Vol. 143, No. 3 The American Naturalist March 1994

MODELING OF DISCONTINUOUS RELATIONSHIPS IN BIOLOGY WITH CENSORED REGRESSION

BERNHARD SCHMID,* WOLFGANG POLASEK,† JACOB WEINER,‡ ANDREAS KRAUSE,† AND PETER STOLL*§

*Program in Conservation Biology (NLU), Botanisches Institut, Universität Basel, Schönbeinstrasse 6, CH-4056 Basel, Switzerland; †Institut für Statistik und Ökonometrie, Universität Basel, Petersgraben 51, CH-4051 Basel, Switzerland; ‡Department of Biology, Swarthmore College, Swarthmore, Pennsylvania 19081

Submitted November 9, 1992; Revised June 7, 1993; Accepted June 11, 1993

Abstract.—Discontinuous relationships between variables are common in biological data. Discontinuities can sometimes give the appearance of curvilinearity, which suggests the data should be analyzed with nonlinear models. Here we show that often a more meaningful analysis can be obtained with censored regression techniques. In a censoring model all points below (or above) a certain threshold are observed only by the value of the threshold (e.g., a baseline temperature). We illustrate the method with an example from plant reproductive biology: plant reproductive mass is never negative but becomes positive only after some "capacity" to flower reaches a threshold. The vegetative mass at which the threshold is reached and the relationship between reproductive mass and vegetative mass above the threshold are estimated from data. Using censored regression with real and simulated data shows that apparent curvilinearity suggested by models that do not account for censoring can be an artifact.

Relationships between biological variables are often discontinuous if a large range of values is considered. For example, a dependent variable may reach an upper (or lower) bound above (or below) certain values of an independent variable; in such a case the dependent variable appears to be censored. Such relationships can be analyzed by censored regression models. Biologists have done this for survival data when the full life span could not be observed for all individuals (Buckley and James 1979; Aitkin and Clayton 1980; Oakes 1986; Petersen 1986, Schneider and Weissfeld 1986; Segal 1988). Censored regression models are, however, rarely used in cases in which there are biological rather than sampling restrictions for a dependent variable (but see Taylor 1973; Wolynetz 1979*a*, 1979*b*; Miller and Halpern 1982). For example, certain characters of organisms only develop when an "ability" or "capacity," whose material basis may be unknown (internal resources, hormones, morphogens), reaches a certain level (fig. 1*A*, *B*); other characters may reach physical limits (fig. 1*C*); and many physiological processes have "baselines" or maximum rates (fig. 1*D*). Further examples include reproductive output (discussed below), canalization of phenotypic charac-

§ E-mail: B.S., schmid3@urz.unibas.ch; W.P., wolfgang@iso.wwz.unibas.ch; A.K., andreas @iso.wwz.unibas.ch; P.S., stoll@urz.unibas.ch.

Am. Nat. 1994. Vol. 143, pp. 494–507.

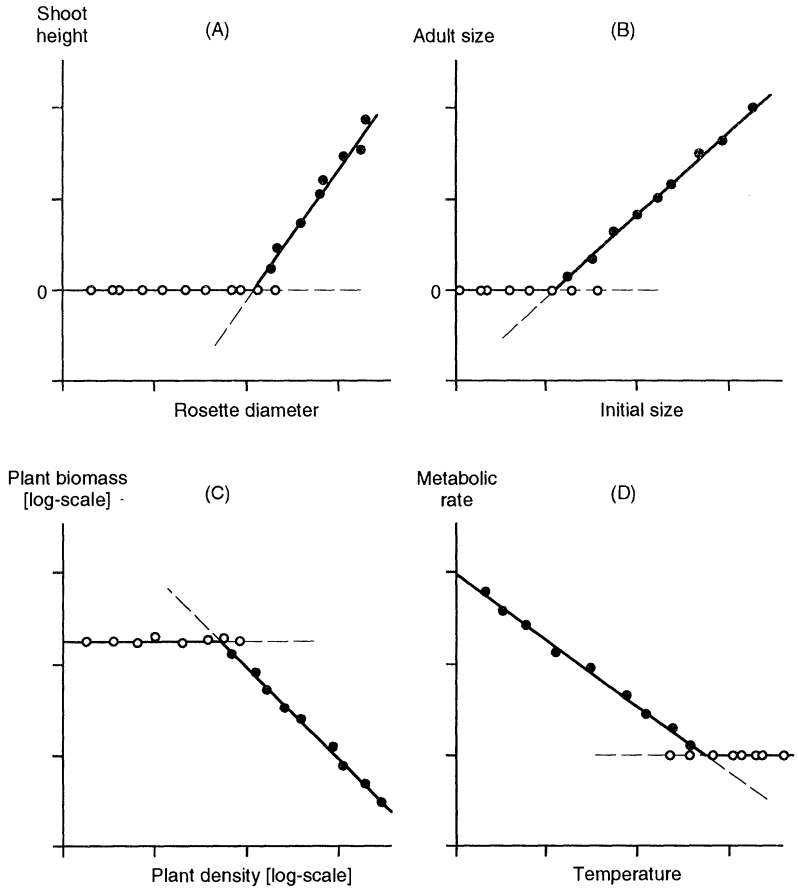

FIG. 1.—Idealized sketches of some discontinuous relationships in biology. *A*, Rosette plants must reach a certain size before the stem bolts and the height can be observed (cf. Werner 1975); *B*, plants or animals of very small initial size (e.g., seed mass or mass at birth) do not have enough starting capital to survive (i.e., to produce positive adult sizes; cf. Thomas and Weiner 1989); *C*, plant size is reduced under intra- or interspecific competition (increasing density), but no competition occurs at low density where all plants can reach the maximum size determined by genotype and competition-free environment (Harper 1977); *D*, basal metabolic rate of homeothermic animals increases with decreasing ambient temperature, but above a certain ambient temperature metabolic rate is constant (cf. West 1972).

ters, prevalence of disease, and behavioral responses. Here we wish to draw attention to statistical methods that have been developed in econometrics for the analysis of censored observation (Amemiya 1985).

We believe that censored regression models can solve many statistical problems that have been encountered in the analysis of discontinuous biological relationships. These statistical problems may have led to unnecessary revisions of biological models. In the case of plant size-density relationships, for example, a linear model on the log-log scale often seems to best represent the underlying biological processes over a certain range of densities (fig. 1*C*; see extensive dis-

cussion of this subject and references in Harper 1977). However, the linear model is now rarely being used because biologically plausible models for curvilinear size-density relationships have been developed (Watkinson 1980, 1986; Vandermeer 1984; Pacala and Silander 1985). In data sets of moderate size a curvilinear model can cover up an existing discontinuity. Below we use another example, in which linear and curvilinear relationships have been proposed, to demonstrate how the censored regression approach can alter our interpretation of data. The example we have chosen is the size dependence of reproductive output in plants.

AN EXAMPLE: PLANT REPRODUCTIVE MASS

Although reproductive allocation in plants has traditionally been defined in terms of the proportion of a plant's total biomass that is in reproductive tissues (see, e.g., Bazzaz and Reekie 1985), the observation that reproductive allocation is often size-dependent has led researchers to study directly the relationship between reproductive output and plant size (Samson and Werk 1986; Weiner 1988; Klinkhamer et al. 1990, 1992). To explain changes in reproductive allocation in response to competition, Weiner (1988) proposed a simple model in which the relationship between reproductive mass (y) and vegetative mass (x) is linear and has a positive x-intercept (i.e., there is a minimum size for reproduction). Evidence in support of this model has been found in four species of clonal composites (Hartnett 1990) and several species of agricultural weeds (Thompson et al. 1991).

Weiner's (1988) model is based on an analogy between a biological plant that produces inflorescences, fruits, and seeds and an industrial plant (factory) that produces goods. The model makes the following predictions about size-dependent reproductive output in plants. First, significant capital investment is required before there can be any sexual reproduction. Therefore, reproductive mass must be zero for very small plants. Second, as capital investment increases with plant size, the capacity to reproduce increases and reaches a threshold at a certain size. Finally, above this minimum size for reproduction, a constant proportion of additional capital can be directly allocated to reproduction, which will lead to a linear relationship between reproductive mass and vegetative mass. This model has been analyzed by normal regression analysis (Samson and Werk 1986; Weiner 1988; Hartnett 1990; Thompson et al. 1991). Curvilinear relationships between size and reproductive output have also been proposed (Reiss 1989; Klinkhamer et al. 1992).

We use one of seven data sets on size-dependent reproduction in the clonal plant *Solidago altissima* L. (tall goldenrod) to evaluate these models under the assumption of a biological censoring mechanism. Figure 2 shows a scatter plot of the reproductive mass r (inflorescences) against the vegetative mass v (stem + leaves) for the 2,545 individually grown shoots. Those with symbol Y belong to the data set arbitrarily chosen for the analysis. It can be seen that the data are effectively split into two parts: censored data points for which only v could be observed and r is zero, and uncensored data points for which both r and v could be observed. If we imagine that the plants for which we set $r = 0$ in fact would have a "reproductive debt" that must be paid back by capital investment, we

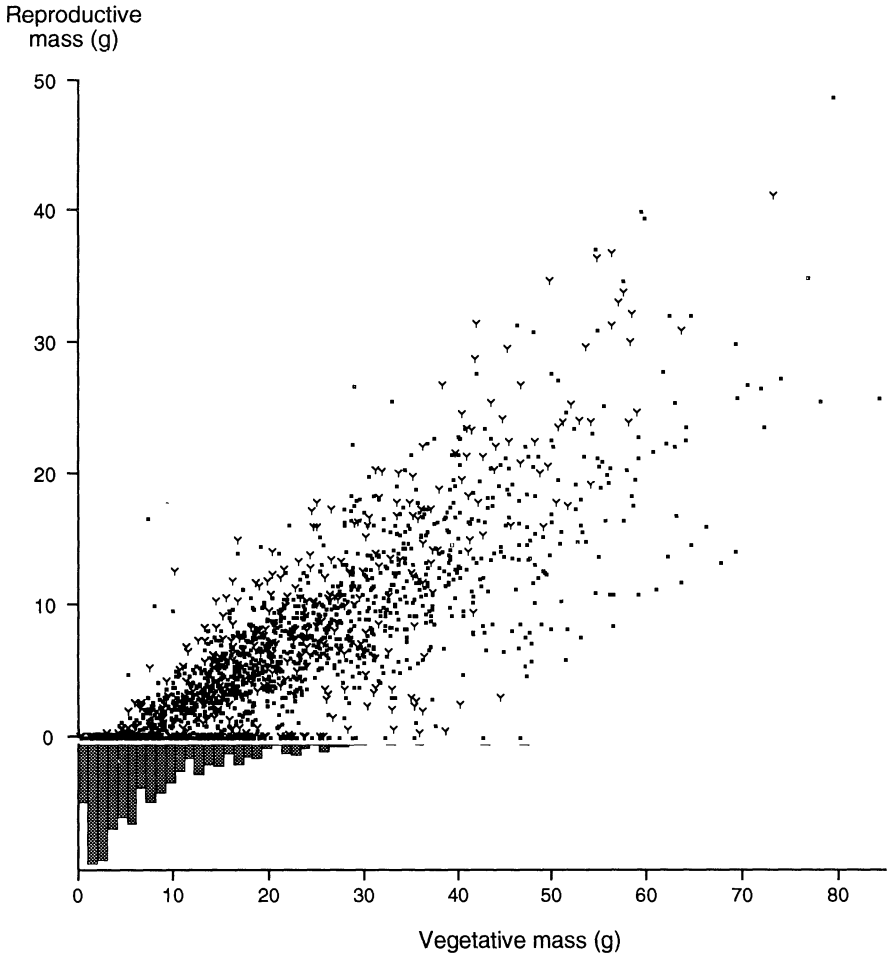

FIG. 2.—Scatter plot of plant reproductive vs. vegetative mass in *Solidago altissima* (seven data sets, total $n = 2{,}545$). The histogram shows the number of plants with zero reproductive mass (688 points clustered on the X-axis); only the points marked with large Y's are analyzed in this article (see Dolt 1991 and Schmid and Weiner 1993 for descriptions of the experiments).

could try to interpret them as plants with a "negative reproductive mass." This is what the censored regression analysis does.

STATISTICAL MODELS

We use uppercase letters for random variables, lowercase letters for fixed data or observed random variables, Greek letters for parameters, and a circumflex ($\hat{}$) for estimates. Before implementing the censored regression model, we briefly review the statistical methods that have previously been used for the analysis of size-dependent reproduction in plants. We begin with the linear regression model

(Samson and Werk 1986):

$$r_i = \alpha + \beta \cdot v_i + u_i = \beta \cdot (v_i - \alpha') + u_i, \quad i = 1, \ldots, n. \tag{1}$$

Here r_i and v_i represent reproductive and vegetative masses of the random variable R and the independent variable v, respectively, α is the intercept on the Y-axis, β is the slope parameter, $\alpha' = -\alpha/\beta$ is the intercept on the X-axis (threshold), and the u_i are the normally distributed errors. Because the zero values of R (censored dependent observations) violate the assumption of a regression model with normally distributed errors, we may simply choose to exclude zeros from the analysis (Weiner 1988; Schmid and Weiner 1993). However, such an exclusion reduces the amount of information in the data. It can be shown analytically (Amemiya 1985) or by simulation (table 1) that this leads to a negative bias in the estimate ($\hat{\beta}$) of β.

If we find a significant lack of fit for model (1) using a general linear test (see, e.g., Neter and Wassermann 1974), we may try an allometric model (Klinkhamer et al. 1992):

$$r_i = \alpha + \beta \cdot v_i^{\gamma} + u_i, \tag{2a}$$

or, reparameterized,

$$r_i = \beta' \cdot (v_i - \alpha')^{\gamma'} + u_i', \quad i = 1, \ldots, n. \tag{2b}$$

If the allometric exponent γ (or γ') is smaller than one, increases in vegetative mass produce diminishing increases in reproductive mass (decreasing returns to scale in the economic analogy); if $\gamma > 1$ (or $\gamma' > 1$), increases in vegetative mass produce accelerating increases in reproductive mass (increasing returns to scale). Because α or α' may differ from zero, model (2) cannot in general be fitted by linear regression of $\log(r_i)$ on $\log(v_i)$ (cf. Klinkhamer et al. 1990). If the model with the intercept α' on the X-axis is chosen, we need the restriction $v_i > \hat{\alpha}'$. (In the allometric model α' cannot, of course, be calculated as $-\alpha/\beta$.) A simple remedy would be to exclude zeros from the analysis. In this case the estimate $\hat{\beta}$ has a strong negative bias, and the estimate $\hat{\gamma}$ has a strong positive bias (see table 1). As a consequence, a true allometric relationship with $\gamma < 1$ will too often appear to be a simple linear relationship, whereas a true linear relationship with $\gamma = 1$ will too often appear to be an allometric relationship with an exponent greater than one.

The observations with $R = 0$ represent the censored data points in the statistical analysis. Obviously, these data points contain valuable information about the reproductive behavior of plants. It may be argued that the minimum size for reproduction should be defined as the vegetative mass θ, at which 50% of all individuals in a population flower, irrespective of the reproductive mass of those that do flower (cf. Werner 1975; Gross 1981; Meagher and Antonovics 1982; Primack and Hall 1990). The parameter θ can be estimated by logistic regression of W (the binary random variable for censoring; see below) on v or, preferably, by probit analysis (Finney 1971; McCullagh and Nelder 1989):

$$p_i = \Phi[\alpha + \beta \cdot v_i] + u_i = \Phi[\beta \cdot (v_i - \theta)] + u_i, \quad i = 1, \ldots, n, \tag{3}$$

TABLE 1

PARAMETER ESTIMATES FOR THE RELATIONSHIP BETWEEN THEORETICAL PLANT REPRODUCTIVE MASS (R) AND VEGETATIVE MASS (v) IN SIMULATED DATA SETS, CALCULATED FROM ORDINARY OR CENSORED LINEAR REGRESSION AND ALLOMETRIC MODELS

	"LINEAR SIMULATION" PROPORTIONALITY FACTOR (SLOPE) IN LINEAR REGRESSION		"NONLINEAR SIMULATION" ALLOMETRIC EXPONENT IN ALLOMETRIC MODEL	
SIMULATED RELATIONSHIP*	Ordinary	Censored	Ordinary	Censored
$R = -4 + .4 \cdot v^{1.0} + U$ (31–49 censored)	.3946 ± .0004	.3999 ± .0004	1.075 ± .004	.997 ± .003
$R = -16 + .8 \cdot v^{1.0} + U$ (75–85 censored)	.7964 ± .0004	.7999 ± .0004	1.032 ± .003	.999 ± .002
$R = -3 + .8 \cdot v^{.8} + U$ (15–29 censored)	. . .	. . .	.867 ± .004	.797 ± .003
$R = -10 + 1.4 \cdot v^{.8} + U$ (40–53 censored)	. . .	. . .	.841 ± .003	.799 ± .002
$R = -4 + .2 \cdot v^{1.2} + U$ (38–57 censored)	. . .	. . .	1.267 ± .003	1.197 ± .003
$R = -19 + .4 \cdot v^{1.2} + U$ (93–104 censored)	. . .	. . .	1.227 ± .003	1.198 ± .003

NOTE.—Six parameter combinations that gave curves that might have been observed in real data are presented; 400 simulations were run for each parameter combination; values for v are designed as 0.125, 0.25, 0.375, 0.5, . . . , 50; U was drawn from random normal numbers with mean zero and variance three; all simulated R values that were not positive were set to zero (i.e., censored) before analyses began (the range of the number of censored data points obtained in the simulations is given in parentheses). Note that all estimates from ordinary analyses deviate by more than 10–20 SE from the true parameters, whereas none of the estimates from censored analyses deviate by more than 1 SE from the true parameters.

* Negative r_i set to zero.

where p_i is the probability for plant i with vegetative mass v_i to reproduce, Φ is the distribution function of the standard normal variable, $\theta = -\alpha/\beta$, and u_i is the binomially distributed error. Here we consider vegetative mass v an independent explanatory variable, although it might be more appropriate to treat it also as a random variable V measured with errors (compare The "Errors-in-Variables" Problem section below).

Probit model (3) can be combined with model (1) or (2) to use all the information in the data for estimating the minimum size for reproduction and the slope of the relationship between reproductive and vegetative mass. This leads to censored regression models referred to as Tobit models in econometrics (Amemiya 1985). We imagine that there exists a normally distributed latent random variable R^* for which no values can be observed at or below zero. In the example the unobserved variable R^* may be viewed as a latent reproductive capacity defined as gross reproductive mass minus gross reproductive mass at the minimum size. The notion of gross reproductive mass implies that even in a plant too small to reproduce a certain amount of its vegetative mass can be viewed as capital invested for reproduction. Only when this capital investment is greater than some threshold, which occurs in plants above the minimum size, is actual reproduction observed. In statistical terms, whenever the dependent latent variable R_i^* is positive, then it is equal to the ("net") reproductive mass r_i. If the latent variable R_i^* is negative, then the reproductive mass r_i equals zero:

$$\begin{aligned} R_i^* &= \alpha + \beta \cdot v_i + u_i,\ u_i \sim N(0, \sigma_u^2), \quad i = 1, \ldots, n. \\ r_i &= R_i \quad \text{if } R_i^* > 0 \quad \text{(uncensored)}, \\ r_i &= 0 \quad \text{if } R_i^* \leq 0 \quad \text{(censored)}. \end{aligned} \tag{4}$$

If we define W as the random variable indicating censoring ($w_i = 1$ for uncensored and $w_i = 0$ for censored data points), then the likelihood function of the censored regression model is

$$L(\alpha, \beta, \sigma) = \prod_i \{1 - \Phi[(\alpha + \beta \cdot v_i)/\sigma]\}^{1-w_i} \cdot \{1/\sigma \cdot \phi[(r_i - \alpha - \beta \cdot v_i)/\sigma]\}^{w_i},$$

where ϕ is the density function of the standard normal variable and $\sigma = \sqrt{\text{var}(u)}$. As in the previous models, the parameters α, β, and σ^2 can be estimated by maximizing the logarithm of the likelihood function. However, in contrast to the previous models or regression models with exponential distributions and right-censoring commonly used in survival analysis, the log-likelihood function of this model has no simple analytic form and therefore must be maximized numerically (Amemiya 1985; cf. Aitkin et al. 1989). This calculation can be done iteratively using the so-called expectation maximization (EM) algorithm (Dempster et al. 1977). The EM algorithm for censored regression uses the mean and the variance of the censored normal distribution:

$$E(R_i^* \mid w_i = 0; \hat{\alpha}, \hat{\beta}, \hat{\sigma}) = \hat{r}_i - \hat{\sigma} \cdot h(\hat{r}_i/\hat{\sigma}),$$

where $h(x)$ is the function $\phi(x)/[1 - \Phi(x)]$ and $\hat{r}_i = \hat{\alpha} + \hat{\beta} \cdot v_i$. Then

$$E[(R_i^* - \hat{r}_i)^2 | w_i = 0; \hat{\alpha}, \hat{\beta}, \hat{\sigma}] = \text{var}(R_i^* | w_i = 0; \hat{\alpha}, \hat{\beta}, \hat{\sigma})$$
$$= \hat{\sigma}^2 + \hat{r}_i \cdot \hat{\sigma} \cdot h(\hat{r}_i/\hat{\sigma}) - [\hat{\sigma} \cdot h(\hat{r}_i/\hat{\sigma})]^2 .$$

The censored values of the dependent variable are replaced by the conditional expectations given the observed data and current estimates (E step). New parameter estimates are obtained by maximizing the log-likelihood for the "corrected" data using the ordinary least-squares method, and the new variance is calculated (M step):

$$\hat{\sigma}^2 = \left[\sum_i (r_i - \hat{r}_i)^2 + \sum_{w_i = 0} \text{var}(R_i^*) \right],$$

where the second summation is only over the originally censored values. The E and M steps are alternated until convergence. The solution converges rapidly because with every step the log-likelihood is increased, or at least not decreased. It has been shown that only one solution exists (Amemiya 1985). Only approximative standard errors of estimates are available when the likelihood is maximized with the EM algorithm or other classical methods. We have also worked out a Bayesian approach using the Gibbs sampler (Gelfand and Smith 1990) to obtain more reliable standard errors (Polasek and Krause 1992).

So far we have introduced the censored linear regression model (eq. [4]) based on the normal regression model (eq. [1]). A straightforward extension is to replace the linear relationship between R and v by a nonlinear one such as the allometric model (eq. [2]). We refer to this as the censored allometric model (eq. [5]):

$$R_i^* = \alpha + \beta \cdot v_i^\gamma + u_i, \; u_i \sim N(0, \sigma_u^2), \quad i = 1, \ldots, n .$$
$$r_i = R_i \quad \text{if } R_i^* > 0 \quad \text{(uncensored)}, \tag{5}$$
$$r_i = 0 \quad \text{if } R_i^* \leq 0 \quad \text{(censored)};$$

with likelihood function

$$L(\alpha, \beta, \gamma, \sigma) = \prod_i \{1 - \Phi[(\alpha + \beta \cdot v_i^\gamma)/\sigma]\}^{1 - w_i} \cdot \{1/\sigma \cdot \phi[(r_i - \alpha - \beta \cdot v_i^\gamma)/\sigma]\}^{w_i} .$$

It should be noted that large sample sizes may be required to reach convergence with commonly used fitting algorithms. For example, the results for the nonlinear relationships presented in table 1 could not be produced with simulated samples of only 100 instead of 400 points because there often exist various, equally well-fitting solutions. A general discussion of this issue, in the context of allometric models without censoring, is given by Klinkhamer et al. (1992).

THE "ERRORS-IN-VARIABLES" PROBLEM

All the models discussed so far assume that reproductive mass R is the dependent random variable and vegetative mass v is an independent explanatory vari-

able (fixed in repeated samples). However, Weiner (1988) proposed a biological model in which the functional relationship between reproductive and vegetative mass is one between two random variables R and V, each with underlying error, due to imprecise measurement or unknown influences of unmeasured explanatory variables (Kendall 1980). Such situations are frequently encountered in biology (see, e.g., LaBarbera 1989). A functional relationship in a bivariate distribution (R, V) such as $R = \alpha + \beta \cdot V$ leads to the statistical model of a structural relationship

$$r_i = \alpha + \beta \cdot (v_i - d_i) + u_i, \quad i = 1, \ldots, n, \tag{6}$$

if the observed variables $r_i = E(R) + u_i$ and $v_i = E(V) + d_i$ are both subject to normal error variation, that is, $u_i \sim N(0, \sigma_u^2)$ and $d_i \sim N(0, \sigma_d^2)$. This is called an errors-in-variables model. For known ratios of the error variances (i.e., $\lambda = \sigma_u^2/\sigma_d^2$), classical statistical methods for estimating α and β for a linear model (eq. [1]) are available (see, e.g., Kendall and Stuart 1973) and can be generalized for nonlinear model (2). Based on Lindley and El-Sayyad, Leamer (1978) gives a simple approximative formula for the estimate β (and its variance) in the linear model (eq. [6]): since $\text{cov}(R, V)$ is positive, we take the positive solution of the quadratic equation $\beta^2 + \tau \cdot \beta - \lambda = 0$,

$$\beta = (-\tau + \sqrt{\tau^2 + 4 \cdot \lambda})/2,$$

where $\tau = (\lambda \cdot \text{var}[V] - \text{var}[R])/\text{cov}(V, R)$. However, the approximative method does not yield a standard error for the estimate of the intercept α. Three special cases follow from the errors-in-variables model (eq. [6]): the direct regression model (eq. [1]) for $\lambda \to \infty$, the major axis or orthogonal regression model for $\lambda = 1$, and the indirect or reverse regression model for $\lambda = 0$: $v_i = \alpha' + 1/\beta \cdot r_i + u_i'$. We are working on extensions of censored regression models (4) and (5) to errors-in-variables models using the Gibbs sampler (Gelfand and Smith 1990; Polasek and Krause 1994). A classical solution to the errors-in-variables problem has recently become available for censored linear model (4) within the MECOSA program (Schepers and Arminger 1992).

IMPLEMENTATION OF THE CENSORED REGRESSION MODEL AND COMPARISON WITH RESULTS OF OTHER MODELS

Models (1)–(5) can be conveniently implemented with general statistical software such as GENSTAT (Payne et al. 1987), GLIM (Payne 1987), or S-plus (Becker et al. 1988). A general outline for implementation of the EM algorithm is presented in the Appendix. As our example, we use a data set of 433 plants (points marked by Y in fig. 2) that were grown in an experimental garden of the University of Basel, Switzerland. A positive reproductive mass could be observed for 310 plants, because 123 individuals had not reached reproductive maturity and their reproductive mass was therefore zero (censored observations). The estimates of statistical models (1)–(5) are shown in table 2, and the model fits are compared in figure 3.

TABLE 2

PARAMETER ESTIMATES FOR THE RELATIONSHIP BETWEEN REPRODUCTIVE AND VEGETATIVE MASS (g) IN *SOLIDAGO ALTISSIMA*

Model*	y-Intercept α	Proportionality Factor (slope) β	x-Intercept Minimum Size α'	Allometric Exponent γ
Linear regression (1)	−3.172	.5068	6.259	(1.000)
Allometric model (2):				
with intercept α (2*a*)	1.782	.03107	NE†	1.671
with intercept α' (2*b*)	1.132	.00459	−14.81	2.044
Probit model (3)‡	(−1.198)	(.1113)	10.76	(1.000)
Censored linear regression (4)	−7.110	.6043	11.77	(1.000)
Censored allometric model (5)	−7.065	.5953	11.89	1.004

NOTE.—The data set includes all points marked by Y's in fig. 2 and redrawn in fig. 3. Estimates are calculated according to the statistical models presented. Except for the allometric model, values in the column "x-Intercept" are negative quotients of the preceding two columns.

* Numbers refer to those used in the text.

† Not estimable because the fitted curve does not cross the X-axis.

‡ Parameters are for probit-transformed data; the x-intercept therefore corresponds to the vegetative mass, at which the probability of reproduction reaches 50% (see text).

We see that models (1) and (2), which do not use the information on the probability of reproduction contained in the censored data points, yield considerably lower estimates of minimum size for reproduction and of reproductive allocation above the minimum size than do models (3)–(5). The problem is particularly severe in the allometric model, which fits a curve across the cloud of points above the X-axis without touching it. Using reparameterization (2*b*) of this model with α' as the intercept on the X-axis still does not produce a positive estimate for minimum size for reproduction (table 2). It could be argued that the problem might have been less severe if the censored data points had been included in the analysis as ordinary zeros. While this would have forced the allometric curve down to the X-axis very close to the origin of the scatter plot, it would have violated the assumption of homoscedasticity and normal distribution of residuals. The results for the simulated data show that, if a data-generating process is linear but includes a censoring mechanism, the scatter of uncensored points may appear curvilinear. It is remarkable that, if censored allometric model (5) is fitted to the data set of the example, any indication of curvature in the relationship between reproductive and vegetative mass disappears. In fact, the lines for models (4) and (5) are practically identical, which supports Weiner's (1988) simple biological model.

CONCLUSION

Statisticians have come to appreciate that the "world is full of censored data problems" (A. F. M. Smith, personal communication). Our goal in this article is to demonstrate with a specific example how censored regression models may be

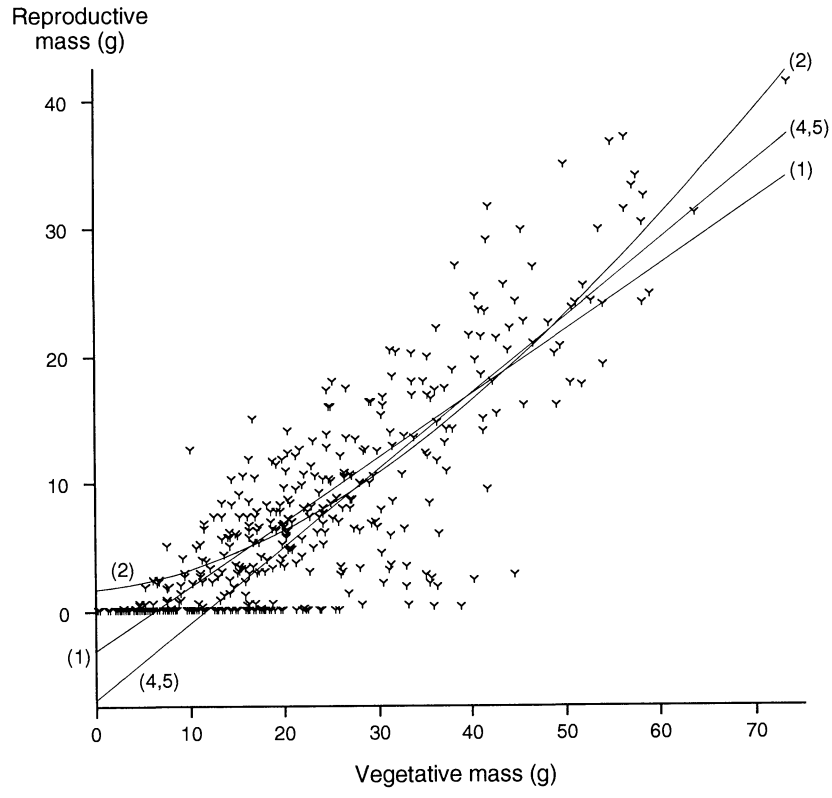

FIG. 3.—The relationship between plant reproductive and vegetative mass in *Solidago altissima* (the data set includes all points marked by large *Y*'s in fig. 2). The fitted values are calculated according to the following statistical models: (1) linear regression, (2) allometric, (4) censored linear regression, and (5) censored allometric; the probit model (3) cannot be shown on these axes.

used and how such models modify the interpretation of biological data. Biologists tend to use data transformations or curvilinear models to avoid truncated variables or discontinous relationships. However, there are clearly cases in which these approaches do not reflect the underlying biological mechanisms. For example, if a random variable has a normal distribution and a substantial proportion of this normal distribution extends beyond a permissible bound, then it is likely that for some units in a sample the values of the variables are censored. Censored regression models can deal with this situation and use all the information contained in the data. We therefore believe that they provide a substantial improvement compared with other estimation procedures. They simplify discontinuous relationships and facilitate their interpretation by introducing latent variables. Fortunately, censored regression models are not difficult to implement with currently available statistical software (see Appendix).

ACKNOWLEDGMENTS

We thank C. Körner and D. Matthies for providing examples of censored regression problems and M. Aitkin, G. Arminger, P. Jordan, P. W. Lane, E. Lüdin, D. Matthies, and T. Scallan for valuable information, discussions, and comments. This research was supported by grants 5001-35229 (Basel Biodiversity Program; Priority Programme Environment) and 31-30041.90 from the Swiss National Science Foundation to B.S.

APPENDIX

Implementation of the Censored Regression Model

The following is a brief recipe for programming the censored regression model using the EM algorithm:

1. Define the vectors **X** and **Y** of length N for the independent and the dependent variable (assign the value of the bound, e.g., zero, where **Y** is not observed).
2. Define a vector **W** and let **W** = 1 if **Y** is observed and $W = 0$ if **Y** is not observed.
3. Fit the linear regression $\mathbf{Y} = a + b \cdot \mathbf{X}$ (by least-squares method) only using the units where **Y** is observed.
4. Define variables A, B, and S and assign to them the initial parameter estimates of step 3 (S is the square root of the residual mean square).
5. Define a vector **FV**, where for all units the fitted values are calculated.
6. Define a vector **Z** with standardized fitted values $\mathbf{Z} = \mathbf{FV}/S$.
7. Define a vector **FZ**, containing the values of the standard normal cumulative distribution function Φ (normal probability integral) for each value in **Z** (e.g., **FZ** = normal [**Z**] in statistical computer programs).
8. Define a vector **PZ** and let $\mathbf{PZ} = 0.3989 \cdot \exp(-[\mathbf{Z} \cdot \mathbf{Z}/2])$, the normal density values of **Z**.
9. Define a vector **HZ** and let $\mathbf{HZ} = S \cdot \mathbf{PZ}/(1 - \mathbf{FZ})$; if division by zero occurs, then set the value in **HZ** to zero.
10. Define a vector **NY** and $\mathbf{NY} = \mathbf{W} \cdot \mathbf{Y} + (1 - \mathbf{W}) \cdot (\mathbf{FV} - \mathbf{HZ})$.
11. Define a vector **V1** and let $\mathbf{V1} = (1 - \mathbf{W}) \cdot (S \cdot S + \mathbf{FV} \cdot \mathbf{HZ} - \mathbf{HZ} \cdot \mathbf{HZ})$.
12. Fit the linear regression $\mathbf{NY} = a + b \cdot \mathbf{X}$ (by least-squares method) using all units.
13. Compare the new parameter estimates with the old ones in A, B; if the absolute difference is larger than some tolerance values (e.g., 0.01%), assign the new parameter to the variables A, B; otherwise convergence is reached.
14. Calculate for all units the new fitted values and assign these to **FV**.
15. Define a vector **V2** and let $\mathbf{V2} = (\mathbf{NY} - \mathbf{FV}) \cdot (\mathbf{NY} - \mathbf{FV})$.
16. Define a variable SS and assign to it the sum of all values in **V1** and **V2** divided by the number of units (N).
17. Replace the old value in S by the square root of the value in SS if this value is greater than zero.
18. Repeat steps 6–17 until convergence is reached at step 13.

The above procedure yields unbiased estimates of the intercept (variable A) and slope (variable B) under the assumptions of the censored linear regression model (4). To fit the censored allometric model (5), steps 3, 4, 12, and 13 have to be modified for a nonlinear estimation procedure. The nonlinear fits in steps 3 and 12 have to be done iteratively; algorithms are usually available in statistical software packages. Note that convergence may be difficult to obtain if steps 6–17 are iterated in the censored allometric model with small to medium-sized samples (e.g., $N < 200$; see text).

LITERATURE CITED

Aitkin, M., and D. Clayton. 1980. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Applied Statistics 29:156–163.

Aitkin, M., D. Anderson, B. Francis, and J. Hinde. 1989. Statistical modelling in GLIM. Clarendon, Oxford.

Amemiya, T. 1985. Advanced econometrics. Blackwell, Oxford.

Bazzaz, F. A., and E. G. Reekie. 1985. The meaning and measurement of reproductive effort in plants. Pages 373–387 *in* J. White, ed. Studies on plant demography: a festschrift for John L. Harper. Academic Press, London.

Becker, R. A., J. M. Chambers, and A. R. Wilks. 1988. The new S language. Wadsworth & Brooks/Cole, Pacific Grove, Calif.

Buckley, J., and I. James. 1979. Linear regression with censored data. Biometrika 66:429–436.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39:1–38.

Dolt, C. 1991. Effects of maternal and paternal environment and genotype on offspring phenotype in the perennial plant *Solidago altissima* L. M.S. thesis. University of Basel, Basel.

Finney, D. J. 1971. Probit analysis. 3d ed. Cambridge University Press, Cambridge.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 85:398–409.

Gross, K. L. 1981. Predictions of fate from rosette size in four "biennial" species: *Verbascum thapsus, Oenothera biennis, Daucus carota,* and *Tragopogon dubius*. Oecologia (Berlin) 48:209–213.

Harper, J. L. 1977. Plant population biology. Academic Press, London.

Hartnett, D. C. 1990. Size-dependent allocation to sexual and vegetative reproduction in four clonal species. Oecologia (Berlin) 84:254–259.

Kendall, M. G. 1980. Multivariate analysis. 2d ed. Griffin, London.

Kendall, M. G., and A. Stuart. 1973. The advanced theory of statistics: inference and relationship. 3d ed. Griffin, London.

Klinkhamer, P. G. L., T. J. de Jong, and E. Meelis. 1990. How to test for proportionality in the reproductive effort of plants. American Naturalist 135:291–300.

Klinkhamer, P. G. L., E. Meelis, T. J. de Jong, and J. Weiner. 1992. On the analysis of size-dependent reproductive output in plants. Functional Ecology 6:308–316.

LaBarbera, M. 1989. Analyzing body size as a factor in ecology and evolution. Annual Review of Ecology and Systematics 20:97–117.

Leamer, E. 1978. Specification searches. Wiley, New York.

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

Meagher, T. R., and J. Antonovics. 1982. The population biology of *Chamaelirium luteum*, a dioecious member of the lily family: life history studies. Ecology 63:1690–1700.

Miller, R., and J. Halpern. 1982. Regression with censored data. Biometrika 69:521–531.

Neter, J., and W. Wassermann. 1974. Applied linear statistical models. Irwin, Homewood, Ill.

Oakes, D. 1986. An approximate likelihood procedure for censored data. Biometrics 42:177–182.

Pacala, S. W., and J. A. Silander. 1985. Neighborhood models of plant population dynamics. I. Single species models of annuals. American Naturalist 125:385–411.

Payne, C. D. 1987. The GLIM system release 3.77 manual. 2d ed. Royal Statistical Society, London.

Payne, R. W., P. W. Lane, A. E. Ainsley, K. E. Bicknell, P. G. N. Digby, S. A. Harding, P. K. Leech, H. R. Simpson, A. D. Todd, P. J. Verrier, and R. P. White. 1987. GENSTAT 5 reference manual. Clarendon, Oxford.

Petersen, T. 1986. Fitting parametric survival models with time-dependent covariates. Applied Statistics 35:281–288.

Polasek, W., and A. Krause. 1992. The hierarchical Tobit model: a case study in Bayesian computing. WWZ-Discussion Papers 9204:1–18.

———. 1994. The Bayesian regression model with simple errors in variables structure. Statistician (in press).

Primack, R. B., and P. Hall. 1990. Costs of reproduction in the pink lady's slipper orchid: a four-year experimental study. American Naturalist 136:638–656.

Reiss, M. 1989. The allometry of growth and reproduction. Cambridge University Press, Cambridge.

Samson, D. A., and K. S. Werk. 1986. Size-dependent effects in the analysis of reproductive effort in plants. American Naturalist 127:667–680.

Schepers, A., and G. Arminger. 1992. MECOSA (mean and covariance structure analysis): user guide. Systeme für Logistik und Informationstechnik, Frauenfeld, Switzerland.

Schmid, B., and J. Weiner. 1993. Plastic relationships between reproductive and vegetative mass in *Solidago altissima*. Evolution 47:61–74.

Schneider, H., and L. Weissfeld. 1986. Estimation in linear models with censored data. Biometrika 73:741–745.

Segal, M. R. 1988. Regression trees for censored data. Biometrics 44:35–47.

Taylor, J. 1973. The analysis of designed experiments with censored observations. Biometrics 29: 35–43.

Thomas, S. C., and J. Weiner. 1989. Growth, death and size distribution change in an *Impatiens pallida* population. Journal of Ecology 77:524–536.

Thompson, B. K., J. Weiner, and S. I. Warwick. 1991. Size-dependent reproductive output in agricultural weeds. Canadian Journal of Botany 69:442–446.

Vandermeer, J. 1984. Plant competition and the yield-density relationship. Journal of Theoretical Biology 109:393–399.

Watkinson, A. R. 1980. Density-dependence in single species populations of plants. Journal of Theoretical Biology 83:345–357.

———. 1986. Plant population dynamics. Pages 137–184 *in* M. J. Crawley, ed. Plant ecology. Blackwell Scientific, Oxford.

Weiner, J. 1988. The influence of competition on plant reproduction. Pages 228–245 *in* J. Lovett Doust and L. Lovett Doust, eds. Plant reproductive ecology: patterns and strategies. Oxford University Press, New York.

Werner, P. A. 1975. Predictions of fate from rosette size in teasel (*Dipsacus fullonom* L.). Oecologia (Berlin) 20:197–201.

West, G. C. 1972. Seasonal differences in resting metabolic rate of Alaskan ptarmigan. Comparative Biochemistry and Physiology A, Comparative Physiology 42:867–876.

Wolynetz, M. S. 1979*a*. AS 138: maximum likelihood estimation from confined and censored normal data. Applied Statistics 28:185–195.

———. 1979*b*. AS 139: maximum likelihood estimation in a linear model from combined and censored normal data. Applied Statistics 28:195–206.

Associate Editor: Thomas R. Meagher