

## ESTIMATING EFFECTS OF LIMITING FACTORS WITH REGRESSION QUANTILES

BRIAN S. CADE, JAMES W. TERRELL, AND RICHARD L. SCHROEDER

Midcontinent Ecological Science Center, Biological Resources Division, U.S. Geological Survey,  
4512 McMurry Avenue, Fort Collins, Colorado, 80525-3400 USA

**Abstract.** In a recent Concepts paper in *Ecology*, Thomson et al. emphasized that assumptions of conventional correlation and regression analyses fundamentally conflict with the ecological concept of limiting factors, and they called for new statistical procedures to address this problem. The analytical issue is that unmeasured factors may be the active limiting constraint and may induce a pattern of unequal variation in the biological response variable through an interaction with the measured factors. Consequently, changes near the maxima, rather than at the center of response distributions, are better estimates of the effects expected when the observed factor is the active limiting constraint. Regression quantiles provide estimates for linear models fit to any part of a response distribution, including near the upper bounds, and require minimal assumptions about the form of the error distribution. Regression quantiles extend the concept of one-sample quantiles to the linear model by solving an optimization problem of minimizing an asymmetric function of absolute errors. Rank-score tests for regression quantiles provide tests of hypotheses and confidence intervals for parameters in linear models with heteroscedastic errors, conditions likely to occur in models of limiting ecological relations. We used selected regression quantiles (e.g., 5th, 10th, . . . , 95th) and confidence intervals to test hypotheses that parameters equal zero for estimated changes in average annual acorn biomass due to forest canopy cover of oak (*Quercus* spp.) and oak species diversity. Regression quantiles also were used to estimate changes in glacier lily (*Erythronium grandiflorum*) seedling numbers as a function of lily flower numbers, rockiness, and pocket gopher (*Thomomys talpoides fossor*) activity, data that motivated the query by Thomson et al. for new statistical procedures. Both example applications showed that effects of limiting factors estimated by changes in some upper regression quantile (e.g., 90–95th) were greater than if effects were estimated by changes in the means from standard linear model procedures. Estimating a range of regression quantiles (e.g., 5–95th) provides a comprehensive description of biological response patterns for exploratory and inferential analyses in observational studies of limiting factors, especially when sampling large spatial and temporal scales.

**Key words:** absolute deviations; limiting factors; linear models; quantiles; rank-score tests; regression; regression quantiles.

### INTRODUCTION

The law of limiting factors (Liebig's law of the minimum) is a basic tenet of ecological science. A limiting factor is the one least available among those factors that affect growth, survival, and reproduction of an organism. Any requisite factor has the potential to limit an organism, but only one will be the active constraint at any given point in time and space (Kaiser et al. 1994). A recent "Concepts" paper in *Ecology* (Thomson et al. 1996) provided a detailed synopsis of the pervasive nature of limiting relationships in ecology and a convincing argument that commonly used statistical methods such as correlation and regression are not well suited for estimating or testing those relationships. Thomson et al. (1996) believed that commonly accepted terminology and statistical methods for estimating functions along the edges of distributions would

enhance the communication of results of descriptive ecological studies where an observed variable acts as a limiting factor and the interior of the distribution is where other factors intervene.

Limiting relationships and the statistical difficulties associated with estimating and testing them have been discussed for a variety of ecological phenomena, including stomatal conductance of tree leaves as a function of photon flux density or leaf water potential (Jarvis 1976), foregut volume as a function of carapace length of lobsters (Maller et al. 1983, Maller 1990), plant growth rates with age (Rabinowitz et al. 1985), animal abundance and body size relationships in macroecology (Brown and Maurer 1987, Blackburn et al. 1992, Griffiths 1992, Blackburn and Gaston 1998), animal responses to habitat (Johnson et al. 1989, Terrell et al. 1996), effects of competition on distribution and abundance of focal species (Goldberg and Scheiner 1993), and algal growth as a function of nutrient availability (Kaiser et al. 1994). Thomson et al. (1996) used

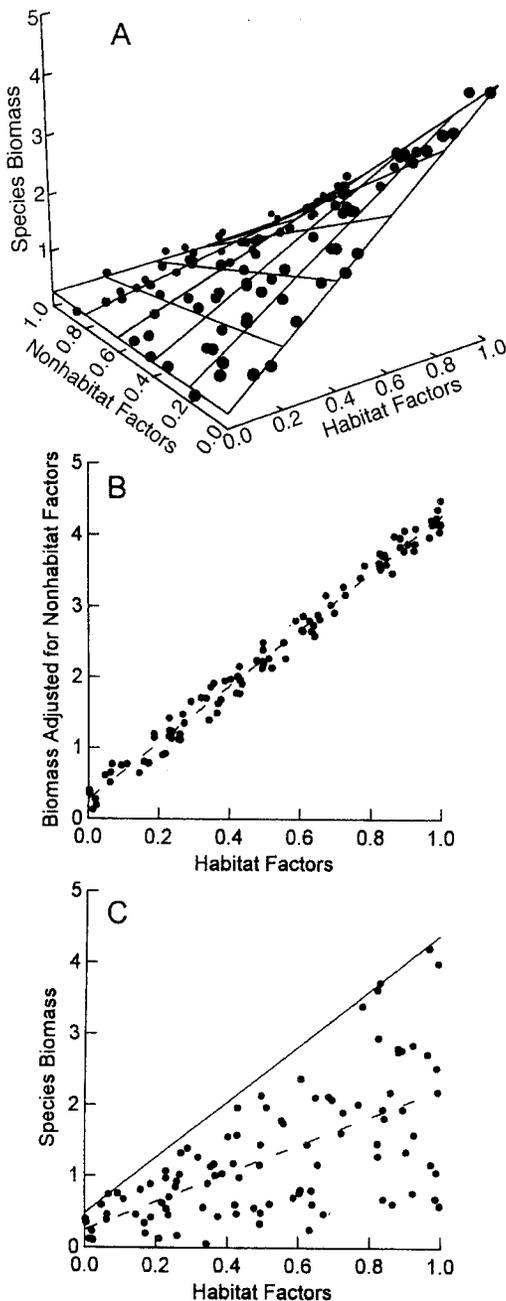


FIG. 1. A sample ( $n = 100$ ) of biomass ( $y$ ), habitat conditions [ $X = \text{uniform}(0, 1)$ ], and nonhabitat factors [ $Z = 0.05(0) + 0.95[\text{uniform}(0, 1)]$ ] from a hypothetical interference interaction model of limiting factors where  $y = 0.25 + 4X - 4XZ + e$ ,  $e$  is uniform  $(-0.25, 0.25)$ , and 5% of the population has no interaction between habitat and nonhabitat factors. The surface plotted in (A) is the least-squares estimate of the mean function ( $b_0 = 0.27$ , 95% CI = 0.22–0.33;  $b_1 = 4.00$ , 95% CI = 3.88–4.12;  $b_2 = -4.02$ , 95% CI = -4.19–-3.86) when nonhabitat factors are modeled. The dashed line in (B) is the OLS regression estimate for biomass adjusted for nonhabitat factors ( $y + 4.02XZ$ ) as a function of habitat. The dashed line in (C) is the OLS estimate for biomass as a function of habitat, and the solid line is the maximum regression quantile.

partitioned regression and logistic slicing as tentative analysis techniques to examine limiting relations between glacier lily (*Erythronium grandiflorum*) seedling numbers and flower numbers, but did not present any statistical theory to justify use of these methods. They discussed weaknesses of their methods and emphasized the need for objective statistical methods to estimate and test slopes along the “edges” of point clouds expected from limiting relationships. Maller (1990) and Kaiser et al. (1994) used methods based on statistical theories associated with estimating missing information (expectation maximization [EM] algorithm) and mixture distributions.

A major challenge of estimating the effects associated with a measured subset of limiting factors is to account for the effects of unmeasured factors in an ecologically realistic manner. Consider a biological response variable (e.g.,  $Y = \text{species biomass}$ ) that changes as a function of some limiting factors (e.g.,  $X = \text{habitat conditions}$ ) that are measured, and as a function of other limiting factors (e.g.,  $Z = \text{nonhabitat factors}$  such as weather and disease) that may not be measured. In this example (other variables could be inserted to describe other ecological phenomena), change in species biomass ( $Y$ ) does not exceed limits imposed by the habitat conditions ( $X$ ), but can be reduced by nonhabitat factors ( $Z$ ). One simple representation of this relation is a linear model in which species biomass is a positive linear function of habitat factors and a negative function of the interaction of habitat and nonhabitat factors, i.e., an interference interaction model (Neter et al. 1996:311–312). We took a sample ( $n = 100$ ) of observations from a hypothetical interference interaction model where a small percentage (5%) of the population is unaffected by the interaction, and we estimated the mean function with least squares regression (Fig. 1A). Change in biomass, due to change in habitat conditions, as the active limiting factor is easily estimated because we can account for the nonhabitat factors with an estimate of the interaction effect; variation about the estimated change in means is small and homogeneous (Fig. 1B). If, however, we do not measure the nonhabitat factors, we cannot estimate the interaction. The resultant distribution of biomass has greater variation that increases with levels of habitat that would support more biomass if habitat were the active constraint (Fig. 1C). Unmeasured nonhabitat factors can reduce biomass more where habitat would limit biomass to higher levels. The linear relation between species biomass and the measured habitat factors most consistent with the relation expected if habitat is the active limiting factor is near the upper “edge,” rather than through the center of the data distribution (compare Figs. 1B and C). Most commonly used regression techniques (linear and nonlinear least squares regression, generalized linear models) estimate functions through the center of data distributions (expected values). We demonstrate how regression quantiles can be

used to model limiting relations and account for the unmeasured ecological factors by estimating changes near the upper extremes of data distributions.

Various points of a univariate one-sample distribution can be described by estimating different quantiles of the cumulative distribution function. The  $\tau$ th one sample quantile estimates have a proportion  $\tau$  of the sample observations less than or equal to the estimate. The median, or 0.50th quantile, describes the center of the distribution such that 50% of the observations are less and 50% greater than the estimate; a 0.90th quantile is an estimate such that 90% of the observations are less and 10% greater than the estimate. Regression quantiles extend this concept of one-sample quantiles to the linear model by expressing the quantile estimates as solutions to an optimization problem of minimizing an asymmetric function of absolute error loss (Koenker and Bassett 1978, 1982, Buchinsky 1991, Koenker and Portnoy 1996). The  $\tau = 0.50$ th regression quantile is equivalent to least absolute deviation (LAD) regression estimates of conditional medians in a linear model (Koenker and Bassett 1978), an alternative to ordinary least squares (OLS) estimates of conditional means for modeling central tendency (Birkes and Dodge 1993, Cade and Richards 1996).

Regression quantiles for linear models with homogeneous (Fig. 2A) or heterogeneous (Fig. 2B) distributions correspond to linear functions such that approximately (this will be defined more precisely in the following sections)  $\tau$  proportion of the observations are below and  $1 - \tau$  proportion of the observations are above the estimated lines. Our example, in which effects of the nonhabitat factors are unknown, results in a heterogeneous distribution where slopes of the regression quantiles increase with higher quantiles (Fig. 2B). The upper regression quantiles ( $\tau > 0.90$ ) of this distribution have slope estimates close to the estimate for change in biomass when habitat is the active constraint, i.e., as if accounting for the interaction of nonhabitat factors (compare Figs. 2A and B). Least squares regression and linear correlation analyses focus on changes through the center of the distribution (conditional means), which underestimate rates of change (slopes) due to the limiting relation of habitat (Fig. 2B). It is possible for change in the center of distributions to be statistically indistinguishable from zero even when estimated changes near the extremes of distributions are nonzero. Concluding that there is no important limiting relation, based on the former evidence, would be incorrect, given the latter evidence.

The statistical theory of regression quantiles has been developed by econometricians during the last 20 years (Koenker and Bassett 1978, 1982, Bassett and Koenker 1982, 1986, Buchinsky 1991, Koenker 1994), but ecological applications have occurred only recently (Koenker et al. 1994, Terrell et al. 1996). We describe statistical properties of the regression quantile estimates and methods to test hypotheses and construct

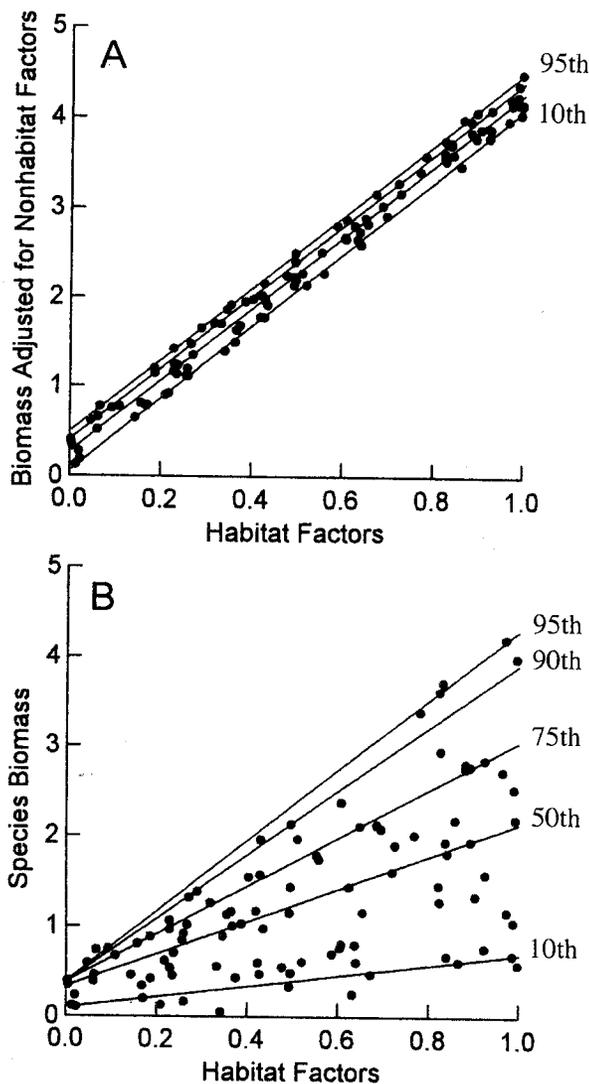


FIG. 2. The same sample of biomass ( $y$ ) and habitat conditions ( $X$ ) as in Fig. 1, but the lines plotted are 95th, 90th, 75th, 50th, and 10th regression quantile estimates (A) for biomass adjusted for nonhabitat factors ( $y + 4.02XZ$ ) as a function of habitat, and (B) for biomass as a function of habitat. Slopes for regression quantiles in (A) range from  $b_1(0.75) = 3.99$  to  $b_1(0.10) = 4.04$ , and slopes for regression quantiles in (B) range from  $b_1(0.10) = 0.60$  to  $b_1(0.95) = 3.91$ .

confidence intervals that are applicable to estimating effects of ecological limiting factors. We present sample applications to demonstrate strengths and weaknesses of various statistical methods for estimating limiting relationships and to motivate other ecologists to explore the analyses possible with regression quantiles. Regression quantiles are a rapidly evolving methodology for linear (and nonlinear) models with an established statistical theory; advances in describing and testing models of limiting relationships should be possible by taking advantage of these procedures.

## REGRESSION QUANTILES

*Properties of the estimates*

The  $\tau$ th quantile ( $0 \leq \tau \leq 1$ ) of a random variable  $Y$  is the inverse of the cumulative distribution function,  $F^{-1}(\tau)$ , which is defined as the smallest real value  $y$  such that the probability of obtaining smaller values of  $Y$  is greater than or equal to  $\tau$ . The one-sample quantile definition is extended to the linear model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \nu(\mathbf{X})\mathbf{e}$  by defining the  $\tau$ th regression quantile as  $Q_Y(\tau | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$  and  $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \nu(\cdot)F_c^{-1}(\tau)$ , where  $\mathbf{y}$  is an  $n \times 1$  vector of dependent responses,  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of unknown regression parameters,  $\mathbf{X}$  is an  $n \times p$  matrix of predictors (first column consists of 1's),  $\nu(\cdot) > 0$  is some known function, and  $\mathbf{e}$  is an  $n \times 1$  vector of random errors that are independent and identically distributed (iid) (Koenker and Bassett 1978, 1982). It is important to note that the term  $\nu(\mathbf{X})$  allows the errors to change as a function of  $\mathbf{X}$  and, thus, various heteroscedastic (inid) and homogeneous (iid) error models are accommodated with regression quantiles (Koenker and Portnoy 1996). If errors are homogeneous, the  $\tau$ th regression quantile simplifies to  $Q_Y(\tau | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + F_c^{-1}(\tau)$  and change in  $y$  across  $X$  is constant for all values of  $\tau$ , but the intercepts ( $\beta_0$ ) in  $\boldsymbol{\beta}(\tau)$  will vary. Any difference in slope estimates for different values of  $\tau$  is due to random sampling variation (Fig. 2A). If errors are heterogeneous with respect to  $X$ , all parameters in  $\boldsymbol{\beta}(\tau)$  may vary with  $\tau$  and regression quantile estimates will reflect this pattern (Fig. 2B). In the context of estimating effects of ecological limiting factors, our attention focuses on higher values of  $\tau$ , but lower values also may provide insight on response patterns.

Estimates,  $\mathbf{b}(\tau)$ , of  $\boldsymbol{\beta}(\tau)$  are obtained by minimizing an asymmetric loss function of absolute values of residuals where positive residuals are given weights equal to  $\tau$  and negative residuals are given weight equal to  $1 - \tau$  (mathematical details are in the Appendix). The term  $\nu(\mathbf{X})$  does not have to be estimated explicitly because it is automatically incorporated in  $\mathbf{b}(\tau)$ . A  $\tau$ th regression quantile with  $p$  estimated parameters passes through at least  $p$  sample observations ( $p$  residuals equal zero). If we denote the number of positive, negative, and zero residuals by  $N^+$ ,  $N^-$ , and  $N^0$ , respectively, and if  $N^0 = p$ , then the proportion of negative residuals is approximately  $\tau$ , ( $N^-/n \leq \tau \leq [N^- + p]/n$ ) and the proportion of positive residuals is approximately  $1 - \tau$ , ( $N^+/n \leq 1 - \tau \leq [N^+ + p]/n$ ). There are at most  $n\tau$  sample observations below ( $N^- \leq n\tau \leq N^- + N^0$ ) and at most  $n(1 - \tau)$  above ( $N^+ \leq n[1 - \tau] \leq N^+ + N^0$ ) a regression quantile estimate (Koenker and Bassett 1978, Koenker and Portnoy 1996). It is in this sense that the proportion of the sample observations less than an estimated regression quantile for specified  $\tau$  is only approximately equal to  $\tau$ . Regression quantiles define an ascending sequence of planes that are above an increasing proportion of observations with

increasing values of  $\tau$  (Fig. 2). We will use the notation for a 100 $\tau$ th (e.g., 50th) rather than for the  $\tau$ th (0.50th) regression quantile in the text when convenient, which is equivalent to using percentages rather than proportions.

Unlike one-sample quantiles where there are at most  $n$  distinct values of  $\tau$  equally spaced on the interval  $[0, 1]$ , in the regression quantile setting, there may be more than  $n$  distinct values of  $\tau$  that are unequally spaced. In practice, there are usually  $<3n$  distinct regression quantile solutions (Koenker and d'Orey 1987, Portnoy 1991). For example, there are 119 distinct regression quantile solutions for the simple linear model and data ( $n = 100$ ) in Fig. 2B that break the interval  $[0, 1]$  for  $\tau$  into 118 unequal intervals; intercept and slopes estimates,  $b_0(0.947, 0.954) = 0.38$  and  $b_1(0.947, 0.954) = 3.91$ , are solutions for the interval  $\tau = (0.947, 0.954)$ . Our focus here is on estimating selected values of  $\tau$  (e.g., 0.50, 0.75, 0.95) rather than solving for all possible values, but any selected value of  $\tau$  will be associated with one of the interval solutions.

Regression quantiles have several important linear model properties that are common to least squares regression estimates of expected values; they are equivariant to (1) scale changes, (2) location shift, and (3) design ( $\mathbf{X}$ ) reparameterization (Koenker and Bassett 1978, Büchinsky 1991, Koenker and Portnoy 1996). Unlike least squares estimates of means, regression quantiles also are (4) equivariant to monotonic transformations, linear or nonlinear (Buchinsky 1991, Koenker and Portnoy 1996). The  $\tau$ th quantile of the transformed data is the transformation of the  $\tau$ th quantile of the original data, i.e., if  $h(\cdot)$  is a nondecreasing function, then for any random variable  $Y$ ,  $Q_{h(Y)}(\tau) = h(Q_Y(\tau))$ . Thus, there is no ambiguity about what is being estimated in the transformed and original (back-transformed) data scales, as there is when estimating means with least squares regression for nonlinear (e.g., logarithmic) monotonic transformations (Bassett 1992, Koenker and Portnoy 1996). The method of Box-Cox transformations can be applied to discover the most suitable transformation to achieve linearity without being concerned about normality or homogeneity of the error distribution (Buchinsky 1995, Koenker and Portnoy 1996).

Regression quantile estimates are insensitive to extreme values of outlying dependent variables. As long as a dependent variable value remains above or below a regression quantile estimate, the estimate will remain unchanged, regardless of the magnitude of the value (Koenker and Portnoy 1996). Including a few extreme observations in an analysis with regression quantiles has less effect on the estimates than it does when using least squares regression, which is notoriously sensitive to even a single outlier (Cade and Richards 1996).

*Confidence intervals and hypothesis tests*

A measure of precision for any statistical estimate of a parameter is desirable to determine values con-

sistent with a selected model and data. Asymptotic sampling theory for regression quantiles implies that the sampling distribution for a specified quantile ( $\tau$ ) is dependent on the density of errors at the estimate (Koenker 1994). For typical unimodal iid errors with greatest density near the center (e.g., normal, lognormal, double exponential), sampling variation of regression quantiles increases for quantiles greater or less than the 50th. For multimodal error distributions where greatest densities are not near the center of the error distribution, it is possible for the sampling distribution of regression quantiles greater or less than the 50th (e.g., 25th and 75th) to have less sampling variation.

Several methods for testing hypotheses and constructing confidence intervals (CI) have been developed for regression quantiles when the error distributions are assumed to be homogeneous (Koenker and Bassett 1978, Koenker 1994, Cade and Richards 1996, Zhou and Portnoy 1996). However, for estimating effects of ecological limiting factors, we are interested in hypothesis-testing procedures that are valid when it is unreasonable to assume the homogeneity model. Koenker (1994) proposed both an  $xy$ -pairs bootstrap procedure and a quantile rank-score test that provided correct test levels under the null hypothesis for heteroscedastic regression models. We used the rank-score test because it is easily inverted to provide confidence intervals when the test statistic is evaluated with reference to a standard normal distribution (an asymptotic approximation). We used a regression quantile rank-score test procedure based on an adaptation of an algorithm described by Koenker and d'Orey (1994) and implemented in S-Plus to test hypotheses and compute confidence intervals (Appendix). The quantile rank-score test is a special case of the more general rank-score tests for regression quantiles developed by Gutenbrunner and Jurečková (1992), Gutenbrunner et al. (1993), and Hušková (1994). The quantile rank-score test can be thought of as an extension of the sign test to quantiles other than the 50th (median) and to the linear model.

When using regression quantiles to estimate changes in the upper edge of distributions associated with limiting factors, it is tempting to consider the maximum ( $\tau = 1.0$ ) as the best possible estimate for the limiting relation. However, the asymptotic variance of the rank-score statistic is zero for  $\tau = 1.0$  because the term  $\tau(1 - \tau)$ , which appears in the variance formula, would equal zero (Appendix). We used values of  $\tau$  less than 1 so that we could estimate precision by calculating a confidence interval. The maximum value of  $\tau$  that can be estimated precisely and still characterize changes in the upper quantiles of heteroscedastic distributions associated with modeling limiting factors will vary depending on sample size and distribution of the data.

The need to evaluate several upper quantiles is demonstrated with the simulated data in Fig. 2B. The slope estimate for  $\tau = 1.0$  is  $b_1(1.0) = 3.89$ , but no confidence

interval can be calculated, whereas the slope estimate for  $\tau = 0.95$  is  $b_1(0.95) = 3.91$  and has a 99% confidence interval of 3.52–4.03. The slope estimate for  $\tau = 0.97$ ,  $b_1(0.97) = 3.90$ , has a 99% confidence interval of -7.16 to 4.01, an estimate similar to  $\tau = 0.95$  (and 1.0), but with greater sampling variation because of lower density of observations. The slope estimate for  $\tau = 0.90$ ,  $b_1(0.90) = 3.53$ , has a 99% confidence interval of 2.29–4.11, a slightly lower estimated change with larger sampling variation. The 95th regression quantile provides the largest estimate of  $\beta_1(\tau)$  with confidence intervals that exclude zero for this sample size ( $n = 100$ ) and simulated data distribution, where 5% of the population is unaffected by interactions with nonhabitat factors. Sampling from a population where a greater proportion of observations of species biomass is unaffected by nonhabitat factors, e.g., 20%, would produce a greater range of upper regression quantiles (e.g.,  $\tau > 0.80$ ) with similar slope estimates, and the less extreme of these quantiles (e.g.,  $0.90 > \tau > 0.80$ ) would have less sampling variation and, hence, smaller confidence intervals.

#### EXAMPLE APPLICATIONS

##### *Acorn abundance and oak forest characteristics*

Acorn production in oak (*Quercus* spp.) forest types is important for wildlife species that depend on mast forage. Schroeder and Vangilder (1997) measured acorn density (average annual numbers per hectare), acorn biomass (average annual kilograms per hectare), oak tree cover, and number of oak species in 43 0.2-ha plots from 1989 to 1993 in Missouri to test mast production relationships in wildlife habitat models. They derived an acorn production suitability index (0–1) from estimates of canopy cover of oaks  $\geq 25$  cm dbh (cc-oaks) and number of oak species (spp-oaks). This index was the arithmetic average of two functions: a piecewise linear function of oak canopy cover (cc-oaks/40 if cc-oaks  $\leq 40$ , 1 if  $40 < \text{cc-oaks} \leq 60$ , or  $1 - [\text{cc-oaks} - 60]/100$  if cc-oaks  $> 60$ ) and a discrete function for the number of oak species (1 if spp-oaks  $\geq 3$ , 0.5 if spp-oaks = 2, and 0.1 if spp-oaks = 1). Maximum suitability for acorn production occurred when canopy cover of oaks  $\geq 25$  cm dbh was 40–60% and there were  $\geq 3$  oak species. The index served as an independent variable on which acorn density and acorn biomass were regressed with least squares regression. Schroeder and Vangilder (1997) noted that the "wedge-shaped" distribution of acorn density and biomass plotted against the acorn suitability index was consistent with the hypothesis that oak cover and species richness act as limiting factors for acorn production.

We analyzed the relationship between average annual acorn biomass and the acorn suitability index (Schroeder and Vangilder 1997) with simple regression models,  $y = \beta_0 + x_1\beta_1 + v(x_1)\mathbf{e}$ , where  $y$  was acorn biomass,

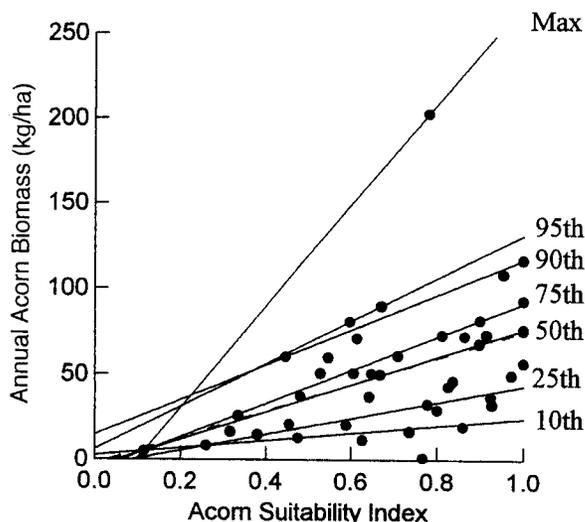


FIG. 3. Average annual biomass of acorns and acorn suitability indices based on oak forest characteristics for  $n = 43$  0.2-ha sample plots in Missouri (data from Schroeder and Vangilder [1997]). Solid lines correspond to the seven selected 100th regression quantile estimates in Table 1, and the dashed line is the OLS regression estimate of the mean.

and  $x_1$  was the acorn suitability index, based on canopy cover and number of oak species. Estimates and 90% confidence intervals were made for 95th, 90th, 75th, 50th, 25th, 10th, and 5th regression quantiles. We also examined all regression quantiles  $\geq 80$ th for similarity of slope estimates (parallelism), length of 90% confidence intervals, and whether confidence intervals excluded zero. Acorn density and biomass were strongly correlated ( $r = 0.92$ ), and either was a reasonable estimate of average annual acorn production.

Average annual acorn biomass increased with increasing acorn suitability indices and increases ( $b_1$ ) were greater for higher quantiles (Fig. 3). Similar patterns were found for acorn density (Schroeder and Vangilder 1997), but we limit our example to biomass. Confidence intervals for  $\beta_1$  calculated with the rank-score tests exclude zero, except for the 95th regression quantile estimate (Table 1). Confidence intervals increase for quantiles farther from the 50th, demonstrating that more extreme regression quantiles (e.g., 5th and 95th) were estimated less precisely than more central quantiles (e.g., 50th). The 90th regression quantile of acorn biomass was the most extreme quantile that could be estimated with any reasonable precision, as indicated by 90% confidence intervals. The estimated slope for the maximum,  $b_1(0.98, 1.00) = 297.1$  kg/ha per unit change in suitability, was far greater than slopes for other upper quantile estimates ( $0.75 < \tau < 0.95$ ), which were in the range 90.1 to 125.2 kg/ha. This extreme estimate fits through the outlying observation of 202.8 kg/ha. If this observation is deleted, our original slope estimate for the 90th regression quantile,  $b_1(0.89, 0.92) = 102.3$ , becomes the estimate

TABLE 1. Estimates of  $\beta_0$  and  $\beta_1$ , 90% confidence intervals for  $\beta_1$ , and  $P$  for  $H_0: \beta_1 = 0$  from rank-score tests for seven selected regression quantiles (100th) for models  $y = \beta_0 + x\beta_1 + v(x_1)e$ , where  $y$  is acorn biomass (kg/ha), and  $x_1$  is the acorn suitability index based on canopy cover.

$\tau$	$b_0$	$b_1$	90% CI for $\beta_1$	$P$
5th	-1.0	23.7	12.5-266.6	0.066
10th	2.4	21.8	18.6-142.5	0.012
25th	-4.2	47.4	38.4-81.7	0.001
50th	-4.3	80.5	69.9-97.1	0.022
75th	-6.3	97.7	29.0-111.5	0.004
90th	14.4	102.3	24.2-145.7	0.040
95th	5.9	125.2	-445.3-166.2	0.203

for the 92-94th regression quantile,  $b_1(0.92, 0.94) = 102.3$ , and the 90% confidence interval narrows to 44.3 to 142.3.

The 90th regression quantile,  $b_1(0.89, 0.92) = 102.3$  kg/ha change in acorn biomass per unit change in the acorn suitability index, is our best approximation of changes in average annual acorn production when forest suitability is the active limiting factor. Changes as great as 145.7 kg/ha are consistent with the model and data, as indicated by the upper endpoint of the 90% CI (Table 1). The least squares regression estimate of the slope for the same model (Schroeder and Vangilder 1997),  $b_1 = 77.6$  (90% CI = 40.0-115.1) was similar to the estimate for the 50th regression quantile (Table 1), but had a longer confidence interval. Neither the 50th regression quantile estimate of the conditional median nor the least squares regression estimate of the conditional mean (and associated 90% CIs) suggested increases in acorn biomass due to increases in forest suitability as high as those indicated by the 90th regression quantile. Larger sample sizes will be required to determine whether greater slopes of more extreme regression quantiles,  $>90$ th, are better approximations of changes in annual acorn production when forest suitability is the active limiting factor. The lower endpoint of 90% confidence intervals for the 10th regression quantile indicates that changes as low as 18.6 kg/ha per unit change in forest suitability may occur when factors unrelated to oak forest attributes are limiting.

Annual acorn production varies due to many factors other than oak forest characteristics, especially weather (Christisen and Kearby 1984). Weather explained 55% and 89% of the variance in acorn production in black oak (*Q. velutina*) and red oak (*Q. rubra*), respectively, over an 8-yr period in east-central Missouri (Sork et al. 1993). Thus, low levels of acorn production at high levels of the acorn suitability index (Schroeder and Vangilder 1997) were consistent with the index being a reasonable quantification of oak forest attributes that limit acorn production, given that other factors (e.g., weather) often were the active limiting constraint. The ecological basis for this relationship is readily understood, because weather can limit acorn production to levels less than those imposed by the number, size, and

species of oaks. High levels of acorn production in forest stands with low acorn suitability indices would have refuted the hypothesized limiting relationship with oak forest suitability. The 90th regression quantile is a more reasonable approximation of the expected increase in acorn production, when oak forest suitability is the active constraint, than is the lower increase in production estimated with the mean of the distribution.

#### *Glacier lilies, gophers, and rocks*

Thomson et al. (1996) counted glacier lily (*Erythronium grandiflorum*) seedlings and flowering plants, computed an index for abundance of surface and sub-surface rocks, and computed an index of pocket gopher (*Thomomys talpoides fossor*) burrowing activity in a square grid of 256  $2 \times 2$  m quadrats in a subalpine meadow in western Colorado. Figures 5–7 in Thomson et al. (1996) suggest a negative limiting relation between flower counts and seedling numbers, in the sense that seedlings were numerous only when flowers were scarce. We reanalyzed the flowering and seedling lily data using regression quantiles, considering several alternative model forms. Thomson et al. (1996) deleted one outlier from their analyses. We included the outlier in our analyses, because most regression quantile estimates were insensitive to the presence of this outlier. We left the outlier off our scatter plots to facilitate visual comparison with graphs of Thomson et al. (1996).

We considered simple linear regression models,  $y = \beta_0 + x_1\beta_1 + v(x_1)e$ , where  $y$  was number of seedlings and  $x_1$  was number of lily flowers; a  $\log_{10}$  transformation of  $y + 1$ ; and a piecewise linear regression model (Neter et al. 1996:474–478) with an indicator variable,  $x_2 = 0$  for  $x_1 \leq 16$  and  $x_2 = 1$  for  $x_1 > 16$ , added to the model. The value of 16 flowers for the breakpoint in the piecewise linear regression was obtained by inspecting the data distribution, but a more refined estimate could be obtained by iterative procedures. These models were based on the discussion of the data pattern in Thomson et al. (1996); i.e., maximum seedling numbers occurred at intermediate numbers of flowers, and the decrease in seedling numbers might be nonlinear. We also estimated a multiple regression model with the index of rockiness added to the model,  $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + v(\mathbf{X})e$ , where  $y$  and  $x_1$  were as previously defined, and  $x_2$  was the index of rockiness. Estimates and 90% confidence intervals were obtained for 5th, 10th, 25th, 50th, 75th, 90th, and 95th regression quantiles for the models. Again, we also examined estimates and 90% confidence intervals for all regression quantiles  $\geq 80$ th.

Regression quantile estimates for glacier lily seedlings as a linear function of flower numbers (Fig. 4A) provided a pattern of lines similar to the partitioned least squares regression estimates in Thomson et al. (1996:Fig. 6). The 95th regression quantile estimate

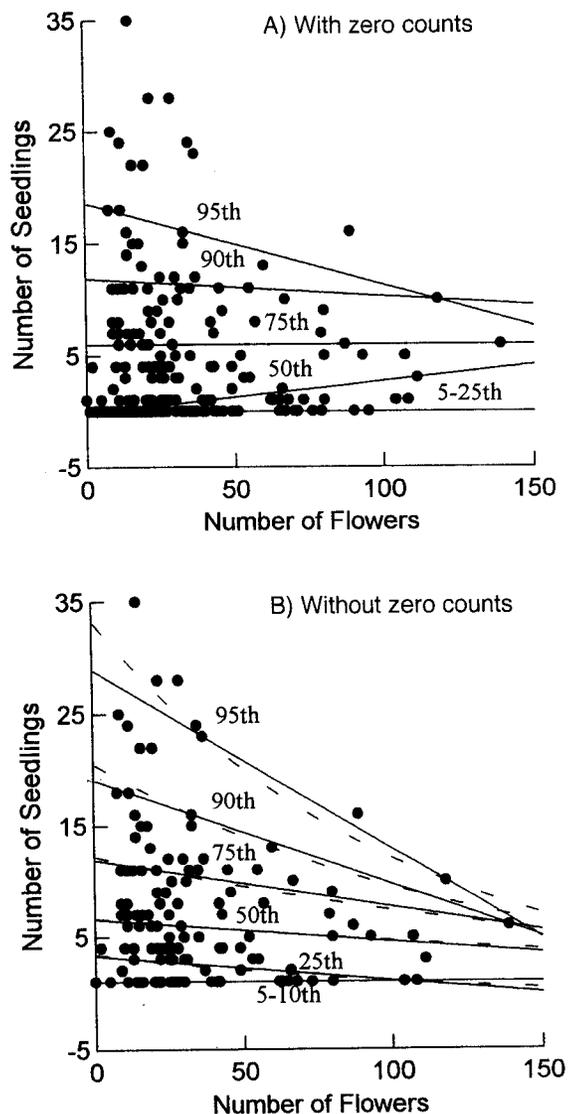


FIG. 4. Glacier lily seedling counts and flower numbers for  $n = 256$  contiguous  $2 \times 2$  m quadrats in subalpine meadow of western Colorado (data from Thomson et al. [1996]). Lines correspond to seven selected 100th regression quantile estimates in Table 2, (A) with zero counts included and (B) without zero counts. Dashed lines in (B) are for models based on  $\log_{10}(\text{seedling counts} + 1)$ . One outlying count of 72 seedlings at 16 flowers is not plotted but was used to estimate regression quantiles.

provided the strongest negative linear relationship between lily seedling and flower numbers, with a 90% CI that (barely) excluded zero (Table 2). Higher quantiles ( $0.95 < \tau < 0.99$ ) had slopes ranging from  $-0.10$  to  $-0.23$  that were consistent with the data pattern, but could not be estimated very precisely (90% confidence intervals overlapped zero to a considerable degree). The maximum regression quantile had slope  $b_1(0.996, 1.000) = -0.53$ , but was driven by the outlying value of 72 seeds at 16 flowers and not consistent with the majority of the data. Piecewise linear regression models

TABLE 2. Estimates of  $\beta_0$  and  $\beta_1$ , 90% confidence intervals for  $\beta_1$ , and  $P$  for  $H_0: \beta_1 = 0$  from rank-score tests for seven selected regression quantiles (100rth) for models  $y = \beta_0 + x_1\beta_1 + \nu(x_1)e$ , where  $y$  represents glacier lily seedling numbers and  $x_1$  represents flower numbers. Models were estimated with ( $n = 256$ ) and without ( $n = 129$ ) the zero seedling counts.

$\tau$	$b_0$	$b_1$	90% CI for $\beta_1$	$P$
With zero seedling counts				
5th	0.00	0.000		
10th	0.00	0.000		
25th	0.00	0.000		
50th	-0.20	0.029	0.002-0.045	<0.001
75th	6.00	0.000	-0.012-0.070	0.347
90th	11.87	-0.016	-0.064-0.137	0.403
95th	18.58	-0.073	-0.092-0.003	0.123
Without zero seedling counts				
5th	1.00	0.000		
10th	1.00	0.000		
25th	3.29	-0.022	-0.043-0.017	0.530
50th	6.57	-0.020	-0.064-0.003	0.186
75th	11.89	-0.042	-0.096-0.014	0.026
90th	19.13	-0.094	-0.164-0.022	0.032
95th	28.94	-0.160	-0.172-0.005	0.145

and nonlinear models obtained by  $\log_{10}$  transformation of seedling counts were consistent with the data pattern, but wider confidence intervals and larger  $P$  values from hypothesis tests indicated that these models were not better alternatives to the simple linear model. We were unable to develop useful confidence intervals from the rank-score test for 5th to 25th regression quantiles because the mass of zeros associated with those estimates violated the positive density assumption of the rank-score test. The first nonzero slope was for the 38th regression quantile. Negative slopes for upper regression quantiles were consistent with the explanation provided by Thomson et al. (1996) that sites where flowers were most numerous, because of lack of pocket gophers (which eat lilies), were rocky sites that provided poor moisture conditions for seed germination; hence, seedling numbers were lower.

The influence of the mass of zero seedling counts (49% of observations) was investigated by estimating regression quantiles for the linear model after truncating zero counts. Differences in estimates and confidence intervals between models with and without zero counts indicated that inclusion of the zeros attenuated the negative slopes for higher regression quantiles, but precision of the estimates was not reduced because confidence intervals were wider for the truncated model (Table 2, Fig. 4B). However, 90% confidence intervals shifted to more negative values when the zeros were truncated. Without the zero counts, 50th, 75th, and 90th regression quantile estimates had greater negative slope estimates and 90% confidence intervals that excluded zero. The nonlinear functions from back-transforming the estimates made with the  $\log_{10}$  transformation also were consistent with the data when the zeros were eliminated, 90% confidence intervals did not overlap zero,

and  $P$  values that slopes equal zero were similar to those for the linear model estimates. There was only a minor improvement in fit for the nonlinear compared to the linear model for the 95th regression quantile (Fig. 4B). Censored regression quantile models (Powell 1986, Buchinsky 1991) are alternative procedures for data with a mass of values at zero, but this will not be explored here.

The interaction between rocks, gophers, and lilies discussed by Thomson et al. (1996) suggested that a multiple regression including number of rocks as an additional independent variable might be informative. Regression quantiles estimated for a model that included rockiness and flower numbers yielded slope estimates for flowers that were positive, rather than negative as in the simple linear model (Table 3, Fig. 5). After accounting for the positive effect of flower numbers on seedling numbers, rockiness had a negative effect on lily seedling numbers. Confidence intervals for all but one coefficient estimate (90th quantile estimate for flowers) did not overlap zero (Table 3). Again, it was not possible to estimate confidence intervals for 5-25th regression quantiles because of the mass of zeros. The positive relation between seedling numbers and flower numbers, after accounting for the negative relation with rockiness, was consistent with the path analysis of Thomson et al. (1996). Greater numbers of flowers provided a greater potential source of seeds and, thus, seedlings, but increasing rockiness decreased moisture availability, which reduced seed germination and, thus, seedling numbers (Thomson et al. 1996). Rockiness had an indirect positive effect on seedling numbers through its positive relation with flower numbers. We considered models that included an interaction between rockiness and flower numbers and that included the gopher activity index, but neither of these terms differed from zero ( $P > 0.10$ ) for regression quantiles  $>50$ th, given that flower numbers and rockiness were already in the model.

Spread in the distribution of the upper 50% of seedling numbers (the nonzero counts) changed twice as much across changes in rockiness at a given level of flower numbers as across changes in flower numbers at a given level of rockiness (Table 3, Fig. 5). Most of the increasing variation in the distribution of seedling numbers as a function of increasing flower numbers (at a given level of rockiness) occurred between the 50th and 75th quantiles, as indicated by a near doubling in estimates from  $b_1(0.50) = 0.04$  to  $b_1(0.75) = 0.09$ , but little difference between estimates from the 75th to 95th regression quantiles. Decreasing variation in the distribution of seedling numbers as a function of increasing rockiness at a given level of flowers occurred for all quantiles  $>50$ th, as indicated by progressively more negative estimates from  $b_2(0.50) = -0.005$  to  $b_2(0.95) = -0.09$ . Limiting factors not included in the model influenced variation in lily seedling numbers more across levels of rockiness than across flower num-

TABLE 3. Estimates of  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$ , 90% confidence intervals for  $\beta_1$  and  $\beta_2$ , and  $P$  for  $H_0: \beta_1 = \beta_2 = 0$  from rank-score tests for seven selected regression quantiles (100th) for models  $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \nu(x_1 + x_2)e$ , where  $y$  is glacier lily seedling numbers,  $x_1$  is flower numbers, and  $x_2$  is rockiness ( $n = 256$ ).

$\tau$	$b_0$	$b_1$	90% CI for $\beta_1$	$b_2$	90% CI for $\beta_2$	$P$
5th	0.00	0.000		0.000		
10th	0.00	0.000		0.000		
25th	0.00	0.000		0.000		
50th	-0.18	0.044	0.020-0.062	-0.005	-0.011--0.001	0.001
75th	4.78	0.092	0.050-0.125	-0.039	-0.045--0.030	0.001
90th	12.09	0.088	-0.023-0.181	-0.058	-0.077--0.014	0.023
95th	20.30	0.085	0.049-0.147	-0.090	-0.112--0.038	0.030

bers. The near parallelism of slope estimates for changes in seedling numbers with changes in flower numbers for quantiles >75th suggests that other limiting factors not included in the model interact more with rockiness to affect seedling numbers.

The 95th regression quantile is our best estimate of the changes in lily seedling numbers that would occur when flower numbers and rockiness are the active limiting factors (Table 3). Regression quantiles for  $\tau > 0.95$  had slope estimates of slightly greater magnitude, but confidence intervals for  $\beta_1$  included zero, although estimates for  $\beta_2$  still differed from zero (e.g.,  $b_1(0.97) = 0.11$ , 90% CI = -0.002-0.353;  $b_2(0.97) = -0.11$ , 90% CI = -0.127--0.046). Upper endpoints of 90% confidence intervals for estimates of the 95th regression quantile indicated that increases in seedling numbers as great as 0.15 per flower and decreases as great as 0.11 per unit of rockiness were consistent with the data and the linear model. For purposes of comparison, the mean function was estimated for this same model by using a generalized linear model assuming a neg-

ative binomial distribution, which is appropriate for count data with a large proportion of zeros (Venables and Ripley 1994, White and Bennetts 1996). Estimates were  $b_1 = 0.05$  (90% CI = 0.024-0.078) and  $b_2 = -0.02$  (90% CI = -0.031--0.016), estimate of dispersion parameter  $\theta = 1.02 \pm (0.10, \text{mean} \pm 1 \text{ SE})$ , and a residual deviance of 256.8 on 253 df indicated a good fit to the negative binomial distribution ( $P = 0.422$ ). These estimated changes in the mean are considerably lower than those for the 95th regression quantile, especially for the effect of rockiness.

#### DISCUSSION

An upper regression quantile may not describe the "correct" limiting function in all situations, but should provide an approximation that is more consistent with the ecological theory of limiting factors than estimates through the center of data distributions. Selecting appropriate upper quantiles to estimate changes in a biological response requires consideration of the statistical properties of the estimates and the underlying ecological processes. Our examples and simulations suggest that, when a larger proportion of a sample is not impacted by interactions with unmeasured factors, then more of the upper quantiles should be parallel. Estimating slopes for the less extreme of these quantiles may provide more precise estimates of change due to limiting factors. If unmeasured factors have additive, rather than interactive, effects with the measured factors, then variation in the response should be homogeneous, all regression quantiles (and OLS regression) should estimate the same slope parameters, and those estimated more precisely should be preferred estimates of rate of change in the biological response to a limiting factor. Confidence intervals calculated by inverting the rank-score test are sensitive to local density of observations around the estimated quantile, such that a quantile slightly less or greater than one initially selected may be estimated with greater or lesser precision depending on the data. Looking at the pattern of estimates and associated confidence intervals for a range of upper regression quantiles is recommended for heteroscedastic distributions. When it is desirable to estimate a regression quantile very close to the maximum (e.g., the 99th), large samples are required to have sufficient density of observations near the estimate to make it precise.

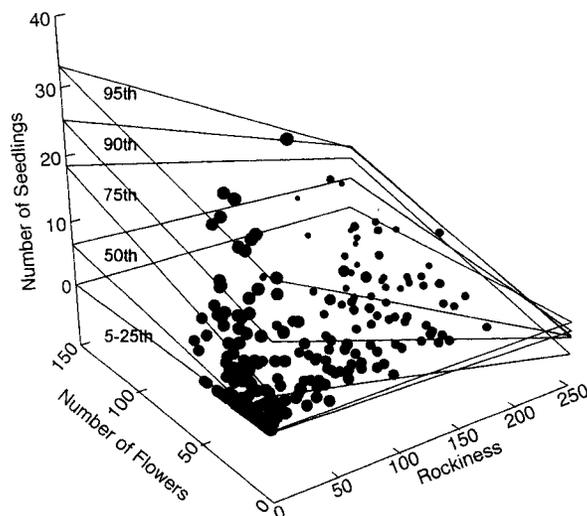


FIG. 5. Glacier lily seedling counts, lily flower numbers, and rockiness index for  $n = 256$  contiguous  $2 \times 2$  m quadrats in subalpine meadow of western Colorado (data from Thomson et al. [1996]). Surfaces correspond to seven selected 100th regression quantile estimates in Table 3. One outlying count of 72 seedlings at 16 flowers was not plotted but was used to estimate regression quantiles.

Regression quantile estimates must fit through sample data points. Therefore, estimates for upper quantiles will be consistent with changes expected when the measured factor is the active limiting constraint only if samples are taken across temporal and spatial scales large enough to include some sample units where the measured factor is the active constraint, or is minimally impacted by factors not measured. When this is an unreasonable assumption, the distributional approach of Kaiser et al. (1994) could be used to provide estimates of limiting relations that occur beyond the range of the sample data. Substantive subject matter theory about the data generation process then must be relied upon to justify distributional assumptions used to derive the estimates (e.g., Kaiser et al. 1994: Figs. 2 and 4), because, inherently, there is no good data-dependent statistical procedure to assess model fit.

Regression quantiles are appropriate for modeling limiting relationships when one variable is clearly the dependent variable (e.g., acorn biomass) and the others are clearly independent variables (e.g., oak forest characteristics), or when, for the sake of a well-developed statistical methodology, the analyst is willing to treat one variable as dependent. Some estimation and hypothesis-testing procedures for regression quantiles may be unfamiliar, but many procedures, such as data transformations to achieve linearity, variable selection, use of indicator variables for categorical variables, and interpretation of parameter estimates, are straightforward extensions of familiar linear modeling theory for least squares regression. Regression quantiles fit smooth functions to data without requiring the subjective groupings to calculate summary values (e.g., the maximum) associated with methods used by Johnson et al. (1989), Blackburn et al. (1992), Griffiths (1992), and Thomson et al. (1996). Regression quantiles readily accommodate multiple predictor variables, something that is more problematic for methods that require binning data into groups. Methods also have been proposed to define functions near the extremes of data distributions based on percentiles of a standard normal distribution for estimating endpoints of a prediction interval in a least squares linear regression model (Bilen 1996, Hubert et al. 1996). These methods are similar to regression quantiles in spirit, but have an assumption of normally distributed errors that is much stronger and more restrictive than when estimating confidence intervals for the conditional mean in least squares regression.

Our applications of regression quantiles stressed estimating changes (slopes) in the upper quantiles of distributions as a function of limiting factors. The rank-score test employing the normality of rank-score statistics provides a method with asymptotic validity for testing hypotheses and constructing confidence intervals for slope parameters. However, more reliable probabilities and confidence interval coverage for small samples and more extreme quantiles might be possible

by evaluating the permutation distribution of the rank-score statistic under the null hypothesis. Rank-score tests can be considered a special case of the distance functions in multiresponse permutation procedures (Mielke and Berry 1983, Zimmerman et al. 1985, Tracy and Tajuddin 1986). The rank-score test does not necessarily provide useful intervals for  $\hat{y}$  at specified  $\mathbf{X}$ , including the intercept ( $\mathbf{X} = 0$ ). Mathematically, it is possible to compute confidence intervals for the intercept with the rank-score test, but current theory does not address testing the intercept (Gutenbrunner et al. 1993). Estimating confidence intervals or confidence bands may be important in some ecological applications when predicted values for limiting relations are desired. Recently, Zhou and Portnoy (1998) extended the direct-order statistic estimates of Zhou and Portnoy (1996) to heteroscedastic regression quantile models for constructing confidence and prediction intervals for  $\hat{y}$  at  $\mathbf{X} = \mathbf{x}$ .

Several recently developed procedures might prove useful in more comprehensive applications of regression quantiles for estimating effects of limiting factors. He (1997) presents a restricted regression quantile estimation approach for location-scale heteroscedastic models that limits the number of unique solutions to at most  $n$ , and prevents crossing of estimated quantiles within the domain of  $\mathbf{X}$ . Nonlinear functions can be estimated with nonparametric smoothing splines (Koenker et al. 1994) and parametric models (Welsh et al. 1994, Koenker and Park 1996). Schwarz (Koenker et al. 1994) or Akaike information criterion (Hurvich and Tsai 1990) can be used with regression quantiles to aid in model selection, similar to applications with other regression estimators. All distinct regression quantile solutions could be used to more completely quantify the structure in data distributions (Bassett and Koenker 1982, Gutenbrunner et al. 1993, Koenker 1994, Koenker and d'Orey 1994).

Regression quantiles are an addition to a small set of statistical procedures (Maller 1990, Kaiser et al. 1994, Thomson et al. 1996) that have been developed to estimate effects of limiting factors when it is known a priori that only a subset of those factors was measured and included in a modeled relationship. These procedures estimate parameters describing changes near the extremes of biological response distributions, which are inherently less precise than estimates of central tendency, regardless of the method employed. The undesirable, low precision of estimates near extremes of distributions is offset by the greater magnitude of estimated effects and increased relevance to ecological limiting factors. The other obvious statistical solutions of measuring all relevant factors in observational studies, or randomizing factors of interest in an experimental design, simply are not possible in many ecological investigations. We believe that expanding data analysis to include estimation of changes in multiple quantiles of response distributions in order to examine

effects of limiting factors will aid development of ecological theory and its application to important resource management issues.

## ACKNOWLEDGMENTS

We thank J. D. Thomson for providing the data on glacier lilies and for reviewing drafts of the manuscript. We also thank R. Koenker, P. W. Mielke, Jr., C. F. Rabeni, and J. E. Roelle for reviewing drafts of the manuscript. The comments of P. Dixon and two anonymous reviewers greatly improved the manuscript. R. Koenker and W. L. Mangus provided invaluable assistance with implementing the rank-score tests for regression quantiles in S-Plus.

## LITERATURE CITED

- Barrodale, I., and F. D. K. Roberts. 1974. Algorithm 478: Solution of an overdetermined system of equations in the  $l_1$  norm. Communications of the Association for Computing Machinery 17:319-320.
- Bassett, G. 1992. The Gauss Markov property for the median. Pages 23-31 in Y. Dodge, editor. *L<sub>1</sub> Statistical analyses*. Elsevier Science (North Holland), Amsterdam, The Netherlands.
- Bassett, G., and R. Koenker. 1982. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association* 77:407-415.
- Bassett, G., and R. Koenker. 1986. Strong consistency of regression quantiles and related empirical processes. *Econometric Theory* 2:191-201.
- Bilen, C. 1996. Computation of confidence bands for percentile lines in the general linear model. Thesis. University of Wyoming, Laramie, Wyoming, USA.
- Birkes, D., and Y. Dodge. 1993. *Alternative methods of regression*. Wiley, New York, New York, USA.
- Blackburn, T. M., and K. J. Gaston. 1998. Some methodological issues in macroecology. *American Naturalist* 151:68-83.
- Blackburn, T. M., J. H. Lawton, and J. N. Perry. 1992. A method of estimating the slope of upper bounds of plots of body size and abundance in natural animal assemblages. *Oikos* 65:107-112.
- Brown, J. H., and B. A. Maurer. 1987. Evolution of species assemblages: effects of energetic constraints and species dynamics on the diversification of the North American avifauna. *American Naturalist* 130:1-17.
- Buchinsky, M. 1991. The theory and practice of quantile regression. Dissertation. Harvard University, Cambridge, Massachusetts, USA.
- . 1995. Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987. *Journal of Econometrics* 65:109-154.
- Cade, B. S., and J. D. Richards. 1996. Permutation tests for least absolute deviation regression. *Biometrics* 52:886-902.
- Christisen, D. M., and W. H. Kearby. 1984. Mast measurement and production in Missouri (with special reference to acorns). Missouri Department of Conservation Terrestrial Series No. 13.
- Goldberg, D. E., and S. M. Scheiner. 1993. ANOVA and ANCOVA: field competition experiments. Pages 69-93 in S. M. Scheiner and J. Gurevitch, editors. *Design and analysis of ecological experiments*. Chapman and Hall, New York, New York, USA.
- Griffiths, D. 1992. Size, abundance, and energy use in communities. *Journal of Animal Ecology* 61:307-315.
- Gutenbrunner, C., and J. Jurečková. 1992. Regression rank scores and regression quantiles. *Annals of Statistics* 20:305-330.
- Gutenbrunner, C., J. Jurečková, R. Koenker, and S. Portnoy. 1993. Tests of linear hypotheses based on regression rank scores. *Nonparametric Statistics* 2:307-331.
- He, X. 1997. Quantile curves without crossing. *American Statistician* 51:186-192.
- Hubert, W. A., T. D. Marwitz, K. G. Gerow, N. A. Binns, and R. W. Wiley. 1996. Estimation of potential maximum biomass of trout in Wyoming streams to assist management decisions. *North American Journal of Fisheries Management* 16:821-829.
- Hurvich, C. M., and C.-L. Tsai. 1990. Model selection for least absolute deviations regression in small samples. *Statistics and Probability Letters* 9:259-265.
- Huškova, M. 1994. Some sequential procedures based on regression rank scores. *Nonparametric Statistics* 3:285-298.
- Jarvis, P. G. 1976. The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philosophy Transactions of Royal Society of London B* 273:593-610.
- Johnson, D. J., M. C. Hammond, T. L. McDonald, C. L. Nustad, and M. D. Schwartz. 1989. Breeding canvasbacks: a test of a habitat model. *Prairie Naturalist* 21:193-202.
- Kaiser, M. S., P. L. Speckman, and J. R. Jones. 1994. Statistical models for limiting nutrient relations in inland waters. *Journal of the American Statistical Association* 89:410-423.
- Koenker, R. 1994. Confidence intervals for regression quantiles. Pages 349-359 in P. Mandl and M. Huškova, editors. *Asymptotic statistics: Proceedings of the Fifth Prague Symposium*. Physica-Verlag, Heidelberg, Germany.
- Koenker, R., and G. Bassett. 1978. Regression quantiles. *Econometrica* 46:33-50.
- Koenker, R., and G. Bassett. 1982. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* 50:43-61.
- Koenker, R., and V. d'Orey. 1987. Computing regression quantiles. *Applied Statistics* 36:383-393.
- Koenker, R., and V. d'Orey. 1994. A remark on Algorithm AS229: computing dual regression quantiles and regression rank scores. *Applied Statistics* 43:410-414.
- Koenker, R., P. Ng, and S. Portnoy. 1994. Quantile smoothing splines. *Biometrika* 81:673-680.
- Koenker, R., and B. J. Park. 1996. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics* 71:265-283.
- Koenker, R., and S. Portnoy. 1996. Quantile regression. University of Illinois at Urbana-Champaign, College of Commerce and Business Administration, Office of Research Working Paper 97-0100.
- Maller, R. A. 1990. Some aspects of a mixture model for estimating the boundary of a set of data. *Journal du Conseil International pour l'Exploration de la Mer* 46:140-147.
- Maller, R. A., E. S. de Boer, L. M. Joll, D. A. Anderson, and J. P. Hinde. 1983. Determination of the maximum foregut volume of western rock lobsters (*Panulirus cygnus*) from field data. *Biometrics* 39:543-551.
- Mielke, P. W., Jr., and K. J. Berry. 1983. Asymptotic clarifications, generalizations, and concerns regarding an extended class of matched pairs tests based on powers of ranks. *Psychometrika* 48:483-485.
- Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. *Applied linear statistical models*. Fourth edition. Richard D. Irwin, Homewood, Illinois, USA.
- Portnoy, S. 1991. Asymptotic behavior of the number of regression quantile breakpoints. *SIAM Journal of Science and Statistical Computing* 12:867-883.
- Powell, J. L. 1986. Censored regression quantiles. *Journal of Econometrics* 32:143-155.
- Rabinowitz, D., J. K. Rapp, P. M. Dixon, and A. T. Khieu. 1985. Separating structural and developmental variability

- in growth rate estimates for *Andropogon scoparius* Michx. Bulletin of the Torrey Botanical Club **112**:403–408.
- Schroeder, R. L., and L. D. Vangilder. 1997. Tests of wildlife habitat models to evaluate oak mast production. Wildlife Society Bulletin **25**:639–646.
- Sork, V. L., J. Bramble, and O. Sexton. 1993. Ecology of mast-fruited in three species of North American deciduous oaks. Ecology **74**:528–541.
- Terrell, J. W., B. S. Cade, J. Carpenter, and J. M. Thompson. 1996. Modeling stream fish habitat limitations from wedged-shaped patterns of variation in standing stock. Transactions of the American Fisheries Society **125**:104–117.
- Thomson, J. D., G. Weiblen, B. A. Thomson, S. Alfaro, and P. Legendre. 1996. Untangling multiple factors in spatial distributions: Lilies, gophers, and rocks. Ecology **77**:1698–1715.
- Tracy, D. S., and I. H. Tajuddin. 1986. Empirical power comparisons of two MRPP rank tests. Communications in Statistics—Theory and Methods **15**:551–570.
- Venables, W. N., and B. D. Ripley. 1994. Modern applied statistics with S-Plus. Springer-Verlag, New York, New York, USA.
- Welsh, A. H., R. J. Carroll, and D. Rupert. 1994. Fitting heteroscedastic regression models. Journal of American Statistical Association **89**:100–116.
- White, G. C., and R. E. Bennetts. 1996. Analysis of frequency count data using the negative binomial distribution. Ecology **77**:2549–2557.
- Zhou, K. Q., and S. L. Portnoy. 1996. Direct use of regression quantiles to construct confidence sets in linear models. Annals of Statistics **24**:287–306.
- Zhou, K. Q., and S. L. Portnoy. 1998. Statistical inference on heteroscedastic models based on regression quantiles. Nonparametric Statistics **9**:239–260.
- Zimmerman, G. W., H. Goetz, and P. W. Mielke, Jr. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. Ecology **66**:606–611.

## APPENDIX

*Regression quantile estimation.*—The right-continuous distribution function for any real-valued random variable  $Y$  is  $F(y) = P(Y \leq y)$  and the  $\tau$ th population quantile of  $Y$  ( $0 < \tau < 1$ ) is  $F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$ . The  $\tau$ th population quantile of  $\mathbf{y}$  conditional on  $\mathbf{X}$  in the linear model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \nu(\mathbf{X})\mathbf{e}$  is defined as  $Q_Y(\tau | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$  and  $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \nu(\cdot)F_e^{-1}(\tau)$ , where  $\mathbf{y}$  is an  $n \times 1$  vector of dependent responses,  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of unknown regression parameters,  $\mathbf{X}$  is an  $n \times p$  matrix of predictors,  $\nu(\cdot) > 0$  is some known function, and  $\mathbf{e}$  is an  $n \times 1$  vector of random errors that are iid as  $F$ . We assume that the first column of  $\mathbf{X}$  consists of 1's (an intercept). If heteroscedastic errors occur as a linear function through the predictors,  $\nu(\mathbf{X}) = (\text{diag}(\mathbf{X}\boldsymbol{\gamma}))$ , where  $\boldsymbol{\gamma}$  is a  $p \times 1$  vector of unknown scale parameters, then we have the familiar location-scale model of heteroscedasticity and  $Q_Y(\tau | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$  and  $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \boldsymbol{\gamma}F_e^{-1}(\tau)$  (Koenker and Bassett 1982, Buchinsky 1991, Gutenbrunner and Jurečková 1992). Homoscedastic regression models are a special case when  $\boldsymbol{\gamma} = (1, 0, \dots, 0)'$  and  $Q_Y(\tau | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$ ,  $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + (F_e^{-1}(\tau), 0, \dots, 0)'$ , because all parameters other than the intercept ( $\beta_0$ ) in  $\boldsymbol{\beta}(\tau)$  are the same for all  $\tau$ . Other forms of heteroscedasticity that are not simple location-scale forms are possible (Koenker and Portnoy 1996). If  $\mathbf{X}$  is a single column of 1's, then  $Q_Y(\tau | \mathbf{X}) = F^{-1}(\tau)$ , i.e., the usual one-sample  $\tau$ th quantile. If  $\mathbf{X}$  is a sequence of 0, 1 indicator variables denoting categorical group membership, then  $Q_Y(\tau | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$ ,  $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + F_e^{-1}(\tau)$  provides the location of the  $\tau$ th quantile for one group ( $\beta_0$ ) and differences between the corresponding  $\tau$ th quantiles of the other groups ( $\beta_p, p \geq 1$ ).

The assumption imposed on  $F_e$  to estimate regression quantiles is that a  $\tau$ th quantile of  $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}(\tau)$  conditional on  $\mathbf{X}$  equals 0,  $F_e^{-1}(\tau | \mathbf{X}) = 0$ . Estimates,  $\mathbf{b}(\tau)$ , of  $\boldsymbol{\beta}(\tau)$  are solutions to the following minimization problem:

$$\min \left[ \sum_{i=1}^n \rho_\tau \left( y_i - \sum_{j=0}^p b_j x_{ij} \right) \right] \quad (\text{A.1})$$

where  $\rho_\tau(e) = e(\tau - I(e < 0))$ , and  $I(\cdot)$  is the indicator function. The estimating equations in A.1 are solved by a modification of the Barrodale and Roberts (1974) simplex linear program for any specified value of  $\tau$  (Koenker and d'Orey 1987). With little additional computation, the entire regression quantile function for all distinct values of  $\tau$  can be estimated (Koenker and d'Orey 1987, 1994). By expanding estimating function (A.1) to

$$\min \left[ \sum_{i \in \{1 | y_i \geq b_j x_{ij}\}} \tau \left| y_i - \sum_{j=0}^p b_j x_{ij} \right| + \sum_{i \in \{1 | y_i < b_j x_{ij}\}} (1 - \tau) \left| y_i - \sum_{j=0}^p b_j x_{ij} \right| \right] \quad (\text{A.2})$$

it can be seen that positive and negative residuals are differentially weighted for regression quantiles other than  $\tau = 0.5$ .

*Rank-score hypothesis tests.*—The essence of the  $\tau$ -quantile rank-score procedure (Koenker 1994) is that rank scores are calculated based on the  $n \times 1$  vector of dual linear programming solutions,  $\mathbf{a}(\tau) = [0, 1]^n$ , from estimating the reduced parameter model  $\mathbf{y} - \mathbf{x}_2 \boldsymbol{\xi}(\tau) = \mathbf{X}_1 \boldsymbol{\beta}_1(\tau) + \nu(\mathbf{X}_1)\mathbf{e}$ , where  $\mathbf{y}$ ,  $\nu(\cdot)$ , and  $\mathbf{e}$  are as previously defined,  $\boldsymbol{\beta}_1(\tau)$  is a  $(p - 1) \times 1$  vector of unknown nuisance regression parameters,  $\mathbf{X}_1$  is an  $n \times (p - 1)$  matrix of predictors,  $\mathbf{x}_2$  is an  $n \times 1$  vector of predictors, and  $\boldsymbol{\beta}_2(\tau)$  is the scalar parameter specified by the null hypothesis  $H_0: \boldsymbol{\beta}_2(\tau) = \boldsymbol{\xi}(\tau)$  (frequently  $\boldsymbol{\xi}(\tau) = 0$ ) for the full parameter model  $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1(\tau) + \mathbf{x}_2 \boldsymbol{\beta}_2(\tau) + \nu(\mathbf{X}_1)\mathbf{e}$ . The  $n \times 1$  vector of rank scores  $\mathbf{s}(\tau) = \mathbf{a}(\tau) - (1 - \tau)\mathbf{1}$  is used in the test statistic  $S(\tau) = n^{-0.5} \mathbf{x}_2' \mathbf{s}(\tau)$ , which is asymptotically normally distributed with  $\mu = 0$  and  $\sigma^2 = \tau(1 - \tau)q^2$ , where  $q^2 = n^{-1} \mathbf{x}_2' (\mathbf{I} - \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1') \mathbf{x}_2$ . The standardized test statistic  $T(\tau) = S(\tau) [\tau(1 - \tau)q^2]^{-0.5}$  is referenced to the standard normal distribution to calculate probabilities under the null hypothesis.

The elements of  $\mathbf{a}(\tau)$  are 1 when the residuals for the reduced model are positive, 0 when the residuals are negative, and in the interval (0, 1) when the residuals are 0 (i.e., the points fit exactly by the  $\tau$ th regression quantile). Rank scores  $\mathbf{s}(\tau)$  are, thus,  $\tau$  for positive residuals,  $\tau - 1$  for negative residuals, and in the interval  $(\tau - 1, \tau)$  when residuals are 0. Validity of the rank-score test requires an assumption of positive density for  $y$  at the estimate,  $f(F^{-1}(\tau)) > 0$ .

Confidence intervals calculated by inverting this test statistic are centered on the estimate, because  $S(\tau) = 0$  for  $H_0$ ;  $\boldsymbol{\beta}_2(\tau) = \boldsymbol{\xi}(\tau) = \mathbf{b}_2(\tau)$ , but are not necessarily symmetric (Koenker 1994). By alternating which independent variables are being tested by the null hypothesis and which are considered nuisance parameters, one can obtain confidence intervals for each independent variable conditioned on the others being in the model. Because the sampling distribution of the rank-score test statistic is discontinuous, we followed Koenker (1994) and interpolated between adjacent hypothesized values of  $\boldsymbol{\beta}_2(\tau) = \boldsymbol{\xi}(\tau)$  for constructing confidence in-

tervals. Although we emphasize testing a single parameter because of its connection to constructing confidence intervals, it is possible to simultaneously test multiple parameters with the quantile rank-score test and to use a  $\chi^2$  distribution with  $p$  degrees of freedom (where  $p$  is number of parameters tested) to approximate the  $P$  value (Gutenbrunner et al. 1993, Koenker 1994).

*Source of computer programs.*—Script files and Fortran

code to implement regression quantiles and the rank-score tests in S-Plus are available in ESA's Electronic Data Archive: *Ecological Archives* E080-001. The BLOSSOM software available at the web site of the Midcontinent Ecological Science Center estimates regression quantiles, but uses a permutation procedure (Cade and Richards 1996) to test hypotheses.