of using binary data in ordinary regression models. Logistic techniques can — and often should — be used even if the actual aim of the study is that of discrimination and not that of prediction. For further details, see Fienberg (1987:105–109) and references therein. Fienberg also comments that when discriminant analysis is used, the independent variables should follow a multinormal distribution with equal covariance matrices within each group. When this assumption is violated, estimation techniques used with ordinary regression models are not optimal. In such cases, we advise the use of the logistic regression technique.

Acknowledgements. We thank Ilkka Hanski for inspiration and comments. Comments by Kenneth Burnham, Luke George, Yrjö Haila, Jari Kouki and by an anonymous referee on the MS are appreciated.

References

Agresti, A. 1984: Analysis of ordinal categorical data. — John Wiley & Sons, New York.

Diamond, J. M. 1975: Assembly of species communities. — In: Cody, M. L. & Diamond, J. M. (eds.), Ecology and evolution of communities: 342–444. Belknap Press of Harvard University Press, Cambridge.

Fienberg, S. E. 1987: The analysis of cross-classified categorial data. 2nd ed. — MIT Press, Cambridge, Massachusetts, USA.

Gilpin, M. E. & Hanski, I. (eds.) 1991: Metapopulation dynamics. — Academic Press, London.

Hanski, I. 1991: Single-species metapopulation dynamics: concepts, models and observations. — Biol. J. Linn. Soc. London 42:17–38.

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 70:180–194.

McCullagh, P. & Nelder, J. A. 1989: Generalized linear models. 2nd ed. — Chapman and Hall, London.

Peltonen, A. & Hanski, I. 1991: Patterns of island occupancy explained by colonization and extinction rates in shrews. — Ecology 72:1698–1708.

Pimm, S. L. 1991: The balance of nature. Ecological issues in the conservation of species and communities. — University of Chicago Press, Chicago.

Taylor, B. 1991: Investigating species incidence over habitat fragments of different areas – a look at error estimation. — Biol. J. Linn. Soc. Lond. 42:177–191.