DP-MS-88-249

PROPER STATISTICAL TREATMENT OF SPECIES-AREA DATA

by

C. Loehle

E. I. du Pont de Nemours and Company
Savannah River Laboratory
Aiken, SC 29808-0001

Joyce

DISCLAIMER

INTRODUCTION

The purpose of this note is to comment on the entire process of analyzing species-area data, particularly as performed by Rydin and Borgegård (1988). They use three different models to test species-area relations for islands over a 100 year period. Several aspects of their analysis of species-area data could be improved, including their comparison of goodness-of-fit and testing of the expected value of z. The reason that these issues are important (their basic conclusions being correct) is that there is acrimonious debate over the best model to use for species-area curves and over whether the slope coefficient is constant or is an artifact, and because the species-area curve is being used for nature reserve design (see Dunn and Loehle 1988 for discussion of these points). The problems pointed out here are common to a large class of allometric-type analyses in ecology. I attempt to show the potential pitfalls inherent in allometric analyses and demonstrate methods for avoiding these problems.

Rydin and Borgegård (1988) document long-term trends in species numbers on islands formed in 1886 when the water level was lowered in Lake Hjälmaren, Sweden. At each remeasurement period, data on species richness was collected for all islands. They fit three different models at each time (Table 1, their Table 2). The three models are

$$\log S = C + z \log A \tag{1}$$

$$S = CA^z \tag{2}$$

$$S = C + z \log A \tag{3}$$

where S is species number, and A is island area. A is usually known with fairly high accuracy. The first mistake is that (1) and (2) are not really the same model. Taking the log transform of both sides of (2),

$$\log S = \log (CA^z)$$
$$\log S = \log C + z \log A$$

Creating new variables from the log of the data values, we get a typical linear equation which can be fit by linear least-squares

$$S' = \beta_0 + \beta_1 A' \qquad (4)$$

The coefficient β_0 obtained from regression is therefore not C but log C. Doing the inverse transform, C can be recovered and the C and z values for (1) and (2) will be identical when there is no error term. The two models do not in general give identical results, however, as shown next. In addition to C in (1) and (2) not having the same meaning, C and z in (3) are not comparable in meaning to either (1) or (2). C and z have come to have specific meanings in the ecological literature (C indicating species richness and other factors and z being related to Preston's canonical hypothesis and other things, Gould, 1979; Sugihara, 1981). Thus it would be much less confusing if β_0 and β_1 were used as designations for arbitrary coefficients to be fit when comparing several equations, rather than C and z.

The second problem is the comparison of R^2 values from models with differently transformed data. It is valid to compare R^2 values for one model under different treatments (years, in this case) and thus their comparisons of slope by sample year are valid. When different models are to be compared, however, the R^2 values are not commensurate because a log-transform may change the R^2 value and in general will increase it (Kvålseth, 1985). This is purely an artifact of the data transformation and results from axis compression. Since (2) has an additive error term, whereas (4) has a multiplicative error term (in the S-A plane), (4) and (2) will not in general give identical values for C and z. In fact, z values for (1) are much larger than for (2) (Table 1). Thus (4) and (2) cannot be considered interchangeable ways of fitting C and z, (as it seems many ecologists assume), and R^2 values for (2), (3), and (4) cannot be compared when R^2 is based on the transformed variables, which is how it appears that Rydin and Borgegård calculated R^2. The question of which error term (additive or multiplicative) is more appropriate for species-area data has not been answered. This leads to a more general problem with the use of R^2 as a measure of goodness of fit. Kvålseth (1985) points out that there are several methods for calculating R^2. For linear models fit by least squares, most of these methods give the exact same results. For linear models with no intercept term and nonlinear models, the various methods give different results. This problem is not generally appreciated nor discussed in regression text books. Kvålseth (1985) recommends a method (his R_1^2) that allows comparison between linear, linear with no intercept, nonlinear (in the parameters) models and those fit by other than ordinary regression. For comparisons, all calculations are done using untransformed values of x and y. His adjusted R^2 is

$$R^2 = 1 - \frac{a\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y}_i)^2} \qquad (5)$$

where summation is over n data points,

$$a = \frac{n - 1}{n - k - 1} \qquad \text{for intercept model, and}$$

$$a = \frac{n}{n-k} \qquad \text{for no-intercept model,}$$

where k is the number of parameters. This adjusted R^2 calculation is not necessarily the one provided by standard computer packages, as shown by Kvålseth (1985). Thus care must be taken when different models are calculated by different packages or by hand or compared and compiled from different published papers. It is not safe to assume that R^2 values are computed from (5) unless the authors specify this. In general R^2 values printed by packages will be based on the transformed (S') values for models such as (1) and (3), rather than on the original axes (Kvålseth, 1985). To emphasize the importance of using (5), Kvålseth (1985) notes that for nonlinear models several of the available R^2 calculations can give values greater than 1. Kvålseth (1985) elaborates (5) for increased resistance to outliers, which may be useful if outliers are not first removed. He also notes that residual analysis should accompany R^2 as a measure of which model is best.

As a final point, Rydin and Borgegård (1988) test whether the confidence limits around z in (2) include 0.25. This test is due to claims that the value of z has underlying significance (Gould, 1979; Sugihara, 1981) or in contrast that z = 0.25 is a statistical artifact (Connor and McCoy, 1979). However, it is not entirely valid to compare 0.25 with the 95% confidence limits around z because the power of the test is unknown. After we fail to reject H_o, further tests are needed to conclude that z is not significantly different from 0.25. Range of the data (on the area axis) can easily be inadequate for species-area studies and small range can lead to low power. Inadequate sample size can also lead to the conclusion of no difference. Finally, very large islands are typically few in number in species-area studies, leading to very high leverage for the values for large islands, particularly if they are much larger than the other islands. Tests are available for examining leverage (Myers, 1986 p.155). These factors must be evaluated and eliminated as causes of the overlap of z with 0.25 before we can conclude that the lack of difference is

meaningful. A method for testing whether the measured z differs from 0.25 for a nonlinear model for species-area data is given in Dunn and Loehle (1988). Using Monte Carlo simulation, sample from a function (2) with z = 0.25 and variance the same as the actual data. This is essentially a bootstrap method. Generate in this way a large series of data sets to which z is fit, giving a distribution of z values against which the sample value can be tested (i.e., is the sample value likely to have been obtained from a population with a true value of z = 0.25?). The effects of factors which give low power can be tested in this context by, e.g., varying the range of the data in the Monte Carlo trial (Dunn and Loehle, 1988).

In conclusion, care must be taken when working with transformed data, particularly for comparing equations and drawing conclusions about the values of parameters. The problems pointed out here are not well known to ecologists and deserve consideration. Particular care should be taken when using results from statistical computer packages.

ACKNOWLEDGMENTS

This work was carried out under contract DE-AC09-76SR00001 with the U. S. Department of Energy, with whom the copyright remains. I would like to thank C. Comiskey and statisticians S. Harris and E. Smith for reviewing the manuscript.

LITERATURE CITED

Connor, E. F. and E. D. McCoy. 1979. The statistics and biology of the species-area relationship. Am. Nat. 113:791-833.

Dunn, C. P., and C. Loehle. 1988. Species-area parameter estimation: testing the null model of lack of relationship. J. Biogeog. In Press.

Gould, S. J. 1979. An allometric interpretation of species-area curves: the meaning of the coefficient. Am. Nat. 114:335-343.

Kvålseth, T. O. 1985. Cautionary note about R^2. Am. Statis. 39:279-285.

Myers, R. H. 1986. Classical and Modern Regression with Applications. Duxbury Press, Boston.

Rydin, H., and S.-O. Borgegård. 1988. Plant species richness on islands over a century of primary succession: Lake Hjälmaren. Ecology 69:916-927.

Sugihara, G. 1981. $S = CA^z$, $z \approx 1/4$: A reply to Connor and McCoy. Am. Nat. 117:790-793.

Table 1. Species-area regressions. In the 1886 figures, islands formed in 1886, just prior to Callmé's survey, are excluded. *n* is the number of islands included in each regression. From Rydin and Borgegård (1988).

	$S = C + z \log A$ (exponential)				$S = CA^z$ (power)				$\log S = C + z \log A$ (transformed power)			
Year	*C*	*z*	*n*	R^2	*C*	*z*	*n*	R^2	*C*	*z*	*n*	R^2
1886	-14.0	13.2	21	0.44	8.6	0.16	21	0.40	0.80	0.20	21	0.47
1892	-53.9	32.7	31	0.65	8.5	0.24	31	0.53	0.16	0.46	30	0.56
1903-1904	-54.6	34.6	35	0.76	6.8	0.28	35	0.69	0.02	0.52	34	0.78
1927-1928	-61.0	39.3	37	0.85	6.3	0.30	37	0.81	-0.02	0.56	37	0.78
1984-1985	-48.0	36.1	37	0.84	0.8	0.24	37	0.78	0.65	0.36	37	0.72