USING LOGISTIC REGRESSION IN ECOLOGY

A. SHANUBHOGUE* and A. P. GORE

Department of Statistics, University of Poona, Pune 411 007, India.

ABSTRACT

Regression analysis as a method of studying relationship between a dependent variable and one or more independent variables is well known. This method fails if the dependent variable is binary i.e. yes-no, present-absent etc. Use of logistic regression is appropriate for such problems. The purpose of this note is to bring to the attention of Indian ecologists this useful statistical tool. The computing aspect is also considered. Two applications in ecology are discussed.

INTRODUCTION

LINEAR regression analysis is a popular statistical method widely used by and taught to biologists in general and ecologists in particular. It is described in detail in most standard textbooks. The computations involved are elementary in case of simple linear regression. Computer programs are widely available for multiple regression (i.e. the case in which there are many independent variables).

These methods are inapplicable if the dependent (response) variable is dichotomous. Many times in ecological and other studies one comes across binary responses such as dead-alive, male-female, sick-healthy, success-failure etc. In toxicological studies, the method of probit analysis is used in such cases. But that requires a highly controlled condition with adequate replication, and hence is not suitable for exploratory studies involving many potential factors, most of which are to be weeded out. The method appropriate in such cases is called logistic regression. This method is well established in statistical literature[1] but has not yet filtered to standard textbooks on applied statistics.

THE MODEL

Suppose y denotes the response or dependent variable, which takes the values zero or one. We denote by $\mathbf{x}$ the collection of all independent variables. Our aim is to know on the basis of $\mathbf{x}$, the chance that y will be unity. Let $P(\mathbf{x})$ denote the probability that y equals one when independent variables assume the values $\mathbf{x}$. Then $P(\mathbf{x})/[1\text{-}P(\mathbf{x})]$ is called the odds ratio and $\ln P(\mathbf{x})/[1-P(\mathbf{x})]$ the log odds (these being natural logarithms). The logistic regression model assumes that log odds can be expressed as a linear combination of values of independent variables. Thus

$$\ln \mathrm{P(x)}/[1\text{–}\mathrm{P(x)}] = \sum_{i=1}^{p} \mathbf{x}_i \ \beta_i \ (= \mathbf{x}\beta \text{ in matrix notation}),$$

where p is the number of independent variables and β_i are called "regression constants". Then the probability of $y = 1$ can be written

$$P(\mathbf{x}) = \exp(\mathbf{x}\beta)/[1+\exp\ (x+\beta)].$$

The problem of interest is estimation of the parameters β, tests of hypotheses about them and checking the suitability of the model.

ESTIMATION AND TESTS OF HYPOTHESES

The parameters are estimated by the standard method of maximum likelihood. This method searches those values of the parameters which make the observed data most likely.

Let $y_1, y_2, \ldots, y_n$ be the n observed binary responses (which have independent Bernoulli distributions) and $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be the associated vectors of independent variables. Then the likelihood for any single response is

$$[P\mathbf{x}_i)]^{y_i}\ [1-P(\mathbf{x}_i)]^{1-y_i},$$

and the joint likelihood is the product of n such terms. The log likelihood is

$$L = \sum\ y_{\mathrm{i}} \ln P(\mathbf{x}_i) + \sum\ (1-y_i) \ln\ [1-P(\mathbf{x}_i)],$$

substituting the value of $P(\mathrm{x_i})$ we get

$$L = \sum_i y_i\ \sum_j x_{ij}\ \beta_j - \sum_i \ln\left[1+\exp\left(\ \sum_j x_{ij'}\,\beta_j\right)\right],$$

* For correspondence.

where exp stands for "e raised to". To maximize the likelihood, partial derivaties of L with respect to β_i, $i = 1, 2, \ldots, p$, are equated to zero. The resulting likelihood equations have to be solved iteratively using the Newton-Raphson method by considering ordinary least squares estimates as initial solutions[1]. This yields the estimates of "regression constants".

It is also of interest to test the significance of these estimated values. For this, the large sample properties of maximum likelihood estimators are invoked. In large samples, these estimates have a joint multivariate normal distribution with mean vector β and covariance matrix given by the inverse of the so-called information matrix[2]. To test whether a particular regression constant is different from zero, z-value is computed by dividing the estimate of that constant by its standard error (i.e. square root of the corresponding diagonal term in the covariance matrix) and it is compared with a suitable cut off point of the standard normal table.

The last question in inference concerns the suitability of the fitted model. Most available procedures[3] check the goodness of fit of the logistic regression model vis-a-vis some other competing model. Recently Hosmer and Lemeshow[4] proposed a method analogous to the standard Pearsonian goodness-of-fit test which considers a general omnibus alternative. We shall omit the details of this procedure.

APPLICATIONS IN ECOLOGY

We shall illustrate the use of logistic regression analysis with two applications. The first concerns egg-laying behaviour of paper wasps *Ropaliadia marginata*. The experimental data in this case were provided by Dr Raghavendra Gadagkar of the Indian Institute of Science, Bangalore.

In natural colonies of Indian paper wasps only one or a small number of females lays eggs while others act as workers and die without ever laying eggs. In an experiment, all the pupae from 19 natural colonies were obtained. The emerging females were isolated into separate cages. They were provided with food with other necessities to allow nest building and egg-laying. Out of 145 females, 75 laid eggs while 70 died without laying eggs. The question of interest is, what distinguishes egg layers from others.

The first exploratory attempt involved study of one independent variable at a time. Various body measurements were taken as independent variables. But their distributions for the egg-layers and non-layers turned out to be very similar. Therefore properties of the parent colony from which an individual came were considered. These included number of empty cells, number of eggs, larvae, pupae, adult males, adult females etc. Here also the egg layers seemed indistiguishable from others. Hence it was decided to adopt logistic regression.

Here the binary response is $y = 1$ (i.e. the individual successfully laid an egg) or $y = 0$ (i.e. the individual died without laying an egg). Various body measurements and nest properties are the independent variables. We first used only the body measurements, suspecting that something in morphology may reveal whether a female will lay an egg. Individual beta coefficients were tested for significance.They all turned out to be essentially zero, implying that morphological characters did not seem to influence the chance of egg laying. Next we considered the nest properties. Here the beta coefficient for the number of empty cells in the parent colony was found to be significantly positive. A summary of this analysis is given in table 1.

The main point that emerged from logistic regression was that other things remaining constant, greater the number of empty cells in a colony, greater is the chance that a female coming from that colony will lay an egg. This could be due to a variety of reasons. One possibility is that empty cells may indicate declining fecundity of the queen or poor health. This may change the larval feeding pattern and result in greater proportion of egg layers. We shall not pursue the biological implications any further. One may wonder as to why this was not revealed in the elementary analysis that preceded logistic regression. The reason is simple. The earlier

Table 1 *Analysis by considering nest properties alone*

Parameters	Estimates	Standard error	z-value
Intercept, β_1	−0.4150	0.3949	−1.0501
Number of eggs, β_2	−1.0153	0.8242	−0.6330
Number of larvae, β_3	0.0103	0.0235	0.4383
Number of pupae, β_4	0.483	0.0412	1.1733
Number of parasitised cells, β_5	−0.1660	0.0424	−0.4848
Number of empty cells, β_6	0.0643	0.0288	2.2353
Number of males, β_7	0.1462	0.3322	0.4400
Number of females, β_8	−0.0268	0.2751	−0.9932

analysis was at the level of an individual, whereas the phenomenon seems to be operative at the colony level. Our pooling all egg layers on one side essentially prevented us from observing patterns at the colony level. Once this was noticed, a simple graph of number of empty cells in a colony against proportion of egg-layers among females from that colony was plotted. It clearly revealed the positive association.

The second application concerns risk taking behaviour of mice *Peromyscus maniculatus* in west Canadian winter. The experimental data were made available by Dr Paul Anderson of the University of Calgary, Canada.

The location of this experiment was a frozen lake surrounded by forest. The mice normally reside on the lake shore and forage in the ecotone under cover. In winter, with lower food availability the animals widen their search area. They experience the risks of over exposure and predation.

The experiment involved keeping food boxes with suitable opening containing sunflower seeds at various perpendicular distances from the ecotone, and checking if the food is eaten. They are moved 1.5 m further away if they have been visited and 1.5 m closer otherwise. Two different directions are involved:towards the lake and towards the forest.

Here the response recorded is $y = 1$ (box visited and seeds eaten) or $y = 0$ (box not visited). The explanatory variables considered were distance of the box and maximum as well as minimum temperatures of the day. Two data sets, one when the boxes were on the lake and the other when the boxes were in the forest, were analysed separately.

The following were the main predictions behind the experiment: (i) As the distance of the food box from ecotone increases, the chance of it being visited should decline since the risk of predation increases, (ii) As the cover on the forest side is better, this decline in visitation should be sharper on the lake side than on the forest side, (iii) As the maximum and minimum temperatures rise, for a given distance, the chance of visitation should increase.

All these can be tested by examining the beta coefficients. It was found that the coefficient for distance was significant and negative. This supports prediction (i). In absolute value, the coefficient for distance in lake data was found to be significantly greater than in the forest data. This supports prediction (ii). The coefficients for temperature were significant and positive. This supports prediction (iii).

COMPUTER PROGRAMS

The three kinds of computations necessary in using logistic regression namely estimation of regression constants, tests of hypotheses and goodness of fit testing, together involve a substantial amount of numerical work and recourse to a computer is almost inevitable. Necessary programs both in BASIC and FORTRAN have been prepared and their listings are available from authors on request. The program in BASIC is prepared for EIKO-II desk top computer while the FORTRAN program is written for ICL 1904 S computer.

29 January 1987

1. Cox, D. R., *The analysis of binary data*, Chapman and Hall, 1970.
2. Rao, C. R., *Linear statistical inference and its applications*, Second Edition, Wiley Eastern, New Delhi, 1973.
3. McCullagh, P. and Nelder, J. A., *Generalized linear models*, 1983, Chapman and Hall.
4. Hosmer, D. W. and Lemeshow, S., *Communications in statistics theory and methods*, 1980, A9(10), 1043.