A gentle introduction to quantile regression for ecologists

Brian S Cade[1,2] and Barry R Noon[3]

Quantile regression is a way to estimate the conditional quantiles of a response variable distribution in the linear model that provides a more complete view of possible causal relationships between variables in ecological processes. Typically, all the factors that affect ecological processes are not measured and included in the statistical models used to investigate relationships between variables associated with those processes. As a consequence, there may be a weak or no predictive relationship between the mean of the response variable (*y*) distribution and the measured predictive factors (*X*). Yet there may be stronger, useful predictive relationships with other parts of the response variable distribution. This primer relates quantile regression estimates to prediction intervals in parametric error distribution regression models (eg least squares), and discusses the ordering characteristics, interval nature, sampling variation, weighting, and interpretation of the estimates for homogeneous and heterogeneous regression models.

Front Ecol Environ 2003; 1(8): 412–420

Regression is a common statistical method employed by scientists to investigate relationships between variables. Quantile regression (Koenker and Bassett 1978) is a method for estimating functional relations between variables for all portions of a probability distribution. Although it has begun to be used in ecology and biology (Table 1), many ecologists remain unaware of it, as it was developed relatively recently and is rarely taught in statistics courses at many universities. We present this introduction both to encourage additional applications in ecology and to educate those who are already contemplating using the method.

In a nutshell:

- Quantile regression is a statistical method that could be used effectively by more ecologists
- Statistical distributions of ecological data often have unequal variation due to complex interactions between the factors affecting organisms that cannot all be measured and accounted for in statistical models
- Unequal variation implies that there is more than a single slope (rate of change) describing the relationship between a response variable and predictor variables measured on a subset of these factors
- Quantile regression estimates multiple rates of change (slopes) from the minimum to maximum response, providing a more complete picture of the relationships between variables missed by other regression methods
- The ecological concept of limiting factors as constraints on organisms often focuses on rates of change in quantiles near the maximum response, when only a subset of limiting factors are measured

[1]*Fort Collins Science Center, US Geological Survey, Fort Collins, CO (brian_cade@usgs.gov);* [2]*Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO;* [3]*Department of Fishery and Wildlife Biology, Colorado State University, Fort Collins, CO*

Typically, a response variable y is some function of predictor variables X, so that $y = f(X)$. Most regression applications in the ecological sciences focus on estimating rates of change in the mean of the response variable distribution as some function of a set of predictor variables; in other words, the function is defined for the expected value of y conditional on X, $E(y|X)$. Mosteller and Tukey (1977) noted that it was possible to fit regression curves to other parts of the distribution of the response variable, but that this was rarely done, and therefore most regression analyses gave an incomplete picture of the relationships between variables. This is especially problematic for regression models with heterogeneous variances, which are common in ecology. A regression model with heterogeneous variances implies that there is not a single rate of change that characterizes changes in the probability distributions. Focusing exclusively on changes in the means may underestimate, overestimate, or fail to distinguish real nonzero changes in heterogeneous distributions (Terrell *et al.* 1996; Cade *et al.* 1999).

The Dunham *et al.* (2002) analyses relating the abundance of Lahontan cutthroat trout (*Oncorhynchus clarki henshawi*) to the ratio of stream width to depth illustrates the value of the additional information provided by quantile regression (Figure 1). The ratio was used as a predictor variable because it was an easily obtained measure of channel morphology that was thought to be related to the integrity of habitat in small streams like those typically inhabited by cutthroat trout. Quantile regression estimates indicated a nonlinear, negative relationship with the upper 30% (≥ 70th percentiles, $P \leq 0.10$ for H_0: $\beta_1 = 0$) of cutthroat densities across 13 streams and 7 years. A weighted least squares regression estimated zero change (90% confidence intervals of –0.014 to 0.012, $P = 0.901$ for H_0: $\beta_1 = 0$) in mean densities with stream width to depth. If the authors

had used mean regression estimates, they would have mistakenly concluded that there was no relation between trout densities and the ratio of stream width to depth.

Many ecological applications have used quantile regression as a method of estimating rates of change for functions along or near the upper boundary of the conditional distribution of responses because of issues raised by Kaiser *et al.* (1994), Terrell *et al.* (1996), Thomson *et al.* (1996), Cade *et al.* (1999), and Huston (2002). These authors suggested that if ecological limiting factors act as constraints on organisms, then the estimated effects for the measured factors were not well represented by changes in the means of response variable distributions, when there were many other unmeasured factors that were potentially limiting. The response of the organism cannot change by more than some upper limit set by the measured factors, but may change by less when other unmeasured factors are limiting (Figure 2). This analytical problem is closely related to the more general statistical issue of hidden bias in observational studies due to confounding with unmeasured variables (Rosenbaum 1995). Multiplicative interactions among measured and unmeasured ecological factors that contribute to this pattern are explored in more detail relative to regression quantile estimates and inferences in Cade (2003).

Self-thinning of annual plants in the Chihuahuan desert of the southwestern US (Cade and Guo 2000) is an example of a limiting factor where conventional mean regression was inappropriate for estimating the reduction in densities of mature plants with increasing germination densities of seedlings (Figure 3). The effects of this density-dependent process were best revealed at the higher plant densities associated with upper quantiles, where competition for resources was greatest and effects of factors other than intraspecific competition were minimal. Changes in the upper quantiles of densities of mature plants were essentially a 1:1 mapping of germination density when it was low (< 100/0.25-m^2), whereas there was a strong decrease in density of mature plants at higher germination densities consistent with density-dependent competition.

Table 1. Applications of quantile regression in ecology and biology

Running speed and body mass of terrestrial mammals	Koenker *et al.* 1994
Global temperature change over the last century	Koenker and Schorfheide 1994
Animal habitat relationships	Terrell *et al.* 1996; Haire *et al.* 2000; Eastwood *et al.* 2001; Dunham *et al.* 2002
Prey and predator size relationships	Scharf *et al.* 1998
Plant self-thinning	Cade and Guo 2000
Vegetation changes associated with agricultural conservation practices	Allen *et al.* 2001
Mediterranean fruit fly survival	Koenker and Geling 2001
Body size of deep-sea gastropods and dissolved oxygen concentration	McClain and Rex 2001
Variation in nuclear DNA of plants across environmental gradients	Knight and Ackerly 2002
Plant species diversity and invasibility	Brown and Peet 2003

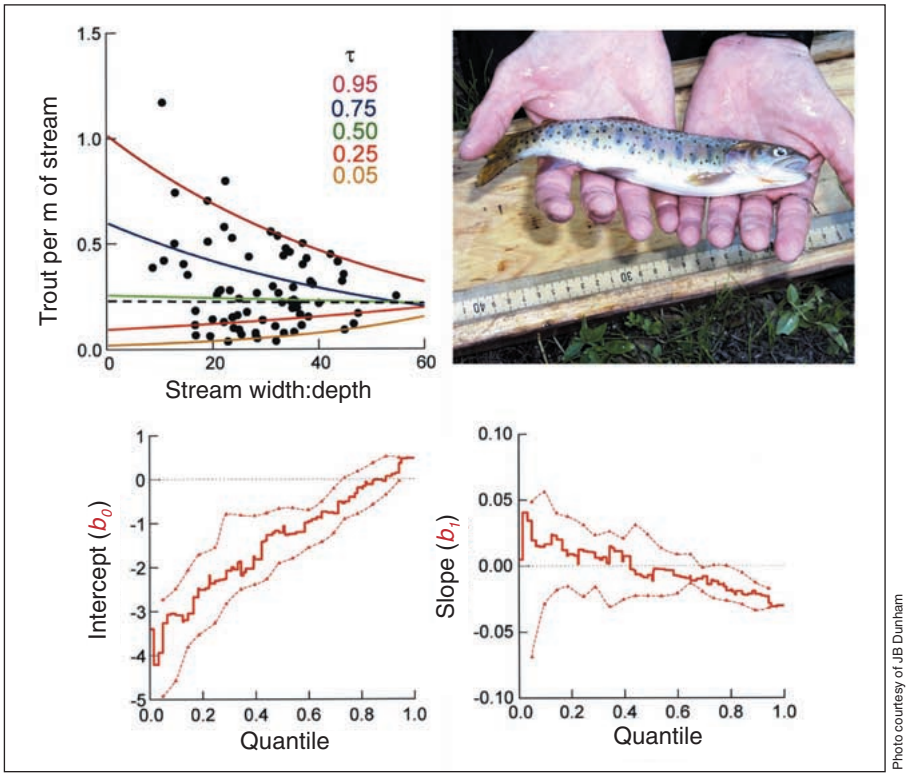

Figure 1. *Quantile regression was used to estimate changes in Lahontan cutthroat trout density (y) as a function of the ratio of stream width to depth (X) for 7 years and 13 streams in the eastern Lahontan basin of the western US. Photo is of a typical adult Lahontan cutthroat from these small streams. (top left) A scatterplot of* n = *71 observations of stream width:depth and trout densities with 0.95, 0.75, 0.50, 0.25, and 0.05 quantile (solid lines) and least squares regression (dashed line) estimates for the model* ln y = $\beta_0 + \beta_1 X + \varepsilon$. *Weighted estimates used here were based on weights = 1/(1.310 –0.017X) but did not differ substantially from unweighted estimates used by Dunham* et al. *(2002). Sample estimates, $b_0(\tau)$ (bottom left) and $b_1(\tau)$ (bottom right), are shown as a red step function. Red dashed lines connect endpoints of 90% confidence intervals.*

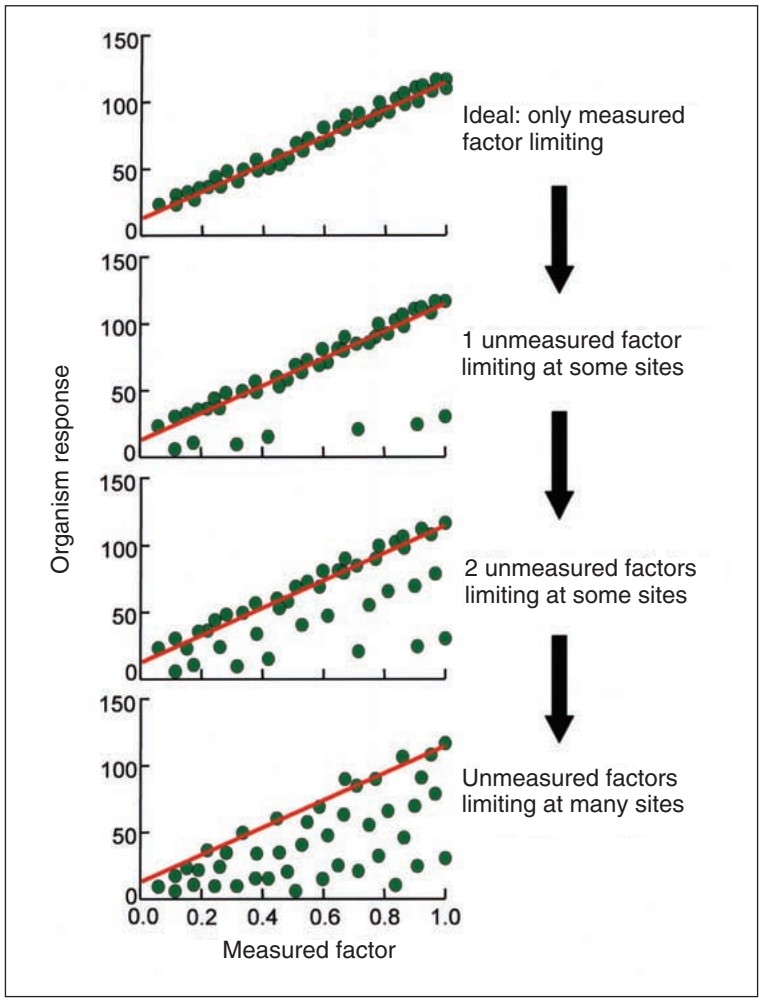

Figure 2. *The top graph represents the ideal statistical situation where an organism response is driven primarily by the measured factor(s) included in the linear regression model; ie all other potential limiting factors are at permissive levels. As we proceed from top to bottom, an increasing number of factors that were not measured become limiting at some sample locations and times, increasing the heterogeneity of organism response with respect to the measured factor(s) included in the regression model.*

Quantile regression was developed in the 1970s by econometricians (Koenker and Bassett 1978) as an extension of the linear model for estimating rates of change in all parts of the distribution of a response variable. The estimates are semiparametric in the sense that no parametric distributional form (eg normal, Poisson, negative binomial, etc) is assumed for the random error part of the model ε, although a parametric form is assumed for the deterministic portion of the model (eg, $\beta_0 X_0 + \beta_1 X_1$). The conditional quantiles denoted by $Q_y(\tau \mid X)$ are the inverse of the conditional cumulative distribution function of the response variable, $F_y^{-1}(\tau \mid X)$, where $\tau \in [0, 1]$ denotes the quantiles (Cade *et al.* 1999; Koenker and Machado 1999). For example, for $\tau = 0.90$, $Q_y(0.90 \mid X)$ is the 90th percentile of the distribution of y conditional on the values of X; in other words, 90% of the values of y are less than or equal to the specified function of X. Note that for symmetric distributions, the 0.50 quantile (or median) is equal to the mean μ.

Here we consider functions of X that are linear in the parameters; eg $Q_y(\tau \mid X) = \beta_0(\tau)X_0 + \beta_1(\tau)X_1 + \beta_2(\tau)X_2 + , \ldots , + \beta_p(\tau)X_p$, where the (τ) notation indicates that the parameters are for a specified τ quantile. The parameters vary with τ due to effects of the τth quantile of the unknown error distribution ε. Parameter estimates in linear quantile regression models have the same interpretation as those in any other linear model. They are rates of change conditional on adjusting for the effects of the other variables in the model, but now are defined for some specified quantile.

In the 1-sample setting with no predictor variables, quantiles are usually estimated by a process of ordering the sample data. The beauty of the extension to the regression model was the recognition that quantiles could be estimated by an optimization function minimizing a sum of weighted absolute deviations, where the weights are asymmetric functions of τ (Koenker and Bassett 1978; Koenker and d'Orey 1987). Currently, the statistical theory and computational routines for estimating and making inferences on regression quantiles are best developed for the linear model (Gutenbrunner *et al.* 1993; Koenker 1994; Koenker and Machado 1999; Cade 2003), but are also available for parametric nonlinear (Welsh *et al.* 1994; Koenker and Park 1996) and nonparametric, nonlinear smoothers (Koenker *et al.* 1994; Yu and Jones 1998).

Quantile regression models present many new possibilities for the statistical analysis and interpretation of ecological data. With those new possibilities come new challenges related to estimation, inference, and interpretation. Here we provide an overview of several of the issues ecologists are likely to encounter when conducting and interpreting quantile regression analyses. More technical discussion is provided in papers cited in References.

Quantiles and ordering in the linear model

Regression quantile estimates are an ascending sequence of planes that are above an increasing proportion of sample observations with increasing values of the quantiles τ (Figure 4). It is this operational characteristic that extends the concepts of quantiles, order statistics, and rankings to the linear model (Gutenbrunner *et al.* 1993; Koenker and Machado 1999). The proportion of observations less than or equal to a given regression quantile estimate – for example, the 90th percentile given by $Q_y(0.90 \mid X)$ in Figure 4 – will not in general be exactly equal to τ. The simplex linear programming solution minimizing the sum of weighted absolute deviations ensures that any regression quantile estimate will fit through at least $p + 1$ of the n sample observations for a model with $p + 1$ predictor

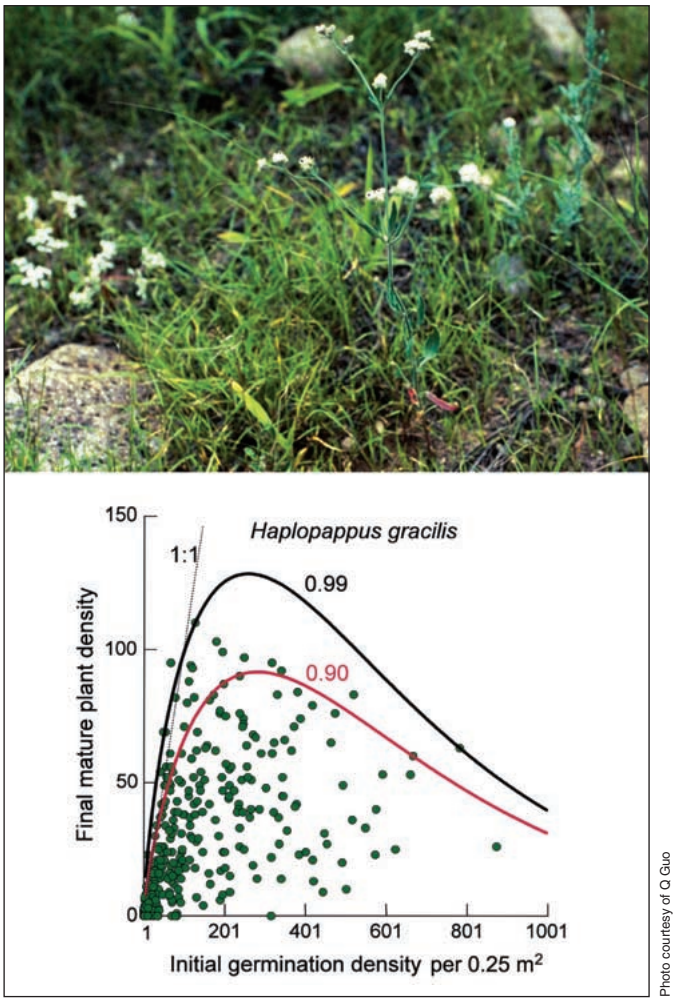

Photo courtesy of Q Guo

Figure 3. *Quantile regression estimates (0.99 and 0.90) were used to describe changes in survival of Chihuahuan desert annuals by modeling changes in mature plant density (y) as a function of germination density of seedlings (X). Here the modified Ricker function,* $y = \beta_0 X^{\beta_1} e^{\beta_2 X} \varepsilon$*, was estimated in its linearized form,* $ln(y + 1) = ln(\beta_0) + \beta_1 ln(X) + \beta_2 X + ln(\varepsilon)$ *for* Haplopappus gracilis *in n = 346 0.25-m² quadrats at Portal, AZ (Cade and Guo 2000). The 1:1 relationship is shown as a dotted black line. Photo is* Eriogonum abertianum*, another common annual plant found in the Chihuahuan desert.*

variables X. This results in a set of inequalities defining a range for the proportion of observations less than or equal to any selected quantile τ, given *n* and *p* (Cade *et al.* 1999; Koenker and Machado 1999).

Regression quantiles, like the usual 1-sample quantiles with no predictor variables, retain their statistical properties under any linear or nonlinear monotonic transformation of y as a consequence of this ordering property; that is, they are equivariant under monotonic transformation of y (Koenker and Machado 1999). Thus it is possible to use a nonlinear transformation (eg logarithmic) of y to estimate linear regression quantiles and then back transform the estimates to the original scale (a nonlinear function) without any loss of information. This, of course, is not true with means, including those from regression models.

Homogeneous and heterogeneous models

The simplest, unconstrained form of regression quantile estimates allows the predictor variables (X) to exert changes on the central tendency, variance, and shape of the response variable (y) distribution (Koenker and Machado 1999). When the only estimated effect is a change in central tendency (for example, means) of the distribution of y conditional on the values of X, we have the familiar homogeneous variance regression model associated with ordinary least squares regression (Figure 4, top). All the regression quantile slope estimates $b_1(\tau)$ are for a common parameter, and any deviation among the regression quantile estimates is simply due to sampling variation (Figure 4, bottom). An estimate of the rate of change in the means from ordinary least squares regression is also an estimate of the same parameter as for the regression quantiles. The intercept estimates $b_0(\tau)$ of the quantile regression model are for the parametric quantile $\beta_0(\tau)$ of y when $X_1, X_2, ..., X_p = 0$, which differ across quantiles and for the mean μ (Figure 4, middle). Here the primary virtue of the regression quantile estimates of the intercept is that they are not dependent on an assumed form of the error distribution as when least squares regression is used, which assumes a normal error distribution.

The properties associated with the intercept translate to any other specified value of the predictors $X_1, X_2, ..., X_p$ as when estimating prediction intervals (Neter *et al.* 1996). The interval between the 0.90 and 0.10 regression quantile estimates in Figure 4 (top) at any specified value of $X_1 = x_1$ is an 80% prediction interval for a single future observation of y. Prediction intervals (for some number of future observations) based on assuming a normal error distribution, as is done in ordinary least squares regression, are sensitive to departures from this assumption (Neter *et al.* 1996). Quantile regression avoids this distributional assumption altogether. Given the skewed response distribution in Figure 4 (top), it is easy to see that a symmetric prediction interval about an estimate of the mean based on a normal error distribution model would not have correct coverage. For example, at $X_1 = 70.5$ the 80% prediction interval for a single new observation is 8.43–10.97, based on the least squares estimate assuming a normal error distribution, whereas the interval based on the 0.90 and 0.10 regression quantile estimates is 8.85–10.88.

When the predictor variables X exert both a change in means and a change in variance on the distribution of y, we have a regression model with unequal variances (a "location-scale model" in statistical terminology). As a consequence, changes in the quantiles of y across X cannot be the same for all quantiles (Figure 5). Slope estimates $b_1(\tau)$ differ across quantiles because the parameters $\beta_1(\tau)$ differ, since the variance in y changes as a function of X (Figure 5, bottom). Note that the pattern of changes in estimates $b_0(\tau)$ mirror those for $b_1(\tau)$ (Figure 5, middle

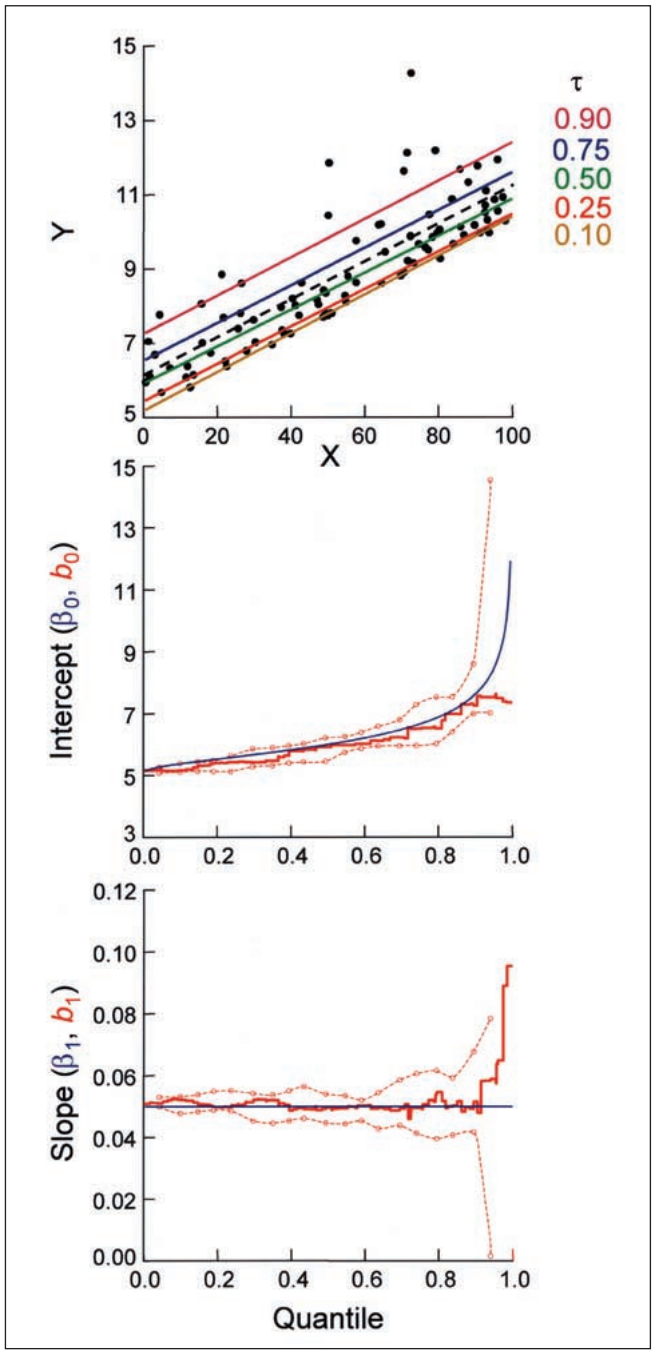

Figure 4. *(top) A sample (n = 90) from a homogenous error (lognormal with median = 0 and σ = 0.75) model, y = β_0 + $\beta_1 X_1$ + ε, β_0 = 6.0 and β_1 = 0.05 with 0.90, 0.75, 0.50, 0.25, and 0.10 regression quantile estimates (solid lines) and least squares estimate of mean function (dashed line). Sample estimates $b_0(\tau)$ (middle) and $b_1(\tau)$ (bottom), are shown as red step functions. Dashed red lines connect endpoints of 90% confidence intervals. Parameters $\beta_0(\tau)$ (middle) and $\beta_1(\tau)$ (bottom), are blue lines.*

and bottom). In this situation, ordinary least squares regression is commonly modified by incorporating weights (which usually have to be estimated) in inverse proportion to the variance function (Neter *et al.* 1996). Typically, the use of weighted least squares is done to improve estimates of the sampling variation for the estimated mean function, and not done specifically to estimate the different rates of change in the quantiles of the distributions of y conditional on X. However, Hubert *et al.* (1996) and Gerow and Bilen (1999) described applications of least squares regression where this might be done. Estimating prediction intervals based on weighted least squares estimates implicitly recognize these unequal rates of change in the quantiles of y (Cunia 1987).

Generalized linear models offer alternative ways to link changes in the variances (σ^2) of y with changes in the mean (μ) based on assuming some specific distributional form in the exponential family – for example, Poisson, negative binomial, or gamma (McCullagh and Nelder 1989). But again, the purpose is usually to provide better estimates of rates of change in the mean (μ) of y rather than estimates in the changes in the quantiles of y that must occur when variances are heterogeneous. Estimating prediction intervals for generalized linear models would implicitly recognize that rates of change in the quantiles of y cannot be the same for all quantiles, and these interval estimates would be linked to and sensitive to violations of the assumed error distribution.

An advantage of using quantile regression to model heterogeneous variation in response distributions is that no specification of how variance changes are linked to the mean is required, nor is there any restriction to the exponential family of distributions. Furthermore, we can also detect changes in the shape of the distributions of y across the predictor variables (Koenker and Machado 1999). Complicated changes in central tendency, variance, and shape of distributions are common in statistical models applied to observational data because of model misspecification. This can occur because the appropriate functional forms are not used (for example, using linear instead of nonlinear), and because all relevant variables are not included in the model (Cade *et al.* 1999; Cade 2003). Failure to include all relevant variables occurs because of insufficient knowledge or ability to measure all relevant processes. This should be considered the norm for observational studies in ecology as it is in many other scientific disciplines.

An example of a response distribution pattern that may involve changes in central tendency, variance, and shape is shown in Figure 6. These data from Cook and Irwin (1985) were collected to estimate how pronghorn (*Antilocapra americana*) densities changed with features of their habitat on winter ranges. Here, shrub canopy cover was the habitat feature used as an indirect measure of the amount of winter forage available. Rates of change in pronghorn densities due to shrub canopy cover (b_1) were fairly constant for the lower 1/3 of the quantiles (0.25 per 1% change in cover), increased moderately for the central 1/3 of the quantiles (0.25–0.50), and doubled (0.50–1.0) in the upper 1/3 of the quantiles (Figure 6, bottom right). Changes in $b_1(\tau)$ do not appear to mirror those for $b_0(\tau)$, indicating that there is more than just a change in central

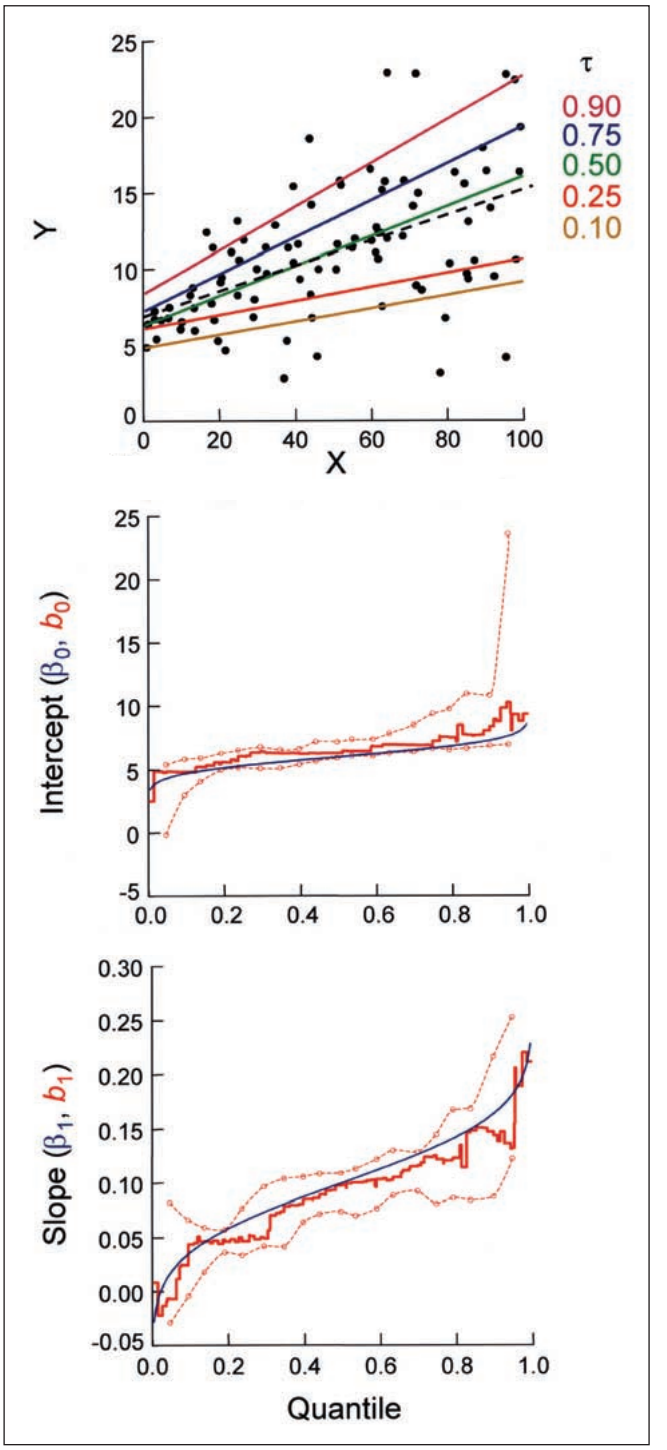

Figure 5. *(top) A sample* (n = 90) *from a heterogeneous error (normal with* $\mu = 0$ *and* $\sigma = 1.0 + 0.05X_1$*) model,* $y = \beta_0 + \beta_1 X_1 + \varepsilon$, $\beta_0 = 6.0$ *and* $\beta_1 = 0.10$ *with 0.90, 0.75, 0.50, 0.25, and 0.10 regression quantile estimates (solid lines) and least squares estimate of mean function (dashed line). Sample estimates* $b_0(\tau)$ *(middle) and* $b_1(\tau)$ *(bottom) are shown as red step function. Red dashed lines connect endpoints of 90% confidence intervals. Parameters* $\beta_0(\tau)$ *(middle) and* $\beta_1(\tau)$ *(bottom) are blue lines.*

tendency and variance of pronghorn densities associated with changes in shrub canopy cover. Clearly, too strong a conclusion is not justified with the small sample (n = 28) and large sampling variation for upper quantiles. But either an ordinary least squares regression estimate (b_1 = 0.483, 90% CI = 0.31–0.66) or more appropriate weighted least squares regression estimate would fail to recognize that pronghorn densities changed at both lower and higher rates as a function of shrub canopy cover at lower and upper quantiles of the density distribution, respectively. Here, regression quantile estimates provide a more complete characterization of an interval of changes in pronghorn densities (0.2–1.0) that were associated with changes in winter food availability as measured by shrub canopy cover. These intervals are fairly large because pronghorn densities on winter ranges are almost certainly affected by more than just food availability.

■ Estimates are for intervals of quantiles

Regression quantile estimates break the interval [0, 1] into a finite number of smaller, unequal length intervals. Thus, while we may refer to and graph the estimated function for a selected regression quantile such as the 0.90, the estimated function actually applies to some small interval of quantiles; for example, [0.894, 0.905] for the 0.90 regression quantile in Figure 4. Unlike the 1-sample quantile estimates, the [0, 1] interval of regression quantile estimates may be broken into more than n intervals that aren't necessarily of equal length $1/n$. The number and length of these intervals are dependent on the sample size, number or parameters, and distribution of the response variable (Portnoy 1991). Estimates plotted as step functions in Figure 4, middle and bottom, are for 101 intervals of quantiles on the interval [0, 1] for which each has an estimate $b_0(\tau)$ and $b_1(\tau)$, corresponding to the intercept and slope. Because the estimates actually apply to a small interval of quantiles, it is appropriate to graph the estimates as a step function.

■ Sampling variation differs across quantiles

It is not surprising that sampling variation differs among quantiles τ. Generally, sampling variation will increase as the value of τ approaches 0 or 1, but the specifics are dependent on the data distribution, model, sample size n, and number of parameters p. Estimates further from the center of the distribution – the median or 50th percentile given by $Qy(0.50|X)$ – usually cannot be estimated as precisely. To display the sampling variation with the estimates (Figure 4, middle and bottom), a confidence band across the quantiles $\tau \in [0, 1]$ was constructed by estimating the pointwise confidence interval for 19 selected quantiles $\tau \in \{0.05, 0.10, ..., 0.95\}$. These intervals were based on inverting a quantile rankscore test (Koenker 1994; Cade *et al.* 1999; Koenker and Machado 1999; Cade 2003). It is possible to compute confidence intervals for all unique intervals of quantiles, but this compu-

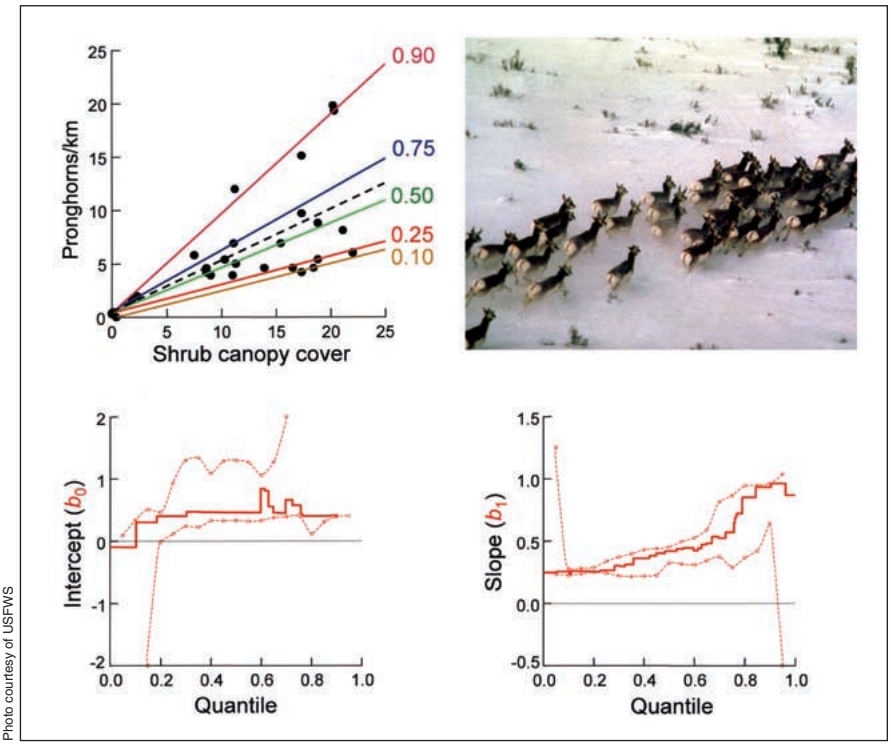

Photo courtesy of USFWS

Figure 6. *(top left) Pronghorn densities* (y) *by shrub canopy cover* (X) *on* n = *28 winter ranges (data from Cook and Irwin 1985) and 0.90, 0.75, 0.50, 0.25, and 0.10 regression quantile estimates (solid lines) and least squares regression estimate (dashed line) for the model* $y = \beta_0 + \beta_1 X + \varepsilon$. *Sample estimates* $b_0(\tau)$ *(bottom left) and* $b_1(\tau)$ *(bottom right) are shown as a red step function. Red dashed lines connect endpoints of 90% confidence intervals. Missing interval endpoints at bottom left were not estimable.*

tational effort is not usually required to obtain a useful picture of the estimates and their sampling variation. The endpoints of the confidence intervals were not connected across quantiles as a step function because they were only estimated for a subset of all possible quantiles.

Other procedures for constructing confidence intervals than the rankscore test inversion exist, including the direct order statistic approach (Zhou and Portnoy 1996, 1998), a drop in dispersion permutation test (Cade 2003), and various asymptotic methods dependent on estimating the variance/covariance matrix and the quantile density function (Koenker and Machado 1999). An advantage of the rankscore test inversion approach is that it turns the quantile regression inference problem into one solved by least squares regression, for which there already exists a wealth of related theory and methods (Cade 2003).

In the example in Figure 4, the 90% confidence intervals for both the intercept (β_0) and slope (β_1) are narrower at lower quantiles, consistent with the fact that the data were generated from a lognormal error distribution (median = 0, σ = 0.75) which has higher probability density, and thus less sampling variation at lower quantiles. Also note that the endpoints of the confidence intervals estimated by inverting the quantile rankscore test are not always symmetric about the estimate (Koenker 1994), which is consistent with the skewed sampling distribution of the estimates for smaller *n* and more extreme quantiles. Sampling variation for the quantiles can change rapidly over a short interval of quantiles, especially near the extremes. For example, in Figure 6, bottom right, the 90% confidence interval for the slope of the 0.90 quantile, $b_1(0.90)$, is narrow enough to exclude zero, whereas for the 0.95 quantile, $b_1(0.95)$, the confidence interval is wide and includes zero. For this reason it is always worthwhile to estimate a range of quantiles rather than basing an analysis on a single selected quantile.

Second order properties of the estimates are useful

Rates of change across quantiles in the slope parameter estimates (for example in Figure 6, bottom right) can be used to provide additional information that can be incorporated into the model to provide estimates with less sampling variation. The sampling variation of a selected τ regression quantile estimate is affected by changes in the parameters in some local interval surrounding the selected quantile, say $\tau \pm h$, where h is some bandwidth (Koenker and Machado 1999). Weighted regression quantile estimates can be based on weights that are inversely proportional to the differences in estimates for some local interval of quantiles, for example 0.90 ± 0.06 (Koenker and Machado 1999; Cade 2003). A variety of methods have been proposed for selecting appropriate bandwidths (Koenker and Machado 1999). The difference between the local interval approach to constructing weights and estimating the variance function to construct weights as for weighted least squares regression (eg Neter *et al.* 1996) is that the former approach allows the weights to vary for different quantiles, whereas the latter approach assumes common weights for all quantiles (Cade 2003). Differential weights by quantiles are appropriate for patterns of response similar to those in Figure 6, where a second order analysis suggested that rates of change in the estimates were probably not just due to changes in means and variances, because the changes in $b_1(\tau)$ across quantiles did not mirror those of $b_0(\tau)$. Common weights for all quantiles are appropriate for patterns of responses similar to those in Figures 1 and 5, where only location and scale changes occurred as indicated by changes in $b_1(\tau)$ across quantiles that mirrored those of $b_0(\tau)$.

Discussion

Estimating quantiles of the response distribution in regression models is not new. This has always been required for constructing prediction and tolerance intervals for future observations, but has usually been done only in a fully parametric model where the error distribution takes some specified form. In the full parametric model, the various quantiles of the response distribution are estimated by a specified multiple of the estimated standard deviation of the parametric error distribution, which is then added to the estimated mean function. Vardeman (1992) stressed the importance of prediction (for some specified number of future observations y) and tolerance intervals (for a proportion of the population and thus any number of future observations y) in statistical applications. The difference between prediction/tolerance intervals and confidence intervals is that the former deal with the sampling variation of individual observations y and the latter with the sampling variation of parameter estimates (which are a function of the n observations). Prediction and tolerance intervals for y are far more sensitive to deviations from an assumed parametric error distribution than are confidence intervals for parameters. Regression quantile estimates can be used to construct prediction and tolerance intervals without assuming any parametric error distribution and without specifying how variance heterogeneity is linked to changes in means.

The additional advantage provided by regression quantiles is that one can directly estimate rate parameters for changes in the quantiles of the distribution of responses conditional on the p predictor variables; ie $\beta_1(\tau)$, $\beta_2(\tau)$, ..., $\beta_p(\tau)$. These cannot be equal for all quantiles τ in models with heterogeneous error distributions. Differences in rates of change at different parts of the distribution are informative in a variety of ecological applications. The concept of limiting factors in ecology often focuses attention on the rates of change near the upper boundary of responses. Complicated forms of heterogeneous response distributions should be expected in observational studies where many important processes may not have been included in the candidate models. From a purely statistical standpoint, rates of change of greater magnitude for more extreme quantiles (eg $\tau > 0.90$ or $\tau < 0.10$) of the distribution may be detected as different from zero in sample estimates more often (ie greater power) than some central estimates such as the mean or median ($\tau = 0.50$). This can occur because greater differences between the parameter estimate and zero (no effect) can offset the greater sampling variation often associated with more extreme quantiles.

Use of regression quantile estimates in linear models with unequal variances will allow us to detect the effects associated with variables that might have been dismissed as statistically indistinguishable from zero based on estimates of means (Terrell *et al.* 1996). The motivation is to address the large variation often found in relationships between ecological variables and the presumed causal factors that is not attributed to random sampling variation. These models are useful when the response variable is affected by more than one factor, when factors vary in their effect on the response, when not all factors are measured, and when the multiple limiting factors interact. However, quantile regression is not a panacea for investigating relationships between variables. It is even more important for the investigator to clearly articulate what is important to the process being studied and why. A search through all possible quantiles on a large number of models with many combinations of variables for those with strong nonzero effects is no more likely to produce useful scientific generalizations than similar unfocused modeling efforts using conventional linear model procedures.

Software is currently available to provide a variety of quantile regression analyses. Scripts and Fortran programs to work with S-Plus are available from the web sites of Roger Koenker (www.econ.uiuc.edu~roger/research/home.html) and the Ecological Archives E080-001 (www.esapubs.org/archive/ecol/E080/001/default.htm). Add-on packages for R are available from the Comprehensive R Archive Network (http://lib.stat.cmu.edu/R/CRAN/). Quantile regression estimates for linear models, quantile rankscore tests, and permutation testing variants are available in the Blossom statistical package available from the US Geological Survey (www.fort.usgs.gov/products/software/blossom.asp). Stata and Shazam are two commercial econometrics programs that provide quantile regression.

Acknowledgments

We thank KD Fausch, CH Flather, JE Roelle, RL Schroeder, and JW Terrell for reviewing drafts of the manuscript.

References

Allen AW, Cade BS, and Vandever MW. 2001. Effects of emergency haying on vegetative characteristics within selected conservation reserve program fields in the northern Great Plains. *J Soil Water Conserv* **56**: 120–25.

Brown RL and Peet RK. 2003. Diversity and invasibility of southern Appalachian plant communities. *Ecology* **84**: 32–39.

Cade BS. 2003. Quantile regression models of animal habitat relationships (PhD dissertation). Fort Collins, CO: Colorado State University. 186 p.

Cade BS and Guo Q. 2000. Estimating effects of constraints on plant performance with regression quantiles. *Oikos* **91**: 245–54.

Cade BS, Terrell JW, and Schroeder RL. 1999. Estimating effects of limiting factors with regression quantiles. *Ecology* **80**: 311–23.

Cook JG and Irwin LL. 1985. Validation and modification of a habitat suitability model for pronghorns. *Wildl Soc Bull 13*: 440–48.

Cunia T. 1987. Construction of tree biomass tables by linear regression techniques. In: Estimating tree biomass regressions and their error. USDA Forest Service, General Technical Report NE-GTR-117. p 27–36.

Dunham JB, Cade BS, and Terrell JW. 2002. Influences of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. *Trans Am Fish Soc* **131**: 86–98.

Eastwood PD, Meaden GJ, and Grioche A. 2001. Modeling spatial variations in spawning habitat suitability for the sole *Solea solea* using regression quantiles and GIS procedures. *Mar Ecol–Prog Ser* **224**: 251–66.

Gerow K and Bilen C. 1999. Confidence intervals for percentiles: an application to estimation of potential maximum biomass of trout in Wyoming streams. *North Am J Fish Mana* **19**: 149–51.

Gutenbrunner C, Jurecková J, Koenker R, and Portnoy S. 1993. Tests of linear hypotheses based on regression rank scores. *J Nonparametr Stat* **2**: 307–31.

Haire SL, Bock CE, Cade BS, and Bennett BC. 2000. The role of landscape and habitat characteristics in limiting abundance of grassland nesting songbirds in an urban open space. *Landscape Urban Plan* **48**: 65–82.

Hubert WA, Marwitz TD, Gerow KG, *et al.* 1996. Estimation of potential maximum biomass of trout in Wyoming streams to assist management decisions. *North Am J Fish Mana* **16**: 821–29.

Huston MA. 2002. Introductory essay: critical issues for improving predictions. In: Scott JM, *et al.* (Eds). Predicting species occurrences: issues of accuracy and scale. Covelo, CA: Island Press. p 7–21.

Kaiser MS, Speckman PL, and Jones JR. 1994. Statistical models for limiting nutrient relations in inland waters. *J Am Stat Assoc* **89**: 410–23.

Knight CA and Ackerly DD. 2002. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecol Lett* **5**: 66–76.

Koenker R. 1994. Confidence intervals for regression quantiles. In: Mandl P and Hušková M (Eds). Asymptotic statistics: proceedings of the 5th Prague Symposium. Physica–Verlag: Heidleberg. p 349–59.

Koenker R and Bassett G. 1978. Regression quantiles. *Econometrica* **46**: 33–50.

Koenker R and d'Orey V. 1987. Computing regression quantiles. *Appl Stat* **36**: 383–93.

Koenker R and Geling O. 2001. Reappraising medfly longevity: a quantile regression survival analysis. *J Am Stat Assoc* **96**: 458–68.

Koenker R and Machado JAF. 1999. Goodness of fit and related inference processes for quantile regression. *J Am Stat Assoc* **94**: 1296–1310.

Koenker R and Park BJ. 1996. An interior point algorithm for nonlinear quantile regression. *J Econometrics* **71**: 265–83.

Koenker R and Schorfheide F. 1994. Quantile spline models for global temperature change. *Climatic Change* **28**: 395–404.

Koenker R, Ng P, and Portnoy S. 1994. Quantile smoothing splines. *Biometrika* **81**: 673–80.

McClain CR and Rex MA. 2001. The relationship between dissolved oxygen concentration and maximum size in deep-sea turrid gastropods: an application of quantile regression. *Mar Biol* **139**: 681–85.

McCullagh P and Nelder JA. 1989. Generalized linear models. New York: Chapman and Hall.

Mosteller F and Tukey JW. 1977. Data analysis and regression. New York: Addison-Wesley.

Neter JM, Kutner H, Nachtsheim CJ, and Wasserman W. 1996. Applied linear statistical models. Chicago, IL: Irwin.

Portnoy S. 1991. Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J Sci Stat Comp* **12**: 867–83.

Rosenbaum PR. 1995. Quantiles in nonrandom samples and observational studies. *J Am Stat Assoc* **90**: 1424–31.

Scharf FS, Juanes F, and Sutherland M. 1998. Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques. *Ecology* **79**: 448–60.

Terrell JW, Cade BS, Carpenter J, and Thompson JM. 1996. Modeling stream fish habitat limitations from wedged-shaped patterns of variation in standing stock. *Trans Am Fish Soc* **125**: 104–17.

Thomson JD, Weiblen G, Thomson BA, *et al.* 1996. Untangling multiple factors in spatial distributions: lilies, gophers and rocks. *Ecology* **77**: 1698–1715.

Vardeman SB 1992. What about the other intervals? *Am Stat* **46**: 193–97.

Welsh AH, Carroll RJ, and Rupert D. 1994. Fitting heteroscedastic regression models. *J Am Stat Assoc* **89**: 100–16.

Yu K and Jones MC. 1998. Local linear quantile regression. *J Am Stat Assoc* **93**: 228–37.

Zhou KQ and Portnoy SL. 1996. Direct use of regression quantiles to construct confidence sets in linear models. *Ann Stat* **24**: 287–306.

Zhou KQ and Portnoy SL. 1998. Statistical inference on heteroscedastic models based on regression quantiles. *J Nonparametr Stat* **9**: 239–60.

