2006/0002 FORT
#21071

*Journal of Statistical Computation and Simulation*
Vol. 76, No. 4, April 2006, 331–355

Taylor & Francis
Taylor & Francis Group

# Rank score and permutation testing alternatives for regression quantile estimates

BRIAN S. CADE*†, JON D. RICHARDS† and PAUL W. MIELKE JR‡

†Fort Collins Science Center, US Geological Survey, 2150 Centre Ave., Bldg. C, Fort Collins, CO 80526-8118, USA
‡Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA

*(Revised 3 March 2004; in final form 16 February 2005)*

Performance of quantile rank score tests used for hypothesis testing and constructing confidence intervals for linear quantile regression estimates ($0 \leq \tau \leq 1$) were evaluated by simulation for models with $p = 2$ and 6 predictors, moderate collinearity among predictors, homogeneous and heterogeneous errors, small to moderate samples ($n = 20$–$300$), and central to upper quantiles (0.50–0.99). Test statisties evaluated were the conventional quantile rank score $T$ statistie distributed as $\chi^2$ random variable with $q$ degrees of freedom (where $q$ parameters are constrained by $H_0$:) and an $F$ statistic with its sampling distribution approximated by permutation. The permutation $F$-test maintained better Type I errors than the $T$-test for homogeneous error models with smaller $n$ and more extreme quantiles $\tau$. An $F$ distributional approximation of the $F$ statistic provided some improvements in Type I errors over the $T$-test for models with >2 parameters, smaller $n$, and more extreme quantiles but not as much improvement as the permutation approximation. Both rank score tests required weighting to maintain correct Type I errors when heterogeneity under the alternative model increased to 5 standard deviations across the domain of $X$. A double permutation procedure was developed to provide valid Type I errors for the permutation $F$-test when null models were foreed through the origin. Power was similar for conditions where both $T$- and $F$-tests maintained correct Type I errors but the $F$-test provided some power at smaller $n$ and extreme quantiles when the $T$-test had no power because of excessively conservative Type I errors. When the double permutation scheme was required for the permutation $F$-test to maintain valid Type I errors, power was less than for the $T$-test with decreasing sample size and increasing quantiles. Confidenee intervals on parameters and tolerance intervals for future predictions were constructed based on test inversion for an example application relating trout densities to stream channel width:depth.

*Keywords*: Ecological limiting factors; Linear model; Permutation procedures; Quantile regression; Rank score statistic

## 1. Introduction

Estimating quantiles ($0 \leq \tau \leq 1$) of a response variable conditional on some set of covariates in a linear model has many applications in the biological and ecological sciences [1]

*Corresponding author. Email: brian_cade@usgs.gov

*Journal of Statistical Computation and Simulation*
ISSN 0094-9655 print/ISSN 1563-5163 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10629360500109168

and econometrics [2]. Quantile regression models allow the entire conditional distribution of a response variable $y$ to be related to some covariates $X$, providing a richer description of functional change than is possible by focusing on just the mean (or other central statistics), yet requiring minimal distributional assumptions [3–5]. Regression quantile estimates are especially enlightening for relationships involving heterogeneous responses where by definition rates of change are not the same across all parts of the response distribution. In many applications in biology and ecology, quantile regression models have been used where scientific considerations suggested that upper quantiles near the maximum better estimated effects of the biological process being measured as a limiting constraint [6–9]. Quantile regression has considerable appeal both for prediction and understanding. It can be used whether interest is in extreme quantiles (*e.g.* 95th–99th percentiles) for estimating rates of change along the boundary of a response distribution associated with some ecological limiting factor or for characterizing an interval (*e.g.* 10th–90th percentiles) of estimated effects associated with heterogeneous distributions. Interpretations and properties of the estimated effects in quantile regression are similar to those from more familiar linear modeling procedures such as least squares regression, but now are made for an interval of quantiles that is selected based on scientific considerations [1, 2, 5, 6].

There is a well-developed theory for estimating covariance matrices to provide inferences with asymptotic validity for linear regression quantile models [3–5]. These covariance methods rely on estimating the reciprocal of the error density function at the quantile of interest, $f(F^{-1}(0))$, *i.e.*, the sparsity function. Performance of asymptotic covariance methods at smaller sample sizes often is poor [10, 11] and the asymptotic theory becomes suspect at more extreme ($<0.3$ and $>0.7$) quantiles [12]. In 1994, Koenker [13] introduced the idea of constructing confidence intervals by inverting a quantile rank score test [14], which does not require estimating the sparsity function and was expected to perform well under linear heteroscedastic regression models and for smaller sample sizes.

We conducted an extensive simulation study to explore many questions related to the performance of the quantile rank score test and potential modifications. We investigated performance of quantile rank score tests across a range of quantiles, sample sizes, error distributions, and model forms likely to be encountered in scientific applications to determine where inferences become unreliable for hypotheses on single to multiple parameters. As expected, test performance was found to erode as quantiles were selected in regions of reduced density of error distributions, as sample size decreased and number of parameters in models increased. However, an $F$ ratio version of the rank score test that we evaluated by permutation arguments was found to perform better than the asymptotic $\chi^2$ form of the quantile rank score test statistic [13] for more extreme quantiles and small samples for hypotheses on parameters other than the intercept. The quantile rank score test was found to perform well for linear heteroscedastic models in the limited simulations of Koenker [13]. We made a more thorough evaluation of the effects of heterogeneity on performance of the quantile rank score test and determined that valid Type I error rates required weighted versions of the test statistics. We established that the quantile rank score tests provided valid inferences for the intercept parameter in quantile regression models, although this hypothesis was excluded under the general theory of rank score tests [14]. This provided the foundation for estimating confidence intervals at any specified value of the covariates to provide prediction and tolerance intervals. Our alternative inference procedures were applied to an analysis of Lahontan cutthroat trout (*Oncorhynchus clarki henshawi*) response to variations in their stream habitat, expanding on the previous quantile regression analyses of Dunham *et al.* [15].

## 2. Quantile regression model

The $\tau$th regression quantile ($0 \leq \tau \leq 1$) for the heteroscedastic linear location-scale model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{\varepsilon}$ is defined as $Q_y(\tau|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + F_\varepsilon^{-1}(\tau)\boldsymbol{\gamma}$; 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 for an intercept term), $\boldsymbol{\gamma}$ is a $p \times 1$ vector of unknown scale parameters, $\boldsymbol{\Gamma}$ is a diagonal $n \times n$ matrix where the $n$ diagonal elements are the $n$ corresponding ordered elements of the $n \times 1$ vector $\mathbf{X}\boldsymbol{\gamma}(\mathrm{diag}(X\boldsymbol{\gamma}))$, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors that are independent and identically distributed (i.i.d), and $F_\varepsilon^{-1}$ is the inverse of the cumulative distribution of the errors [4, 5, 11, 16]. Homoscedastic regression models are a special case of the linear location-scale model when $\boldsymbol{\gamma} = (1, 0, \ldots, 0)'$ and $Q_y(\tau|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$, $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + (F_\varepsilon^{-1}(\tau), 0, \ldots, 0)'$, where all parameters other than the intercept ($\beta_0$) in $\boldsymbol{\beta}(\tau)$ are the same for all $\tau$. More general forms of heteroscedastic errors can be accommodated with regression quantiles [5, 17] but were not considered here.

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

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

where $\rho_\tau(e) = e(\tau - I(e < 0))$ and $I(\cdot)$ is the indicator function. The estimating equation (1) yields primal solutions in a modification of the Barrodale and Roberts [18] simplex linear program for any specified value of $\tau$ [19]. With little additional computation the entire regression quantile process for all distinct values of $\tau$ can be estimated [19, 20].

Consistent estimates with reduced sampling variation for heteroscedastic linear models can be obtained by implementing weighted versions of the regression quantile estimators, where weights are based on the sparsity function at a given quantile and covariate value [5, 21]. In the linear location-scale model the weights simplify to the $n \times n$ matrix, $\boldsymbol{W} = \boldsymbol{\Gamma}^{-1}$, where the $p \times 1$ vector of scale parameters $\boldsymbol{\gamma}$ usually would have to be estimated in applications [5, 16, 22]. The weighted regression quantile estimates then are given by

$$\min\left[\sum_{i=1}^{n} \rho_\tau\left(y_i - \sum_{j=0}^{p} b_j x_{ij}\right) w_i\right] \tag{2}$$

where $\rho_\tau(e) = e(\tau - I(e < 0))$, $I(\cdot)$ is the indicator function and $w_i$ is a weight inversely proportional to the scale parameters, which is easily implemented by multiplying $\mathbf{y}$ and $\mathbf{X}$ by $\mathbf{W}$ and then using estimator (1).

## 3. Rank score test statistics

The primal linear programming solution for estimator (1) has as its corresponding dual solution

$$\max\{y'a | X'a = (1 - \tau)X'\mathbf{1}, \quad a \in [0, 1]^n\} \tag{3}$$

where $\mathbf{1}$ denotes an $n$-vector of 1's, which serves as the basis for constructing rank score tests using the regression quantile estimates [13, 14, 17, 20]. The $\tau$-quantile rank score test

uses the $\tau$-quantile score function, $\varphi_\tau(t) = \tau - I(t < \tau)$, on the $n \times 1$ vector of dual linear programming solutions, $\mathbf{a}(\tau) = [0, 1]^n$, associated with estimating the reduced parameter model corresponding to constraints imposed by the null hypothesis on the full parameter model. The reduced parameter model, $\mathbf{y} - \mathbf{X}_2\boldsymbol{\xi}(\tau) = \mathbf{X}_1\boldsymbol{\beta}_1(\tau) + \boldsymbol{\Gamma}_\varepsilon$, is constructed by partitioning $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is $n \times (p - q)$ and $\mathbf{X}_2$ is $n \times q$; and by partitioning $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1(\tau)$ is a $(p - q) \times 1$ vector of unknown nuisance parameters under the null, and $\boldsymbol{\beta}_2(\tau)$ is a $q \times 1$ vector of parameters specified by the null hypothesis $H_0$: $\boldsymbol{\beta}_2(\tau) = \boldsymbol{\xi}(\tau)$ (frequently $\boldsymbol{\xi}(\tau) = \mathbf{0}$) for the full parameter model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1(\tau) + \mathbf{X}_2\boldsymbol{\beta}_2(\tau) + \boldsymbol{\Gamma}\boldsymbol{\varepsilon}$; and $\mathbf{y}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\varepsilon}$ are as stated earlier. The $n \times 1$ vector of rank scores $\mathbf{r}(\tau) = \mathbf{a}(\tau) - (1 - \tau)\mathbf{1}$, where $\mathbf{1}$ denotes an $n \times 1$ vector of 1's, is regressed on the design matrix and the test statistic

$$T = \frac{\mathbf{S}(\tau)'\mathbf{Q}^{-1}\mathbf{S}(\tau)}{\tau(1 - \tau)}, \tag{4}$$

where $\mathbf{Q} = n^{-1}\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2$ and $\mathbf{S}(\tau) = n^{-0.5}(\mathbf{X}_2 - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2)'\mathbf{r}(\tau)$, is asymptotically distributed under $H_0$: as $\chi^2$ with $q$ degrees of freedom. Rank scores $\mathbf{r}(\tau)$ are $\tau$ for positive residuals, $\tau - 1$ for negative residuals, and in the interval $(\tau - 1, \tau)$ when residuals are 0 (observations fit exactly by the estimate). The rank scores, $\mathbf{r}(\tau)$, correspond to the quantile weights used in estimating the reduced parameter null model in estimator (1). Validity of the rank score test assumes a positive density for $y$ at the point $F^{-1}(\tau)$.

If $\mathbf{X}_2 = \mathbf{x}_2$ and $\boldsymbol{\beta}_2(\tau)$ is a scalar, *i.e.*, a single predictor is being tested, then the quantile rank score statistic has 1 degree of freedom and can be referenced to a standard normal distribution [13, 17]. This construction allows confidence intervals to be easily estimated by inversion with a modification of the linear program used to estimate regression quantiles [13]. Because the sampling distribution of the rank score test statistic is discrete, Koenker [13] recommended interpolating between adjacent hypothesized values of $\boldsymbol{\beta}_2(\tau) = \boldsymbol{\xi}(\tau)$ for constructing confidence intervals when inverting quantile rank score tests. Confidence intervals estimated by inverting the quantile rank score test may be asymmetric.

The $\tau$-quantile rank score test is based on a nondecreasing, square integrable scoring function with mean $\mu(\varphi) = 0$ and variance $\sigma_2(\varphi) = \tau(1 - \tau)$ and, thus is similar in form to the aligned rank transform statistic considered by Mansouri [23]. Note that $\mathbf{S}(\tau)'Q^{-1}\mathbf{S}(\tau)$ in test statistic (4) is the sum of squares of regression for $\mathbf{X}_2$, $\mathrm{SSReg}(\tau) = \mathrm{SSE}(\tau)_{\mathrm{red}} - \mathrm{SSE}(\tau)_{\mathrm{full}}$, where $\mathrm{SSE}(\tau)_{\mathrm{red}} = \mathbf{r}(\tau)'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{r}(\tau)$ and $\mathrm{SSE}(\tau)_{\mathrm{full}} = \mathbf{r}(\tau)'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{r}(\tau)$. Mansouri [23] proved that a test statistic like (4) was just the limiting ($n \to \infty$) form of an $F$ statistic,

$$F_{q,n-p} = \frac{\mathrm{SSE}(\tau)_{\mathrm{red}} - \mathrm{SSE}(\tau)_{\mathrm{full}}}{q\mathrm{MSE}(\tau)}, \tag{5}$$

where $\mathrm{MSE}(\tau) = \mathrm{SSE}(\tau)_{\mathrm{full}}/(n - p) \to \sigma^2(\varphi)$, and established via simulation that (5) had better small sample Type I error rates than (4). Because the sampling distribution of the $\tau$-quantile rank score test increases in discreteness and discontinuity as $\tau$ approaches 0 or 1, we expected that there might be some small sample performance advantages to using (5) over (4) for hypothesis tests or constructing confidence intervals by inverting the quantile rank score test.

The $F$ statistic for the quantile rank score test (5) is based on a regression with a dependent variable, $\mathbf{r}(\tau)$, that is a function of residuals under the reduced parameter null model. This test statistic is amenable to evaluation by permutation arguments that have been developed for testing subhypotheses in least squares regression [24–26]. The permutation distribution computed for (5) might yield more reliable Type I error rates at smaller sample sizes and more extreme quantiles than the $F$-distribution approximation with $q$ and $n - p$ degrees of

freedom. The quantile rank score $F$-test evaluated by permutation arguments is defined by the slightly simpler form of the observed value of the statistic

$$F_o = \frac{\mathrm{SSE}(\tau)_{\mathrm{red}} - \mathrm{SSE}(\tau)_{\mathrm{full}}}{\mathrm{SSE}(\tau)_{\mathrm{full}}}, \tag{6}$$

where $\mathrm{SSE}(\tau)_{\mathrm{red}}$ and $\mathrm{SSE}(\tau)_{\mathrm{full}}$ are as stated earlier, because the degrees of freedom in (5) are unnecessary as they are invariant under permutation. Note that $F_o \times \mathrm{SSE}(\tau)_{\mathrm{full}}/\tau(1-\tau) = T$ and $F_o \times ((n-p)/q) = F_{q,n-p}$.

Following Kennedy and Cade [24], Anderson and Legendre [25], and Anderson and Robinson [26], the observed value of the rank score test statistic, $F_o$, is evaluated under the null hypothesis by permuting the $\tau$-quantile rank scores, $\mathbf{r}(\tau)$, among the rows of the design matrix ($\mathbf{X}$) with equal probability, $(n!)^{-1}$. A large random sample of size $m$ is used to approximate the $n!$ possible permutations. Probability under the null hypothesis is approximated by (the number of $F \geq F_o + 1)/(m+1)$. We used a minimum of $m + 1 = 10{,}000$ to achieve probability approximations with minimal variation due to the Monte Carlo resampling.

Although permuting residuals ($\mathbf{e} = \mathbf{y} - \mathbf{X}_1\mathbf{b}_1$) under the reduced parameter null model does not in general yield exact permutation probabilities except when the null parameter is just an intercept ($\beta_0$), this permutation approach due to Freedman and Lane [27] converges in distribution and has asymptotic correlation of 1 with the exact test where $\boldsymbol{\beta}_1$ is known [26] and has performed well in simulation studies [24, 25, 28, 29]. There is some correlation ($-(n-1)^{-1}$) among residuals and they do not have constant variance ($E[\mathbf{e}\,\mathbf{e}'] = \sigma^2(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')$), implying that they are not exactly exchangeable. Dependency among the residuals decreases with increasing sample size providing some asymptotic justification for treating them as exchangeable random variables [30]. The $\tau$-quantile rank score transformation of residuals to $[\tau - 1, \tau]$ under the null model should approach constant variance more rapidly than raw residuals. There are at most $n - p + q$ residuals with rank scores of $\tau$ or $\tau - 1$, and at least $p - q$ rank scores in the interval $(\tau - 1, \tau)$. Together these conditions should reduce dependency among the transformed residuals and improve exchangeability under the null model.

An obvious modification of the quantile rank score tests $T$ and $F$ is to incorporate a weights matrix, $\mathbf{W}$, in estimating the reduced parameter null model and in constructing the test statistics (4), (5), and (6). In the weighted version of the $T$- and $F$-rank score tests, the rank scores $\mathbf{r}(\tau)$ are estimated from the weighted null model (2) and the matrices $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}$ are replaced by their weighted counterparts $\mathbf{WX}_1$, $\mathbf{WX}_2$, and $\mathbf{WX}$ in the test statistics (4), (5), and (6).

A potential problem with the permutation scheme for linear models occurs when null models are constrained through the origin as occurs when testing the intercept or when a null weighted model does not include all the variables contributing to the weight function. There is additional sampling variation not accounted for by the usual permutation distribution of the test statistic because a null model that is constrained through the origin no longer has residuals with $\tau$-quantile $= 0$. If the number of positive, negative, and zero residuals are denoted by $N^+$, $N^-$, $N^0$, respectively, and if $N^0 = p - q$ under a null model that includes an intercept, then there are at most $n\tau$ negative residuals ($N^- \leq n\tau \leq N^- + N^0$) and at most $n(1-\tau)$ positive residuals ($N^+ \leq n[1-\tau] \leq N^+ + N^0$) [3, 21]. When the null model does not include an intercept, the limits on the number of positive (negative) residuals can exceed these values by amounts consistent with binomial random variation with probability $1 - \tau$ (or $\tau$ for negative residuals). This is similar to the property that the mean of the residuals $\neq 0$ when least squares regression is constrained through the origin. Consequently, we modified a recently proposed double permutation scheme for least squares regression through the origin [31] for the quantile rank score test. The first step randomly determines the number of the $n$ rank scores, $\mathbf{r}(\tau)$, to have value $\tau - 1$ (negative residuals) as a random number from a binomial random variable

with 'success' probability $\tau$, and conversely the number to have value $\tau$ (positive residuals) with probability $1 - \tau$. As with the observed vector of rank scores, at least $p - q$ observations corresponding to the zero residuals were constrained to have scores in the interval $(\tau - 1, \tau)$ in the random vector of rank scores, $\mathbf{r}^*(\tau)$. The second step randomly permutes the random vector of rank scores, $\mathbf{r}^*(\tau)$, with respect to $\mathbf{X}$. Both steps are repeated $m$ times. So unlike the conventional permutation approach, the vector of rank scores, $\mathbf{r}^*(\tau)$, in the double permutation procedure has a randomly varying number of scores ($\tau$ and conversely $\tau - 1$) for each of the $m$ permutations to $\mathbf{X}$.

## 4. Simulation experiment

Although our primary interest was in performance of the test statistics for quantile regression models estimated with heterogeneous responses, we first conducted a set of Monte Carlo simulations with homogeneous errors to establish performance for models with simpler error structure. Normal ($\mu = 0, \sigma = 1$), uniform (min $= -2$, max $= 2$), and lognormal (median $= 0. \sigma = 0.75$) error distributions were used to provide responses with symmetric, unimodal variation with greatest density at the center, symmetric variation with constant density, and asymmetric variation with low density in a long upper tail. Error distributions were centered on the quantiles $\tau = \{0.50, 0.75, 0.90, 0.95, \text{and } 0.99\}$ so that $F_\varepsilon^{-1}(\tau|\mathbf{X}) = 0$, providing a range of central to extreme regression quantiles.

Simple 2 parameter and 6 parameter multiple regression models were simulated for $n = 20$, 30, 60, 90, 150, and 300. Independent variables were structured to have a range of values and correlation structure similar to what might be expected in many biological and ecological investigations: $X_0$ was a column of 1's for the intercept; $X_1$ was uniformly distributed (0, 100); $X_2$ was negatively correlated ($r = -0.89$) with $X_1$ specified by the function $X_2 = 4000 - 20X_1 + N(\mu = 0, \sigma = 300)$; $X_3$ was positively correlated ($r = 0.94$) with $X_1$ specified by the function $X_3 = 10 + 0.4X_1 + N(\mu = 0, \sigma = 16)$; $X_4$ was a 0, 1 indicator variable randomly assigning half the sample to each of two groups; and $X_5$ was the multiplicative interaction of $X_3$ and $X_4$. Thus, $X_1$ ranged 0–100, $X_2$ had most values in the range 0–5000 and was inversely related to $X_1$, and $X_3$ had most values in the range 0–60 and was positively related to $X_1$. Variables $X_2$ and $X_3$ were negatively correlated ($r = -0.85$) with each other through their indirect functional relation with $X_1$. The indicator variable ($X_4$) and its interaction with $X_3$ ($X_5$) allowed us to simulate the effect of $X_3$ for the regression quantile function with differing slopes, intercepts, or both terms for the two groups.

Each combination of conditions (quantile, error distribution, sample size, and model structure) was sampled 1000 times, and the test statistics $T$ and $F_0$ were computed for each sample. Probabilities for the permutation $F$-test were evaluated with separate $m + 1 = 10{,}000$ random samples of the permutation distribution. Cumulative distribution function (CDF) plots of the Type I error probabilities under the null hypothesis were graphed and compared with the expected uniform CDF. Point estimates for $\alpha = 0.05$ and 0.10, corresponding to coverage for 95% and 90% confidence intervals, were graphed across the combination of model conditions. The 99% binomial confidence intervals for 1000 simulations are 0.076–0.124 for $\alpha = 0.10$ and 0.032–0.068 for $\alpha = 0.05$, and are provided on graphs to serve as a guide to judge when the estimated error rates exceeded variation expected from the sampling simulations. Power under the alternative hypotheses was graphed only for $\alpha = 0.05$ across all combinations of conditions, although CDF plots were initially examined.

All data for the simulation studies were generated with functions in S-Plus 2000 (Mathsoft, Inc., Seattle, WA, USA). Regression quantile estimates and test statistics

were computed by a static memory compilation of Fortran 95 routines implemented in the Blossom software available from the US Geological Survey (available from: www.fort.usgs.gov/products/software/blossom/blossom.asp). Regression quantile estimates and $T$-rank score tests from the software used in simulations were compared with estimates from the S-Plus scripts developed by Koenker (available from: www.econ.uiuc.edu~roger/research/home.html) for selected models both before and after simulations were completed and found to agree to at least seven decimal places.

### 4.1 *Homogeneous error structure – simple regression*

The simple 2 parameter regression model, $y = \beta_0 + \beta_1 X_1 + \varepsilon$ was evaluated for $H_0$: $\beta_1 = 0$ with $\beta_0 = 6.0$ and $\beta_1 = 0.0, 0.01, 0.05, 0.10$, and 0.20. Estimated Type I error rates ($\beta_1 = 0.0$) for the permutation $F$-test maintained nominal rates across all conditions whereas the $T$-test became excessively conservative for $n \leq 30$ at $\tau = 0.95$ and for $n \leq 150$ at $\tau = 0.99$ (figure 1). Type I errors for the permutation test were consistent with exact exchangeability for this hypothesis. Results for $\tau = 0.75$ were nearly identical to those for $\tau = 0.50$ and, therefore, were not graphed for this or subsequent simulations. Type I error rates were similar for all error distributions for most conditions so only results of the lognormal error distribution are given in the figures for this and subsequent simulations. Results for normal and uniform error distributions are in Cade [ref. 32, Appendix 2].

The $F$ distribution approximation of the $F$ statistic controlled Type I errors under the same conditions where the $\chi^2$ approximation of the $T$-test statistic was well behaved and provided some improvement for smaller samples and more extreme quantiles. However, the $F$ distribution approximation did not maintain Type I errors as well as the permutation approximation at smaller $n$ and more extreme quantiles. A comparison of the CDF's for $\tau = 0.99$, $n = 30$ and 90, and the lognormal error distribution are in figure 2. In the simple regression model, the permutation $F$-test is identical to a permutation version of the $T$-test for $H_0$: $\beta_1 = 0$, indicating that the improvements in Type I error rates for the permutation $F$-test is due to the permutation approximation and not the alternative form of test statistic (figure 2).

The simple 2 parameter regression model also was evaluated for $H_0$: $\beta_0 = 0$ with $\beta_1 = 0.10$ and $\beta_0 = 0.0, 0.5, 1.0, 2.0$, and 3.0. Type I error rates for the intercept under the null hypothesis ($\beta_0 = 0.0$) were slightly better maintained by the double permutation $F$-test than the $T$-test for $n \geq 150$ and $\tau = 0.99$, with both tests tending to be a little conservative (figure 1). Both tests maintained Type I error rates well for $\tau = 0.50$–$0.95$, with the double permutation $F$-test becoming more conservative than the $T$-test at smaller sample sizes ($n = 20$ and 30). Typical improvement in Type I error rates for the double permutation $F$-test for $H_0$: $\beta_0 = 0$ compared to the conventional permutation test are shown in figure 3.

Power for nonzero slopes ($\beta_1 = 0.01, 0.05, 0.10$, and 0.20) was similar for the $F$- and $T$-tests for $\tau = 0.50$–$0.90$ but the $F$-test had greater power (relative power = 0.98–1.35) at $\tau = 0.95$ and 0.99 at smaller $n$ (figure 4). The $F$-test provided effective power down to $n = 30$ for $\tau = 0.95$ and $n = 150$ for $\tau = 0.99$, whereas the $T$-test only provided effective power down to $n = 60$ and 300, respectively, because of very conservative Type I error rates at smaller sample sizes (figure 4). The drop in power with increasing quantiles was greatest for the lognormal (figure 4), less for normal, and least for the uniform error distributions [ref. 32, Appendix 2]. The uniform and normal error distributions had considerably better power for $\tau = 0.95$ and 0.99 than the lognormal distribution, consistent with the greater density at upper quantiles of these distributions. Power for detecting nonzero intercepts ($\beta_0 = 0.5, 1.0, 2.0$, and 3.0) was slightly greater for the $T$ compared with the double permutation $F$-test as quantiles increased and sample size decreased (figure 4). Normal and uniform error distributions had

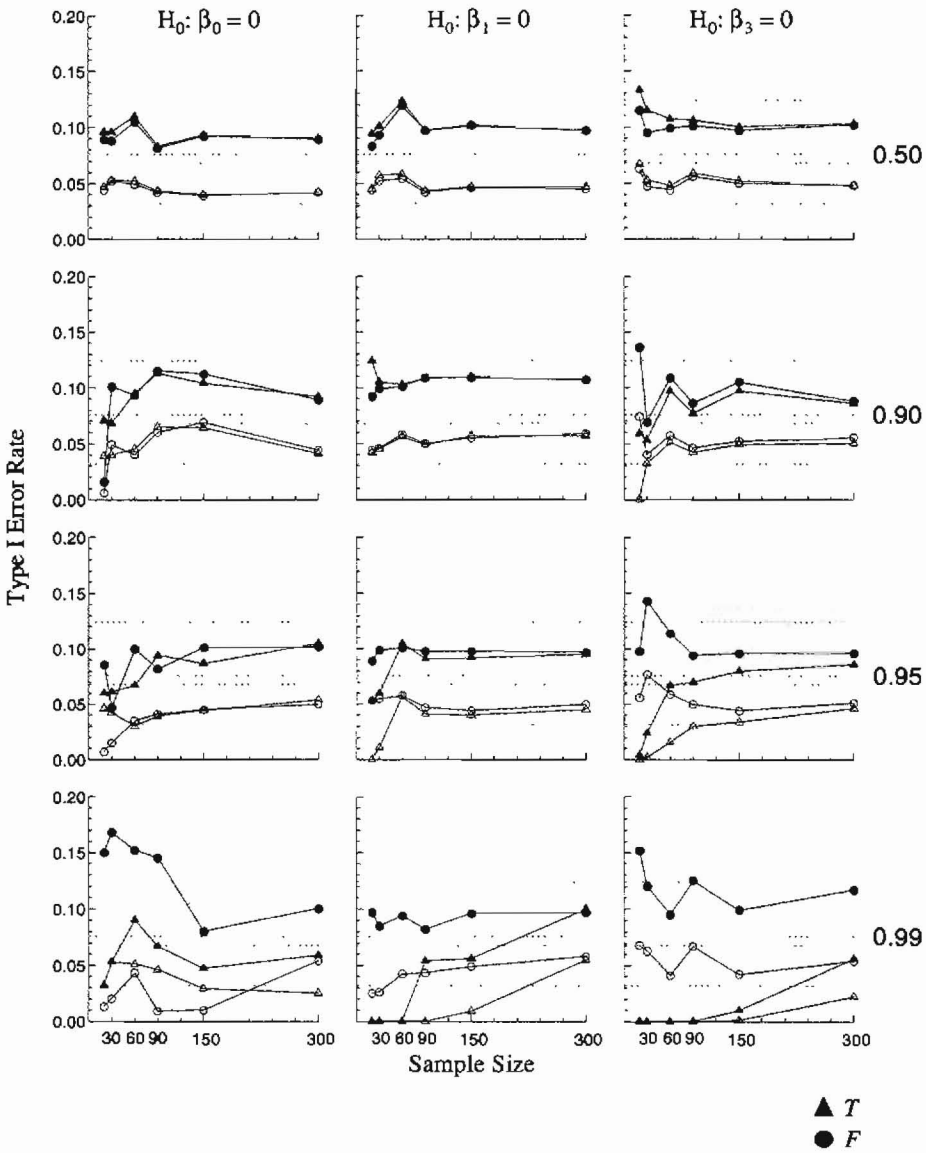


Figure 1. Estimated Type I error rates for $\alpha = 0.05$ (open) and 0.10 (solid); for the permutation $F$- (circles) and $\chi^2$-distributed $T$- (triangles) rank score tests; for homogeneous lognormal error distributions; for $H_0$: $\beta_0 = 0$ (double permutation) and $H_0$: $\beta_1 = 0$ in the model $y = \beta_0 + \beta_1 X_1 + \varepsilon$, and $H_0$: $\beta_3 = 0$ in the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$; for $\tau = [0.50, 0.90, 0.95, \text{and } 0.99]$; and for $n = 20, 30, 60, 90, 150, \text{and } 300$. Fine dotted lines are 99% binomial confidence intervals around $\alpha = 0.05$ and 0.10 for 1000 random samples used at each combination of $H_0$, $n$, and quantile.

considerably less reduction in power than the lognormal distribution for $\tau = 0.90$–$0.99$ [ref. 32, Appendix A].

### 4.2 *Homogeneous error structure – multiple regression*

Simulations for multiple regression models were limited to testing subhypotheses because these are the tests for which permutation approximations have been controversial [24, 25].

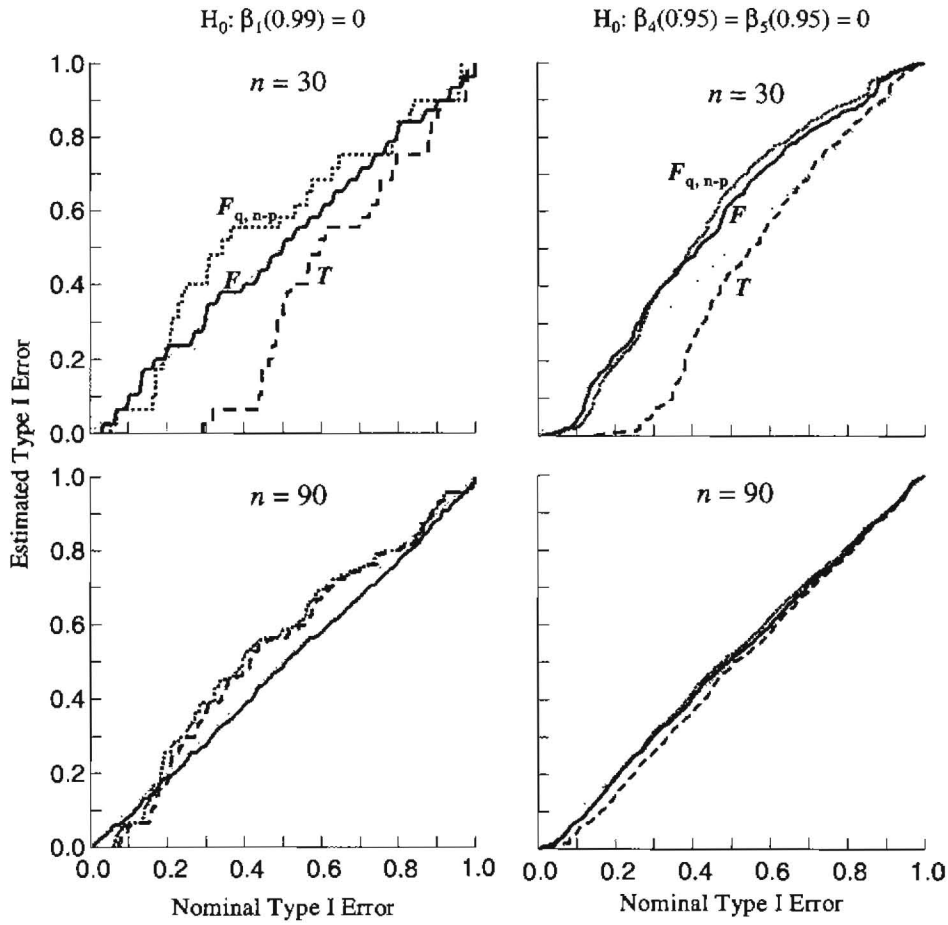


Figure 2. Cumulative distributions of 1000 estimated Type I errors for permutation approximation of $F$ (solid black line), $F_{q,n-p}$ distribution approximation of (red square dot) of $F$, and $\chi^2$-distribution approximation of $T$- (green dashes) rank score tests for $H_0$: $\beta_1 = 0$ for $\tau = 0.99$ in the model $y = \beta_0 + \beta_1 X_1 + \varepsilon$ and for $H_0$: $\beta_4 = \beta_5 = 0$ for $\tau = 0.95$ in the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + (1 + \gamma_{X_1})\varepsilon$ for $n = 30$ and 90 for the lognormal error distribution.

The 6-parameter model, $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$, was evaluated for $H_0$: $\beta_3 = 0$ with $\beta_0 = 36.0$, $\beta_1 = 0.10$, $\beta_2 = -0.005$, $\beta_4 = 2.0$, and $\beta_3 = \beta_5 = 0.0$. The permutation $F$-test maintained better Type I error rates than the $T$-test for $n < 150$ for $\tau = 0.95$ and $n < 300$ for $\tau = 0.99$ (figure 1). Otherwise, both tests similarly maintained valid Type I error rates. The 6-parameter model also was evaluated for $H_0$: $\beta_4 = 0$ with $\beta_0 = 36.0$, $\beta_1 = 0.10$, $\beta_2 = -0.005$, $\beta_3 = 0.05$, and $\beta_4 = \beta_5 = 0.0$. Type I error rates were similar to those for $H_0$: $\beta_3 = 0$. Type I error rates also were similar for normal and uniform error distributions [ref. 32, Appendix 2]. Power was not estimated for multiple regression models with homogeneous errors.

### 4.3 *Heterogeneous error structure – simple regression*

The 2 parameter regression model with heterogeneous errors, $y = \beta_0 + \beta_1 X_1 + (1 + \gamma X_1)\varepsilon$, was evaluated with $\gamma = 0.025$, 0.05, and 0.10 for $H_0$: $\beta_1 = 0$ with $\beta_0 = 6.0$ and $\beta_1 = 0.0$ to evaluate the effects of increasing heterogeneity on Type I error rates for the rank score

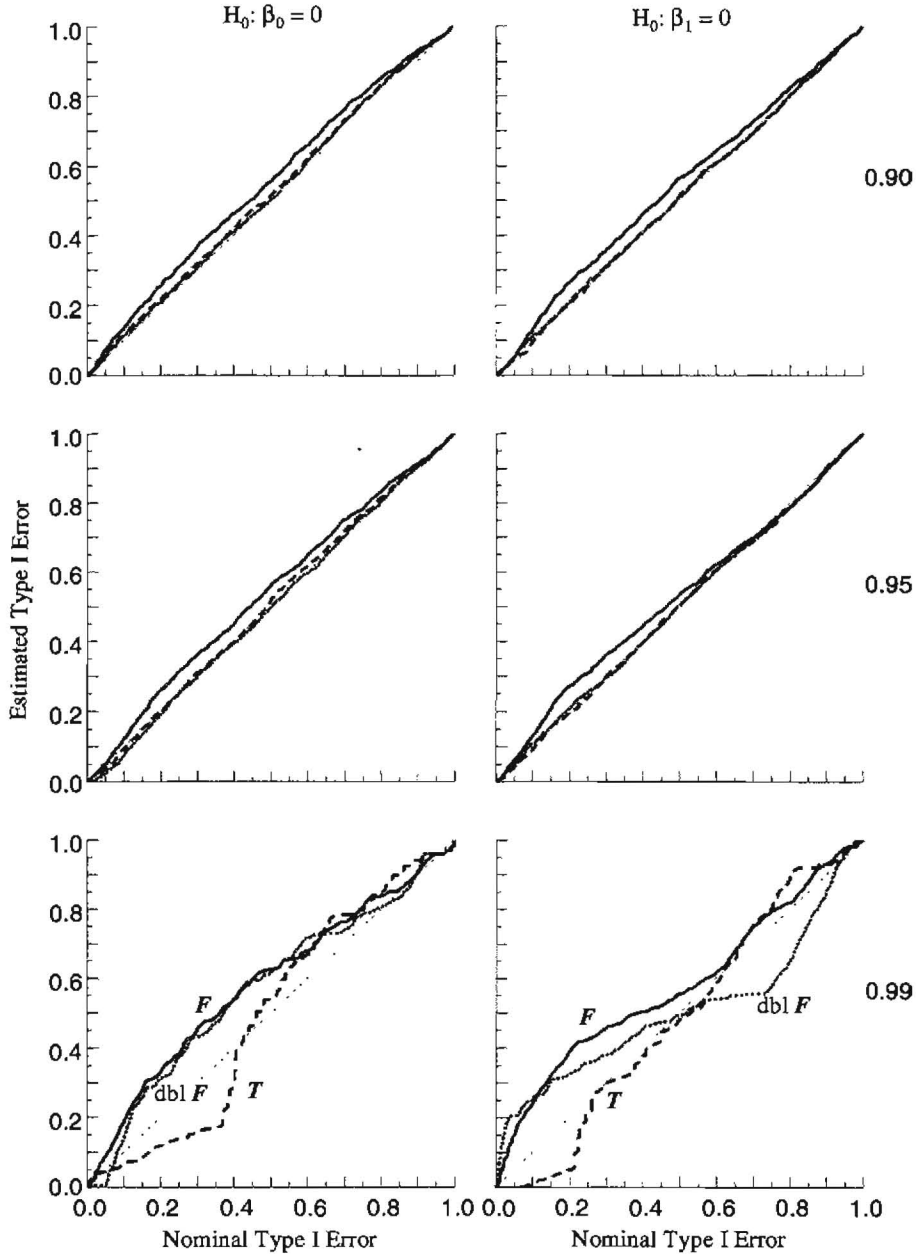


Figure 3. Cumulative distributions of 1000 estimated Type I errors for permutation $F$- (black solid), double permutation $F$- (purple square dots), and $T$- (green dashed) rank score tests of $H_0$: $\beta_0 = 0$ for the model $y = \beta_0 + \beta_1 X_1 + \varepsilon$ and of $H_0$: $\beta_1 = 0$ for the weighted model $wy = w\beta_0 + w\beta_1 X_1 + w(1 + \gamma X_1)\varepsilon$, with $\gamma = 0.05$ and $w = (1 + \gamma X_1)^{-1}$; for $\tau = \{0.90, 0.95, \text{and } 0.99\}$; for the lognormal error distribution and $n = 90$.

tests. Type I error rates became increasingly liberal for the $F$- and $T$-tests with increasing heterogeneity, except that the $T$-test became excessively conservative at $n < 60$ for $\tau = 0.95$ and at $n < 150$ for $\tau = 0.99$ (figure 5). Results were similar for normal and uniform error distributions [ref. 32, Appendix 2]. Type I error rates when $\gamma = 0.10$, corresponding to a

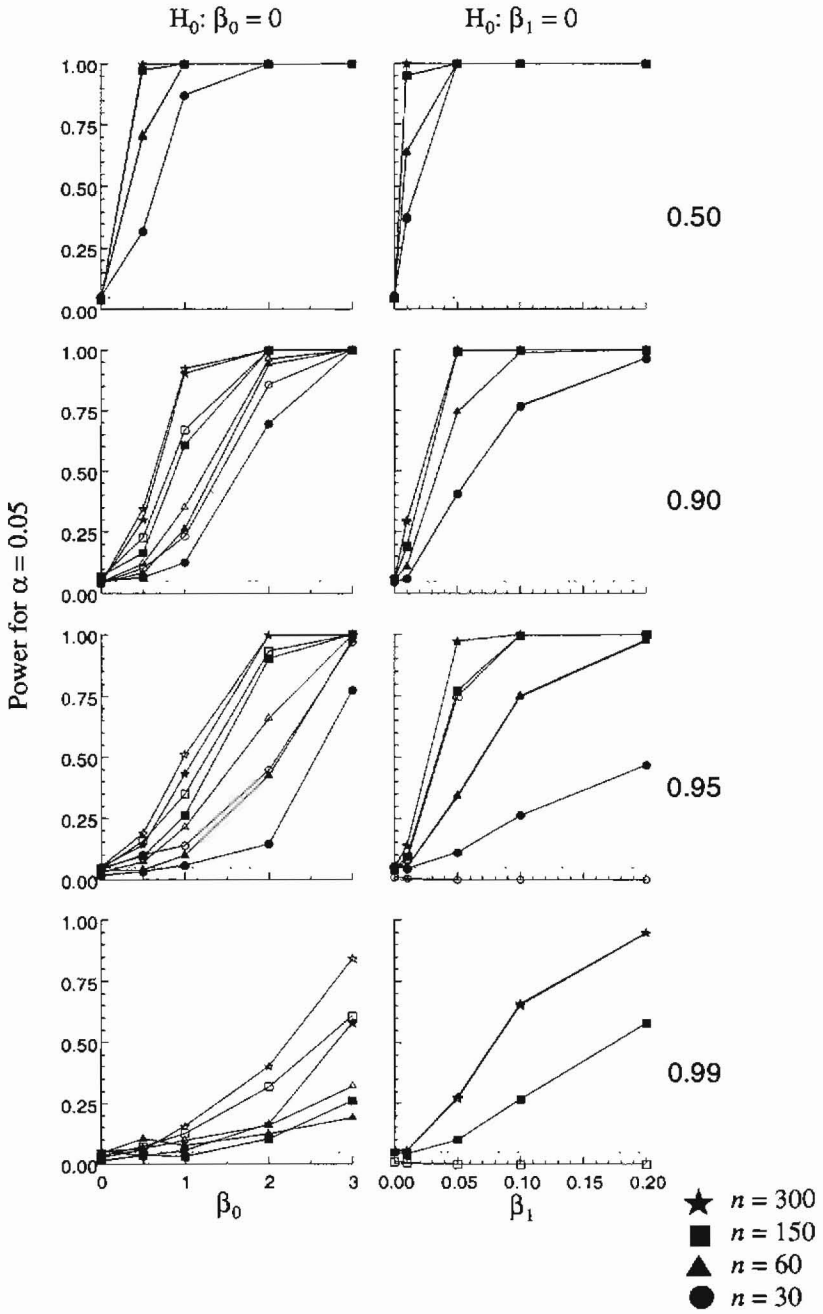


Figure 4. Estimated power for $\alpha = 0.05$ for the permutation $F$- (solid) and $\chi^2$-distributed $T$- (open) rank score tests; for homogeneous lognormal error distributions; for $H_0$: $\beta_0 = 0$ (double permutation) and $H_0$: $\beta_1 = 0$ in the model $y = \beta_0 + \beta_1 X_1 + \varepsilon$; for $\beta_0 = 0.0$, 0.5, 1.0, 2.0, and 3.0 and for $\beta_1 = 0.0$, 0.01, 0.05, 0.10, and 0.20; for $\tau = \{0.50, 0.90, 0.95, \text{and } 0.99\}$; and for $n = 30$ (circle), 60 (triangle), 150 (square), and 300 (star). Open symbols often are hidden behind solid symbols when equal. Sample sizes that had no power $>\alpha$ for either test were not graphed. 1000 random samples were used at each combination of effect size, $n$, and quantile.

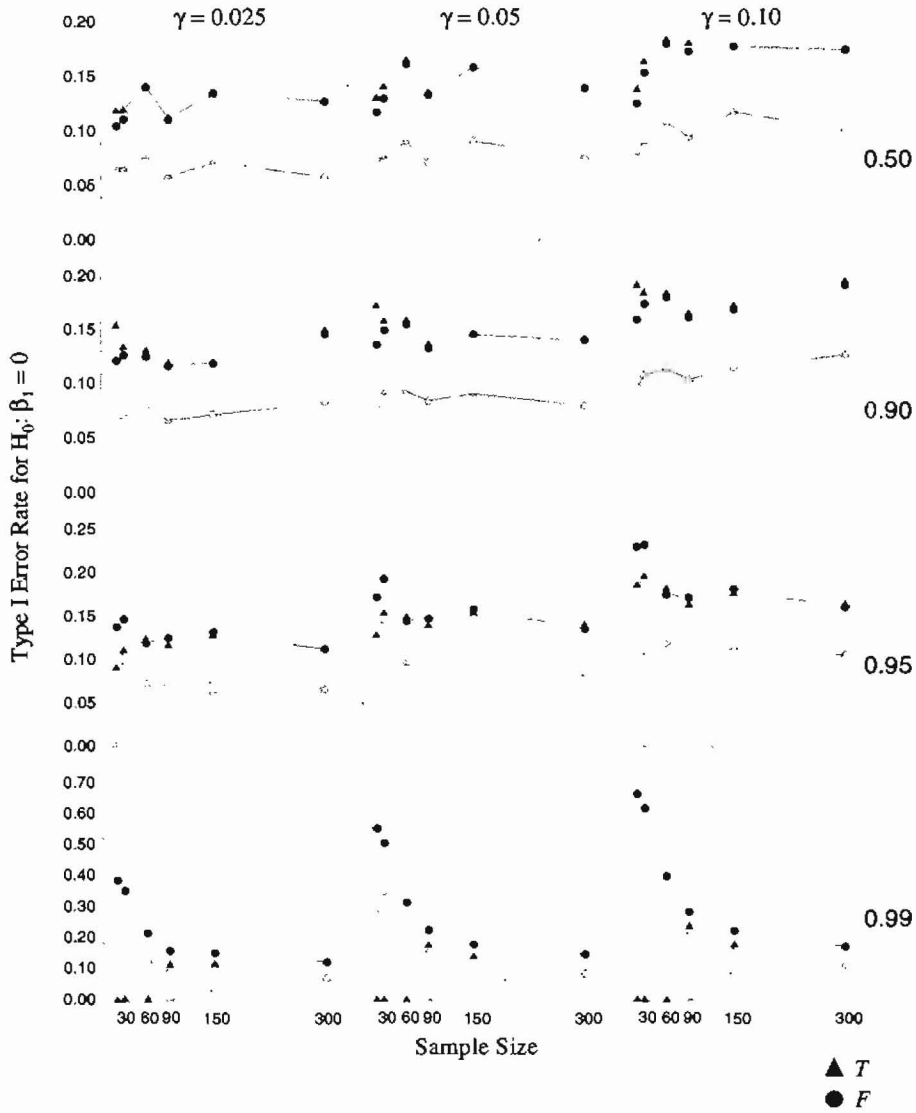


Figure 5. Inflation of Type I error rates for $\alpha = 0.05$ (open) and 0.10 (solid) with increasing heterogeneity; for the permutation $F$- (circles) and $\chi^2$-distributed $T$- (triangles) rank score tests for $H_0$: $\beta_1 = 0$; for lognormal error distributions with $\gamma = 0.025$, 0.05, and 0.10 in the model $y = \beta_0 + \beta_1 X_1 + (1 + \gamma X_1)\varepsilon$; for $\tau = \{0.50, 0.90, 0.95, \text{and } 0.99\}$; and for $n = 20, 30, 60, 90, 150$, and 300. Fine dotted lines are 99% binomial confidence intervals around $\alpha = 0.05$ and 0.10 for 1000 random samples used at each combination of $\gamma$, $n$, and quantile.

10-fold increase in $\sigma$ across the domain of $X_1$ (0–100), were inflated so that nominal 95% confidence intervals based on inverting the tests would have actual coverage of only 90%.

Weighted versions of the regression quantile estimates and the rank score tests for $\gamma = 0.05$ were simulated using known weights, $w = (1 + 0.05X_1)^{-1}$, in estimate (2). Type I error rates for $H_0$: $\beta_1 = 0$ were improved for the weighted versions of both tests (figure 6), except for $\tau = 0.95$ at $n < 90$ and $n = 0.99$ at $n < 300$. Double permutation $F$-tests were not as liberal for $\tau = 0.99$ at $n < 300$ for the normal and uniform error distributions as for the lognormal error distribution but the $T$-tests had similar conservative Type I error rates. The

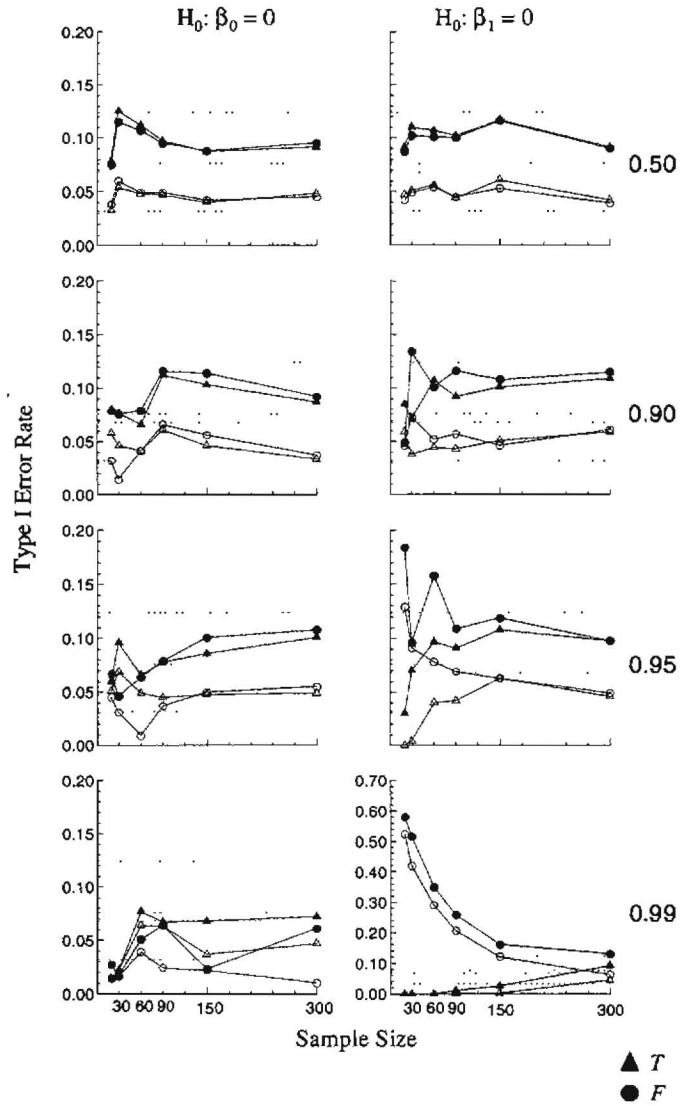


Figure 6. Estimated Type I error rates for $\alpha = 0.05$ (open) and 0.10 (solid); for the double permutation $F$- (circles) and $\chi^2$-distributed $T$- (triangles) rank score tests for $H_0$: $\beta_0 = 0$ and $H_0$: $\beta_1 = 0$; for the weighted model $wy = w(\beta_0 + \beta_1 X_1 + (1 + \gamma X_1)\varepsilon)$, $\gamma = 0.05$, and $w = (1 + \gamma X_1)^{-1}$; with lognormal error distributions; for $\tau = [0.50, 0.90, 0.95, \text{and } 0.99]$; and for $n = 20, 30, 60, 90, 150, \text{and } 300$. Fine dotted lines are 99% binomial confidence intervals around $\alpha = 0.05$ and 0.10 for 1000 random samples used at each combination of $H_0$, $n$, and quantile.

double permutation $F$-test was required to maintain correct Type I error rates because rank scores for the weighted null model $[\mathbf{r}(\tau) = \mathbf{W}\mathbf{X}_1$, where $\mathbf{X}_1$ is a column of 1's] forced the estimate through the origin. Here, again the standard permutation $F$-test had slightly more liberal Type I errors than the double permutation $F$- or the $T$-test, except at $\tau = 0.99$ and $n < 300$ where none of the weighted statistics worked well (figure 3). Power for detecting $\beta_1 = 0.01, 0.05, 0.10$, and 0.20 for $\gamma = 0.05$ for the weighted rank score tests was less for the double permutation $F$-compared with the $T$-test for $\tau = 0.90$–$0.99$, with the difference becoming greater with increasing quantiles with decreasing $n$ (figure 7). A similar pattern was

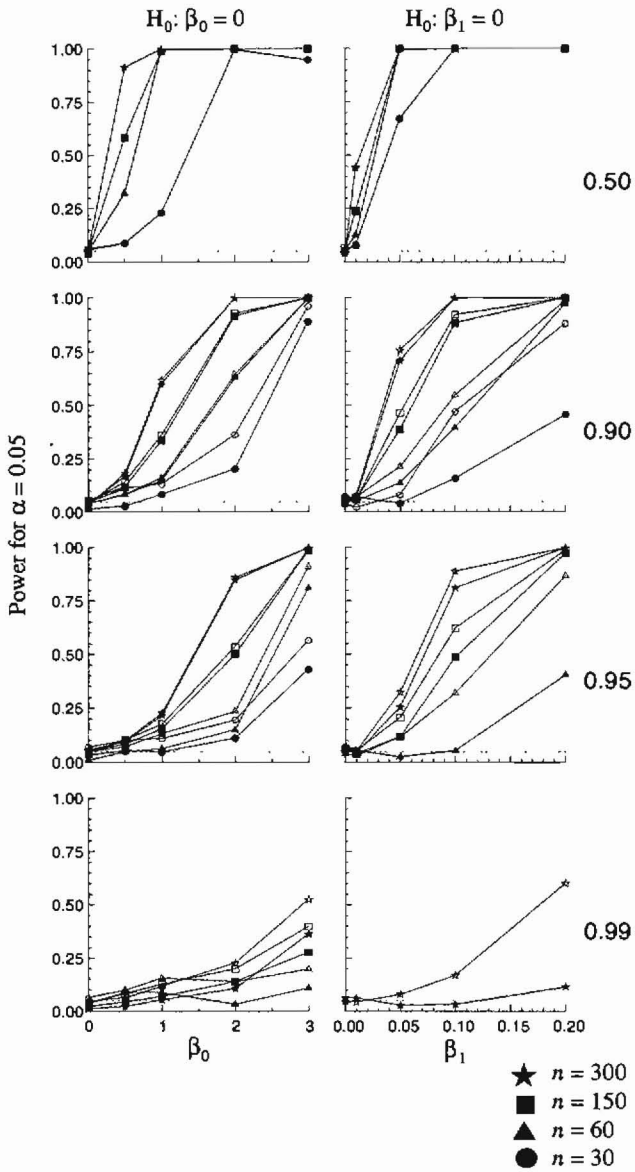


Figure 7. Estimated power for $\alpha = 0.05$ for the double permutation $F$- (solid) and $\chi^2$-distributed $T$- (open) rank score tests; for heterogeneous lognormal error distributions; for $H_0$: $\beta_0 = 0$ and $H_0$: $\beta_1 = 0$ in the weighted model $wy = w(\beta_0 + \beta_1 X_1 + (1 + \gamma X_1)\varepsilon)$, $\gamma = 0.05$, and $w = (1 + \gamma X_1)^{-1}$; for $\beta_0 = 0.0$, 0.5, 1.0, 2.0, and 3.0 and for $\beta_1 = 0.0$, 0.01, 0.05, 0.10, and 0.20; for $\tau = \{0.50, 0.90, 0.95, \text{and } 0.99\}$; and for $n = 30$ (circle), 60 (triangle), 150 (square), and 300 (star). Open symbols often are hidden behind solid symbols when equal. Sample sizes that had no power $>\alpha$ for either test were not graphed. 1000 random samples were used at each combination of effect size, $n$, and quantile.

observed for normal and uniform error distributions except that the reduction in power with increasing quantiles was not as great as for the lognormal error distribution.

The $H_0$: $\beta_0 = 0$ also was evaluated in the 2 parameter weighted regression model, $y = \beta_0 + \beta_1 X_1 + (1 + \gamma X_1)\varepsilon$, with $\gamma = 0.05$, known weights $w = (1 + 0.05X_1)^{-1}$, $\beta_1 = 0.10$, and $\beta_0 = 0.0, 0.5, 1.0, 2.0$, and 3.0. Type I error rates for both the double permutation $F$- and $T$-tests became conservative with decreasing $n$ and increasing $\tau > 0.90$ (figure 6). Power

to detect $\beta_0 = 0.5$, 1.0, 2.0, and 3.0 was slightly greater for the $T$ compared with the double permutation $F$-weighted rank score tests, again with the discrepancy increasing with increasing $\tau$ and decreasing $n$ (figure 7), similar to homogeneous error distribution models. Power decreased to very low levels for $\tau = 0.99$ for the lognormal error distribution. The decrease in power with increasing quantiles was not as great for uniform and normal error distributions.

### 4.4 *Heterogenous error structure – multiple regression*

The 6-parameter model, $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + (1 + \gamma X_1)\varepsilon$, with $\gamma = 0.05$ and known weights $w = (1 + 0.05X_1)^{-1}$ was evaluated for the full model hypothesis $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ for $\beta_0 = 36.0$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ for Type I error rates, and with $\beta_3 = 0.10$, 0.15, 0.20, and 0.25 for power. Here, again, the double permutation scheme was required because the null hypothesis with the weighted model ($wy = w\beta_0 + w\varepsilon$) forced the fit through the origin as all covariates ($X_1$) in the weights function were not in the null model. Type I error rates were well maintained by both tests until $n \leq 60$ for $\tau = 0.95$ and $n \leq 150$ for $\tau = 0.99$, where the double permutation $F$-test became liberal and the $T$-test became conservative (figure 8). Similar patterns were observed for heteroscedastic normal and uniform error distributions except that the Type I error rates for the double permutation $F$ were not as excessively liberal for $\tau = 0.99$ and small $n$. Power estimated with one of the five slope parameters ($\beta_3$) allowed to be nonzero was slightly less for the double permutation $F$ compared with the $T$-test (figure 10). Power was low for $\tau = 0.95$ to nonexistent for $\tau = 0.99$. Power for this and other hypotheses evaluated for the multiple regression models was only evaluated for the lognormal error distribution.

Type I error rates for subhypotheses involving continuous variables in the 6-parameter model with $\gamma = 0.05$ and known weights $w = (1 + 0.05X_1)^{-1}$ were evaluated for $H_0$: $\beta_3 = 0$ and $H_0$: $\beta_3 = \beta_5 = 0$ with $\beta_0 = 36.0, \beta_1 = 0.10, \beta_2 = -0.005, \beta_4 = 2.0$, and $\beta_3 = \beta_5 = 0.0$. Subhypotheses involving categorical predictors also were evaluated for weighted estimates and test statistics for $H_0$: $\beta_4 = 0$ and $H_0$: $\beta_4 = \beta_5 = 0$ with $\beta_0 = 36.0, \beta_1 = 0.10, \beta_2 = -0.005, \beta_3 = 0.05$, and $\beta_4 = \beta_5 = 0.0$. Both tests became increasingly conservative with decreasing $n$ for $\tau = 0.95$ and 0.99, with the $F$-test maintaining Type I error rates better than the $T$-test at higher quantiles and smaller $n$ (figures 8 and 9). Patterns were similar for normal and uniform error distributions. Power estimated for $H_0$: $\beta_3 = 0$ with $\beta_3 = 0.10$, 0.15, 0.20, and 0.25 for the lognormal error distribution was similar for the tests and became exceedingly low to nonexistent for $\tau = 0.90$–$0.99$ (figure 10). Power for the subhypothesis $H_0$: $\beta_4 = 0$ for $\beta_4 = 1.5$, 3.0, 6.0, and 12.0 and the lognormal error distribution were similar for the tests with a slight advantage for the permutation $F$-test for $\tau = 0.90$ and 0.95 (figure 10). Power was nonexistent for $\tau = 0.99$.

The $F$-distribution approximation of the $F_{q,n-p}$ rank score statistic maintained Type I error rates well under similar sample sizes and quantiles where the $\chi^2$-distributional approximation of the $T$-rank score statistic worked well when testing subhypotheses in multiple regression models. However, probabilities for the $F_{q,n-p}$ statistic and those provided by the permutation approximation of the $F_0$ statistic were closer to nominal error rates for smaller $n$ for more extreme quantiles than those for the $\chi^2$-distributional approximation of the $T$ statistic. An example for $H_0$: $\beta_4 = \beta_5 = 0$ for the lognormal error distribution and $\tau = 0.95$ is in figure 2.

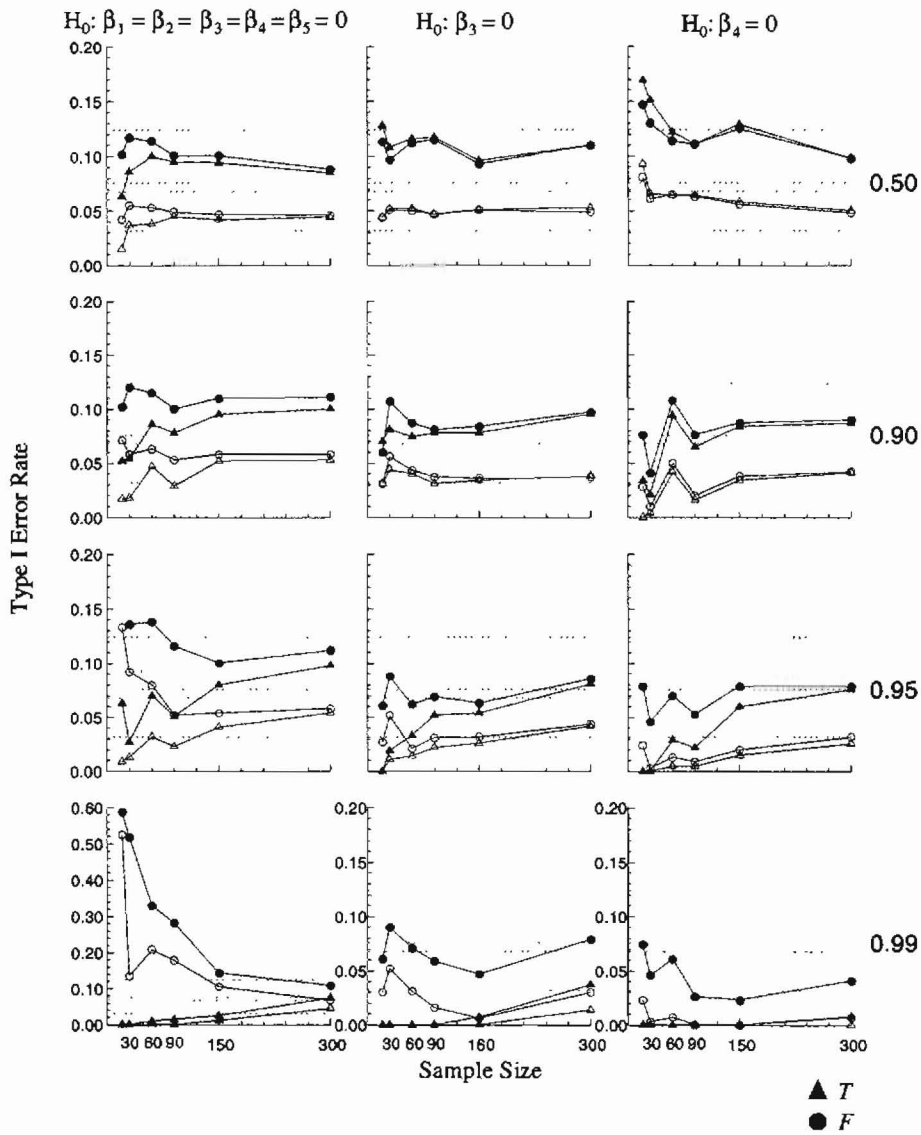


Figure 8. Estimated Type I error rates for $\alpha = 0.05$ (open) and 0.10 (solid); for the permutation $F$- (circles) and $\chi^2$-distributed $T$- (triangles) rank score tests for $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (double permutation), $H_0$: $\beta_3 = 0$, and $H_0$: $\beta_4 = 0$; for heterogeneous lognormal error distributions in the weighted model $wy = w(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + (1 + \gamma X_1)\varepsilon)$, $\gamma = 0.05$, $w = (1 + \gamma X_1)^{-1}$; for $\tau = \{0.50, 0.90, 0.95, \text{and } 0.99\}$; and for $n = 20$, 30, 60, 90, 150, and 300. Fine dotted lines are 99% binomial confidence intervals around $\alpha = 0.05$ and 0.10 for 1000 random samples used at each combination of $H_0$, $n$, and quantile.

### 4.5 *Summary of constraints on inferences*

Bounds on the sample sizes and quantiles providing both valid Type I error rates and power $>\alpha = 0.05$ were delineated for the $T$- and permutation $F$-tests based on the previous simulations for 11 models and hypothesis tests (figure 11). Contours corresponding to effective ranks also were graphed, where effective rank $= ((1 - \tau) \times n)/p$ for $\tau \geq 0.5$ following

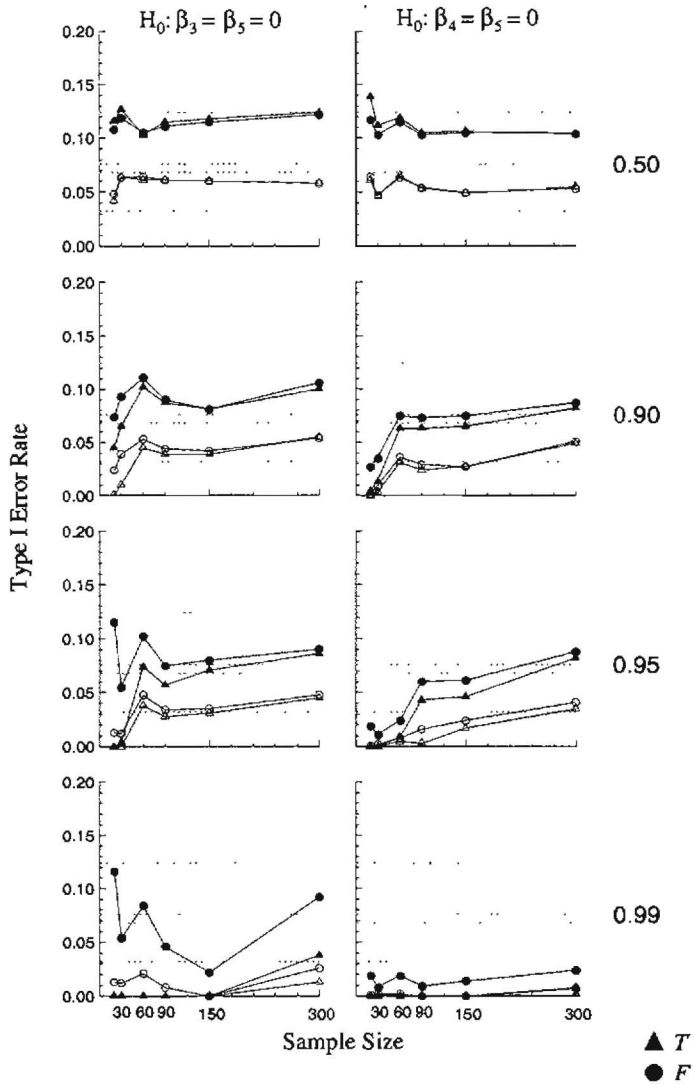


Figure 9. Estimated Type I error rates for $\alpha = 0.05$ (open) and 0.10 (solid): for the permutation $F$- (circles) and $\chi^2$-distributed $T$- (triangles) rank score tests for $H_0$: $\beta_3 = \beta_5 = 0$ and $H_0$: $\beta_4 = \beta_5 = 0$ for heterogeneous lognormal error distributions in the weighted model $wy = w(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + (1 + \gamma X_1)\varepsilon)$, $\gamma = 0.05$, $w = (1 + \gamma X_1)^{-1}$; for $\tau = \{0.50, 0.90, 0.95, \text{and } 0.99\}$; and for $n = 20, 30, 60, 90, 150$, and 300. Fine dotted lines are 99% binomial confidence intervals around $\alpha = 0.05$ and 0.10 for 1000 random samples used at each combination of $H_0$, $n$, and quantile.

Chernozhukov and Umantsev [12]. Effective rank combines the number of parameters ($p$), samples size ($n$), and quantile ($\tau$) into a ratio of order statistic to number of parameters that has been used for comparing inference procedures for quantile regression [12]. Valid inferences always were obtained for effective ranks $\geq 2$ for homogeneous error models, with the permutation $F$-test extending valid inferences to effective ranks as small as 0.05–0.75, except when testing the intercept where the $T$-test was valid for smaller effective ranks (figure 11). Valid inferences always were obtained for effective ranks $\geq 4$ for heterogeneous error models, and for effective ranks as small as 2–4 for most models. The permutation $F$-test

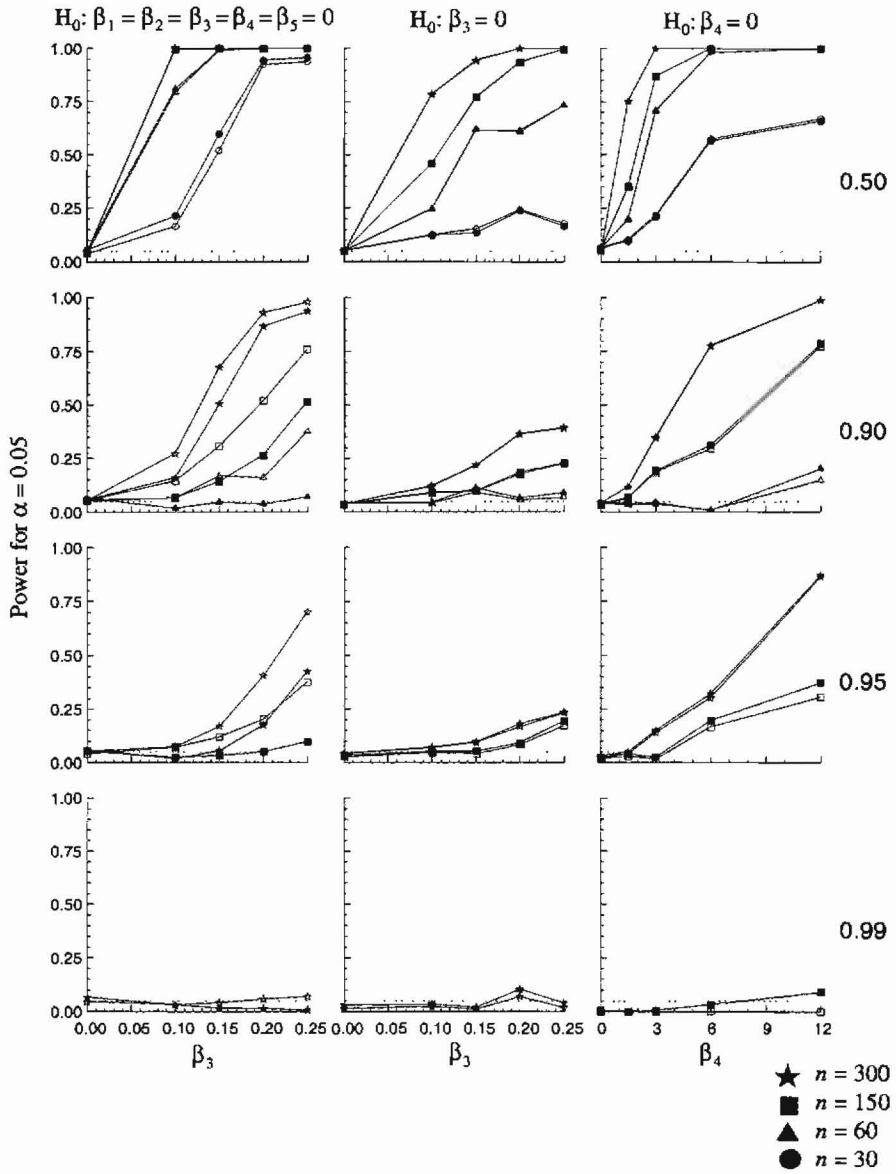


Figure 10. Estimated power for $\alpha = 0.05$ for the permutation $F$- (solid) and $\chi^2$- distributed $T$- (open) rank score tests for $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (double permutation), $H_0$: $\beta_3 = 0$, and $H_0$: $\beta_4 = 0$; for heterogeneous lognormal error distributions in the weighted model $wy = w(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + (1 + \gamma X_1)\varepsilon)$, $\gamma = 0.05$, $w = (1 + \gamma X_1)^{-1}$; for $\tau = \{0.50, 0.90, 0.95, \text{and } 0.99\}$: and for $n = 30$ (circle), 60 (triangle), 150 (square), and 300 (star). Open symbols often are hidden behind solid symbols when equal. Sample sizes with no power $>\alpha$ for either test were not graphed. 1000 random samples were used at each combination of effect size, $n$, and quantile.

extended valid inferences to effective ranks as small as 0.5–1.0, except when testing the intercept where the $T$-test was valid for smaller effective ranks. Larger effective ranks were required when testing hypotheses on categorical variables in heterogeneous multiple regression models.

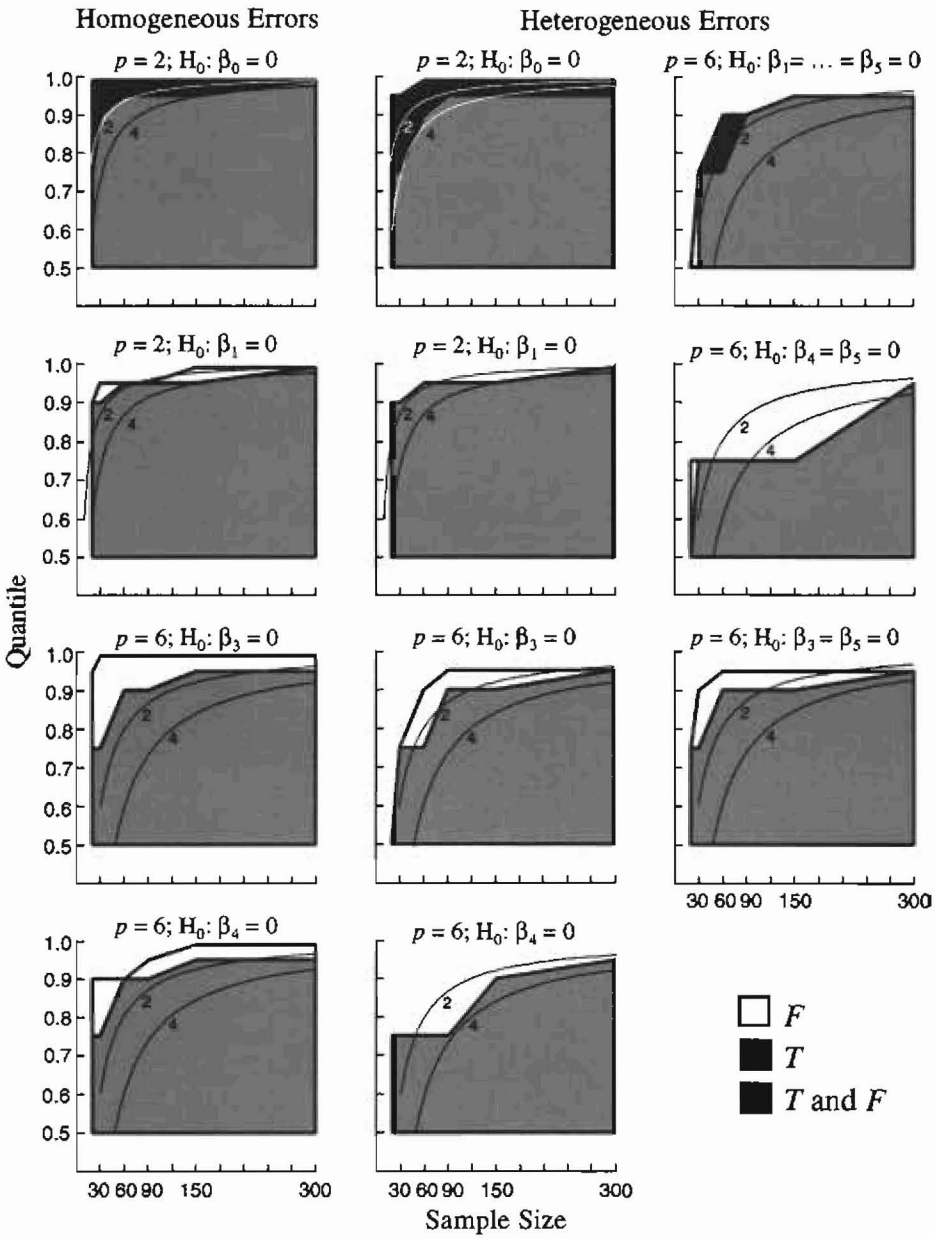


Figure 11. Samples sizes and quantiles providing valid Type I errors and power $> \alpha = 0.05$ for at least one nonzero effect for $T$ and permutation $F$ (double permutation where appropriate) $\tau$-quantile rank score tests based on simulations for $n = 20, 30, 60, 90, 150$, and 300; $\tau = \{0.50, 0.75, 0.90, 0.95,$ and $0.99\}$; for lognormal, normal, and uniform homogeneous and heterogeneous error distributions; and for $p = 2$- and 6-parameter models and hypotheses in figures 1–10. Black shaded regions are where only $T$ was valid, white regions are where only permutation $F$ was valid, and gray shaded regions are where both were valid. Thin contour lines are effective ranks of 2 and 4, where effective rank $= ((1 - \tau) \times n)/p$ for $\tau \geq 0.5$.

## 5. Example application

We constructed confidence intervals for regression quantile estimates of Lahontan cutthroat trout (*Oncorhynchus clarki henshawi*) density (trout $m^{-1}$) as a function of stream channel morphology (width:depth ratio) for $n = 71$ observations of streams sampled from 1993 to 1999 in Nevada [15]. Width:depth ratio is a measure that integrates stream channel characteristics thought to be related to small stream integrity and, thus, fish populations and is easily measured for assessing fish habitat conditions and land use impacts over large regions, such as the Lahontan basin. Lahontan cutthroat trout are a threatened species of special interest to federal land management agencies, inhabiting streams in a desert basin that are particularly sensitive to impacts of overgrazing and channel alteration due to road construction.

Here we considered the nonlinear model $y = \exp(\beta_0 X_0 + \beta_1 X_1 + \varepsilon)$, where $y$ is trout $m^{-1}$ and $X_1$ is width:depth ratio, used by Dunham *et al.* [15] but weighted to account for heteroscedasticity. We also provide confidence intervals for estimates of $\beta_0$ and prediction intervals for the lower 90% of the population that were not provided by Dunham *et al.* [15] because the validity of the quantile rank score test for the intercept had not been established. The model was estimated in the linear form $w(\ln y) = w(\beta_0 X_0 + \beta_1 X_1 + \varepsilon)$ and estimates for selected regression quantiles were plotted by exponentiating to back transform to the nonlinear form (figure 12). Estimates for all quantiles were plotted as a step function with 90% confidence intervals for 19 quantiles $\tau = \{0.05, 0.10, \ldots, 0.95\}$ (figure 12). The unweighted estimates served as a basis for developing the weighting function. The decrease in estimates of $\beta_1$ mirrored the increase in estimates of $\beta_0$ with increasing $\tau$, suggesting a linear location-scale model with error variation of the form $\gamma_0 - \gamma_1 X_1$ as a reasonable approximation. The weight function was estimated by the average pairwise difference between the 76 unweighted regression quantile estimates for $b_0(\tau)$ to estimate $\gamma_0$ and for $b_1(\tau)$ to estimate $\gamma_1$. The average pairwise difference is used as an estimate of dispersion among empirical quantiles in multiresponse permutation procedures and has been found to be relatively insensitive to effects of a few outlying values [33]. The estimated standard deviation function was $1.310 - 0.017X_1$ and its reciprocal provided weights for the weighted regression quantile estimate (2).

Confidence interval endpoints were estimated from a linear interpolation between hypothesized parameter values that had $T$-test statistics that bracketed the standard normal test statistic $= 1.645$ associated with $\alpha = 0.10$ [13]. Confidence intervals also were constructed based on inverting the double permutation $F$-test for the same quantiles (figure 12). The possible boundary values for the estimated confidence interval endpoints were obtained from the linear programming implementation used to construct intervals by inverting the $T$-test statistic [13]. These values were then used as hypothesized parameter values of $\xi(\tau)$ in the transformation $\mathbf{y} - \mathbf{X}_2\xi(\tau)$ to test the $H_0$: $\boldsymbol{\beta}_2(\tau) = \xi(\tau)$ with (6), where $\boldsymbol{\beta}_2$ was either $\beta_0$ or $\beta_1$ and $\mathbf{X}_2$ was either $X_0$ or $X_1$ depending on the parameter being tested. We used $m + 1 = 100{,}000$ permutations to compute probabilities for the $F$-tests associated with confidence interval endpoints. We used a linear interpolation based on the $P$-values to estimate the endpoints rather than the more conservative approach of using the closest estimated confidence interval endpoint with $P \leq \alpha$. This had a similar effect as the linear interpolation for the $T$-test.

There was little difference in the estimated 90% confidence intervals across $\tau = \{0.05, 0.10, \ldots, 0.95\}$ for the double permutation $F$- and $T$-test inversion approaches, except for intervals for $\beta_0(0.05)$ where the upper endpoint for the $F$-test inversion was closer to zero (figure 12). The overall pattern and width of intervals for the weighted estimates were similar to their unweighted counterparts [15, 32], which was consistent with the rather weak (<1 standard deviation change) pattern of heterogeneity across width:depth ratios. Both weighted and

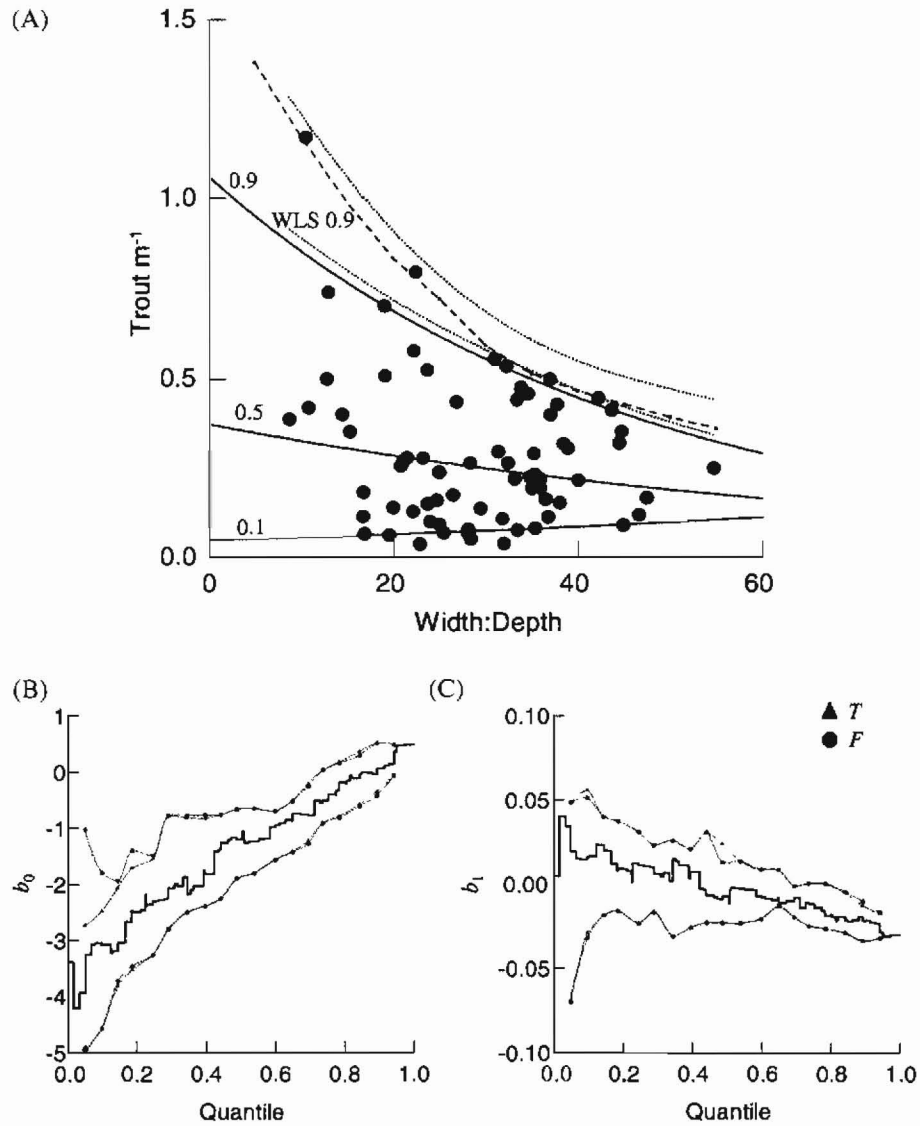


Figure 12. (A) Lahontan cutthroat trout $m^{-1}$ and width:depth ratios for 13 small streams sampled over 7 years ($n = 71$); exponentiated estimates for 0.90, 0.50, and 0.10 regression quantiles (solid lines) and weighted least squares 0.90 percentile (dotted line) estimates in the model $w(\ln y) = w(\beta_0 + \beta_1 X_1 + (\gamma_0 - \gamma_1 X_1)\varepsilon)$, $w = (1.310 - 0.017X_1)^{-1}$ (dotted lines). Dashed line is nonsimultaneous one-sided upper 90% confidence intervals for 0.90 regression quantile for selected width:depth ratios between 5 and 55. Upper dotted line is nonsimultaneous one-sided upper 90% confidence interval for 0.90 percentile estimate based on the weighted least squares model. Solid lines in (B) and (C) are step functions for estimates of $\beta_0$ and $\beta_1$ by $\tau = [0, 1]$ and dashed lines connect pointwise 90% confidence intervals for $\tau \in \{0.05, 0.10, 0.15, \ldots, 0.95\}$ based on inverting the $T$- (triangles) and double permutation $F$- (circles) rank score tests (circles) with linear interpolation between estimated endpoints.

unweighted confidence bands supported an interpretation that increasing stream width:depth ratios from 15 to 45 decreased the highest 20% of trout densities ($\tau \geq 0.80$) by 11–64%, *i.e.*, $\exp(-0.004 \times 30) = 0.887$ and $\exp(-0.034 \times 30) = 0.361$.

A one-sided upper 90% confidence band for the 0.90 quantile that was not simultaneous in $X_1$ was estimated for 11 equally spaced width:depth values between 5 and 55 corresponding to the range of ratios in the sample (figure 12). This was done by forming confidence intervals

for $\beta_0$ with a two-sided $\alpha = 0.20$ after shifting the width:depth ratios by the 11 selected values. Obviously, more values of width:depth ratio could have been used to obtain a smoother band. For comparison, a 90th percentile line based on a weighted least squares regression of the log transformed trout densities and corresponding nonsimultaneous one-sided upper 90% confidence intervals were estimated following Vardeman [34] and Gerow and Bilen [35]. Both the quantile regression and weighted least squares intervals are interpreted as upper tolerance intervals for an individual value of width:depth [34]. The latter estimates assumed a normal distribution for the log transformed data, resulting in slightly wider intervals (figure 12). A lower confidence interval (*e.g.* for 0.10 quantile) was of little interest with this data as it was effectively 0 for all width:depth ratios.

## 6. Discussion

The permutation $F$-rank score test showed some promise for providing better control of Type I error rates than the Chi-square distributional approximation of the $T$-rank score test at more extreme quantiles when sample sizes were small. However, some of this advantage was diminished by having reduced power when weighted estimates and double permutation schemes were required. The random binomial variation introduced to the rank scores by the double permutation scheme at extreme quantiles and smaller samples reduced power by creating a mass of permutations with extreme patterns of rank scores, *e.g.*, $n - p + q$ of the $\mathrm{r}(\tau) = \tau - 1$ for $\tau$th quantile. Our double permutation scheme is conceptually similar to the resampling procedure proposed for quantile regression by Parzen *et al.* [36], except that their procedure does not maintain scores corresponding to the zero residuals as does ours and requires augmenting the data with an additional observation $(n + 1)$ to solve the estimating function. Parzen *et al.* [36] found their resampling procedure maintained correct coverage probabilities for parameters but Koenker [13] found liberal coverage probabilities.

The example application with the Lahontan cutthroat trout data suggested that differences between the $T$- and $F$-rank score tests may not always be of sufficient magnitude to affect the interpretation of an analysis when quantiles used are not too extreme (*e.g.*, $0.05 \leq \tau \leq 0.95$). When estimating models for more extreme quantiles (*e.g.*, $\tau = 0.99$), fairly large samples ($n > 300$) will be required for models with more than just a few parameters to ensure reliable confidence intervals by either test, especially if weighted estimates and tests are desired. The $F$-distribution approximation of the $F_{q,n-p}$ form of the rank score statistic offered some advantages over the $\chi^2$ distribution approximation of the $T$-rank score statistic at smaller $n$ and more extreme quantiles when testing subhypotheses in multiple regression models. But there often was greater improvement by going to the permutation approximation of the $F$-rank score statistic. Lower power to detect alternative hypotheses for more extreme quantiles ($\tau = 0.95$ and 0.99) in the lower density tails of the lognormal error distribution compared to the normal and uniform error distributions was as expected.

Part of the motivation for using the quantile rank score test for inference was a belief that it was insensitive to and eliminated the need to formally model error heterogeneity [13]. However, our simulations established that the rank score tests were not immune to the effects of heterogeneity. When variation across a single independent variable was >2.5 standard deviations, tests and confidence interval estimates benefitted from using weighted estimates and weighted versions of the rank score tests. Without the use of weights, nominal 95% confidence intervals would have only 92% coverage when variation was 5 standard deviations across the independent variable. Power of unweighted rank score tests was only slightly lower [32] than the more appropriately weighted tests. Thus, when heterogeneity is not too great, inferences

made without weighted rank score tests will not depart much from inferences made with the more appropriately weighted tests. Furthermore, simulations with unweighted estimates in multiple regression models indicated that Type I error rates were often well maintained for subhypotheses on variables not contributing to heterogeneity under the alternative [32]. We used a simple pairwise difference approach based on the initial unweighted estimates to construct weights in our example application. Other approaches for estimating weights include regressing absolute values of residuals from an unweighted fit of the 0.5 quantile on the independent variables for linear location-scale models [37] and the sparsity estimation approach for more general heteroscedastic models [5].

Although the simulations presented here were limited to continuous dependent variables, other simulations with discrete count data from Poisson distributions established that the rank score tests maintained correct Type I errors even though the integer count data violated the continuity assumption [32]. As with simulations for continuous dependent variables, the permutation $F$-test maintained Type I error rates better for more extreme quantiles at smaller $n$ than the $T$-test in simulations with Poisson distributed counts.

A comparison of our simulation results with evaluations by Chernozhukov and Umantsev [12] indicated that the quantile rank score tests provided valid inferences for smaller effective ranks (more extreme quantiles and smaller sample size relative to number of parameters) than standard asymptotic methods based on the variance/covariance matrix which required effective ranks $\geq 25$. The rank score tests were valid for effective ranks $>2$ except for testing categorical variables in multiple regression models which had minimum effective ranks of 2–4. Valid inferences were possible for effective ranks $<1$ for the permutation $F$-rank score tests on parameters other than the intercept and for the $T$-rank score test on the intercept, *e.g.*, for $n \geq 30$ for $\tau = 0.95$ and for $n \geq 150$ for $\tau = 0.99$ in simple regression models with homogeneous errors. Thus, the quantile rank score tests provide valid inferences well into regions of extreme rank behavior (effective ranks $< 10$) where specialized asymptotic approaches for extreme values are required [12].

One of the potential benefits of analyzing data with regression quantiles is to focus attention on the utility of prediction and tolerance intervals in the linear model [34]. Our simulations established the validity of the quantile rank score tests for constructing confidence intervals for $\beta_0$, and, therefore, by implication for other values of $\mathbf{X} = \mathbf{x}$. Inverting tests on appropriate regression quantile estimates allows construction of prediction and tolerance intervals without assuming a specific form of the error distribution. Zhou and Portnoy [37] provided alternative order statistic based approaches to constructing such intervals with regression quantiles. Quantile regression based tolerance intervals estimated in our example application were slightly narrower than comparable intervals based on weighted least squares estimates that assumed a normal error distribution. Parametric distributional approaches for setting prediction and tolerance intervals will provide narrower intervals only when distributional assumptions are well founded. This will not be common in most ecological and biological applications. Recall that the assumed parametric error distributional form is of less consequence when estimating parameters and intervals associated with the conditional mean than it is when trying to estimate parameters associated with other parts of the probability distribution, as is required for constructing prediction and tolerance intervals.

The tolerance intervals constructed in our example application provide probability statements for an individual value of the independent variable. Simultaneous intervals for all $X$ might be estimated by emulating computations for the Working–Hotelling procedure for simultaneous confidence bands [38, pp. 156–157]. This would provide a distribution-free alternative to simultaneous tolerance bands based on weighted least squares estimates and normal error distributions [35, 39].

## Acknowledgements

M. J. Anderson, K. D. Fausch, C. H. Flather, P. Good, R. Koenker, B. R. Noon, and J. E. Roelle reviewed earlier drafts of the manuscript. We thank J. B. Dunham for providing the Lahontan cutthroat trout data.

## References

[1] Cade, B.S. and Noon, B.R., 2003, A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, **1**, 412–420.
[2] Koenker, R. and Halloek, K.F., 2001, Quantile regression: an introduction. *Journal Economic Perspectives*, **15**, 143–156.
[3] Koenker, R. and Bassett, G., 1978, Regression quantiles. *Econometrica*, **46**, 33–50.
[4] Koenker, R. and Bassett, G., 1982, Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, **50**, 43–61.
[5] Koenker, R. and Machado, J.A.F., 1999, Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**, 1296–1310.
[6] Cade, B.S., Terrell, J.W. and Sehroeder, R.L., 1999, Estimating effects of limiting factors with regression quantiles. *Ecology*, **80**, 311–323.
[7] Cade, B.S. and Guo, Q., 2000, Estimating effects of constraints on plant performance with regression quantiles. *Oikos*, **91**, 245–254.
[8] Huston, M.A., 2002, Introductory essay: critical issues for improving predictions. In: J.M. Scott, P.J. Heglund and M.L. Morrison (Eds) *Predicting Species Occurrences: Issues of Accuracy and Scale* (Covelo, CA: Island Press), pp. 7–21.
[9] Knight, C.A. and Ackerly, D.D., 2002, Variation in nuelear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters*, **5**, 66–76.
[10] Koenker, R., 1987, A comparison of asymptotic testing methods for $l_1$-regression. In: Y. Dodge (Ed.) *Statistical Data Analysis Based on the $L_1$-Norm and Related Methods* (Amsterdam: Elsevier Seience Publishers B. V. (North-Holland)), pp. 287–295.
[11] Buchinsky, M., 1991, The theory and practice of quantile regression. PhD dissertation. Harvard University, Cambridge. MA, USA, 211 pp.
[12] Chernozhukov, V. and Umantsev, L.. 2001, Conditional value-at-risk: aspects of modeling and estimation. *Empirical Economics*, **26**, 271–292.
[13] Koenker, R., 1994, Confidence intervals for regression quantiles. In: P. Mandl and M. Hušková (Eds) *Asymptotic Statistics: Proceedings of the 5th Prague Symposium* (Heildelburg: Physica-Verlag), pp. 349–359.
[14] Gutenbrunner, C., Jurečková, J., Koenker, R. and Portnoy, S., 1993, Tests of linear bypotheses based on regression rank scores. *Nonparametric Statistics*, **2**, 307–331.
[15] Dunham, J.B., Cade, B.S. and Terrell. J.W., 2002, Influences of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. *Transactions of the American Fisheries Society*, **131**, 86–98.
[16] Gutenbrunner, C. and Jureěková, J., 1992, Regression rank scores and regression quantiles. *Annals of Statisitics*, **20**, 305–330.
[17] Koenker, R., 1997. Rank tests for linear models. In: G.S. Maddala and C.R. Rao (Eds) *Handbook of Statistics*, (Amsterdam: Elsevier). Vol. 15, pp. 175–199.
[18] Barrodale, I. and Roberts, F.D.K., 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.
[19] Koenker, R. and d'Orey, V., 1987, Computing regression quantiles. *Applied Statistics*, **36**, 383–393.
[20] Koenker, R. and d'Orey, V., 1994, A remark on algorithm AS229: eomputing dual regression quantiles and regression rank scores. *Applied Statistics*, **43**, 410–414.
[21] Koenker. R. and Portnoy, S., 1996, Quantile regression. University of Illinois at Urbana-Champaign, College of Commeree and Business Administration, Office of Research Working Paper 97-0100, 77 pp.
[22] Koenker, R. and Zhao, Q., 1994, $L$-estimation for linear heteroscedastic models. *Nonparametric Statistics*, **3**, 223–235.
[23] Mansouri, H., 1999, Aligned rank transform tests in linear models. *Journal of Statistical Planning and Inference*, **79**, 141–155.
[24] Kennedy, P.E. and Cade, B.S., 1996, Randomization tests for multiple regression. *Communications in Statistics – Simulation and Computation*, **25**, 923–936.
[25] Anderson, M.J. and Legendre, P., 1999, An empirieal comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal Statistical Computation and Simulation*, **62**, 271–303.
[26] Anderson, M.J. and Robinson, J., 2001, Permutation tests for linear models. *Australian New Zealand Journal of Statistics*, **43**, 75–88.
[27] Freedman, D. and Lane, D., 1983, A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics*, **1**, 292–298.
[28] Cade, B.S. and Richards, J.D., 1996, Permutation tests for least absolute deviation regression. *Biometrics*, **52**, 886–902.

[29] Legendre, P., 2000, Comparison of permutation methods for the partial correlation and partial Mantel tests. *Journal of Statistical Computation and Simulation*, **67**, 37–73.
[30] Randles, R.H., 1984, On tests applied to residuals. *Journal of the American Statistical Association*, **79**, 349–354.
[31] Legendre, P. and Desdevises, Y., 2002, Independent contrasts and regression through the origin. Avaible at: www.fas.umontreal.ca/BIOL/legendre/reprints/index.html.
[32] Cade, B.S., 2003, Quantile regression models of animal habitat relationships. PhD dissertation, Colorado State University, Fort Collins, CO, 186 pp.
[33] Mielke, P.W., Jr. and Berry, K.J., 2001, Permutation methods: a distance function approach (New York, Inc.: Springer-Verlag), 352 pp.
[34] Vardeman, S.B., 1992, What about the other intervals? *The American Statistician*, **46**, 193–197.
[35] Gerow, K. and Bilen, C., 1999, Confidence intervals for percentiles: An application to estimation of potential maximum biomass of trout in Wyoming streams. *North American Journal of Fisheries Management*, **19**, 149–151.
[36] Parzen, M.I., Wei, L.J. and Ying, Z., 1994, A resampling method based on pivotal estimating functions. *Biometrika*, **81**, 341–350.
[37] Zhou, K.G. and Portnoy, S.L., 1998, Statistical inference on heteroscedastic models based on regression quantiles. *Nonparametric Statistics*, **9**, 239–260.
[38] Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W., 1996, *Applied Linear Statistical Models* (Chicago, IL: Irwin), 1408 pp.
[39] Turner, D.L. and Bowden, D.C., 1977, Simultaneous confidence bands for percentile lines in the general linear model. *Journal of the American Statistical Association*, **72**, 886–889.