

ON MODELS PREDICTING ABUNDANCE OF SPECIES AND ENDEMIC  
FOR THE DARWIN FINCHES IN THE GALÁPAGOS ARCHIPELAGO

T. H. HAMILTON AND I. RUBINOFF

*Department of Zoology, The University of Texas, Austin, and  
Museum of Comparative Zoology, Cambridge, Massachusetts*

Reprinted from *EVOLUTION*, Vol. 18, No. 2, June 12, 1964  
pp. 339-342

*Made in United States of America*

ON MODELS PREDICTING ABUNDANCE OF SPECIES AND ENDEMICIS FOR THE DARWIN FINCHES IN THE GALÁPAGOS ARCHIPELAGO<sup>1</sup>

T. H. HAMILTON AND I. RUBINOFF

Department of Zoology, The University of Texas, Austin, and  
Museum of Comparative Zoology, Cambridge, Massachusetts

Accepted April 15, 1964

In a previous report to this journal (1963) we made conclusions on the processes of speciation and production of endemics among the Darwin finches in the Galápagos Archipelago. The conclusions were inferred from multiple-regression analysis wherein four independent variables of the insular environment were quantified and tested for respective abilities to predict species data. The analysis was an attempt by linear regression to discover (a) the environmental determinants of insular variation in the species numbers ( $Y_1$ ) and (b) the determinants of such variation for numbers of endemic subspecies ( $Y_2$ ). Insular area ( $X_1$ ) and number of land plant species ( $X_2$ )—presumably indexing ecologic and floristic diversity—were found lacking in power to “push” or “move” the dependent variables (the  $Y$ 's). Isolation, measured by distance between islands ( $X_3$ ) and by distance from a given island to Indefatigable Island near the center of the archipelago ( $X_4$ ), was the major predictor of number of species and endemics for islands, with  $X_3$  being more important than  $X_4$  in terms of contribution to  $Y$  variances ( $\Sigma y^2$ ). For both problem *a* and *b*, however, a component ( $\Sigma dy^2$ ) of the variance remained unaccounted for by regression of the  $Y$ 's on the  $X$ 's, and this was attributed to error inherent to the analysis, to factors not considered, or to both.

<sup>1</sup>Publication supported by the U. S. Public Health Service.

To decrease the values of  $\Sigma dy^2$ , we have (i) shifted from use of desk calculator to digital computer analysis; (ii) included in the analysis new environmental factors ( $X_5$ ,  $X_6$ ) not previously considered; and (iii) utilized new models or estimating equations involving transformations of the arithmetic values of the primary measurements of the  $Y$ 's or  $X$ 's to their logarithmic counterparts in nonlinear or joint linear-nonlinear variation. Below we summarize results of such a comparative multiple-regression analysis, and, following this, a statement is made of the bearing of the new information on our previous conclusions (Hamilton and Rubinoff, 1963). Data for the  $Y$ 's and the four  $X$ 's are from our previous report, and the new factors ( $X_5$ ,  $X_6$ ) are from a comparable study of plant species abundance in the archipelago (Hamilton et al., 1963).

*Theory and results from computer analysis.*—Some standard estimating equations (where *a* and *b* represent constants set by solution of simultaneous equations; cf. Mordecai, 1941) are as follows for analysis with 2 independent variables:

$$Y = a_{y.i,j} + b_{y.i,j} X_i + b_{y.j,i} X_j; \quad (1)$$

$$Y = a_{y.i,j} + b_{y.i,j} \log X_i + b_{y.j,i} \log X_j; \quad (2)$$

$$\log Y = \log a_{y.i,j} + \log b_{y.i,j} X_i + \log b_{y.j,i} X_j; \quad (3)$$

$$\log Y = \log a_{y.i,j} + b_{y.i,j} \log X_i + b_{y.j,i} \log X_j. \quad (4)$$

The first is multiple linear regression and the last three are examples of multiple nonlinear analyses. With inclusion of new factors, the number of

TABLE 1. Coefficients of multiple determinations ( $R^2$ ) and contributions of the  $X$ 's to variance

Dependent variable	Independent variables	Variance (= $R^2$ )	Major contributor(s) to $R^2$
<i>Linear analysis</i>			
$Y_1$ (species)	$X_1 - X_6$	0.866	$X_3$ , 0.481; $X_2$ , 0.308
$Y_2$ (endemics)	$X_1 - X_6$	0.873	$X_3$ , 0.813
<i>Nonlinear analysis</i>			
$Y_1$	$\log X_1 - \log X_6$	0.821	$\log X_1$ , 0.301; $\log X_3$ , 0.257; $\log X_2$ , 0.171
$Y_2$	$\log X_1 - \log X_6$	0.774	$\log X_3$ , 0.705
$\log Y_1$	$X_1 - X_6$	0.799	$X_3$ , 0.514; $X_2$ , 0.158; $X_1$ , 0.109
$\log Y_2$	$X_1 - X_6$	0.710	$X_3$ , 0.550; $X_2$ , 0.109
$\log Y_1$	$\log X_1 - \log X_6$	0.759	$\log X_1$ , 0.247; $\log X_3$ , 0.236; $\log X_2$ , 0.165
$\log Y_2$	$\log X_1 - \log X_6$	0.624	$\log X_3$ , 0.518
<i>Joint linear-nonlinear analysis</i>			
$Y_1$	$6 X$ 's + $6 \log X$ 's	0.891	$\log X_5$ , 0.489; $X_4$ , 0.244
$Y_2$	$6 X$ 's + $6 \log X$ 's	0.934	$X_3$ , 0.813
$\log Y_1$	$6 X$ 's + $6 \log X$ 's	0.831	$X_3$ , 0.453; $\log X_2$ , 0.288
$\log Y_2$	$6 X$ 's + $6 \log X$ 's	0.878	$X_3$ , 0.684

simultaneous equations to be solved for the necessary constants ( $a_{y,ij} \dots, b_{y,ij} \dots$ ) increases, and the usefulness of computer resolution for problems involving more than 2 factors is obvious. For joint linear-nonlinear analysis, either  $Y$  or  $\log Y$  is estimated by various  $X$ 's and  $\log X$ 's (cf., e.g., Croxton and Cowden, 1939).

Table 1 summarizes results of application of the formulae listed (equations 1-4) to problems a and b. In addition to  $X_1, X_2, X_3$ , and  $X_4$ , the new factors considered are elevation ( $X_5$ ) and area of adjacent island ( $X_6$ ).  $X_5$  is thought to positively index ecologic diversity, and  $X_6$ , on a negative basis, to index isolation, particularly for prediction of endemics (cf. Hamilton et al., 1964). Table 2 is a simple correlation matrix for the 2  $Y$ 's, the 6  $X$ 's, and their 8 logarithmic counterparts; it will be noted that none of the  $X$ 's is strongly correlated (i.e.,  $> 0.80$ ) with another  $X$ . "Goodness of fit" for the formulae used is based on increases in value of coefficients for multiple determination ( $R^2$ ) for  $Y$  or  $\log Y$  (see Snedecor, 1957), and percentages of variation explained by each independent variable are based on contribu-

tions to  $R^2$ , independent of other variables considered in multiple regression. Thus the linear and nonlinear analyses utilize 6 variables each in multiple regression, and the joint linear-nonlinear analyses have the use of 12 variables (6 arithmetic, 6 logarithmic).

By multiple linear analysis (equation 1) nearly 87% of the insular variation in species numbers or  $Y_1$  (= variance =  $R^2 \times 100$ ) can be accounted for by variation in the 6  $X$ 's, with distance from nearest island ( $X_3$ ) being the major contributor (48%). The other "mover" of  $Y_1$  is insular number of land plant species ( $X_2$  contributing 31%; see table 1). The remaining  $X$ 's account for negligible, insignificant components of the variation. Values for  $R^2$  resulting from application of equations 2, 3, and 4 to the data are all smaller than the preceding, indicating poorer fits. Equation 4 (log-log analysis), explaining 82% of the  $Y_1$  variation, provides the next best fit. In addition to having a larger  $R^2$  value, equation 1 may also be favored over equation 4 because the latter's major contributor ( $\log X_1$ , area: 30%) is less than that for  $X_3$  in equation 1, which explains 48% of the 87% variation. Nonlinear equations have smaller  $R^2$  values than the linear one for  $Y_2$  or  $\log Y_2$ , but  $X_3$  or  $\log X_3$  is in each case the major "mover" of the dependent variable.

In joint linear-nonlinear analysis, a problem of redundancy arises since each environmental factor is included both as an arithmetic and a logarithmic value. While the meaning of the findings is still uncertain, 89% of the variation in  $Y_1$  can be attributed to variation in the 12 factors, with the logarithm of elevation ( $\log X_5$ ) and arithmetic distance from the center of the archipelago ( $X_4$ ) being the major variance contributors (50%, 24%). Of possible interest here is that in our study (*op. cit.*) of abundance of land plant species in this archipelago, elevation was the major predictor of insular number of species. However,  $X_3$  continues to be the major statistical "mover" for variation in  $\log Y_1, Y_2$ , and  $\log Y_2$  (table 1), and this finding supports that noted above and previously reported (Hamilton and Rubinoff, 1963).

*Concluding discussion.*—In this study, a major problem is that of curve-fitting and the determination of linear or nonlinear variation for the dependent and independent variables. The transformation of arithmetic to logarithmic values poses a problem. For example, the correlation of small measurements and their logarithmic counterparts (cf.,  $Y_1$  versus  $\log Y_1$  in table 2 where  $r = +0.99$ ) is very good, but for large measurements (area, elevation) differences appear (e.g.,  $X_1$  versus  $\log X_1$ :  $r = +0.69$ ). This may be an important aspect of the redundancy problem in joint linear-nonlinear analysis. As yet we cannot determine whether the high  $R^2$  value (0.89) for linear-nonlinear prediction of  $Y_1$  by the 12 variables (with  $\log X_5$  and  $X_4$  being major

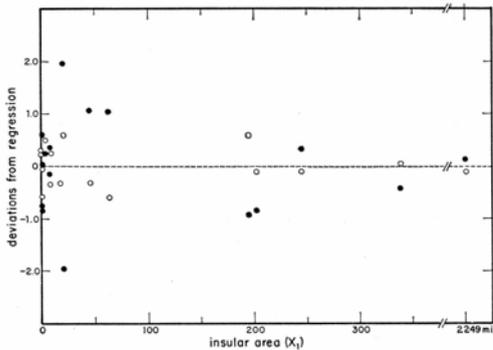


FIG. 1. Deviations from regression ( $d_{y.xxx}$ ) plotted against insular area. To show that the multiple-regression equation  $\hat{Y}_1 = 8.234 - 0.001X_1 + 0.015X_2 - 0.165X_3$  predicts number of Darwin finch species more accurately for the larger than for the smaller islands ( $n = 16$ ) in the Galápagos Archipelago. The  $d_{y.xxx}$  values (closed circles: ●) represent  $Y_1 - \hat{Y}_1$ . Contributions to the variance ( $R^2 = 0.836$ ) here are:  $X_1, 0.047$ ;  $X_2, 0.308$ ;  $X_3, 0.481$ . Note the negative  $d_{y.xxx}$  value regression of  $Y_1$  for area ( $X_1$ ); its variance contribution is, however, negligible. Open circles (○) denote deviations from regression by the equation  $\hat{Y}_2 = -0.313 + 0.001X_1 + 0.095X_3$ . To show the primarily linear (= horizontal) spread of  $d_{y.xxx}$  points resulting from prediction of insular numbers of endemic subspecies. For the latter equation, variance is 0.863, receiving 0.050 from area and 0.813 from isolation ( $X_3$ ). For both  $\hat{Y}_1$  and  $\hat{Y}_2$ , the computer "selected" the best 2 or 3  $X$ 's from the 12 predicting ones.

TABLE 2. Simple correlation matrix<sup>1</sup>

		Logarithms														
$Y_1$	$Y_2$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$Y_1$	$Y_2$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	
$Y_1$	1.00	-0.60	0.38	0.55	-0.69	-0.64	0.23	0.63	0.99	-0.52	0.55	0.69	-0.66	-0.55	0.19	0.70
	$Y_2$	1.00	-0.86	0.13	0.90	0.44	-0.34	-0.22	0.59	0.93	-0.06	-0.08	0.83	0.61	-0.15	-0.25
		$X_1$	1.00	0.57	-0.33	-0.22	0.03	0.73	0.33	-0.22	0.66	0.45	-0.32	-0.12	0.11	0.54
			$X_2$	1.00	-0.00	-0.42	-0.75	0.66	0.51	0.15	0.79	0.90	-0.08	-0.09	0.00	0.71
				$X_3$	1.00	0.51	-0.31	-0.39	-0.67	0.83	-0.20	-0.18	0.94	0.67	-0.16	-0.34
					$X_4$	1.00	-0.09	-0.23	-0.62	0.34	-0.39	-0.68	0.57	0.80	-0.44	-0.23
						$X_5$	1.00	0.47	0.23	-0.26	0.21	0.07	-0.27	0.03	0.63	0.35
							$X_6$	1.00	0.59	-0.23	0.87	0.65	-0.36	-0.06	0.16	0.92
								$\log Y_1$	1.00	-0.52	0.50	0.64	-0.63	-0.54	0.19	0.67
									$\log Y_2$	1.00	-0.08	-0.02	0.71	0.48	-0.11	-0.26
										$\log X_1$	1.00	0.81	-0.22	-0.12	0.02	0.85
											$\log X_2$	1.00	-0.25	-0.38	0.21	0.73
												$\log X_3$	1.00	0.79	-0.11	-0.29
													$\log X_4$	1.00	-0.14	-0.09
														$\log X_5$	1.00	0.03
															$\log X_6$	1.00

<sup>1</sup> Respective values for correlation coefficients ( $r$ ) where for 16 islands of the Galápagos Archipelago  $Y_1$  = number of species,  $Y_2$  = number of endemic subspecies,  $X_1$  = area (in square miles),  $X_2$  = number of land plant species,  $X_3$  = distance (in miles) from nearest island,  $X_4$  = distance (in miles) from center of archipelago,  $X_5$  = elevation (in feet), and  $X_6$  = area of adjacent island (see text). With 14 degrees of freedom, 5% and 1% levels of significance are respectively achieved by values of 0.50 and 0.62 for any single  $r$  (cf. Snedecor, 1957, p. 174).

“movers”) is meaningful or a result of some variables “masking” others. In any event, the two “movers” here are one of elevation (indexing ecologic diversity) and one of isolation. In spite of an increase in the number of  $X$  variables from 4 to 6 (or 12!),  $X_3$  remains the best predictor of  $Y_2$ . For predictions of  $Y_1$  new variance contributors appear in the form of log area (30%), land plant species numbers (31%), log elevation (49%), and distance from archipelago center (24%), with  $X_3$  also being a significant contributor in 5 of 6 equations (table 1).

Deviations from regression, plotted against the  $X$ 's, give new information about problems a and b, and for prediction of  $Y_2$ , the scatter plots are mostly horizontal along (for example) the  $X_1$  axis (see open circles of fig. 1). This confirms the linear variation in insular number of endemic subspecies, with isolation, measured by least interisland distance, being the major predictor ( $b_{y_2x_3 \dots}$  values varying from +0.09 to +0.11). For prediction of  $Y_1$ , such residual variation is curved, and fig. 1 (by closed circles) shows that equation 1 predicts numbers of finch species for the larger Galápagos islands with greater fidelity than for the smaller islands. This hints of non-linear regression of  $Y_1$  on  $X_1$ , but the same trend is true for  $\log Y_1$  on  $\log X_1$ . Thus knowledge of isolation, plant species numbers, and area *over-* and *underpredicts* the size of the finch species faunas for the smaller islands of the Galápagos. To an equal or lesser extent, the same trend for misprediction is true for the more isolated islands

( $b_{y_1x_3 \dots}$  . . . varying from -0.16 to -0.17 for arithmetic values). This nonobvious finding holds true for each of the linear, nonlinear, and joint linear-nonlinear predictions.

The question arises for the finches in the Galápagos: why does the species-environmental variant analysis, but not the endemics-environmental analysis, depart from normality? Disregarding inadequacies of the  $X$  and  $Y$  data, we surmise that this is a result in part of reduced insular area and isolation, and that the outer, smaller islands of the archipelago have unstable species communities, but relatively stable endemic communities. The significance of our inference, if truly valid, is uncertain, but it may relate to MacArthur and Wilson's (1963) equilibrium theory for fluctuations in insular species numbers. We are also reminded of Snow's (1950) finding, for the islands of São Tomé and Príncipe in the Gulf of Guinea, that the endemic avifaunal members are common and seemingly stable, and that it is among the nonendemics that species are extinct or in danger of becoming extinct. At the minimum, our findings hint of undescribed problems in increased or decreased ecological competition, as well as in species equilibrium phenomena, for the smaller, outer islands of the Galápagos. Isolation, influencing the dispersal of individuals and of their ecological requirements in part, must be an aspect of the problem not yet unraveled.

*Summary.*—Area, elevation, land plant species numbers, area of adjacent island, and two measures of isolation are used in arithmetic and

logarithmic quantifications to predict insular variation in number of species and endemics for the Darwin finches in the Galápagos Archipelago. Roughly 80% of the variation in numbers of endemics is accounted for by isolation (positive regression) measured as the distance from the nearest island. For variation in species numbers, this measure of isolation (negative regression) and plant species numbers (positive regression) are major predictors (48%, 31%) by linear analysis, but for linear-nonlinear analysis, log elevation and distance from archipelago center are also important (49%, 26%). For species numbers, each estimating equation mispredicts for the small, isolated islands more than for the larger, neighboring ones. This hints of special problems for increased or decreased ecological competition, and for instability or stability of species or endemics communities, in the peripherally isolated, smaller islands.

#### ACKNOWLEDGMENTS

We are indebted to the Harvard Computing Center for programming assistance and statistical advice, and to R. H. Barth, Jr., and the National Science Foundation for further assistance and financial support. We are also indebted to F. B. May, Bureau of Statistical Research, The University of Texas, for advice in interpreting the results of the computer analysis.

#### LITERATURE CITED

- CROXTON, F. E., AND D. G. COWDEN. 1939. Applied general statistics. Prentice-Hall, Inc., New York.
- HAMILTON, T. H., AND I. RUBINOFF. 1963. Isolation, endemism, and multiplication of species in the Darwin finches. *EVOLUTION*, **17**: 388-404.
- , —, R. H. BARTH, JR., AND G. BUSH. 1963. Species abundance: natural regulation of insular variation. *Science*, **142**: 1575-1577.
- , R. H. BARTH, JR., AND I. RUBINOFF. 1964. The environmental control of insular variation in bird species numbers. *Proc. U. S. Nat. Acad. Sci.*, **51** (in press).
- MACARTHUR, R. H., AND E. O. WILSON. 1963. An equilibrium theory of insular zoogeography. *EVOLUTION*, **17**: 373-387.
- MORDECAI, E. 1941. Methods of correlation analysis (first revision). John Wiley and Sons, New York.
- SNEDECOR, G. W. 1957. Statistical methods (fifth ed., reprinted). Iowa State College Press, Ames.
- SNOW, D. W. 1950. The birds of São Tomé and Príncipe in the Gulf of Guinea. *Ibis*, **92**: 579-595.