Nontraditional Regression Analyses

Joel C. Trexler; Joseph Travis

Ecology, Volume 74, Issue 6 (Sep., 1993), 1629-1637.

Stable URL:
http://links.jstor.org/sici?sici=0012-9658%28199309%2974%3A6%3C1629%3ANRA%3E2.0.CO%3B2-J

Ecology, 74(6), 1993, pp. 1629–1637

NONTRADITIONAL REGRESSION ANALYSES[1]

JOEL C. TREXLER
Department of Biological Sciences, Florida International University, Miami, Florida 33199 USA

JOSEPH TRAVIS
*Department of Biological Science B-142, Florida State University,
Tallahassee, Florida 32306-2043 USA*

Abstract. Least-squares linear regression and multiple regression are among the most commonly used analytical techniques of ecologists. However, these techniques only address a portion of the possible applications of regression methods. We discuss two less commonly used regression analyses that could find wide application in ecology, logistic regression and LOWESS regression. Logistic regression is appropriate in cases where the dependent variable is categorical, dichotomous, or polychotomous. It can be used with continuous and/or discrete independent variables. Logistic regression is motivated by the underlying binomial or multinomial distribution of dichotomous and polychotomous dependent variables and transforms the data to explicitly model these distributions. Locally weighted regression scatterplot smoothing or LOWESS regression is used to model the relationship between a dependent variable and independent variable when no single functional form will do. LOWESS regression is motivated by the assumption that neighboring values of the independent variable are the best indicators of the dependent variable in that range of independent values.

INTRODUCTION

In many contexts the ecologist examines the proposition that the expected value of a dependent variable, Y, is a function of an associated value of an independent variable, X. The most familiar case is linear regression, in which the expected value of Y is a linear function of X. If one accepts that the relationship is linear throughout the range of X, then one can estimate the parameters of the regression function with the formulae found in every statistical textbook. If one wishes to test the significance of the relationship, and one is willing to accept the proposition that Y has a normal distribution around its expected value with a variance that is constant throughout the range of X, then one employs the standard statistical tests also found in every textbook. However, in some contexts, the assumptions of linearity and normal errors are unacceptable. In these cases the ecologist may still wish to examine the proposition that the values of Y are somehow determined by the values of X but is forced to do so without the familiar companion of normal linear regression.

How best to proceed from this point will depend upon the nature of the biological problem at hand and what the investigator is willing to accept. In some cases the function that relates Y to X can be specified from considerations that are extrinsic to the data to be analyzed, for example when a theoretical model specifies a particular function or when other data have indicated a particular function. In these cases, a variety of nonlinear regression models are available to estimate the parameters of the specified functions and, in some cases, to test the fit of the data to the model (see Juliano and Williams 1987 for an example). In still other cases, the distribution of the dependent variable is known in advance, but it is not normal; in many cases linear-regression methods for other distributions of the dependent variable are available (McCullagh and Nelder 1989). Finally, a few specialized cases have been designed for particular types of data for which linear-regression estimators have been derived from first principles (e.g., Pacala and Dobson 1988).

In this paper, we highlight two approaches that may be used profitably on a broad range of ecological data. *Logistic regression,* which we describe in the next section, examines the functional relationship between a binomial dependent variable and an independent variable that may be either discrete or continuous in its distribution. In the final section we highlight an approach in which the investigator is inclined to believe that the functional nature of the relationship between

[1] For reprints of this Special Feature, see footnote 1, p. 1615.

Y and X changes within the range of X that is being examined in such a way that no single functional form will suffice to describe the pattern.

Logistic Regression

The regression methodologies most familiar to ecologists are intended for use with continuous or numerical dependent variables (Shanubhogue and Gore 1987). However, the analysis of a discrete dependent variable (proportions, rates, or odds arising from dichotomous or polychotomous dependent variables) requires special treatment because the errors associated with such variables may not be normally distributed. A likely distribution for dichotomous dependent variables is the binomial. A regression model that explicitly assumes binomially distributed errors is logistic regression (Agresti 1990), which addresses the binomial nature of the errors through the use of the logit transformation. The logit can be generalized for polychotomous dependent variables with a multinomial distribution (Fienberg 1980, Hosmer and Lemeshow 1989, Agresti 1990). Logistic regression is used in situations analogous to the use of regression or analysis of covariance when analyzing normally distributed continuous dependent variables. Regression, analysis of variance, analysis of covariance, logistic regression, log-linear models, and probit analysis are all related as special cases of generalized linear models derived for different error distributions and links of mean to variance (McCullagh and Nelder 1989).

The model.—Modelling the effect of discrete and continuous independent variables on discrete dependent variables such as survival (dead vs. alive), size (large vs. small), or color (green vs. not green) requires that the dependent variable be transformed to a scale amenable to analysis. The model used for logistic regression is:

$$\ln(\pi(x)/1 - \pi(x)) = \alpha + \beta_1 R + \beta_2 x + \beta_3 Rx, \qquad (1)$$

which can be re-written to define $\pi(x)$ as:

$$\pi(x) = \frac{\exp(\alpha + \beta_1 R + \beta_2 x + \beta_3 Rx)}{1 + \exp(\alpha + \beta_1 R + \beta_2 x + \beta_3 Rx)}. \qquad (2)$$

$\pi(x)$ is the logistic regression function (see Cox and Snell 1989:18–23 for the motivation of this function), R is a categorical variable, x is a continuous variable, and α, β_1, β_2, and β_3 are parameters to be estimated. $\pi(x)$ is approximated by the proportion of successes (p) where success is defined as the occurrence of an arbitrarily defined outcome (McCullagh and Nelder 1989). Multiple categorical and/or continuous independent variables can be incorporated into the model. Categorical independent variables with more than two levels are modelled with dummy variables (Anderson et al. 1980:173–174). β_3 in this model refers to a linear interaction of the continuous and discrete variables. Nonlinearity in the continuous variable can be introduced by adding polynomials of that variable. When possible, it is best to use maximum likelihood methods to fit the model to data.

For logistic regression, the dependent variable (Y) takes the value 0 or 1 with $\Pr(Y = 1 \mid$ the independent variables $x) = \pi(x)$. Y is transformed as an odds ratio, the probability that an event occurs relative to its converse. Thus, an odds ratio of one indicates that the probability of an event and its converse are equal (i.e., $p = 0.5$). The natural log of the odds ratio is the logit transformation, which has several desirable characteristics. These include that the zero-to-one range of p is expanded to $-\infty$ and $+\infty$ and that the binomial distribution of errors is modelled. Thus, information is not lost at extreme values of p, as is the case when a normal-theory model is fit to dichotomous data (Agresti 1990:84–87). In general, it is not necessary to have several observations at each x to estimate Y.

Alternatives.—An alternative approach to logistic regression commonly used by ecologists is the angular transformation, which expands the range of values taken by p and serves to stabilize the variance. When sample sizes are unequal for various predictor values, weighted analysis is most appropriate. This approach is effective for large sample sizes and p ranging from 0.25 to 0.75. When n is small, the angular transform approximation overestimates the true variance with the extent of the overestimation increasing as p deviates from 0.5. The angular transformation loses information at extreme values of p ($p < 0.10$ and $p > 0.90$) (Cox 1970). Thus, the angular transformation should be avoided when small sample sizes or small and/or large proportions are analyzed (Finney 1964, Cox 1970).

The probit transformation, based on the cumulative normal distribution, is almost identical to the logit (McCullagh and Nelder 1989) over the interval $0.1 < p < 0.9$ and they can seldom be differentiated based on tests of the validity of their fit to data (Finney 1964: 466). Logit and probit regression differ in the shape of the function relating $\Pr(Y = 1 \mid \underset{\sim}{x})$ to $\underset{\sim}{x}$. Most authors argue for the logit based on its ease of application and interpretation (Finney 1964, Cox 1970, Fleiss 1973, Anderson et al. 1980, Cox and Snell 1989, McCullagh and Nelder 1989, Agresti 1990). The logistic function, which motivates the logit transformation, may be considered more general in its potential applications than the probit. The logistic function yields equivalent log odds ratios for several sampling models, while the probit does not (see McCullagh and Nelder 1989, Agresti 1990).

There are many transforming functions possible other than logit and probit. For example, both the logit and probit assume that p approaches 0.0 and 1.0 symmetrically. However, asymmetric distributions are possible. For example, Agresti (1990:104–107) discusses the complementary log-log function for cases where p approaches 1.0 more rapidly than 0.0. When logit or probit models fail to provide a good fit to data, other sampling distributions should be explored.

Logits can be fit using weighted least-squares, called empirical logistic regression (or minimum logit chi-squared when used in slope and point estimation for bioassays). Fitting empirical logits is outlined in Cox (1970) and Trexler et al. (1988). However, when many zeros are present, this transformation can yield incorrect results and it is best to use a maximum likelihood method.

Model selection.—Logistic regression requires fitting a hierarchy of models and settling on the most parsimonious one. The "adequacy" of a logit model is assessed in two ways, the significance of its parameters in explaining variation in the dependent variable and the fit of its predictions to data (goodness-of-fit). The contribution of parameters in a logistic regression model is assessed by comparing the predictive value of models with and without them. If the removal of a parameter does not noticeably decrease the predictive power of a model, the parameter can be excluded. Hosmer et al. (1989) describe a technique applying Mallow's statistic (C_p) to logit analysis in a stepwise model selection routine, and BMDP-LR is a statistical program that uses stepwise regression techniques in model building.

The ratio of the likelihood of a given model to that of the "saturated" model is called the likelihood ratio, where the saturated model has as many parameters as data points, comparable to a linear regression through two points (Hosmer and Lemeshow 1989). In order to assess the size (significance) of the likelihood ratio relative to a standard distribution, it must be re-expressed as a quantity called the deviance (D), given by:

$$D = -2 \cdot \ln(\text{likelihood ratio}). \qquad (8)$$

The difference of the deviance between two models, one with and the other without a parameter, tests the hypothesis that the excluded parameter is equal to zero and has a chi-square distribution with one degree of freedom (see Hosmer and Lemeshow [1989:224] for cases with polychotomous dependent variable). This hypothesis test is called the likelihood ratio test. If only the log likelihoods for the two models are known, the model with and the model without the parameter in question, the difference of the log likelihoods times -2 is equivalent to the likelihood ratio chi-square test.

Other tests are also used to evaluate the contribution of specific parameters to a model. The most common is Wald's test, which is the ratio of a parameter estimate to its standard error. The significance of this value can be assessed by comparison to a standard normal distribution. However, the validity of this test is in doubt (Hauk and Donner 1977, Jennings 1986) and it is best used in a qualitative way. The likelihood ratio test is the preferred method of parameter evaluation (Hosmer and Lemeshow 1989).

Goodness-of-fit assesses the relationship of the observed and predicted values of the dependent variables and the role of individual observations in fitting the model (i.e., the identification of outliers). The deviance statistic D (Eq. 8) is distributed as a chi-square with $J - (p_* + 1)$ degrees of freedom to test the hypothesis that the fitted model is adequate. In this case, J is the number of distinct combinations of all independent variables and p_* is the number of parameters in the model. There should be replication at some combinations of independent variables to apply this test (but see Hosmer and Lemeshow 1989:139). Analyses comparable to assessment of the "leverage" of individual observations in a least squares analysis are possible (e.g., Cook's statistic [Draper and Smith 1981]). In logistic regression, leverage assesses the contribution of individual independent variable combinations to the fit of a model. The change in a summary statistic such as D that results from deleting all subjects with a particular independent variable pattern yields a clear illustration of its contribution to the model's explanatory power (see Hosmer and Lemeshow 1989:154–168).

An example.—We illustrate the use of logit analysis with data on the probability of multiple paternity of broods of offspring produced by female livebearing fish (sailfin mollies, *Poecilia latipinna*) caught from natural populations. Pregnant females were collected and the genotypes of 24 of their embryos were determined for three allozymes. These data permitted a conservative assignment of females into two groups, those with broods sired by more than one male and those with broods sired by a single male. A discussion of the possible sources of error from this approach and estimation of the proportion of females misidentified as singly mated when actually multiply mated is provided in Travis et al. (1990). We analyzed those data with least-squares regression of brood size as the dependent variable and female size and multiple/single sire status as continuous and discrete independent variables, respectively. That analysis indicated that females carrying multiply sired broods tended to have larger brood sizes than same-sized females with singly sired broods (Travis et al. 1990). Here we report a re-analysis of those data addressing the question, "Is the probability of multiple paternity related to female size and/or fecundity?" This analysis demonstrates a discrete dependent variable, mating status as single or multiple,

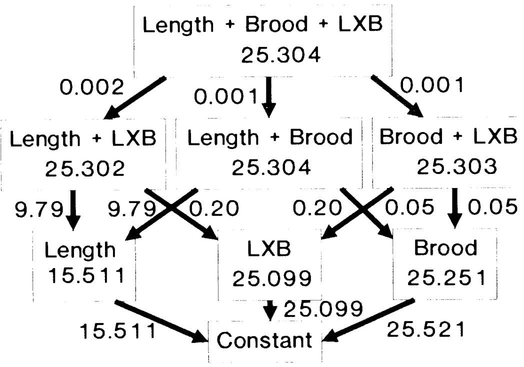

FIG. 1. A flow chart of the hierarchy of models fit to data on the odds of multiple paternity for broods of offspring from female sailfin mollies. The likelihood ratio chi-square for deletion of each parameter is listed on the arrow connecting models. Each of these chi-square values has one degree of freedom. The log likelihood for each model is enclosed within the box with its name. Length refers to female standard length, brood refers to brood size, and L × B refers to the interaction of length with brood size. Each model included a constant.

and continuous independent variables, brood size and female size.

Analysis of the 1985 data began by examining the distribution of the two independent variables for normality and outliers. In this case, both continuous variables were log transformed to normalize their distributions. Our logit analysis was conducted using the LOGIT module of SYSTAT and the BMDP package LR for comparison. The BMPD package provides a stepwise analysis of parameters, while models must be specified one at a time by the researcher with SYSTAT.

A hierarchy of models sequentially eliminating each variable was tested using SYSTAT to determine the most parsimonious list of variables needed to model the 1985 data. The log likelihood ratio chi-square test was used to evaluate the contribution of each parameter by taking differences of the log likelihood estimate for each model and multiplying by -2. The log likelihood ratio chi-square printed by SYSTAT compares the log likelihood of a given model to a model with its only parameter a constant. Note that the SYSTAT log likelihood ratio chi-square will tend to overestimate the significance of a given parameter if multicollinearity is present. Multicollinearity occurs when two independent variables are correlated and there is no way to distinguish their effects separately. Female size and brood size illustrate this in our 1985 data. When female size is removed from a model including both female size and brood size, the decrease in χ^2 is not significant (Length + Brood → Brood, Fig. 1). However, if only female size is in the model, it makes a significant contribution to predicting the mating status of females (χ^2_1, = 15.511, $P < .001$). Dropping either brood size or the length × brood interaction from a model including both has virtually identical effects suggesting multicollinearity for these factors (Brood + L × B → L × B and Brood + L × B → Brood, Fig. 1).

While both female size and brood size can explain differences in the probability that a female is singly or multiply mated, brood size is better at predicting mating status than is female size. When brood size is removed from the model including both brood size and female size, the decrease in log likelihood is significant ($\chi^2_1 = 9.79$, $P = .008$). However, when female size is removed from the model including both brood size and female size, the decrease in log likelihood is not significant ($\chi^2_1 = 0.052$, $P > .75$). This indicates that brood size explains variation in mating status beyond that explained by female size, but that the converse is not true.

The Wald's statistics for the model BROOD + CONSTANT support our conclusion that brood size is an important variable in explaining the probability of multiple mating. The t statistic for the comparison of the multiple paternity category to the single paternity category is -1.42 ($P = .08$). SYSTAT does not permit a test of the goodness-of-fit of each model or assessment of the influence of individual covariate patterns. We did this using the BMDP logistic regression package (LR). Neither the deviance chi-square test nor the other goodness-of-fit tests reported by BMDPLR suggested a lack of fit by a model containing a constant and brood size ($P \gg .5$ for tests of goodness-of-fit). Only 2 of 16 covariance patterns were found to yield a change in the deviance chi-square by 10% or more when excluded from analysis; in only one case did the change in deviance exceed 1 ($D = 5.248$ for the CONSTANT + BROOD model including all covariance patterns). In both of these cases the highly influential independent variable pattern is in the center of the brood size range [ln(brood size) = 3.75 and 3.80], near the point where the probability changes from high for single paternity to high for multiple paternity (Fig. 2). In one case, only one observation was present, rendering it impossible to observe an intermediate value for probability of single mating (predicted probability = 0.48, observed = one singly mated). In the second case, only two observations were present (predicted probability of single mating = 0.75, observed = one singly mated and one multiply mated). There is no biological reason to exclude these points so we leave them in the final analysis.

Further reading and applications.—Logistic regression has been proposed for several applications relevant to ecologists. These include analysis of selection experiments (Heisey 1985, Murtaugh 1988), capture–recapture data (Cormack 1981, Pollock et al. 1984, Green and Macdonald 1987), habitat selection (McCullagh and Nelder 1989:128–135), functional responses (Trexler et al. 1988), and competition (Schoener and Adler 1991). Other authors have discussed the use of log-linear models for applications that may be expanded to include covariates (Fienberg 1970, Holford 1980). Some recent ecological articles using logits include Horton et al. (1988), Hepp et al. (1989), Kaplan (1992), Rebertus et al. (1989), Stoddard (1987), Trexler (1985), and Wauters and Dhondt (1985). Several texts provide clear and informative discussions of logistic log-linear models and logistic regression. We have found Anderson et al. (1980), Hosmer and Lemeshow (1989), and Agresti (1990) to provide clear and informative discussions of this topic for nonstatisticians. Statistical programs including SPSSX, SAS, BMDP, and SYSTAT provide options permitting the use of logistic regression with parameter estimation by maximum likelihood.

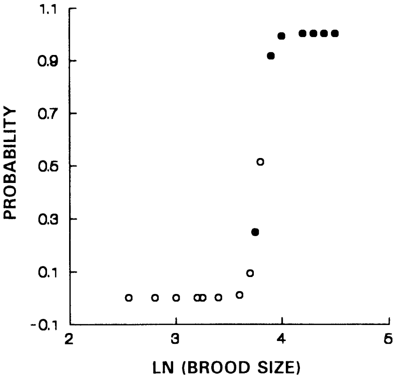

FIG. 2. Probability of multiple paternity predicted from logistic regression plotted against observed brood size. ● indicate broods shown to result from multiple matings and ○ indicate broods identified as resulting from a single mating.

LOCALLY WEIGHTED REGRESSIONS

In many cases it is unreasonable to expect a priori that a single functional description of the dependence of Y on X will suffice throughout the range of X that is being examined. Consider as an example the situation in which anuran larvae are exposed to size-limited predators. Within some range of larval growth rates, the faster the larva grows, the sooner it attains a size beyond which it cannot be eaten. As a result of this process, there is directional selection on larval growth rate through some range of growth rates such that the probability of survival will be an increasing function of larval growth rate. However, there may be a range of growth rates that are all so low that there is no chance of survival, and there may be another range of growth rates that are all sufficiently high that they convey equivalently high probabilities of survival. Thus a single functional relationship may not suffice through the full range of observable larval growth rates. It is of real interest to identify such heterogeneity if it exists because we would then be able to identify regions of growth-rate variation that are effectively neutral with respect to natural selection. Extensive ranges of such effective neutrality can alter the conclusions one draws from such data on the intensity of natural selection. How can we perform such a diagnosis?

A number of methods are available for this purpose, but the skeptic may first wonder why one would not fit a polynomial model to such data, particularly because higher order polynomials can be bent into almost any shape to mimic almost any pattern of data. Four answers can be offered. First, data shaped oddly (e.g., sigmoidal trends in Y) can usually be fit by a wide variety of models that differ in number of parameters and functional forms, leading to a crisis of confidence in what interpretation can be drawn from any one such fitted model (see Trexler et al. 1988 for examples). Second, polynomial regressions do not fit values locally and are not very flexible in shape; the shape of the data at the larger values of X determines the fitted estimates at the smaller values of X. Higher order polynomials can often be employed to obtain a function that fits the local shape of the data, but this method does not always work. Third, the shape of polynomial regressions can be quite sensitive to outliers, points with unusually high leverage or influence, and therefore are not very robust. Fourth, better methods are available that circumvent these problems.

Rationale.—The method that we consider here derives the predicted value of Y (denoted $\hat{y}_i$) at a given X value (x_i) from a function of the Y values that are associated with the neighboring values of that X value. This procedure assumes that the Y values of neighboring X values will be the best indicators of what the Y value should be at the given X value. This is a sensible assumption; the key problems with making the assumption operational are how many neighboring X values should be considered and how the Y values that correspond to neighboring X values should be weighted in the calculation of the predicted value of Y.

Many readers will be familiar with this general premise through their familiarity with "running averages" in time series. Time series are smoothed by replacement of the observed value y_i at x_i by the average of y_{i-1}, $y_{i.}$, and y_{i+1}. In this case, we use the Y values at only immediately adjacent X values and weight them evenly in calculating the "predicted" value of Y. We could give more weight to y_i and proportionately less weight to y_{i-1} and y_{i+1}, for example using weights of 0.25, 0.50, and 0.25 for the sequential Y values. We might prefer a "smoother" fit in some cases, in which we would use more points adjacent to our focal value

x_i in the calculation of $\hat{y}_i$ and devise an appropriate weighting scheme.

This specific method will usually be inapplicable to most bivariate plots that are not time series because the X values will not be ordinal, equally spaced indices. A number of methods are designed for this specific problem; they take into account the uneven spacing of X values and the information contained by the X values themselves. We highlight the use of one such technique called "LOWESS" regression (Locally Weighted Regression Scatterplot Smoothing: Cleveland 1979). Other methods, such as cubic splines, may be more appropriate in many contexts (see Schluter 1988). Our purpose here is to highlight the utility of this class of models so that readers will be motivated to investigate them more widely.

Method.—The fitting of a line to a bivariate scatter of n points in LOWESS regression is done in a series of iterated steps. First, the user must decide how "smooth" the fitted relationship should be. A smoother fit uses more of the nearby points in its estimation of $\hat{y}_i$ for a given x_i, whereas a less smooth fit uses fewer points and is thereby more sensitive to local variation in the shape of the relationship. Let q be the number of adjacent points to be used in the estimation procedure such that $q/n = f$, the fraction of all points used to derive an estimate of $\hat{y}_i$ for each x_i. As f increases, the fitted line will be smoother; $f = 1$ corresponds to standard linear regression. In the second step, each point to be used in the estimation is given a weight. The focal point has the largest weight, and the weights decrease symmetrically around the focal X value; points outside the "frame" (i.e., those points not among the q nearest neighbors to the focal point along the X axis) have weight 0. The weight given to a point x_k for a focal point x_i, denoted $w_i(x_k)$, is computed as

$$w_i(x_k) = T[(x_i - x_k)/d],$$

where

$T(u) = (1 - |u|^3)^3$ for $|u| < 1$
$T(u) = 0$ elsewhere
d = distance from x_i to its q^{th} nearest neighbor, calculated only along the X-axis.

In the third step, a simple linear regression is fitted to the q points by weighted least squares, using the weights derived in step 2. Fourth, the estimate $\hat{y}_i$ is computed for the focal point x_i as the fitted value of Y from the regression. This procedure is then repeated for the next point until all n points have estimated Y values. Note that points at either extreme of the independent value have most of their q neighbors only on one side; this pattern prevents the fitted values from forcibly describing a flat line at the extremes.

A "first fit" is now in hand, but the fit is not necessarily robust; outliers may be exercising inordinate influence on the line. The first step in the next stage, that of obtaining a robust fit, is to calculate the residuals from the fitted values

$$r_i = y_i - \hat{y}_i.$$

Let m designate the median of the absolute values of the residuals. If the residuals are normally distributed, m is $\approx 2/3$ the value of the root mean square error around the regression, and $6m$ is approximately 4 times the root mean square error. Second, robustness weights, denoted rw_i, are calculated for each point as

$$rw_i = B(r_i/6m),$$

where

$B(u) = (1 - u^2)^2$ for $|u| < 1$
$B(u) = 0$ elsewhere.

These values give greater weight to points with low residuals and deemphasize those points with high residuals. A residual value in excess of $6m$ produces a robustness weight that is nearly 0, and values $\ll 6m$ have weights nearly equal to 1. Third, the LOWESS procedure is repeated, but this time the weights used for the weighted least squares for each point are provided by the product of the robustness weights and the original neighborhood weights. This procedure can be repeated until there is minimal change in the final product.

Example.—In order to illustrate the technique, we reexamine data from an experimental study of natural selection (Travis 1983). Each datum consists of the average body size of full-sib tadpoles of *Hyla gratiosa* at age 4 wk from a single enclosure in the field and the fraction of the original number of tadpoles placed in that enclosure that survived to age 4 wk. Tadpoles were exposed to insect predators that cannot prey on tadpoles beyond a specific body size and that therefore should take a greater toll on groups of more slowly growing tadpoles. A number of full-sib families were represented in the experiment, and the published result (repeated with another species by Travis et al. 1985) indicated that there was a positive relationship between the average size of tadpoles in an enclosure and the survival rate that occurred in that enclosure, in other words that when a cohort had a higher growth rate it suffered lower cumulative mortality.

The bivariate plot of arcsine-transformed survival values and larval body masses is illustrated with a linear regression line in Fig. 3A. The linear regression is significant ($F_{1,16} = 19.6$, $P < .001$) and accounts for $\approx 52\%$ of the variance in transformed survival values. Visual inspection suggests that, although there is a good relationship through the lower half of the larval body masses, there is little relationship at the higher masses.

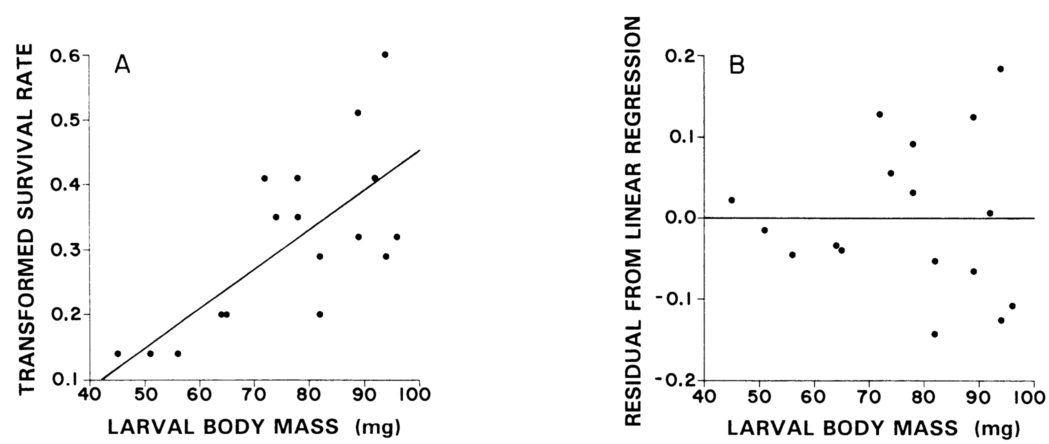

FIG. 3. (A) Survival rate (angularly transformed) of tadpoles in a single enclosure plotted as a function of the average body mass of the survivors in the enclosure. Data from Travis (1983). Line indicates the normal least-squares regression. (B) Residuals from the linear regression depicted in Part A plotted as a function of the independent variable, average body mass.

The graph suggests that at higher body masses variance in survival rates is large and reflects the influence of factors other than insect predation. The linear regression captures the major trend in the data but does not indicate a region of effective neutrality.

The inadequacy of the fit of the linear regression can be seen in a plot of residuals against the independent variable (Fig. 3B); the flat regression line drawn through that plot indicates an unbiased fit but one in which the variance around the line increases with increasing mean body masses. No transformation was found to alleviate this problem. A quadratic regression (Fig. 4A) provides a hint that the fitness function will become flatter at higher body masses but does not indicate much of a neutral region within the range of body masses exhibited in the study. The fit of a LOWESS regression with $f = 0.67$ (Fig. 4B) shows that survival rates increase steadily with increases in larval body mass up to the 75-mg level, beyond which the increase is minimal at best. This fit indicates that growth rates that produce body masses >75 mg at age 4 wk convey no increases in survival rates, indicating the region of effective neutrality of growth rate with respect to insect predation. A LOWESS regression with more local sensitivity, $f = 0.33$ (Fig. 4C), produces the same general impression; the fitted line becomes irregular and inconsistent in direction at growth rates >75 mg.

Choosing the parameters of LOWESS.—The LOWESS method appears to carry the burden of considerable subjectivity. However, analysis and simulation studies offer substantial comfort for those who are skeptical on these grounds (Cleveland 1979). First, it appears that the use of linear regression within each set of neighboring points to estimate $\hat{y}_i$ provides an effective balance between ease of computation and the ability to reproduce known patterns. Second, two it-

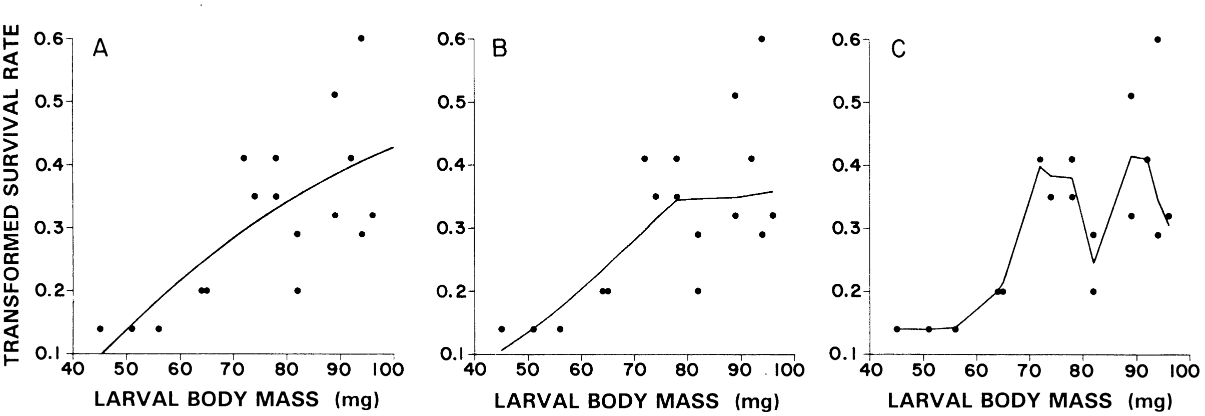

FIG. 4. (A) Data from Fig. 3A with a line depicting a least-squares quadratic model. (B) Data from Fig. 3A with a line depicting a LOWESS regression model with $f = 0.67$. (C) Data from Fig. 3A with a line depicting a LOWESS regression model with $f = 0.33$.

erations (as illustrated above) have proven adequate to provide a sufficient convergence on a set of $\hat{y}_i$ for a large number of real and simulated data sets. Although other weighting functions could be used, the "tricube" used above enhances the accuracy of the distributional approximation for calculating the standard errors of the $\hat{y}_i$.

The greatest subjectivity lies in choosing f: how "smooth" should the fit be made? Either of two methods for choosing f seems effective. The first is a least-squares criterion (Cleveland 1979). An initial value of f, f_0, is chosen to minimize the sum of the squared deviations from the set of $\hat{y}_i$ before the robustness step is employed. The robustness weights are then calculated, and a new value of f is chosen so as to minimize the sum of the product of the robustness weights and the residuals for the set of $\hat{y}_i$ values. This procedure is repeated until a negligible change in f is observed.

The second method is much simpler (Cleveland 1985). First, choose an arbitrary value of f and calculate the robust LOWESS estimates $\hat{y}_i$ and the residuals from those estimates, $r_i = y_i - \hat{y}_i$. If the plot of the residuals, r_i, against the x_i shows any dependence of r_i on x_i, then f is too large and should be decreased. In practice, one ought therefore to begin with a small value of f and increase it slowly to the point at which the residuals show a pattern, from which a slightly smaller f is finally chosen.

Statistical testing.—The standard errors of the $\hat{y}_i$ can be estimated in three ways. First, Cleveland (1979) offers analytical formulas for the error that are based on either a chi-square approximation for the error distribution or a modification of other robust methods. Second, one can employ maximum likelihood methods derived explicitly for this problem (Cleveland 1981). Third, error estimation and more customized statistical testing can be achieved through computer-intensive nonparametric methods such as bootstrapping and cross-validation (Chambers et al. 1983, Efron and Gong 1983, Efron and Tibshirani 1991).

Further reading.—Chambers et al. (1983) and Cleveland (1985) offer a very accessible introduction to LOWESS regression and a variety of other graphical methods. The general problem of robustness is addressed in detailed fashion in the treatments of Huber (1981) and Hampel et al. (1986). Although some treatments of robust regressions are available in a number of computer software packages, some of the computations in some packages do not always give the correct results (Street et al. 1988).

Acknowledgments

We thank Drs. Charles McCulloch and Deborah Burr for clarifying some issues associated with logistic regression analysis.

Literature Cited

Agresti, A. 1990. Categorical data analysis. John Wiley & Sons, New York, New York, USA.

Anderson, S., A. Auquier, W. W. Hauck, D. Oakes, W. Vandaele, and H. I. Weisberg. 1980. Statistical methods for comparative studies. John Wiley & Sons, New York, New York, USA.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. Graphical methods for data analysis. Wadsworth International Group, Belmont, California, USA.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association **74**:829–836.

———. 1981. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. American Statistician **35**:54.

———. 1985. The elements of graphing data. Wadsworth, Monterey, California, USA.

Cormack, R. M. 1981. Log-linear models for capture–recapture experiments on open populations. Pages 197–215 *in* R. W. Hiorns and D. Cooke, editors. The mathematical theory of the dynamics of biological populations II. Academic Press, London, England.

Cox, D. R. 1970. The analysis of binary data. Methuen, London, England.

Cox, D. R., and E. J. Snell. 1989. Analysis of binary data. Second edition. Chapman and Hall, New York, New York, USA.

Draper, N., and H. Smith. 1981. Applied regression analysis. Second edition. John Wiley & Sons, New York, New York, USA.

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, jackknife, and cross-validation. American Statistician **37**:36–48.

Efron, B., and R. Tibshirani. 1991. Statistical data analysis in the computer age. Science **253**:390–395.

Fienberg, S. E. 1970. The analysis of multidimensional contingency tables. Ecology **51**:419–433.

———. 1980. The analysis of cross-classified categorical data. Second edition. MIT Press, Cambridge, Massachusetts, USA.

Finney, D. J. 1964. Statistical method in biological assay. Hafner, New York, New York, USA.

Fleiss, J. L. 1973. Statistical methods for rates and proportions. John Wiley & Sons, New York, New York, USA.

Green, P. E. J., and P. D. M. Macdonald. 1987. Analysis of mark–recapture data from hatchery-raised salmon using log-linear models. Canadian Journal of Fisheries and Aquatic Sciences **44**:316–326.

Hampel, F. R., E. Ronchetti, P. J. Rousseeuw, and W. Stahel. 1986. Robust statistics: the approach based on influence functions. John Wiley & Sons, New York, New York, USA.

Hauck, W. W., and A. Donner. 1977. Wald's test as applied to hypotheses in logit analysis. JASA **72**:851–853.

Heisey, D. M. 1985. Analyzing selection experiments with log-linear models. Ecology **66**:1744–1748.

Hepp, G. R., R. A. Kennamer, and W. F. Harvey, IV. 1989. Recruitment and natal philopatry of Wood Ducks. Ecology **70**:897–903.

Holford, T. R. 1980. The analysis of rates and survivorship using log-linear models. Biometrics **36**:299–305.

Horton, D. R., J. L. Capinera, and P. L. Chapman. 1988. Local differences in host use by two populations of the Colorado potato beetle. Ecology **69**:823–831.

Hosmer, D. W., Jr., B. Jovanovic, and S. Lemeshow. 1989. Best subsets logistic regression. Biometrics **45**:1265–1270.

Hosmer, D. W., Jr., and S. Lemeshow. 1989. Applied logistic regression. John Wiley & Sons, New York, New York, USA.

Huber, P. J. 1981. Robust statistics. John Wiley & Sons, New York, New York, USA.

Jennings, D. E. 1986. Judging inference adequacy in logistic regression. JASA **81**:471–476.

Juliano, S. A., and F. M. Williams. 1987. A comparison of methods for estimating the functional response parameters of the random predator equation. Journal of Animal Ecology **56**:641–653.

Kaplan, R. H. 1992. Greater maternal investment can decrease offspring survival in the frog *Bombina orientalis.* Ecology **73**:280–288.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Chapman and Hall, New York, New York, USA.

Murtaugh, P. A. 1988. Use of logistic regression in modelling prey selection by *Neomysis mercedis.* Ecological Modelling **43**:225–233.

Pacala, S. W., and A. P. Dobson. 1988. The relation between the number of parasites/host and host age: population dynamic causes and maximum likelihood estimation. Parasitology **96**:197–210.

Pollock, K. W., J. E. Hines, and J. D. Nichols. 1984. The use of auxiliary variables in capture recapture and removal experiments. Biometrics **40**:325–340.

Rebertus, A. J., G. B. Williamson, and E. B. Moser. 1989. Longleaf pine pyrogenicity and turkey oak mortality in Florida xeric sandhills. Ecology **70**:60–70.

Schluter, D. 1988. Estimating the form of natural selection on a quantitative trait. Evolution **42**:849–861.

Schoener, T. W., and G. H. Adler. 1991. Greater resolution of distributional complementarities by controlling for habitat affinities: a study with Bahamian lizards and birds. American Naturalist **137**:669–692.

Shanubhogue, A., and A. P. Gore. 1987. Using logistic regression in ecology. Current Science **56**:933–935.

Stoddard, J. L. 1987. Microcrustacean communities of high-elevation lakes in the Sierra Nevada, California. Journal of Plankton Research **9**:631–650.

Street, J. O., R. J. Carroll, and D. Ruppert. 1988. A note on computing robust regression estimates via iteratively reweighted least squares. American Statistician **42**:152–154.

Travis, J. 1983. Variation in growth and survival of *Hyla gratiosa* larvae in experimental enclosures. Copeia 1983: 232–237.

Travis, J., W. H. Keen, and J. Juilianna. 1985. The effects of multiple factors on viability selection in *Hyla gratiosa* tadpoles. Evolution **39**:1087–1099.

Travis, J., J. C. Trexler, and M. Mulvey. 1990. Multiple paternity and its correlates in female *Poecilia latipinna* (Poeciliidae). Copeia 1990:722–729.

Trexler, J. C. 1985. Density-dependent parasitism by a eulophid parasitoid: tests of an intragenerational hypothesis. Oikos **44**:415–422.

Trexler, J. C., C. E. McCulloch, and J. Travis. 1988. How can the functional response best be determined? Oecologia (Berlin) **76**:206–214.

Wauters, L., and A. A. Dhondt. 1985. Body weight, longevity and reproductive success in red squirrels (*Sciurus vulgaris*). Journal of Animal Ecology **58**:837–851.