Paper 213-30

An Introduction to Quantile Regression and the QUANTREG Procedure

Colin (Lin) Chen, SAS Institute Inc., Cary, NC

ABSTRACT

Ordinary least-squares regression models the relationship between one or more covariates X and the conditional mean of a response variable Y given $X = x$. In contrast, quantile regression models the relationship between X and the conditional quantiles of Y given $X = x$, so it is especially useful in applications where extremes are important, such as environmental studies where upper quantiles of pollution levels are critical from a public health perspective. Quantile regression also provides a more complete picture of the conditional distribution of Y given $X = x$ when both lower and upper or all quantiles are of interest, as in the analysis of body mass index where both lower (underweight) and upper (overweight) quantiles are closely watched health standards. This paper describes the new QUANTREG procedure in SAS 9.1, which computes estimates and related quantities for quantile regression by solving a modification of the least-squares criterion.

INTRODUCTION

This paper introduces the QUANTREG procedure, which computes estimates and related quantities for quantile regression. For SAS 9.1, an experimental version of the procedure can be downloaded from Software Downloads at support.sas.com.

Ordinary least-squares regression models the relationship between one or more covariates X and the *conditional mean* of the response variable Y given $X = x$. Quantile regression, which was introduced by Koenker and Bassett (1978), extends the regression model to *conditional quantiles* of the response variable, such as the 90th percentile. Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile.

As an example of data with this structure, consider the scatterplot in Figure 1 of body mass index (BMI) against age for 8,250 men from a four-year (1999–2002) survey by the National Center for Health Statistics. More details about the data can be found in Chen (2004). Body mass index, defined as the ratio of weight (kg) to squared height (m^2), is a measure of overweight or underweight. The percentiles of BMI for specified ages are of particular interest. As age increases, these percentiles provide growth patterns of BMI not only for the majority of the population, but also for underweight or overweight extremes of the population. In addition, the percentiles of BMI for a specified age provide a reference for individuals at that age with respect to the population.

The curves in Figure 1 represent fitted conditional quantiles of BMI, including the median, computed with the QUANTREG procedure for a polynomial regression model in age. During the quick growth period (ages 2 to 20), the dispersion of BMI increases dramatically; it becomes stable during middle age, and then it contracts after age 60. This pattern suggests that an effective way to control overweight in a population is to start in childhood.

Note that ordinary least-squares regression can be used to estimate conditional percentiles by making a distributional assumption such as normality for the error term in the model. However, it would not be appropriate here since the difference between each fitted percentile curve and the mean curve would be constant with age. Least-squares regression assumes that the covariates affect only the location of the conditional distribution of the response, and not its scale or any other aspect of its distributional shape.

The main advantage of quantile regression over least-squares regression is its flexibility for modeling data with heterogeneous conditional distributions. Data of this type occur in many fields, including econometrics, survival analysis, and ecology; refer to Koenker and Hallock (2001). Quantile regression provides a complete picture of the covariate effect when a set of percentiles is modeled, and it makes no distributional assumption about the error term in the model.

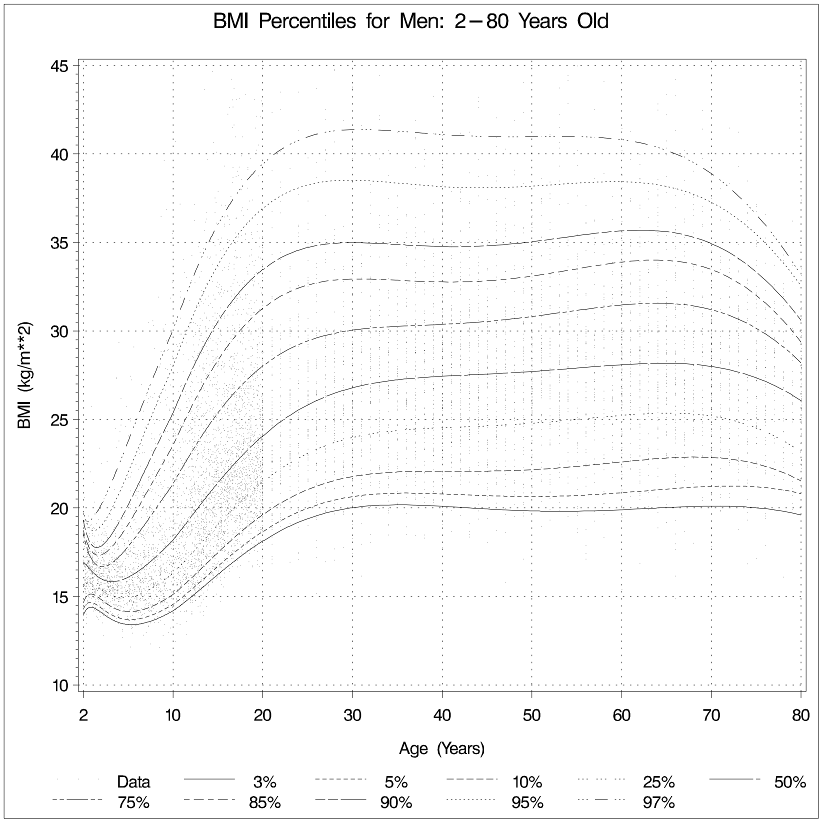


Figure 1. BMI with Growth Percentile Curves

The next section provides a more formal definition of quantile regression, followed by a closer look at the use of the QUANTREG procedure in the BMI example. A second example introduces nonparametric quantile regression. Subsequent sections discuss various aspects of quantile regression, including algorithms for estimating regression coefficients, confidence intervals, statistical tests, detection of leverage points and outliers, and quantile process plots. These aspects are illustrated with a third example using economic growth data. The last section discusses the scalability of the QUANTREG procedure.

QUANTILE REGRESSION

Quantile regression generalizes the concept of a univariate quantile to a conditional quantile given one or more covariates.

For a random variable Y with probability distribution function

$$F(y) = \text{Prob } (Y \leq y)$$

the τth quantile of Y^* is defined as the inverse function

$$Q(\tau) = \text{inf } \{y : F(y) \geq \tau\}$$

*Recall that a student's score on a test is at the τth quantile if his (or her) grade is better than $100\tau\%$ of the students who took the test. The score is also said to be at the 100τth percentile.

where $0 < \tau < 1$. In particular, the median is $Q(1/2)$.

For a random sample $\{y_1, ..., y_n\}$ of Y, it is well known that the sample median is the minimizer of the sum of absolute deviations

$$\min_{\xi \in \boldsymbol{R}} \sum_{i=1}^{n} |y_i - \xi|$$

Likewise, the general τth sample quantile $\xi(\tau)$, which is the analogue of $Q(\tau)$, may be formulated as the solution of the optimization problem

$$\min_{\xi \in \boldsymbol{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

where $\rho_\tau(z) = z(\tau - I(z < 0))$, $0 < \tau < 1$. Here $I(\cdot)$ denotes the indicator function.

Just as the sample mean, which minimizes the sum of squared residuals

$$\hat{\mu} = \text{argmin}_{\mu \in \boldsymbol{R}} \sum_{i=1}^{n} (y_i - \mu)^2$$

can be extended to the linear conditional mean function $E(Y|X = x) = x'\beta$ by solving

$$\hat{\beta} = \text{argmin}_{\beta \in \boldsymbol{R}^p} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

the linear conditional quantile function, $Q(\tau|X = x) = x'\beta(\tau)$, can be estimated by solving

$$\hat{\beta}(\tau) = \text{argmin}_{\beta \in \boldsymbol{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta)$$

for any quantile $\tau \in (0, 1)$. The quantity $\hat{\beta}(\tau)$ is called the τth **regression quantile**. The case $\tau = 1/2$, which minimizes the sum of absolute residuals, corresponds to median regression, which is also known as L_1 regression.

USING THE QUANTREG PROCEDURE

The QUANTREG procedure computes the quantile function $Q(\tau|X = x)$ and conducts statistical inferences on the estimated parameters $\hat{\beta}(\tau)$. This section introduces the QUANTREG procedure by revisiting the body mass index example and by applying nonparametric quantile regression to ozone data.

Growth Charts with Body Mass Index

Smooth quantile curves have been widely used for reference charts in medical diagnosis to identify unusual subjects, whose measurements lie in the tails of the reference distribution. This example explains how to use the QUANTREG procedure to create growth charts for BMI.

A SAS data set named *bmimen* was created by merging and cleaning the 1999–2000 and 2001–2002 survey results for men published by the National Center for Health Statistics. This data set contains the variables WEIGHT (kg), HEIGHT (m), BMI(kg/m^2), AGE (year), and SEQN (respondent sequence number) for 8,250 men.

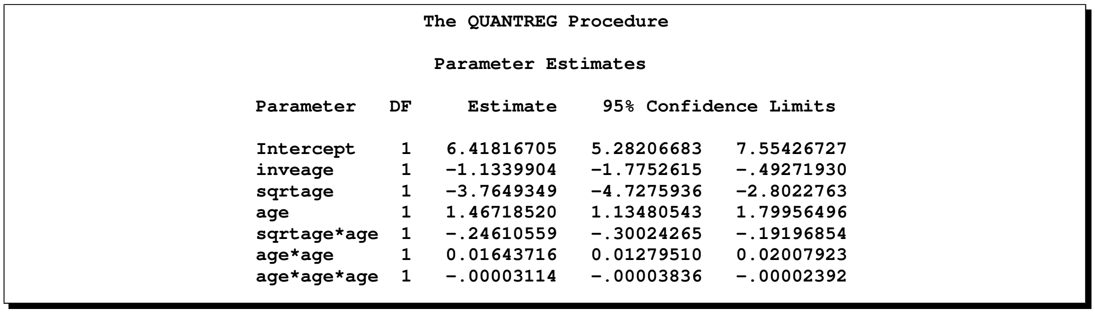
The logarithm of BMI is used as the response (although this does not help the quantile regression fit, it helps with statistical inference.) A preliminary median regression is fitted with a parametric model, which involves six powers of AGE.

The following statements invoke the QUANTREG procedure:

```
proc quantreg data=bmimen algorithm=interior ci=resampling;
    model  logbmi = inveage sqrtage age sqrtage*age age*age age*age*age
                    / diagnostics cutoff=4.5 quantile=.5;
    id seqn age weight height bmi;
    test_age_cubic: test age*age*age / wald lr;
run;
```

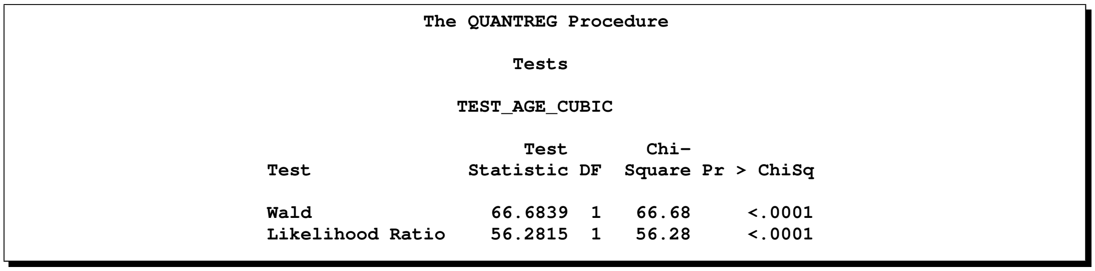
The MODEL statement provides the model, and the option QUANTILE=0.5 requests median regression, which computes $\hat{\beta}(\frac{1}{2})$ using the interior point algorithm as requested with the ALGORITHM= option (see the next section for details about this algorithm).

Figure 2 displays the estimated parameters and 95% confidence intervals, which are computed by the resampling method as requested by the CI= option. All of the parameters are considered significant since the confidence intervals do not contain zero.



```
                         The QUANTREG Procedure

                          Parameter Estimates

          Parameter    DF     Estimate    95% Confidence Limits

          Intercept     1   6.41816705   5.28206683   7.55426727
          inveage       1   -1.1339904   -1.7752615   -.49271930
          sqrtage       1   -3.7649349   -4.7275936   -2.8022763
          age           1   1.46718520   1.13480543   1.79956496
          sqrtage*age   1   -.24610559   -.30024265   -.19196854
          age*age       1   0.01643716   0.01279510   0.02007923
          age*age*age   1   -.00003114   -.00003836   -.00002392
```

Figure 2. Parameter Estimates with Median Regression: Men



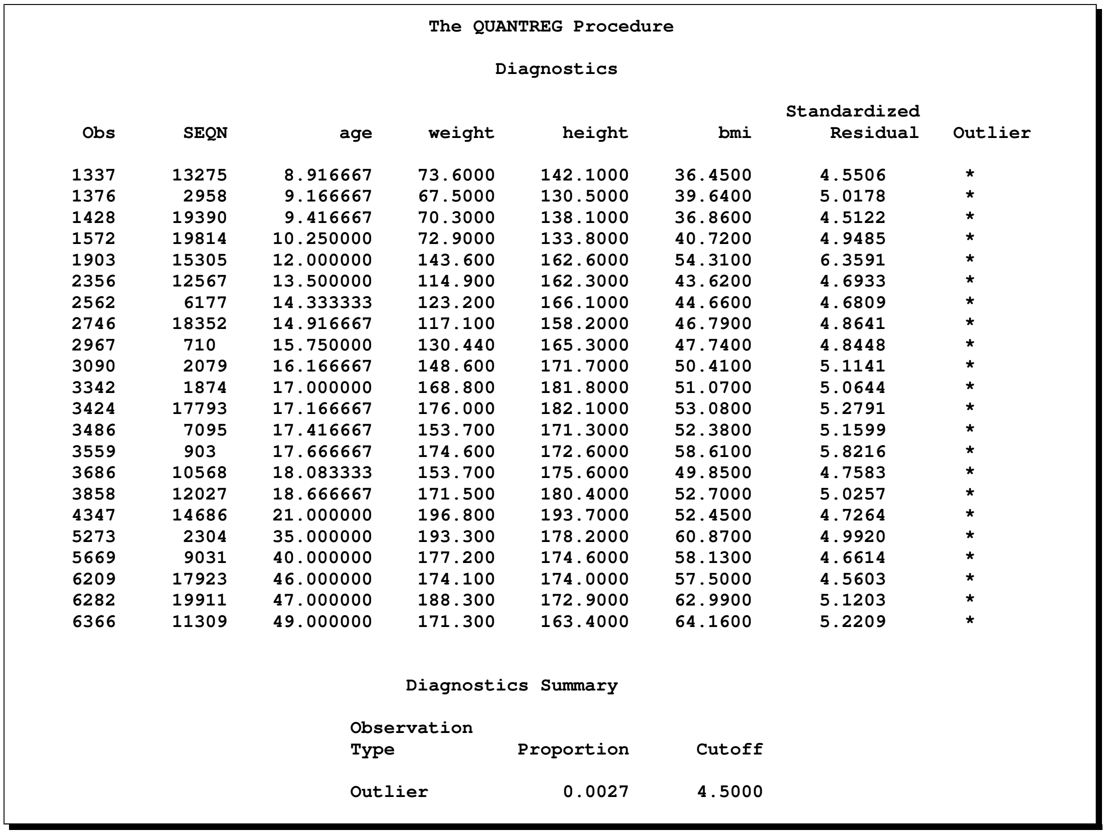
```
                         The QUANTREG Procedure

                                 Tests

                             TEST_AGE_CUBIC

                                 Test       Chi-
           Test              Statistic  DF  Square Pr > ChiSq

           Wald                66.6839   1   66.68     <.0001
           Likelihood Ratio    56.2815   1   56.28     <.0001
```

Figure 3. Test of Significance for Cubic Term

The TEST statement requests Wald and likelihood ratio tests for the significance of the cubic term in AGE. The test results, shown in Figure 3, indicate that this term is significant. Higher-order terms are not significant.

Median regression and, more generally, quantile regression are robust to extremes of the response variable. The DIAGNOSTICS option in the MODEL statement requests a diagnostic table of outliers, shown in Figure

4, which uses a cutoff value specified with the CUTOFF= option. The variables specified in the ID statement are included in the table.



```
                              The QUANTREG Procedure

                                   Diagnostics

                                                                 Standardized
   Obs       SEQN          age      weight      height         bmi      Residual   Outlier

  1337      13275     8.916667     73.6000    142.1000     36.4500      4.5506        *
  1376       2958     9.166667     67.5000    130.5000     39.6400      5.0178        *
  1428      19390     9.416667     70.3000    138.1000     36.8600      4.5122        *
  1572      19814    10.250000     72.9000    133.8000     40.7200      4.9485        *
  1903      15305    12.000000     143.600    162.6000     54.3100      6.3591        *
  2356      12567    13.500000     114.900    162.3000     43.6200      4.6933        *
  2562       6177    14.333333     123.200    166.1000     44.6600      4.6809        *
  2746      18352    14.916667     117.100    158.2000     46.7900      4.8641        *
  2967      710      15.750000     130.440    165.3000     47.7400      4.8448        *
  3090       2079    16.166667     148.600    171.7000     50.4100      5.1141        *
  3342       1874    17.000000     168.800    181.8000     51.0700      5.0644        *
  3424      17793    17.166667     176.000    182.1000     53.0800      5.2791        *
  3486       7095    17.416667     153.700    171.3000     52.3800      5.1599        *
  3559      903      17.666667     174.600    172.6000     58.6100      5.8216        *
  3686      10568    18.083333     153.700    175.6000     49.8500      4.7583        *
  3858      12027    18.666667     171.500    180.4000     52.7000      5.0257        *
  4347      14686    21.000000     196.800    193.7000     52.4500      4.7264        *
  5273       2304    35.000000     193.300    178.2000     60.8700      4.9920        *
  5669       9031    40.000000     177.200    174.6000     58.1300      4.6614        *
  6209      17923    46.000000     174.100    174.0000     57.5000      4.5603        *
  6282      19911    47.000000     188.300    172.9000     62.9900      5.1203        *
  6366      11309    49.000000     171.300    163.4000     64.1600      5.2209        *


                             Diagnostics Summary

                          Observation
                          Type            Proportion      Cutoff

                          Outlier             0.0027      4.5000
```

Figure 4. Diagnostics with Median Regression: Men

With CUTOFF=4.5, 22 men are identified as outliers. All of these men have large positive standardized residuals, which indicates that they are overweight for their age. The cutoff value 4.5 is ad hoc; it corresponds to a probability less than 0.5E−5 if normality is assumed, but the standardized residuals for median regression usually do not meet this assumption.

In order to construct the chart shown in Figure 1, the same model used for median regression is used for other quantiles. Note that the QUANTREG procedure computes fitted values only for a single quantile at a time. When fitted values are required for multiple quantiles, you can use the following macro.

```
%macro quantiles(NQuant, Quantiles);
  %do i=1 %to &NQuant;
    proc quantreg data=bmimen ci=none algorithm=interior;
        model logbmi = inveage sqrtage age sqrtage*age age*age age*age*age
                        / quantile=%scan(&Quantiles,&i,’,’);
        output out=outp&i pred=p&i;
    run;
  %end;
%mend;
```

The following statements request fitted values for 10 quantiles ranging from 0.03 to 0.97.

```
%let quantiles = %str(.03,.05,.10,.25,.5,.75,.85,.90,.95,.97);
%quantiles(10,&quantiles);
```

The 10 output data sets are merged, and the fitted BMI values together with the original BMI values are plotted against AGE to create the display shown in Figure 1.

The fitted quantile curves reveal important information. Compared to the 97th percentile in reference growth charts published by CDC in 2000, the 97th percentile for 10-year-old boys in Figure 1 is 6.4 BMI units higher (an increase of 27%). This can be interpreted as a warning of overweight or obesity. Refer to Chen (2004) for a detailed analysis.

Ozone Levels in Pittsburgh, Pennsylvania

Tracing seasonal trends in the level of tropospheric ozone is essential for predicting high-level periods, observing long-term trends, and discovering potential changes in pollution. Traditional methods for modeling seasonal effects are based on the conditional mean of ozone concentration; however, the upper conditional quantiles are more critical from a public health perspective. In this example, the QUANTREG procedure fits conditional quantile curves for seasonal effects using nonparametric quantile regression with cubic B-splines.

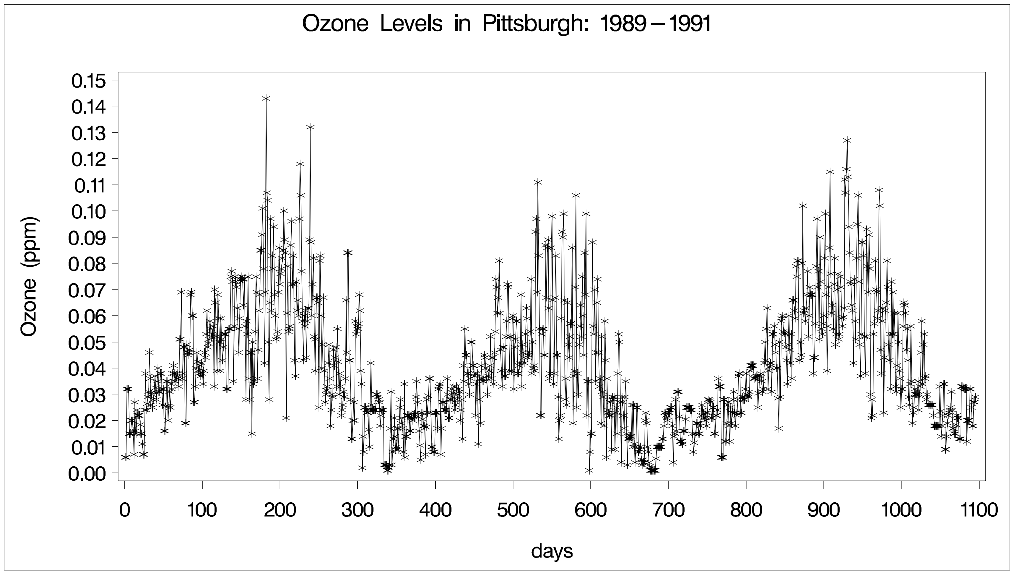


Figure 5. Time Series of Ozone Levels in Pittsburgh, Pennsylvania

The data used here are from Chock, Winkler, and Chen (2000), who studied the association between daily mortality and ambient air pollutant concentrations in Pittsburgh, Pennsylvania. The data set *ozone* contains the following variables: OZONE (daily-maximum one-hour ozone concentration (ppm)), T (daily maximum temperature (^{o}C)), and DAY (index of 1095 days (3 years)).

Figure 5, which displays the time series plot of ozone concentration for the three years, shows a clear seasonal pattern. Cubic B-splines are used to fit the seasonal effect. These splines are generated with 11 knots, which split the 3 years into 12 seasons.

The following statements construct 15 basis functions for DAY using the TRANSREG procedure.

```
proc transreg design data=ozone details;
   model bspline(day / knots=90 182 272 365 455 547 637 730 820 912 1002);
   output out=bs(drop=_: int:);
run;
```

The 15 basis functions (include the implicit intercept) are saved in the output data set *bs*, which is merged with *ozone* to create a data set named *ozbs*.

The following statements fit the conditional mean using the REG procedure by least-squares regression.

```
proc reg data=ozbs;
    model  ozone = x1-x15 / noint;
    output out=outp0 pred=p0;
run;
```

From the conditional mean, parallel conditional quantile curves can be generated based on a distributional assumption (such as normality). However, these parallel curves provide a poor fit of the heteroscedasticity in the data.

The conditional quantiles can be fitted with quantile regression by using the QUANTREG procedure. You can use the following macro to compute fitted values for multiple quantiles.

```
%macro quantiles(NQuant, Quantiles);
  %do i=1 %to &NQuant;
    proc quantreg data=ozbs algorithm=smooth;
        model  ozone = x1-x15 / noint quantile=%scan(&Quantiles,&i,'',''); 
        output out=outp&i pred=p&i;
    run;
  %end;
%mend;
```

The following statements request fitted values for the median and three upper quantiles.

```
%let quantiles = %str(.5,.75,.90,.95);
%quantiles(4,&quantiles);
```

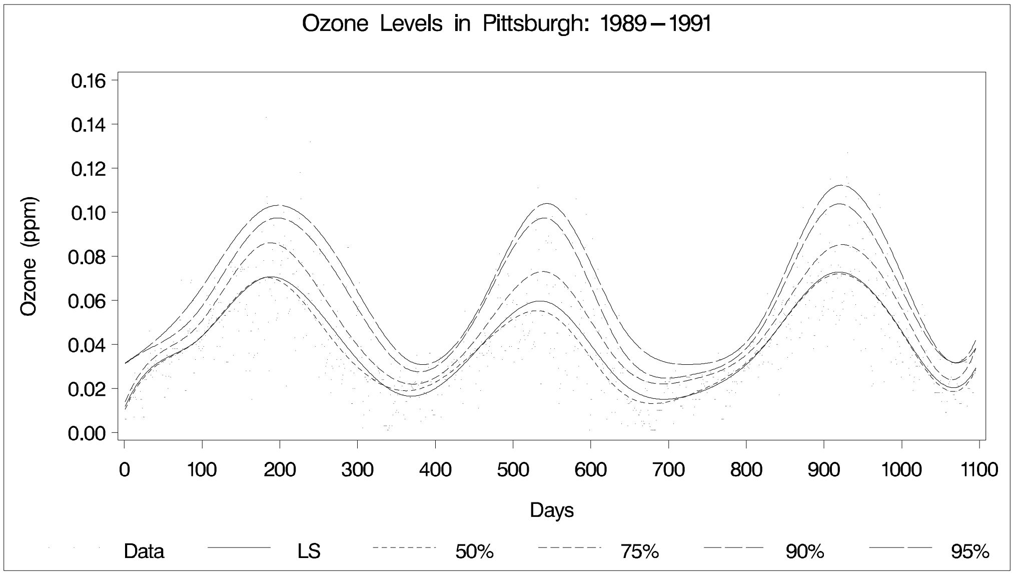


Figure 6. Quantiles and Mean Ozone Levels in Pittsburgh, Pennsylvania

Figure 6 displays the conditional mean curve obtained with the REG procedure and the quantile curves obtained with the QUANTREG procedure.

The curves show that peak ozone levels occur in the summer. The median curve (labeled 50%) and the mean curve (labeled LS) are close. This indicates that the distribution of ozone concentration is roughly symmetric. For the three years (1989–1991), these two curves do not cross the 0.08 ppm line, which is the 1997 EPA 8-hour standard. These two curves and the 75% curve show a drop for the ozone concentration levels in 1990. However, with the 90% and 95% curves, peak ozone levels tend to increase. This indicates that there might have been more low ozone concentration days in 1990, but the top 10% and 5% tend to have higher ozone concentration levels.

The quantile curves also show that high ozone concentration in 1989 had a longer duration than in 1990 and 1991. This is indicated by the wider spread of the quantile curves in 1989.

The following section provides some theoretical background for regression quantile estimates and inferences.

REGRESSION QUANTILE ESTIMATES

Let $A = (x_1, ..., x_n)$ denote the matrix consisting of n observed vectors of the random vector X, and let $y = (y_1, ..., y_n)$ denote the n observed responses. The model for linear quantile regression is

$$y = A'\beta + \epsilon$$

where $\theta = (\theta_1, ..., \theta_p)'$ is the unknown p-dimensional vector of parameters and $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ is the n-dimensional vector of unknown errors.

As discussed earlier, the τth regression quantile is a solution of

$$\min_{\beta \in \boldsymbol{R}^p} [\sum_{i \in \{i: y_i \geq x_i'\beta\}} \tau |y_i - x_i'\beta| + \sum_{i \in \{i: y_i < x_i'\beta\}} (1-\tau)|y_i - x_i'\beta|]$$

The special case $\tau = \frac{1}{2}$ is equivalent to L_1 (median) regression.

Since the early 1950s it has been recognized that median regression can be formulated as a linear programming (LP) problem and solved efficiently with some form of the simplex algorithm. In particular, the algorithm of Barrodale and Roberts (1973) has been used extensively.

The simplex algorithm is computationally demanding in large statistical applications. In theory, the number of iterations can increase exponentially with the sample size. However, this algorithm is still popularly used when the data set contains less than tens of thousands of observations.

Several alternatives have been developed to handle L_1 regression for larger data sets. The interior point approach of Karmarkar (1984) solves a sequence of quadratic problems in which the relevant interior of the constraint set is approximated by an ellipsoid. The worst-case performance of the interior point algorithm has been proved to be better than that of the simplex algorithm. More important, experience has shown that the interior point algorithm is advantageous for larger problems.

Like L_1 regression, general quantile regression fits nicely into the standard primal-dual formulations of linear programming.

Besides the interior point method, various heuristic approaches have been provided for computing L_1-type solutions. Among these, the finite smoothing algorithm of Madsen and Nielsen (1993) is the most useful. It approximates the L_1-type objective function with a smoothing function, so that the Newton-Ralphon algorithm can be used iteratively to obtain the solution after a finite number of loops. The smoothing algorithm extends naturally to general quantile regression.

The QUANTREG procedure implements the simplex, interior point, and smoothing algorithms. The details are described in the Appendix.

CONFIDENCE INTERVALS

The QUANTREG procedure provides three methods to compute confidence intervals for the regression quantile parameter $\beta(\tau)$: sparsity, rank, and resampling. The sparsity method is the most direct and the fastest, but it involves estimation of the sparsity function, which is not robust for data that are not independently and identically distributed. To deal with this problem, the QUANTREG procedure computes a Huber sandwich estimate using a local estimate of the sparsity function. The rank method, which computes confidence intervals by inverting the rank score test, does not suffer from this problem, but it uses the simplex algorithm and is computationally expensive with large data sets. The resampling method, which uses the bootstrap, can overcome all of these problems, but it is unstable for small data sets. Based on these properties, the QUANTREG uses a combination of the rank method and the resampling method as the default. The three methods are described in the Appendix.

COVARIANCE AND CORRELATION OF PARAMETER ESTIMATES

The QUANTREG procedure provides two methods to compute the covariance and correlation matrices of the estimated parameters: an asymptotic method and a bootstrap method. Bootstrap covariance and correlation matrices are computed when resampling confidence intervals are computed. Otherwise, asymptotic covariance and correlation matrices are computed.

Asymptotic Covariance and Correlation

This method corresponds to the SPARSITY method for the confidence intervals. For the sparsity function in the computation of the asymptotic covariance and correlation, both i.i.d. and non i.i.d. estimates are computed. By default, the QUANTREG procedure computes non i.i.d. estimates. Since the rank method does not provide a covariance-correlation estimate, the asymptotic covariance-correlation is computed when the confidence intervals are computed using this method.

Bootstrap Covariance-Correlation

This method corresponds to the resampling method for the confidence intervals. The Markov chain marginal bootstrap (MCMB) method is used.

LINEAR TEST

Two tests are available in the QUANTREG procedure for the linear null hypothesis $H_0 : \beta_2 = 0$. Here β_2 denotes a subset of the parameters, where the parameter vector $\beta(\tau)$ is partitioned as $\beta'(\tau) = (\beta_1'(\tau), \beta_2'(\tau))$, and the covariance matrix Ω for the parameter estimates is partitioned correspondingly as Ω_{ij} with $i = 1, 2; j = 1, 2$; and $\Omega^{22} = (\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12})^{-1}$.

The Wald test, which is based on the estimated coefficients for the unrestricted model, is given by

$$T_W(\tau) = \hat{\beta}_2'(\tau)\hat{\Sigma}(\tau)^{-1}\hat{\beta}_2(\tau)$$

where $\hat{\Sigma}(\tau)$ is an estimator of the covariance of $\hat{\beta}_2(\tau)$. The QUANTREG procedure provides two estimators for the covariance as described in the previous section. The estimator based on the asymptotic covariance is

$$\hat{\Sigma}(\tau) = \frac{1}{n}\hat{\omega}(\tau)^2\Omega^{22}$$

where $\hat{\omega}(\tau) = \sqrt{\tau(1-\tau)}\hat{s}(\tau)$ and $\hat{s}(\tau)$ is the estimated sparsity function. The estimator based on the bootstrap covariance is the empirical covariance of the MCMB samples.

The likelihood ratio test is based on the difference between the objective function values in the restricted and unrestricted models. Let $D_0(\tau) = \sum \rho_\tau(y_i - x_i\hat{\beta}(\tau))$, $D_1(\tau) = \sum \rho_\tau(y_i - x_{1i}\hat{\beta}_1(\tau))$, and set

$$T_{LR}(\tau) = 2(\tau(1-\tau)\hat{s}(\tau))^{-1}(D_1(\tau) - D_0(\tau))$$

where $\hat{s}(\tau)$ is the estimated sparsity function. Refer to Figure 3 for an example of these tests.

Koenker and Machado (1999) prove that these two tests are asymptotically equivalent and that the distributions of the test statistics converge to χ^2_q under the null hypothesis, where q is the dimension of β_2.

LEVERAGE POINT AND OUTLIER DETECTION

The QUANTREG procedure uses robust multivariate location and scale estimates for leverage point detection.

Mahalanobis distance is defined as

$$MD(x_i) = [(x_i - \bar{x})'\bar{C}(A)^{-1}(x_i - \bar{x})]^{1/2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{C} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})'(x_i - \bar{x})$. Here, $x_i = (x_{i1}, ..., x_{i(p-1)})'$ does not include the intercept variable. The relationship between the Mahalanobis distance $MD(x_i)$ and the hat matrix $H = (h_{ij}) = A'(AA')^{-1}A$ is

$$h_{ii} = \frac{1}{n-1}MD_i^2 + \frac{1}{n}$$

Robust distance is defined as

$$RD(x_i) = [(x_i - T(A))'C(A)^{-1}(x_i - T(A))]^{1/2}$$

where $T(A)$ and $C(A)$ are robust multivariate location and scale estimates computed with the minimum covariance determinant (MCD) method of Rousseeuw and Van Driessen (1999).

These distances are used to detect leverage points. You can use the DIAGNOSTICS and LEVERAGE options in the MODEL statement to request leverage point and outlier diagnostics. Two new variables, LEVERAGE and OUTLIER, are created and saved in an output data set specified in the OUTPUT statement.

Let $C(p) = \sqrt{\chi^2_{p;1-\alpha}}$ be the cutoff value. The variable LEVERAGE is defined as

$$\text{LEVERAGE } = \begin{cases} 0 & \text{if } RD(x_i) \leq C(p) \\ 1 & \text{otherwise} \end{cases}$$

You can specify a cutoff value with the LEVERAGE option in the MODEL statement.

Residuals $r_i, i = 1, ..., n$ based on quantile regression estimates are used to detect vertical outliers. The variable OUTLIER is defined as

$$\text{OUTLIER } = \begin{cases} 0 & \text{if } |r_i| \leq k\sigma \\ 1 & \text{otherwise} \end{cases}$$

You can specify the multiplier k of the cutoff value with the CUTOFF= option in the MODEL statement. You can specify the scale σ with the SCALE= option in the MODEL statement. By default, $k = 3$ and the scale σ is computed as the corrected median of the absolute residuals $\sigma = \text{median}\{|r_i|/\beta_0, i = 1, ..., n\}$, where $\beta_0 = \Phi^{-1}(.75)$ is an adjustment constant for consistency with the normal distribution.

Refer to Figure 4 for an example of outlier detection.

The following example illustrates how the QUANTREG procedure provides statistical inference. It also illustrates how the procedure computes quantile processes and creates graphical displays.

ECONOMETRIC GROWTH STUDY

This example uses a SAS data set named *growth*, which contains economic growth rates for countries during two time periods, 1965–1975 and 1975–1985. The data come from a study by Barro and Lee (1994) and have also been analyzed by Koenker and Machado (1999).

There are 161 observations and 15 variables in the data set. The variables, which are listed in the following table, include the national growth rates (GDP) for the two periods, 13 covariates, and a name variable (Country) for identifying the countries in one of the two periods.

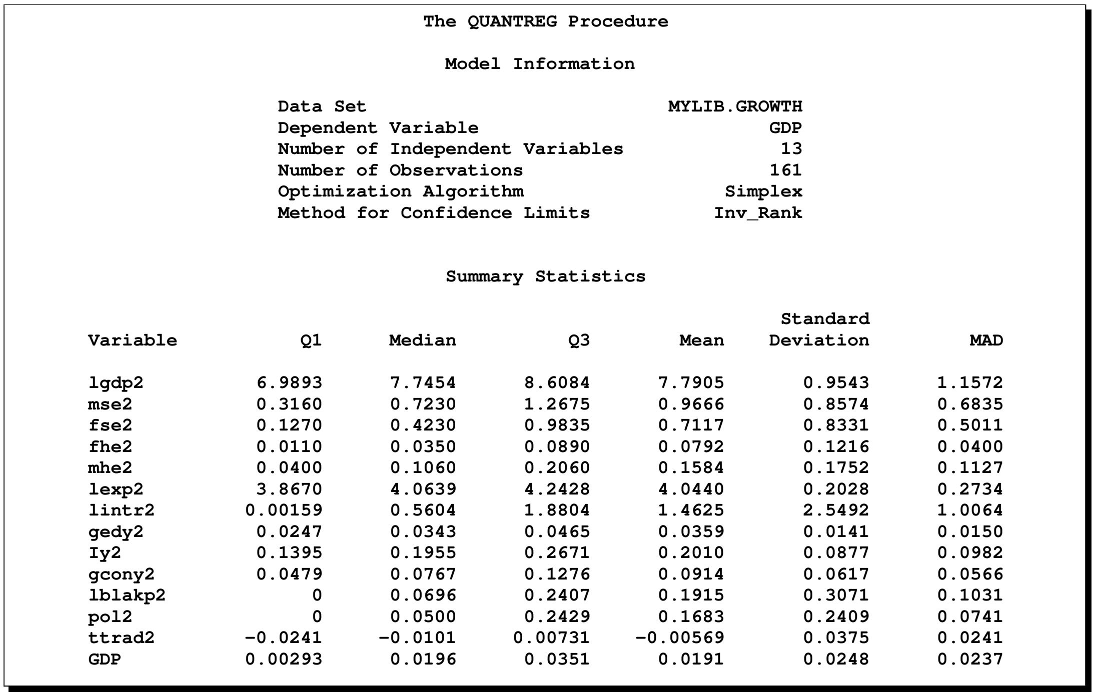
Variable	Description
Country	Country's Name and Period
GDP	Annual Change Per Capita GDP
lgdp2	Initial Per Capita GDP
mse2	Male Secondary Education
fse2	Female Secondary Education
fhe2	Female Higher Education
mhe2	Male Higher Education
lexp2	Life Expectancy
lintr2	Human Capital
gedy2	Education/GDP
Iy2	Investment/GDP
gcony2	Public Consumption/GDP
lblakp2	Black Market Premium
pol2	Political Instability
ttrad2	Growth Rate Terms Trade

The goal is to study the effect of the covariates on GDP. First, median regression is used for a preliminary exploration.

```
ods graphics on;
proc quantreg data=growth;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2
                lintr2 gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
                / quantile=.5 diagnostics leverage(cutoff=8)
                  plots=(rdplot ddplot reshistogram);
    id Country;
    test_lgdp2: test lgdp2 / lr wald;
run;
ods graphics off;
```

The QUANTREG procedure employs the default simplex algorithm to estimate the parameters. Since this is a relatively small data set, the rank method is used to compute confidence limits.

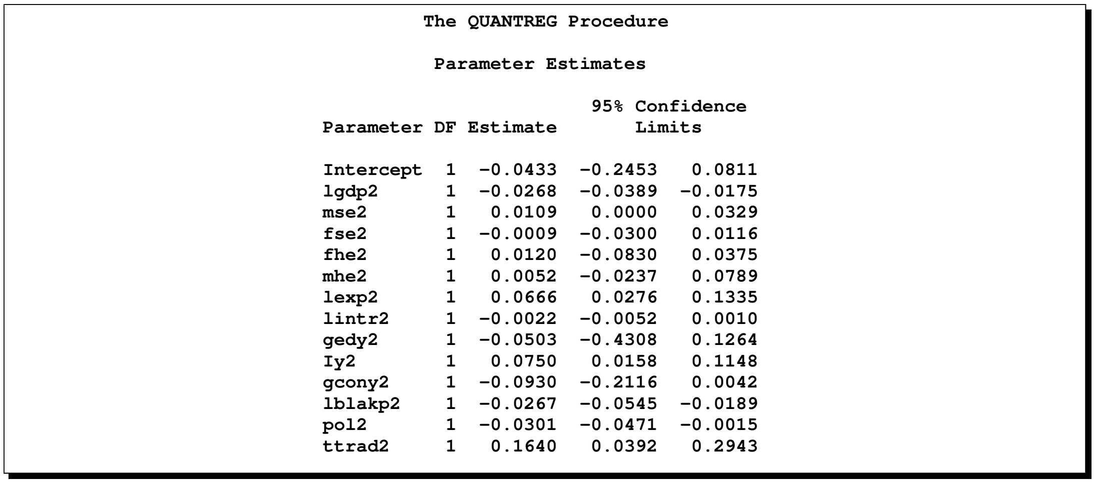
Figure 7 displays model information and summary statistics for the variables in the model. Six summary statistics are computed, including the median and the median absolute deviation (MAD), which are robust measures of univariate location and scale, respectively. For the variable lintr2 (Human Capital), both the mean and standard deviation are much larger than the corresponding robust measures, median and MAD. This indicates that this variable may have outliers.



```
                              The QUANTREG Procedure

                                Model Information

                    Data Set                            MYLIB.GROWTH
                    Dependent Variable                           GDP
                    Number of Independent Variables               13
                    Number of Observations                       161
                    Optimization Algorithm                   Simplex
                    Method for Confidence Limits            Inv_Rank


                                Summary Statistics

                                                                   Standard
       Variable         Q1     Median         Q3       Mean     Deviation          MAD

       lgdp2        6.9893     7.7454     8.6084     7.7905        0.9543       1.1572
       mse2         0.3160     0.7230     1.2675     0.9666        0.8574       0.6835
       fse2         0.1270     0.4230     0.9835     0.7117        0.8331       0.5011
       fhe2         0.0110     0.0350     0.0890     0.0792        0.1216       0.0400
       mhe2         0.0400     0.1060     0.2060     0.1584        0.1752       0.1127
       lexp2        3.8670     4.0639     4.2428     4.0440        0.2028       0.2734
       lintr2      0.00159     0.5604     1.8804     1.4625        2.5492       1.0064
       gedy2        0.0247     0.0343     0.0465     0.0359        0.0141       0.0150
       Iy2          0.1395     0.1955     0.2671     0.2010        0.0877       0.0982
       gcony2       0.0479     0.0767     0.1276     0.0914        0.0617       0.0566
       lblakp2           0     0.0696     0.2407     0.1915        0.3071       0.1031
       pol2              0     0.0500     0.2429     0.1683        0.2409       0.0741
       ttrad2      -0.0241    -0.0101    0.00731   -0.00569        0.0375       0.0241
       GDP         0.00293     0.0196     0.0351     0.0191        0.0248       0.0237
```

Figure 7. Model Information and Summary Statistics

Figure 8 displays parameter estimates and 95% confidence intervals computed with the rank method.



```
                              The QUANTREG Procedure

                               Parameter Estimates

                                            95% Confidence
                   Parameter DF Estimate        Limits

                   Intercept  1  -0.0433  -0.2453   0.0811
                   lgdp2      1  -0.0268  -0.0389  -0.0175
                   mse2       1   0.0109   0.0000   0.0329
                   fse2       1  -0.0009  -0.0300   0.0116
                   fhe2       1   0.0120  -0.0830   0.0375
                   mhe2       1   0.0052  -0.0237   0.0789
                   lexp2      1   0.0666   0.0276   0.1335
                   lintr2     1  -0.0022  -0.0052   0.0010
                   gedy2      1  -0.0503  -0.4308   0.1264
                   Iy2        1   0.0750   0.0158   0.1148
                   gcony2     1  -0.0930  -0.2116   0.0042
                   lblakp2    1  -0.0267  -0.0545  -0.0189
                   pol2       1  -0.0301  -0.0471  -0.0015
                   ttrad2     1   0.1640   0.0392   0.2943
```

Figure 8. Parameter Estimates

Diagnostics for the median regression fit are displayed in Figure 9 and Figure 10, which are requested with the PLOTS= option. Figure 9 plots the standardized residuals from median regression against the robust MCD distance. This display is used to diagnose both vertical outliers and horizontal leverage points. Figure 10 plots the robust MCD distance against the Mahalanobis distance. This display is used to diagnose leverage points.

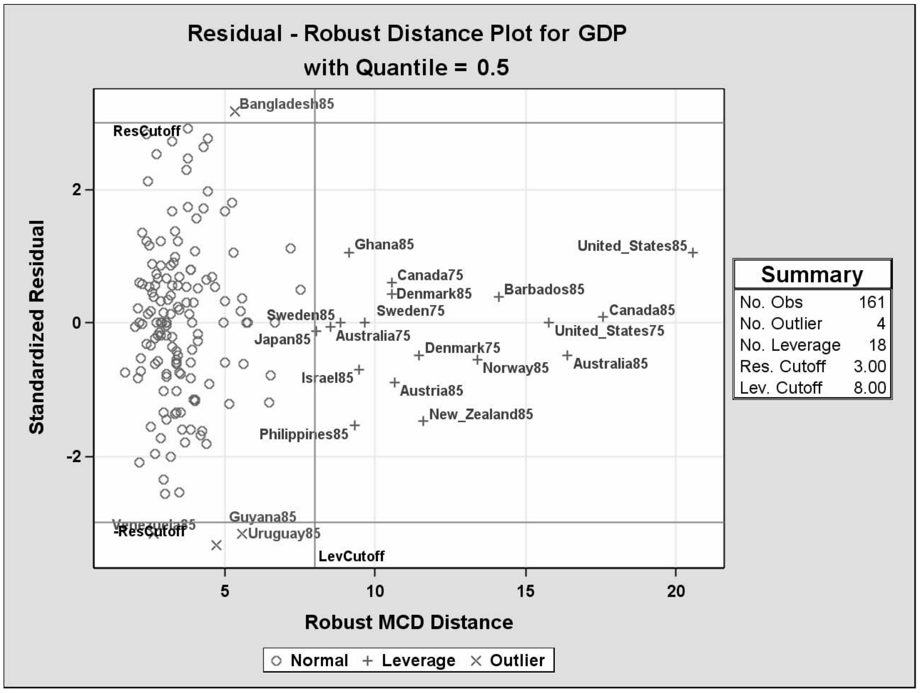


Figure 9. Residual-Robust Distance Plot

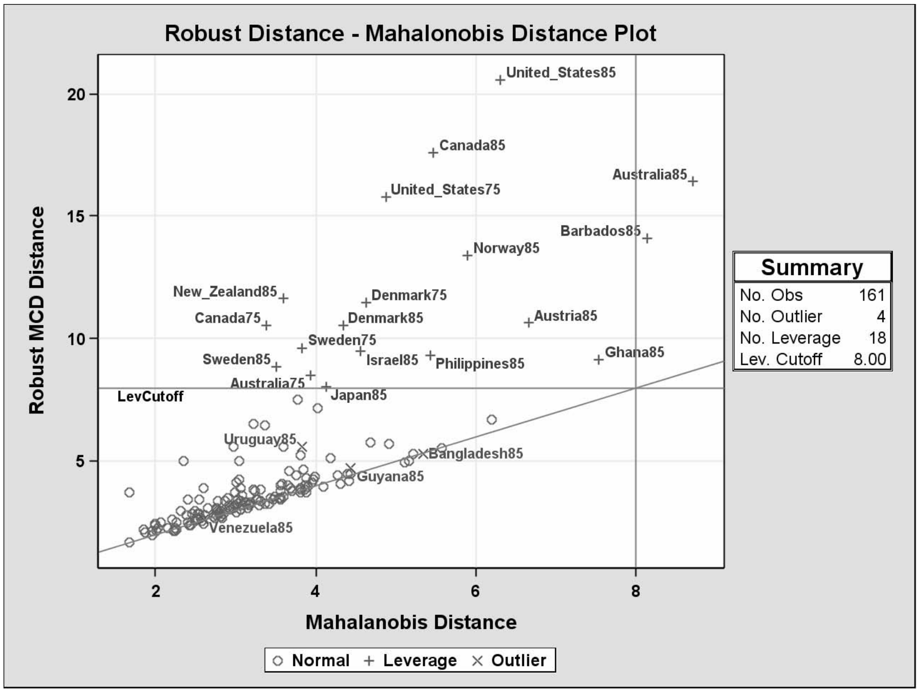


Figure 10. Robust Distance-Mahalanobis Distance Plot

The cutoff value 8 specified with the LEVERAGE option is close to the maximum of the Mahalanobis distance. Eighteen points are diagnosed as high leverage points, and almost all are countries with high Human Capital, which is the major contributor to the high leverage as observed from the summary statistics. Four points are diagnosed as outliers using the default cutoff value of 3. However, these are not extreme outliers.

A histogram of the standardized residuals from median regression and two fitted density curves are displayed in Figure 11. This shows that median regression fits the data well.

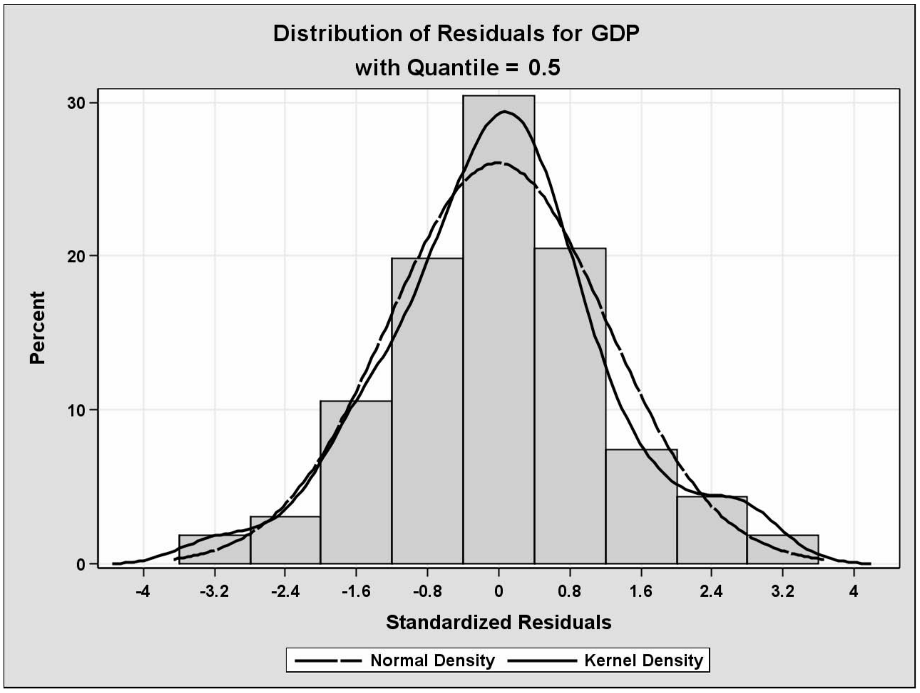
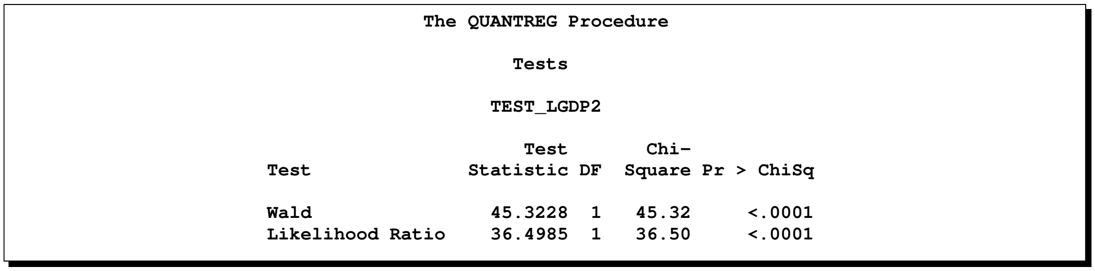


Figure 11. Histogram for Residuals

Tests of significance for the initial per-capita GDP (LGDP2) are shown in Figure 12.



```
                          The QUANTREG Procedure

                                  Tests

                                TEST_LGDP2

                                Test       Chi-
Test                       Statistic DF   Square Pr > ChiSq

Wald                         45.3228  1    45.32     <.0001
Likelihood Ratio             36.4985  1    36.50     <.0001
```

Figure 12. Tests for Regression Coefficient

The QUANTREG procedure computes entire quantile processes for covariates when the option QUANTILE=ALL is specified in the MODEL statement. The regression quantile $\hat{\beta}(\tau)$, as a function of τ, is called a quantile process when τ varies continuously in $(0, 1)$. Confidence intervals for quantile processes can be computed with the sparsity or resampling methods, but not the rank method because the computation would be prohibitively expensive.

```
ods graphics on;
proc quantreg ci=sparsity data=growth;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2 lintr2 gedy2 Iy2 gcony2
                lblakp2 pol2 ttrad2  / quantile=all plot=quantplot;
run;
ods graphics off;
```

A total of 14 quantile process plots are computed. Figure 13 and Figure 14 display two panels of eight selected process plots. The 95% confidence bands are shaded.

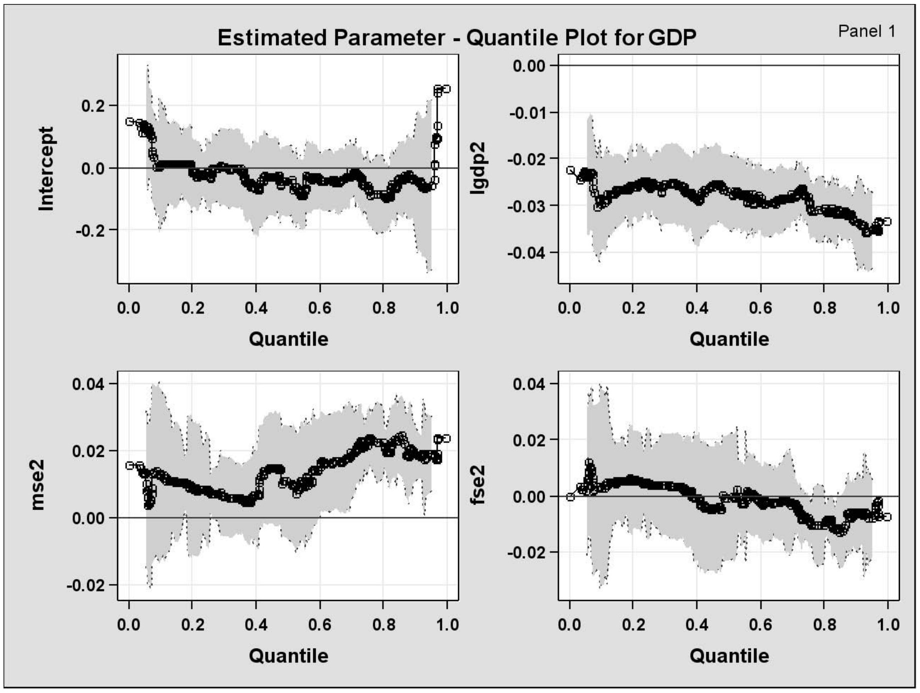


Figure 13. Quantile Processes with 95% Confidence Bands

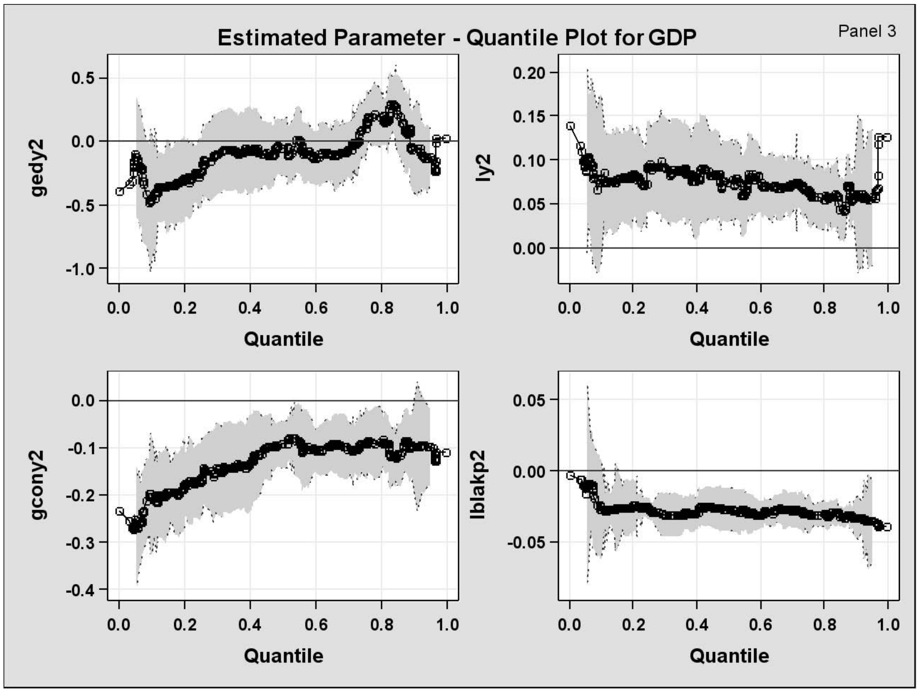


Figure 14. Quantile Processes with 95% Confidence Bands

As pointed out by Koenker and Machado (1999), previous studies of the Barro growth data have focused on the effect of the initial per-capita GDP on the growth of this variable (annual change per-capita GDP). A single process plot for this effect can be requested with the following statements:

```
ods graphics on;
proc quantreg ci=sparsity data=growth;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2 lintr2 gedy2 Iy2 gcony2
                lblakp2 pol2 ttrad2 / quantile=all plot=quantplot(lgdp2);
run;
ods graphics off;
```

The plot is shown in Figure 15.

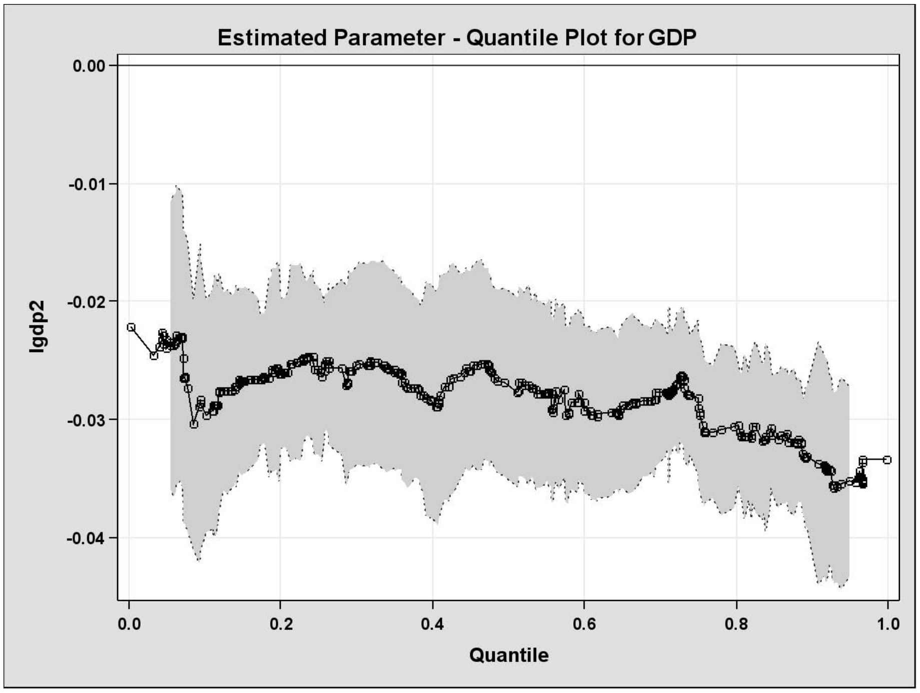


Figure 15. Quantile Process Plot for LGDP2

The confidence bands here are computed using the sparsity method with the non i.i.d. assumption, unlike Koenker and Machado (1999), who used the rank method for a few selected points. The figure suggests that the effect of the initial level of GDP is relatively constant over the entire distribution, with a slightly stronger effect in the upper tail.

The effects of other covariates are quite varied. An interesting covariate is public consumption/GDP (gcony2) (first plot in second panel), which has a constant effect over the upper half of the distribution and a larger effect in the lower tail. For the analysis of effects of other covariates, refer to Koenker and Machado (1999).

SCALABILITY

The QUANTREG procedure implements algorithms for parallel computing, which can be used when you are running on a machine with multiple processors. You can use the global SAS option CPUCOUNT to specify the number of threads. For example, the following statement specifies eight threads:

```
options cpucount = 8;
```

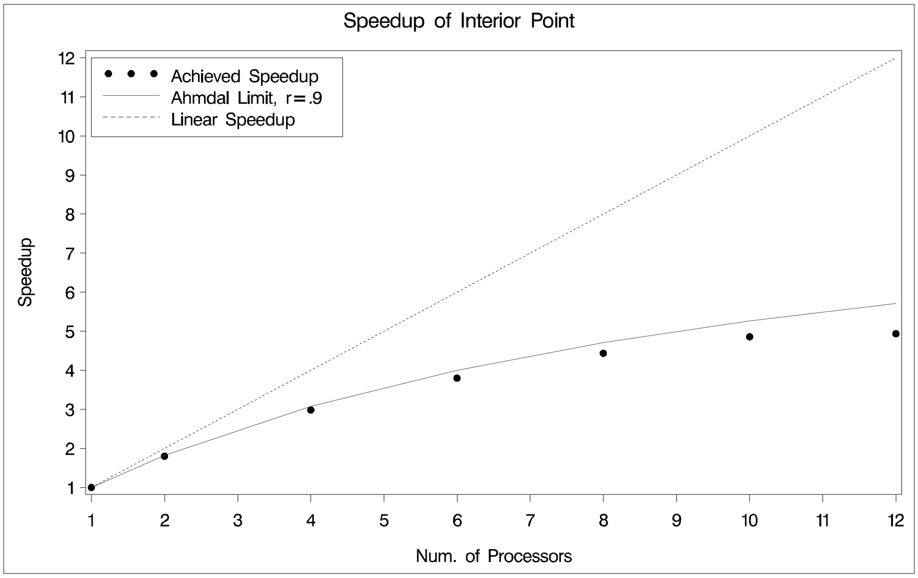


Figure 16. Speedup of the Interior Point Algorithm

Figure 16 displays the speedup achieved in the implementation of a parallel version of the interior point algorithm, together with the Amdahl limit and the linear speedup. The simulation was run on a Sun Solaris 8 configured with 12 processors for data sets with 10,000 observations and 200 variables. The Amdahl limit is based on the parallelizable fraction of 0.9, which is close to the empirical percentage of time spent on cross products. The achieved speedup is close to the Amdahl limit.

More details about optimization and multithreading in computation for quantile regression can be found in Chen and Wei (2005).

APPENDIX

Simplex Algorithm

Let $\mu = [y - A'\beta]_+$, $\nu = [A'\beta - y]_+$, $\phi = [\beta]_+$, and $\varphi = [-\beta]_+$, where $[z]_+$ is the nonnegative part of z.

Let $D_{LAR}(\beta) = \sum_{i=1}^{n} |y_i - x_i'\beta|$. For the L_1 problem, the simplex approach solves $\min_\beta D_{LAR}(\beta)$ by the reformulation

$$\min_{\beta}\{e'\mu + e'\nu | y = A'\beta + \mu - \nu, \{\mu, \nu\} \in \boldsymbol{R}_+^n\}$$

where e denotes an n-vector of ones.

Let $B = [A' \ \ -A' \ I \ \ -I]$, $\theta = (\phi' \ \varphi' \ \mu' \ \nu')'$, and $d = (\mathbf{0}' \ \mathbf{0}' \ e' \ e')'$ where $\mathbf{0}' = (0 \ 0 \ ... \ 0)_p$. The reformulation presents a standard LP problem:

$$\begin{array}{ll} (P) & \min_{\theta} d'\theta \\ \text{subject to} & B\theta = y \\ & \theta \geq 0 \end{array}$$

This problem has the dual formulation

$$\begin{array}{ll} (D) & \max_z y'z \\ \text{subject to} & B'z \leq d \end{array}$$

which can be simplified as $\max_z\{y'z|Az = 0, z \in [-1,1]^n\}$. By setting $\eta = \frac{1}{2}z + \frac{1}{2}e, b = \frac{1}{2}Ae$, it becomes $\max_\eta\{y'\eta|A\eta = b, \eta \in [0,1]^n\}$. For quantile regression, the minimization problem is $\min_\beta \sum \rho_\tau(y_i - x_i'\beta)$, and a similar set of steps lead to the dual formulation

$$\max_z\{y'z|Az = (1-\tau)Ae, z \in [0,1]^n\}$$

The QUANTREG procedure solves this LP problem using the simplex algorithm of Barrodale and Roberts (1973). This algorithm solves the primary LP problem (*P*) by two stages, which exploit the special structure of the coefficient matrix B. The first stage only picks the columns in A' or $-A'$ as pivotal columns. The second stage only interchanges the columns in I or $-I$ as basis or nonbasis columns. The algorithm obtains an optimal solution by executing these two stages interactively. Moreover, because of the special structure of B, only the main data matrix A is stored in the current memory.

This special version of the simplex algorithm for median regression can be naturally extended to quantile regression for any given quantile, even for the entire quantile process (Koenker and d'Orey 1993). It greatly reduces the computing time required by a general simplex algorithm, and it is suitable for data sets with less than 5,000 observations and 50 variables.

Interior Point Algorithm

There are many variations of interior point algorithms. The QUANTREG procedure uses the Primal-Dual with Predictor-Corrector algorithm as implemented in Lustig, Marsden, and Shanno (1992). More information about this particular algorithm and related theory can also be found in the text by Roos, Terlaky, and Vial (1997).

To be consistent with the conventional LP setting, let $c = -y$, $b = (1-\tau)Ae$, and let u be the general upper bound. The linear program to be solved is

$$\begin{array}{ll} & min\{c'z\} \\ \text{subject to} & Az = b \\ & 0 \leq z \leq u \end{array}$$

To simplify the computation, this is treated as the *primal* problem. The problem has n variables. The index i denotes a variable number, k denotes an iteration number, and if used as a subscript or superscript it denotes "of iteration k".

Let v be the primal slack so that $z + v = u$. Associate dual variables w with these constraints. The Interior Point solves the system of equations to satisfy the Karush-Kuhn-Tucker (KKT) conditions for optimality:

$$\begin{array}{ll} & Az = b \\ & z + v = u \\ & A't + s - w = c \\ & ZSe = 0 \\ & VWe = 0 \\ & z, s, v, w \geq 0 \\ \text{where} & W = diag(w), \text{ (that is, } W_{i,j} = w_i \text{ if } i = j, W_{i,j} = 0 \text{ otherwise)} \end{array}$$

$$V = diag(v), Z = diag(z), S = diag(s)$$

These are the conditions for feasibility, with the addition of *complementarity* conditions $ZSe = 0$ and $VWe = 0$. $c'z = b't - u'w$ must occur at the optimum. Complementarity forces the optimal objectives of the primal and dual to be equal, $c'z_{opt} = b't_{opt} - u'w_{opt}$.

The *duality gap*, $c'z - b't + u'w$, is used to measure the convergence of the algorithm. You can specify a tolerance for this convergence criterion with the TOLERANCE= option in the PROC statement.

The Interior Point algorithm works by using Newton's method to find a direction $(\Delta z^k, \Delta t^k, \Delta s^k, \Delta v^k, \Delta w^k)$ to move from the current solution $(z^k, t^k, s^k, v^k, w^k)$ toward a better solution.

To do this, two steps are used. The first step is called an *affine* step, which solves a linear system using Newton's method to find a direction (Δz^k_{aff}, Δt^k_{aff}, Δs^k_{aff}, Δv^k_{aff}, Δw^k_{aff}) to reduce the complementarity toward zero. The second step is called a *centering* step, which solves another linear system to determine a centering vector (Δz^k_c, Δt^k_c, Δs^k_c, Δv^k_c, Δw^k_c) to further reduce the complementarity. The centering step may not reduce too much of the complementarity; however, it builds up the *central path* and makes substantial progress toward the optimum in the next iteration. With these two steps, then

$$(\Delta z^k, \Delta t^k, \Delta s^k, \Delta v^k, \Delta w^k) = (\Delta z_{aff}, \Delta t_{aff}, \Delta s_{aff}, \Delta v_{aff}, \Delta w_{aff}) + (\Delta z_c, \Delta t_c, \Delta s_c, \Delta v_c, \Delta w_c)$$

$$(z^{k+1}, t^{k+1}, s^{k+1}, v^{k+1}, w^{k+1}) = (z^k, t^k, s^k, v^k, w^k) + \kappa(\Delta z^k, \Delta t^k, \Delta s^k, \Delta v^k, \Delta w^k)$$

where κ is the *step length* assigned a value as large as possible but not so large that a z_i^{k+1}, s_i^{k+1}, v_i^{k+1}, or w_i^{k+1} is "too close" to zero. You can control the step length with a parameter specified by the KAPPA= option in the PROC statement.

Although the Predictor-Corrector variant entails solving two linear systems instead of one, fewer iterations are usually required to reach the optimum. The additional overhead of calculating the second linear system is small, as the factorization of the $(A\Theta^{-1}A')$ matrix has already been performed to solve the first linear system.

You can specify the starting point with the INEST= option in the PROC statement. By default, the starting point is set to be the least-squares estimate.

Smoothing Algorithm

The finite smoothing algorithm was used by Clark and Osborne (1986) and by Madsen and Nielsen (1993) for the L_1 regression. It can be naturally extended to compute regression quantiles. What follows is a brief description of this algorithm; more details can be found in Chen (2003).

The nondifferentiable function

$$D_{\rho_\tau}(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta)$$

can be approximated by the smooth function

$$D_{\gamma,\tau}(\beta) = \sum_{i=1}^{n} H_{\gamma,\tau}(r_i(\beta))$$

where $r_i(\beta) = y_i - x_i'\beta$ and

$$H_{\gamma,\tau}(t) = \begin{cases} t(\tau-1) - \frac{1}{2}(\tau-1)^2\gamma & \text{if } t \le (\tau-1)\gamma \\ \frac{t^2}{2}\gamma & \text{if } (\tau-1)\gamma \le t \le \tau\gamma \\ t\tau - \frac{1}{2}\tau^2\gamma & \text{if } t \ge \tau\gamma \end{cases}$$

See Figure 17 for a plot of $H_{\gamma,\tau}$ and ρ_τ.

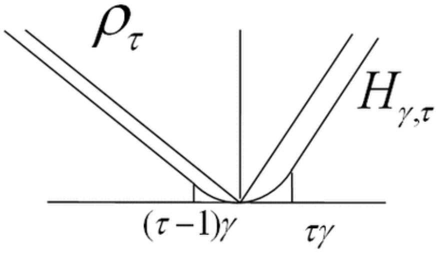


Figure 17. Objective Functions $H_{\gamma,\tau}$ and ρ_τ

The function $H_{\gamma,\tau}$ is determined by whether $r_i(\beta) \le (\tau-1)\gamma$, $r_i(\beta) \ge \tau\gamma$, or $(\tau-1)\gamma \le r_i(\beta) \le \tau\gamma$. These inequalities divide $\boldsymbol{R}^p$ into subregions separated by the parallel hyperplanes $r_i(\beta) = (\tau-1)\gamma$ and $r_i(\beta) = \tau\gamma$. The set of all such hyperplanes is denoted by $B_{\gamma,\tau}$:

$$B_{\gamma,\tau} = \{\beta \in \boldsymbol{R}^p | \exists i : r_i(\beta) = (\tau-1)\gamma \text{ or } r_i(\beta) = \tau\gamma\}.$$

Define the sign vector $s_{\gamma,\tau}(\beta) = (s_1(\beta), ..., s_n(\beta))'$ by

$$s_i = s_i(\beta) = \begin{cases} -1 & \text{if } r_i(\beta) \le (\tau-1)\gamma \\ 0 & \text{if } (\tau-1)\gamma \le r_i(\beta) \le \tau\gamma \\ 1 & \text{if } r_i(\beta) \ge \tau\gamma \end{cases}$$

and introduce $w_i = w_i(\beta) = 1 - s_i^2(\beta)$. Thus,

$$D_{\gamma,\tau}(\beta) = \frac{1}{2}\gamma r' W_{\gamma,\tau} r + v'(s) r + c(s),$$

where $W_{\gamma,\tau}$ is the diagonal n by n matrix with diagonal elements $w_i(\beta)$, $v'(s) = (s_1((2\tau-1)s_1+1)/2, ..., s_n((2\tau-1)s_n+1)/2)$, $c(s) = \sum[\frac{1}{4}(1-2\tau)\gamma s_i - \frac{1}{4}s_i^2(1-2\tau+2\tau^2)\gamma]$, and $r(\beta) = (r_1(\beta), ..., r_n(\beta))'$. The gradient of $D_{\gamma,\tau}$ is given by

$$D^{(1)}_{\gamma,\tau}(\beta) = -A[\frac{1}{\gamma} W_{\gamma,\tau}(\beta) r(\beta) + g(s)]$$

and for $\beta \in \boldsymbol{R}^p \backslash B_{\gamma,\tau}$ the Hessian exists and is given by

$$D^{(2)}_{\gamma,\tau}(\beta) = \frac{1}{\gamma} A W_{\gamma,\tau}(\beta) A'.$$

The gradient is a continuous function in $\boldsymbol{R}^p$, whereas the Hessian is piecewise constant.

The smoothing algorithm for minimizing D_{ρ_τ} is based on minimizing $D_{\gamma,\tau}$ for a set of decreasing γ. The essential advantage of the smoothing algorithm is that the solution $\beta_{0,\tau}$ can be detected when $\gamma > 0$ is small enough, i.e., it is not necessary to let γ converge to zero in order to find a minimizer of D_{ρ_τ}. For every new value of γ, information from the previous solution is utilized. The algorithm stops before going through the whole sequence of γ, which is generated by the algorithm itself. The convergence is indicated by no change of the status while γ goes through this sequence. Therefore, the algorithm is usually called the finite smoothing algorithm.

For a given threshold γ, a modified Newton iteration is used. Starting from an initial estimator $\beta_\gamma(\tau)$, the search direction **h** is found by solving the equation

$$D^{(2)}_{\gamma,\tau}(\beta_\gamma(\tau))\mathbf{h} = D^{(1)}_{\gamma,\tau}(\beta_\gamma(\tau))$$

Another advantage of the smoothing algorithm is that when solving the above equation for a new threshold γ', the factorization only needs a partial update instead of a full update as in the interior point algorithm. This saves time with large p.

Each of the previous three algorithms has its own advantages. None of them can fully dominate the others. Although the simplex algorithm is slow with a large number of observations, it is the most stable of the algorithms. For various kinds of data, especially data with a large portion of outliers and leverage points, the simplex algorithm always find a solution, while the other two algorithms might fail with floating point errors. The interior point algorithm is very fast for *slender* data sets, which have a large number of observations and a small number of covariates. The algorithm has a simple structure and might be easily adopted to other situations, e.g., constrained quantile regression. The finite smoothing algorithm is simple in theory for quantile regression, and has the advantage in computing speed with a large number of covariates.

Sparsity

Consider the linear model

$$y_i = x_i'\beta + \epsilon_i$$

and assume that $\{\epsilon_i\}$, $i = 1, ..., n$, are i.i.d. with a distribution F and a density $f = F'$, where $f(F^{-1}(\tau)) > 0$ in a neighborhood of τ. Under some mild conditions

$$\sqrt{n}(\hat\beta(\tau) - \beta(\tau)) \rightarrow N(0, \omega^2(\tau, F)\Omega^{-1})$$

where $\omega^2(\tau, F) = \tau(1-\tau)/f^2(F^{-1}(\tau))$ and $\Omega = \lim_{n\rightarrow\infty} n^{-1}\sum x_i x_i'$. See Koenker and Bassett (1982).

This asymptotic distribution for the regression quantile $\hat\beta(\tau)$ can be used to construct confidence intervals. However, the reciprocal of the density function

$$s(\tau) = [f(F^{-1}(\tau))]^{-1}$$

which is called the *sparsity function*, must be estimated first.

Available estimators of the sparsity function are very sensitive to the i.i.d. assumption. Alternately, Koenker and Machado (1999) considered the non i.i.d. case. By assuming the local linearity of the conditional quantile function $Q(\tau|x)$ in x, they proposed a local estimator of the density function using difference quotient. A Huber sandwich estimate of the covariance and standard error is used to construct the confidence intervals.

By default, the QUANTREG procedure computes the non i.i.d. confidence intervals. The i.i.d. confidence intervals can be requested with the IID option in the PROC statement.

Inversion of Rank Tests

The classical theory of rank tests can be extended to the test of the hypothesis H_0: $\beta_2 = \eta$ in the linear regression model $y = X_1\beta_1 + X_2\beta_2 + \epsilon$. Here $(X_1, X_2) = A'$. See Gutenbrunner and Jureckova (1992). By inverting this test, confidence intervals for the regression quantile estimates of β_2 may be computed.

The rankscore function $\hat a_n(t) = (\hat a_{n1}(t), ..., \hat a_{nn}(t))$ can be solved from the dual

$$\max\{(y - X_2\eta)'a | X_1'a = (1-t)X_1'e,\ a \in [0,1]^n\}$$

For a fixed quantile τ, integrating $\hat{a}_{ni}(t)$ with respect to the τ-quantile score function

$$\varphi_\tau(t) = \tau - I(t < \tau)$$

yields the τ-quantile scores:

$$\hat{b}_{ni} = -\int_0^1 \varphi_\tau(t) d\hat{a}_{ni}(t) = \hat{a}_{ni}(\tau) - (1-\tau)$$

Under the null hypothesis H_0: $\beta_2 = \eta$

$$S_n(\eta) = n^{-1/2} X_2' \hat{b}_n(\eta) \rightarrow N(0, \tau(1-\tau)\Omega_n)$$

where $\Omega_n = n^{-1} X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2$.

Let

$$T_n(\eta) = \frac{1}{\sqrt{\tau(1-\tau)}} S_n(\eta)\Omega_n^{-1/2}$$

then $T_n(\hat{\beta}_2(\tau)) = 0$ from the constraint $A\hat{a} = (1-\tau)Ae$ in the full model. A critical value can be specified for T_n. The dual vector $\hat{a}_n(\eta)$ is a piecewise constant in η and η may be altered without compromising the optimality of $\hat{a}_n(\eta)$ as long as the signs of the residuals in the primal quantile regression problem do not change. When η gets to such a boundary the solution does change, but may be restored by taking one simplex pivot. The process may continue in this way until $T_n(\eta)$ exceeds the specified critical value. Since $T_n(\eta)$ is piecewise constant, interpolation can be used to obtain the desired level of confidence interval; see Koenker and d'Orey (1993).

Resampling

The bootstrap can be implemented to compute confidence intervals for regression quantile estimates. As in other regression applications, both the residual bootstrap and the xy-pair bootstrap can be used. The former assumes i.i.d. random errors and resamples from the residuals, while the later resamples xy pairs and accommodates some forms of heteroscedasticity. Koenker (1994) considered a more interesting resampling mechanism, resampling directly from the full regression quantile process, which he called the Heqf bootstrap.

Unlike these bootstrap methods, Parzen, Wei, and Ying (1994) observed that

$$S(b) = n^{-1/2} \sum_{i=1}^{n} x_i(\tau - I(y_i \le x_i'b))$$

which is the estimating equation for the τth regression quantile, is a pivotal quantity for the true τth quantile regression parameter β_τ, i.e., its distribution may be generated exactly by a random vector U which is a weighted sum of independent, re-centered Bernoulli variables. They further showed that for large n the distribution of $\hat{\beta}(\tau) - \beta_\tau$ can be approximated by the conditional distribution of $\hat{\beta}_U - \hat{\beta}_n(\tau)$, where $\hat{\beta}_U$ solves an augmented quantile regression problem with $n+1$ observation and $x_{n+1} = -n^{-1/2}u/\tau$ and y_{n+1} is sufficiently large for a given realization of u. This approach, by exploiting the asymptotically pivot role of the quantile regression "gradient condition," also achieves some robustness to certain heteroscedasticity.

Although the bootstrap method by Parzen, Wei, and Ying (1994) is much simpler, it is still too time consuming for relatively large data sets, especially for high-dimensional data sets. He and Hu (2002) developed a new general resampling method, referred to as the Markov chain marginal bootstrap (MCMB). For quantile regression, the MCMB method has the advantage that it solves p one-dimensional equations instead of p-dimensional equations, as do the previous bootstrap methods. This greatly improves the feasibility of the resampling method in computing confidence intervals for regression quantiles. Since resampling methods achieve stability only for relatively large data sets, they are not recommended for small data sets ($n < 5000$ and $p < 20$).

The QUANTREG procedure implements the MCMB resampling methods due to the work of He and Hu (2002).

REFERENCES

Barro, R. and Lee, J. W. (1994), "Data Set for a Panel of 138 Countries," discussion paper, National Bureau of Econometric Research. <http://www.nber.org/pub/barro.lee>.

Barrodale, I. and Roberts, F. D. K. (1973), "An Improved Algorithm for Discrete l_1 Linear Approximation," *SIAM J. Numer. Anal.,* 10, 839-848.

Chen, C. (2003), "A Finite Smoothing Algorithm for Quantile Regression," submitted, preprint available from the author.

Chen, C. (2004), "Growth Charts of Body Mass Index (BMI) with Quantile Regression," MS. available from the author.

Chen, C. and Wei, Y. (2005), "Computational Issues on Quantile Regression," *Special Issue on Quantile Regression and Related Methods*, *Sankhya*, forthcoming.

Chock, D. P., Winkler, S. L., and Chen, C. (2000), "A Study of the Association between Daily Mortality and Ambient air Pollutant Concentrations in Pittsburgh, Pennsylvania," *Journal of the Air and Waste Management Association*, 50, 1481–1500.

Clark D. I. and Osborne, M. R. (1986), "Finite Algorithms for Huber's M-estimator," *SIAM J. Sci. Statist. Comput.*, 6, 72–85.

Gutenbrunner, C. and Jureckova, J. (1992), "Regression Rank Scores and Regression Quantiles." *Annals of Statistics*, 20, 305–330.

He, X. and Hu, F. (2002), "Markov Chain Marginal Bootstrap," *Journal of the American Statistical Association*, 97, 783–795.

Karmarkar, N. (1984), "A New Polynomial-time Algorithm for Linear Programming," *Combinatorica*, 4, 373–395.

Koenker, R. (1994), "Confidence Intervals for Quantile Regression," *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, P. Mandl and M. Huskova eds., Heidelberg: Physica-Verlag.

Koenker, R. and Bassett, G. W. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R. and Bassett, G. W. (1982), "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50, 43–61.

Koenker, R. and d'Orey, V. (1993), "Computing Regression Quantiles," *Applied Statistics*, 43, 410–414.

Koenker, R. and Hallock K. (2001), "Quantile Regression: An Introduction," *Journal of Economic Perspectives*, 15, 143–156.

Koenker, R. and Machado, A. F. (1999), "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310.

Lustig, I. J., Marsden, R. E., and Shanno, D. F. (1992), "On Implementing Mehrotra's Predictor-Corrector Interior-Point Method for Linear Programming," *SIAM J. Optimization*, 2, 435–449.

Madsen, K. and Nielsen, H. B. (1993), “A Finite Smoothing Algorithm for Linear L_1 Estimation,” *SIAM J. Optimization*, 3, 223–235.

Parzen, M. I., Wei, L. J., and Ying, Z. (1994), “A Resampling Method Based on Pivotal Estimating Functions,” *Biometrika*, 81, 341–350.

Roos, C., Terlaky, T., and Vial, J.-Ph. (1997), “Theory and Algorithms for Linear Optimization,” Chichester, England: John Wiley & Sons.

Rousseeuw, P. J. and Van Driessen, K. (1999), “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” *Technometrics*, 41, 212–223.

ACKNOWLEDGMENTS The author would like to thank Dr. Robert Rodriguez, Professor Xuming He, and Professor Roger Koenker for reviews, comments, and discussions.

CONTACT INFORMATION Colin (Lin) Chen, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919) 531-6388, FAX (919) 677-4444, Email Lin.Chen@sas.com.