SAS/STAT® 9.22 User's Guide

The QUANTREG Procedure

(Book Excerpt)

SAS® Documentation

This document is an individual chapter from *SAS/STAT® 9.22 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2010. *SAS/STAT® 9.22 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, May 2010

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/publishing** or call 1-800-727-3228.

Chapter 73

The QUANTREG Procedure

Contents

Overview: QUANTREG Procedure

The QUANTREG procedure models the effects of covariates on the conditional quantiles of a response variable by means of quantile regression.

Ordinary least squares (OLS) 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 median or 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.

Figure 73.1 Trout Density in Streams

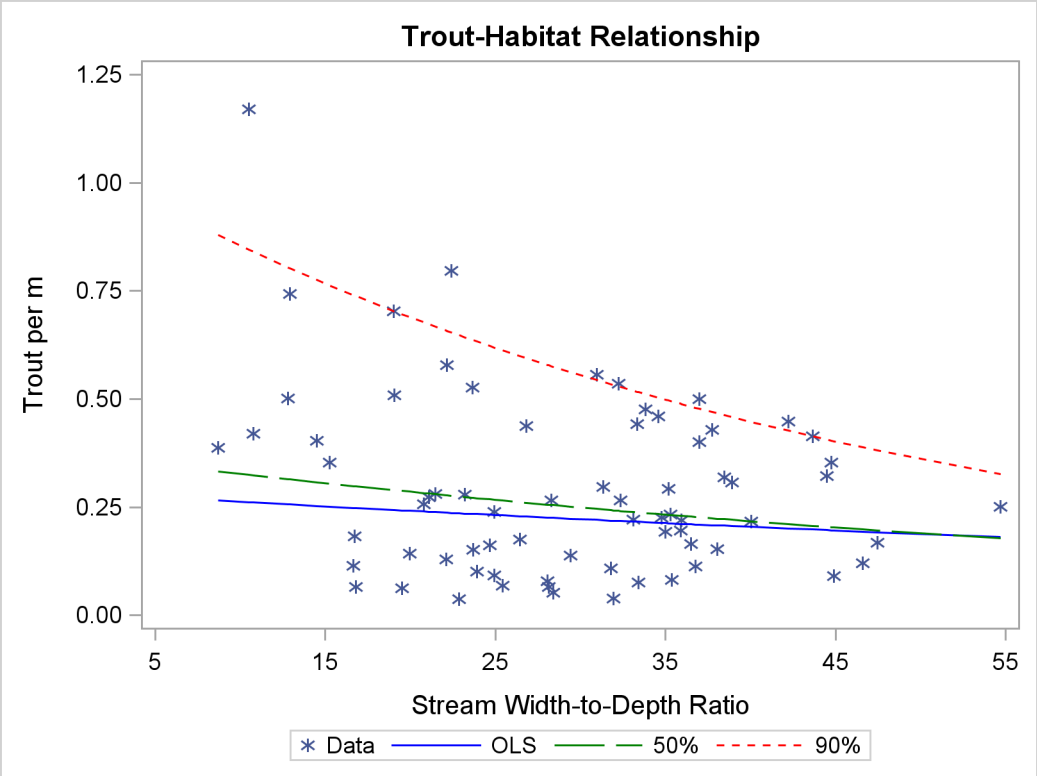


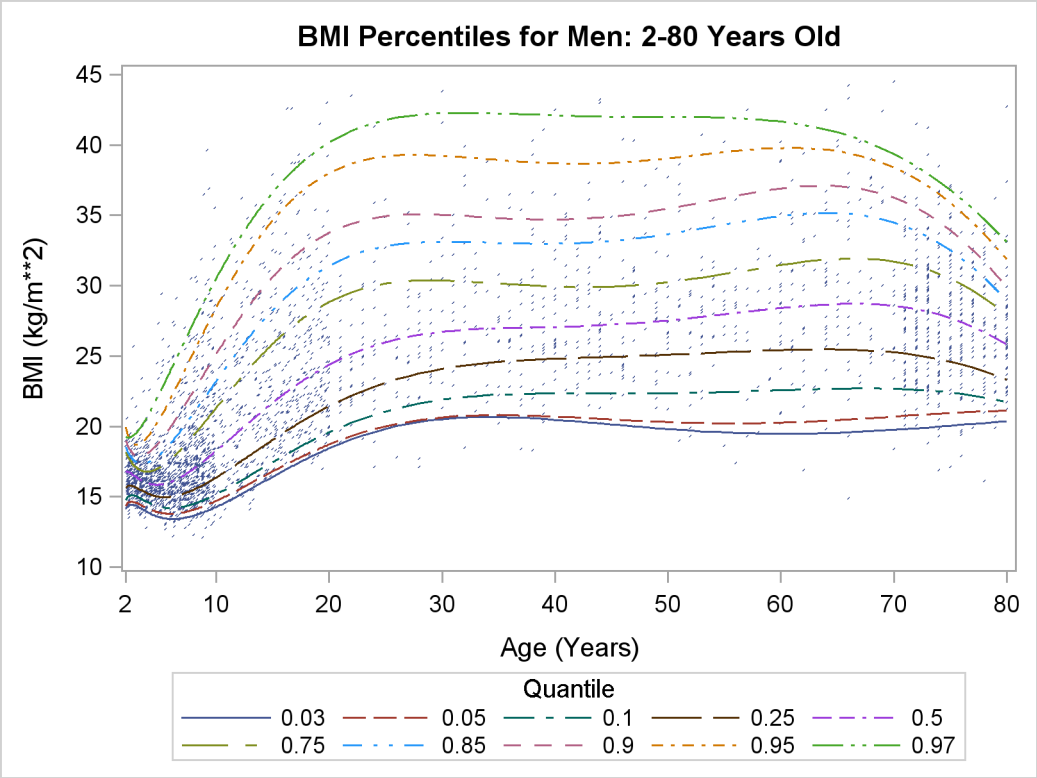
Figure 73.1 illustrates an ecological study in which it is revealing to model upper conditional quantiles. The points represent measurements of trout density and stream width-to-depth ratio taken at 13 streams over seven years.

As analyzed by Dunham, Cade, and Terrell (2002), in addition to the ratio, trout density depends on

a number of unmeasured limiting factors related to the integrity of stream habitat. The interaction of these factors results in unequal variances for the conditional distributions of density given the ratio. When the ratio is the "active" limiting effect, changes in the upper conditional percentiles of density provide a better estimate of this effect than changes in the conditional mean.

The two dashed curves represent the conditional 90th and 50th percentiles of density as determined with the QUANTREG procedure. The analysis was done by using a simple linear regression model for the logarithm of density. (The curves in Figure 73.1 were obtained by transforming the fitted lines back to the original scale. For more details, see the section "Analysis of Fish-Habitat Relationships" on page 6075.) The slope parameter for the 90th percentile has an estimated value of −0.0215 and is significant with a p-value less than 0.01. On the other hand, the slope parameter for the 50th percentile is not significantly different from zero. Similarly, the slope parameter for the mean, obtained with OLS regression, is not significantly different from zero.

Figure 73.2 Quantiles for Body Mass Index



Quantile regression is especially useful with data that are heterogeneous in the sense that the tails and the central location of the conditional distributions vary differently with the covariates. An even more pronounced example of heterogeneity is shown in Figure 73.2, which plots the body mass index of 8,250 men versus their age.

Here, both upper (overweight) and lower (underweight) conditional quantiles are important because they provide the basis for developing growth charts and establishing health standards. The curves in Figure 73.2 were determined with the QUANTREG procedure by using polynomial quantile regression; details are provided in the section "Growth Charts for Body Mass Index" on page 6080. Clearly, the rate of change with age (as expressed by the regression coefficients), particularly for ages less than 20, is different for each conditional quantile.

Heterogeneous data occur in many fields, including biomedicine, econometrics, survival analysis, and ecology. Quantile regression, which includes median regression as a special case, provides a complete picture of the covariate effect when a set of percentiles is modeled, and so it offers the capability to capture important features of the data that might be missed by models that average over the conditional distribution.

Because it makes no distributional assumption about the error term in the model, quantile regression offers considerable model robustness. The assumption of normality, which is often made with OLS regression in order to compute conditional quantiles as offsets from the mean, forces a common set of regression coefficients for all the quantiles. Obviously, quantiles with common slopes would be inappropriate in the preceding examples.

Quantile regression is also flexible in the sense that it does not involve a link function that relates the variance and the mean of the response variable. Generalized linear models, which you can fit with the GENMOD procedure, require both a link function and a distributional assumption such as the normal or Poisson distribution. The goal of generalized linear models is inference about the regression parameters in the linear predictor for the mean of the population. In contrast, the goal of quantile regression is inference on regression coefficients for the conditional quantiles of a response variable that is usually assumed to be continuous.

Quantile regression also offers a degree of data robustness. Unlike OLS regression, it is robust to extreme points in the response direction (outliers). However, it is not robust to extreme points in the covariate space (leverage points). When both types of robustness are of concern, you should consider using the ROBUSTREG procedure (Chapter 75, "The ROBUSTREG Procedure.")

Also, unlike OLS regression, quantile regression is equivariant to monotone transformations of the response variable. For instance, as illustrated in the trout example, the logarithm of the 90th conditional percentile of trout density is the 90th conditional percentile of the logarithm of density.

Note that quantile regression cannot be carried out simply by segmenting the unconditional distribution of the response variable and then obtaining least squares fits for the subsets. This approach leads to disastrous results when, for example, the data include outliers. In contrast, quantile regression uses *all* of the data for fitting quantiles, even the extreme quantiles.

Features

The main features of the QUANTREG procedure are as follows:

- offers simplex, interior point, and smoothing algorithms for estimation
- provides sparsity, rank, and resampling methods for confidence intervals
- provides asymptotic and bootstrap methods for covariance and correlation matrices of the estimated parameters
- provides the Wald, rank, and likelihood ratio tests for the regression parameter estimates
- provides outlier and leverage-point diagnostics

- enables parallel computing when multiple processors are available
- provides row-wise or column-wise output data sets with multiple quantiles
- provides regression quantile spline fits
- produces fit plots, diagnostic plots, and quantile process plots by using ODS Graphics

The next section provides notation and a formal definition for quantile regression.

Quantile Regression

Quantile regression generalizes the concept of a univariate quantile to a conditional quantile given one or more covariates. Recall that a student's score on a test is at the τth quantile if his or her score is better than that of $100\tau\%$ of the students who took the test. The score is also said to be at the 100τth percentile.

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

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

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

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

$$\text{median} = \arg\min_{\xi \in \mathbf{R}} \sum_{i=1}^{n} |y_i - \xi|$$

Likewise, the general τth sample quantile $\xi(\tau)$, which is the analog of $Q(\tau)$, is formulated as the minimizer

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

where $\rho_\tau(z) = z(\tau - I(z < 0))$, $0 < \tau < 1$, and where $I(\cdot)$ denotes the indicator function. The loss function ρ_τ assigns a weight of τ to positive residuals $y_i - \xi$ and a weight of $1 - \tau$ to negative residuals.

Using this loss function, the linear conditional quantile function extends the τth sample quantile $\xi(\tau)$ to the regression setting in the same way that the linear conditional mean function extends the sample mean. Recall that OLS regression estimates the linear conditional mean function $E(Y|X = x) = x'\beta$ by solving for

$$\hat{\beta} = \arg\min_{\beta \in \mathbf{R}^p} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

The estimated parameter $\hat{\beta}$ minimizes the sum of squared residuals in the same way that the sample mean $\hat{\mu}$ minimizes the sum of squares:

$$\hat{\mu} = \arg\min_{\mu \in \mathbf{R}} \sum_{i=1}^{n} (y_i - \mu)^2$$

Likewise, quantile regression estimates the linear conditional quantile function, $Q(\tau|X = x) = x'\beta(\tau)$, by solving

$$\hat{\beta}(\tau) = \arg\min_{\beta \in \mathbf{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta)$$

for $\tau \in (0, 1)$. The quantity $\hat{\beta}(\tau)$ is called the τth *regression quantile*. The case $\tau = 0.5$, which minimizes the sum of absolute residuals, corresponds to median regression, which is also known as L_1 regression.

The set of regression quantiles

$$\{\beta(\tau) : \tau \in (0, 1)\}$$

is referred to as the *quantile process*.

The QUANTREG procedure computes the quantile function $Q(\tau|X = x)$ and conducts statistical inference on the estimated parameters $\hat{\beta}(\tau)$.

Getting Started: QUANTREG Procedure

The following examples demonstrate how you can use the QUANTREG procedure to fit linear models for selected quantiles or for the entire quantile process. The first example explains the use of the procedure in the fish-habitat example, and the second example explains the use of the procedure to construct growth charts for body mass index.

Analysis of Fish-Habitat Relationships

Quantile regression is used extensively in ecological studies (Cade and Noon 2003). Recently, Dunham, Cade, and Terrell (2002) applied quantile regression to analyze fish-habitat relationships for Lahontan cutthroat trout in 13 streams of the eastern Lahontan basin, which covers most of northern Nevada and parts of southern Oregon. The density of trout (number of trout per meter) was measured by sampling stream sites from 1993 to 1999. The width-to-depth ratio of the stream site was determined as a measure of stream habitat.

The goal of this study was to explore the relationship between the conditional quantiles of trout density and the width-to-depth ratio. The scatter plot of the data in Figure 73.1 indicates a nonlinear relationship, and so it is reasonable to fit regression models for the conditional quantiles of the log of density. Since regression quantiles are equivariant under any monotonic (linear or nonlinear) transformation (Koenker and Hallock 2001), the exponential transformation converts the conditional quantiles to the original density scale.

The data set trout, which follows, includes the average numbers of Lahontan cutthroat trout per meter of stream (Density), the logarithm of Density (LnDensity), and the width-to-depth ratios (WDRatio) for 71 samples.

```
data trout;
   input Density WDRatio LnDensity @@;
   datalines;
0.38732     8.6819    -0.94850    1.16956    10.5102     0.15662
0.42025    10.7636    -0.86690    0.50059    12.7884    -0.69197
0.74235    12.9266    -0.29793    0.40385    14.4884    -0.90672
0.35245    15.2476    -1.04284    0.11499    16.6495    -2.16289
0.18290    16.7188    -1.69881    0.06619    16.7859    -2.71523
0.70330    19.0141    -0.35197    0.50845    19.0548    -0.67639
0.06279    19.4959    -2.76796    0.14190    19.9446    -1.95265
0.25725    20.7852    -1.35772    0.27240    21.0870    -1.30048
0.27983    21.4564    -1.27357    0.12860    22.0917    -2.05105
0.57867    22.1627    -0.54702    0.79667    22.4070    -0.22731
0.03730    22.8553    -3.28880    0.27897    23.2003    -1.27666
0.52587    23.6662    -0.64270    0.15075    23.6937    -1.89215
0.10071    23.9129    -2.29548    0.16128    24.6643    -1.82461
0.09254    24.9451    -2.38011    0.23937    24.9492    -1.42974
0.06914    25.4138    -2.67158    0.17586    26.4412    -1.73805
0.43725    26.8025    -0.82725    0.07812    28.0558    -2.54945
0.06576    28.1194    -2.72174    0.26539    28.3045    -1.32654
0.05159    28.3949    -2.96440    0.13779    29.4083    -1.98202
0.55589    30.9569    -0.58719    0.29714    31.3376    -1.21354
0.10857    31.7868    -2.22035    0.03897    31.9464    -3.24485
0.53572    32.2492    -0.62414    0.26580    32.3725    -1.32500
0.22114    33.1017    -1.50896    0.44212    33.3530    -0.81618
0.07646    33.4036    -2.57099    0.47616    33.8079    -0.74200
0.45934    34.5639    -0.77796    0.22627    34.7844    -1.48603
0.19356    35.0004    -1.64215    0.29216    35.1803    -1.23045
0.23243    35.2959    -1.45917    0.08155    35.3704    -2.50654
0.19528    35.9115    -1.63332    0.22023    35.9382    -1.51308
0.16411    36.4884    -1.80723    0.11296    36.7694    -2.18072
```

```
0.49981    36.9893    -0.69353    0.40012    37.0055    -0.91599
0.42912    37.7344    -0.84601    0.15294    38.0394    -1.87770
0.31935    38.4524    -1.14147    0.30667    38.9076    -1.18198
0.21722    40.0388    -1.52685    0.44777    42.2364    -0.80347
0.41371    43.6465    -0.88259    0.32136    44.4753    -1.13520
0.35369    44.7545    -1.03935    0.09101    44.9001    -2.39684
0.11982    46.6135    -2.12175    0.16831    47.4509    -1.78197
0.25125    54.6916    -1.38129
;
```

The following statements use the QUANTREG procedure to fit a simple linear model for the 50th and 90th percentiles of LnDensity:

```
ods graphics on;

proc quantreg data=trout alpha=0.1 ci=resampling;
   model LnDensity = WDRatio / quantile=0.5 0.9
                               CovB seed=1268;
   test WDRatio / wald lr;
run;
```

The MODEL statement specifies a simple linear regression model with LnDensity as the response variable Y and WDRatio as the covariate X. The QUANTILE= option requests that the regression quantile function $Q(\tau|X = x) = x'\beta(\tau)$ be estimated by solving

$$\hat{\beta}(\tau) = \arg\min_{\beta \in \mathbf{R}^2} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta)$$

where $\tau = (0.5, 0.9)$.

By default, the regression coefficients $\hat{\beta}(\tau)$ are estimated with the simplex algorithm, which is explained in the section "Simplex Algorithm" on page 6098. The ALPHA= option requests 90% confidence limits for the regression parameters, and the option CI=RESAMPLING specifies that the intervals are to be computed with the MCMB resampling method of He and Hu (2002). By specifying the CI=RESAMPLING option, the QUANTREG procedure also computes standard errors, t values, and p-values of regression parameters with the MCMB resampling method. The SEED= option specifies a seed for the resampling method. The COVB option requests covariance matrices for the estimated regression coefficients, and the TEST statement requests tests for the hypothesis that the slope parameter (the coefficient of WDRatio) is zero.

Figure 73.3 displays model information and summary statistics for the variables in the model. The summary statistics include the median and the standardized median absolute deviation (MAD), which are robust measures of univariate location and scale, respectively. See Huber (1981, p. 108) for more details about the standardized MAD.

Figure 73.3 Model Fitting Information and Summary Statistics

The QUANTREG Procedure

Model Information

Data Set	WORK.TROUT
Dependent Variable	LnDensity
Number of Independent Variables	1
Number of Observations	71
Optimization Algorithm	Simplex
Method for Confidence Limits	Resampling

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
WDRatio	22.0917	29.4083	35.9382	29.1752	9.9859	10.4970
LnDensity	-2.0511	-1.3813	-0.8669	-1.4973	0.7682	0.8214

Figure 73.4 and Figure 73.5 display the parameter estimates, standard errors, 95% confidence limits, t values, and p-values that are computed by the resampling method.

Figure 73.4 Parameter Estimates at QUANTILE=0.5

Parameter Estimates

Parameter	DF	Estimate	Standard Error	90% Confidence Limits		t Value	Pr > |t|
Intercept	1	-0.9811	0.3952	-1.6400	-0.3222	-2.48	0.0155
WDRatio	1	-0.0136	0.0123	-0.0341	0.0068	-1.11	0.2705

Figure 73.5 Parameter Estimates at QUANTILE=0.9

Parameter Estimates

Parameter	DF	Estimate	Standard Error	90% Confidence Limits		t Value	Pr > |t|
Intercept	1	0.0576	0.2606	-0.3769	0.4921	0.22	0.8257
WDRatio	1	-0.0215	0.0075	-0.0340	-0.0091	-2.88	0.0053

The 90th percentile of trout density can be predicted from the width-to-depth ratio as follows:

$$y_{90} = \exp(0.0576 - 0.0215x)$$

This is the upper dashed curve plotted in Figure 73.1. The lower dashed curve for the median can be obtained in a similar fashion.

The covariance matrices for the estimated parameters are shown in Figure 73.6. The resampling method used for the confidence intervals is used to compute these matrices.

Figure 73.6 Covariance Matrices of the Estimated Parameters

Estimated Covariance Matrix for Quantile = 0.5

	Intercept	WDRatio
Intercept	0.156191	−.004653
WDRatio	−.004653	0.000151

Estimated Covariance Matrix for Quantile = 0.9

	Intercept	WDRatio
Intercept	0.067914	−.001877
WDRatio	−.001877	0.000056

The tests requested with the TEST statement are shown in Figure 73.7. Both the Wald test and the likelihood ratio test indicate that the coefficient of width-to-depth ratio is significantly different from zero at the 90th percentile, but the deference is not significant at the median.

Figure 73.7 Tests of Significance

Test Results

Quantile	Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
0.5	Wald	1.2339	1	1.23	0.2666
0.5	Likelihood Ratio	1.1467	1	1.15	0.2842
0.9	Wald	8.3031	1	8.30	0.0040
0.9	Likelihood Ratio	9.0529	1	9.05	0.0026

In many quantile regression problems it is useful to examine how the estimated regression parameters for each covariate change as a function of τ in the interval $(0, 1)$. The following statements use the QUANTREG procedure to request the estimated quantile processes $\hat{\beta}(\tau)$ for the slope and intercept parameters:

```
proc quantreg data=trout alpha=0.1 ci=resampling;
   model LnDensity = WDRatio / quantile=process seed=1268
                               plot=quantplot;
run;

ods graphics off;
```

The QUANTILE=PROCESS option requests an estimate of the quantile process for each regression parameter. The options ALPHA=0.1 and CI=RESAMPLING specify that 90% confidence bands for the quantile processes are to be computed with the resampling method.

Figure 73.8 displays a portion of the objective function table for the entire quantile process. The objective function is evaluated at 77 values of τ in the interval $(0, 1)$. The table also provides predicted values of the conditional quantile function $Q(\tau)$ at the mean for WDRatio, which can be used to estimate the conditional density function.

Figure 73.8 Objective Function

Objective Function for Quantile Process

Label	Quantile	Objective Function	Predicted at Mean
t0	0.005634	0.7044	−3.2582
t1	0.020260	2.5331	−3.0331
t2	0.031348	3.7421	−2.9376
t3	0.046131	5.2538	−2.7013
.	.	.	.
.	.	.	.
.	.	.	.
t73	0.945705	4.1433	−0.4361
t74	0.966377	2.5858	−0.4287
t75	0.976060	1.8512	−0.4082
t76	0.994366	0.4356	−0.4082

Figure 73.9 displays a portion of the table of the quantile processes for the estimated parameters and confidence limits.

Figure 73.9 Objective Function

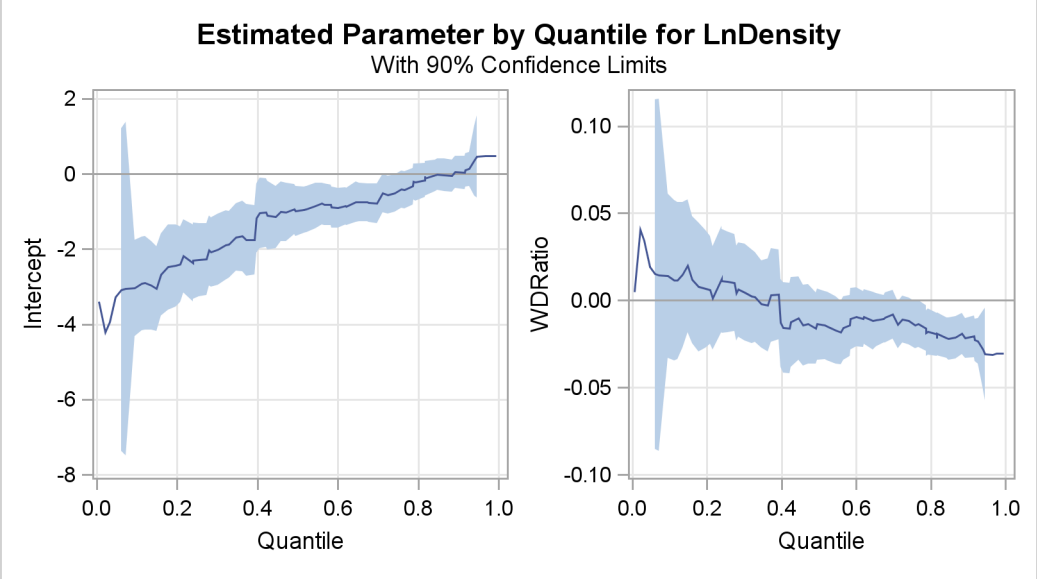
Parameter Estimates for Quantile Process

Label	Quantile	Intercept	WDRatio
.	.	.	.
.	.	.	.
.	.	.	.
t57	0.765705	−0.42205	−0.01335
lower90	0.765705	−0.91952	−0.02682
upper90	0.765705	0.07541	0.00012
t58	0.786206	−0.32688	−0.01592
lower90	0.786206	−0.80883	−0.02895
upper90	0.786206	0.15507	−0.00289
.	.	.	.
.	.	.	.
.	.	.	.

The PLOT=QUANTPLOT option in the MODEL statement, together with the ODS GRAPHICS statement, requests a plot of the estimated quantile processes. The left side of Figure 73.10 displays the process for the intercept, and the right side displays the process for the coefficient of WDRatio.

The process plot for WDRatio shows that the slope parameter changes from positive to negative as the quantile increases, and it changes sign with a sharp drop at the 40th percentile. The 90% confidence bands show that the relationship between LnDensity and WDRatio (expressed by the slope) is not significant below the 78th percentile. This situation can also be seen in Figure 73.9, which shows that 0 falls between the lower and upper confidence limits of the slope parameter for quantiles below 0.78. Since the confidence intervals for the extreme quantiles are not stable due to insufficient data, the confidence band is not displayed outside the interval (0.05, 0.95).

Figure 73.10 Quantile Processes for Intercept and Slope



Growth Charts for Body Mass Index

Body mass index (BMI) is defined as the ratio of weight (kg) to squared height (m^2) and is a widely used measure for categorizing individuals as 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.

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. More details can be found in Chen (2005).

The data set used in this example is a subset of the original data set of Chen (2005). It contains the two variables BMI and AGE with 3264 observations.

```
data bmimen0;
   input bmi age @@;
datalines;
18.6  2.0 17.1  2.0 19.0  2.0 16.8  2.0 19.0  2.1  15.5   2.1
16.7  2.1 16.1  2.1 18.0  2.1 17.8  2.1 18.3  2.1  16.9   2.1
15.9  2.1 20.6  2.1 16.7  2.1 15.4  2.1 15.9  2.1  17.7   2.1
15.7  2.1 16.8  2.1 15.6  2.1 18.1  2.1 15.7  2.1  17.2   2.1
14.5  2.2 17.2  2.2 16.3  2.2 15.4  2.2 16.0  2.2  15.8   2.2

   ... more lines ...

29.0 80.0 24.1 80.0 26.6 80.0 24.2 80.0 22.7 80.0  28.4  80.0
26.3 80.0 25.6 80.0 24.8 80.0 28.6 80.0 25.7 80.0  25.8  80.0
22.5 80.0 25.1 80.0 27.0 80.0 27.9 80.0 28.5 80.0  21.7  80.0
33.5 80.0 26.1 80.0 28.4 80.0 22.7 80.0 28.0 80.0  42.7  80.0
;
```

The logarithm of BMI is used as the response (although this does not improve 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.

```
data bmimen;
   set bmimen0;
   sqrtage = sqrt(age);
   inveage = 1/age;
   logbmi  = log(bmi);
run;
```

The following statements invoke the QUANTREG procedure:

```
proc quantreg data=bmimen algorithm=interior(tolerance=1e-5) ci=resampling;
   model logbmi = inveage sqrtage age sqrtage*age
                  age*age age*age*age
                  / diagnostics cutoff=4.5 quantile=.5 seed=1268;
   id age 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 $\hat{\beta}(\frac{1}{2})$ by using the interior point algorithm as requested with the ALGORITHM= option. See the section "Interior Point Algorithm" on page 6099 for details about this algorithm.

Figure 73.11 displays the estimated parameters, standard errors, 95% confidence intervals, t values, and p-values that are computed by the resampling method as requested by the CI= option. All of the parameters are considered significant since the p-values are smaller than 0.001.

Figure 73.11 Parameter Estimates with Median Regression: Men

The QUANTREG Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	7.8909	0.8168	6.2895	9.4924	9.66	<.0001
inveage	1	−1.8354	0.4350	−2.6884	−0.9824	−4.22	<.0001
sqrtage	1	−5.1247	0.7135	−6.5237	−3.7257	−7.18	<.0001
age	1	1.9759	0.2537	1.4785	2.4733	7.79	<.0001
sqrtage*age	1	−0.3347	0.0424	−0.4179	−0.2515	−7.89	<.0001
age*age	1	0.0227	0.0029	0.0170	0.0284	7.77	<.0001
age*age*age	1	−0.0000	0.0000	−0.0001	−0.0000	−7.40	<.0001

The TEST statement requests Wald and likelihood ratio tests for the significance of the cubic term in AGE. The test results, shown in Figure 73.12, indicate that this term is significant. Higher-order terms are not significant.

Figure 73.12 Test of Significance for Cubic Term

Test test_age_cubic Results

Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
Wald	54.7417	1	54.74	<.0001
Likelihood Ratio	56.9473	1	56.95	<.0001

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

With CUTOFF=4.5, 14 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 73.2, the same model used for median regression is used for other quantiles. Note that the QUANTREG procedure can compute fitted values for multiple quantiles.

Figure 73.13 Diagnostics with Median Regression

Diagnostics

Obs	age	bmi	Standardized Residual	Outlier
1337	8.900000	36.500000	5.3575	*
1376	9.200000	39.600000	5.8723	*
1428	9.400000	36.900000	5.3036	*
1505	9.900000	35.500000	4.8862	*
1764	14.900000	46.800000	5.6403	*
1838	16.200000	50.400000	5.9138	*
1845	16.300000	42.600000	4.6683	*
1870	16.700000	42.600000	4.5930	*
1957	18.100000	49.900000	5.5053	*
2002	18.700000	52.700000	5.8106	*
2016	18.900000	48.400000	5.1603	*
2264	32.000000	55.600000	5.3085	*
2291	35.000000	60.900000	5.9406	*
2732	66.000000	14.900000	-4.7849	*

The following statements request fitted values for 10 quantile levels ranging from 0.03 to 0.97:

```
proc quantreg data=bmimen ci=none algorithm=interior(tolerance=1e-5);
model logbmi = inveage sqrtage age sqrtage*age
                age*age age*age*age
                / quantile=0.03,0.05,0.1,0.25,0.5,0.75,
                           0.85,0.90,0.95,0.97;
output out=outp pred=p/columnwise;
run;

data outbmi;
   set outp;
   pbmi = exp(p);
run;

proc sgplot data=outbmi;
   title 'BMI Percentiles for Men: 2-80 Years Old';
   yaxis label='BMI (kg/m**2)' min=10 max=45
         values=(10 15 20 25 30 35 40 45);
   xaxis label='Age (Years)' min=2 max=80
         values=(2 10 20 30 40 50 60 70 80);
   scatter x=age y=bmi /markerattrs=(size=1);
   series  x=age y=pbmi/group=QUANTILE;
run;
```

The fitted values are stored in the OUTPUT data set outp. The COLUMNWISE option arranges these fitted values for all quantiles in the single variable p by groups of the quantiles. After the exponential transformation, the fitted BMI values together with the original BMI values are plotted against AGE to create the display shown in Figure 73.2.

The fitted quantile curves reveal important information. 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 effective population weight control should start in childhood.

Compared to the 97th percentile in reference growth charts published by CDC in 2000 (Kuczmarski et al. 2002), the 97th percentile for 10-year-old boys in Figure 73.2 is 6.4 BMI units higher (an increase of 27%). This can be interpreted as a warning of overweight or obesity. See Chen (2005) for a detailed analysis.

Syntax: QUANTREG Procedure

PROC QUANTREG < *options* > ;
BY *variables* ;
CLASS *variables* ;
EFFECT *name* = *effect-type* (*variables* < / *options* >) ;
ID *variables* ;
MODEL *response* = *independents* < / *options* > ;
OUTPUT < *OUT*= *SAS-data-set* > < *options* > ;
PERFORMANCE < *options* > ;
TEST *effects* < / *options* > ;
WEIGHT *variable* ;

The PROC QUANTREG statement invokes the procedure. The CLASS statement specifies which explanatory variables are treated as categorical. The ID statement names variables to identify observations in the outlier diagnostics tables. The MODEL statement is required and specifies the variables used in the regression. Main effects and interaction terms can be specified in the MODEL statement, as in the GLM procedure (Chapter 39, "The GLM Procedure.") The OUTPUT statement creates an output data set containing predicted values, residuals, and estimated standard errors. The PERFORMANCE statement tunes the performance of PROC QUANTREG by using single or multiple processors available in the hardware. The TEST statement requests linear tests for the model parameters. The WEIGHT statement identifies a variable in the input data set whose values are used to weight the observations. In one invocation of PROC QUANTREG, multiple OUTPUT and TEST statements are allowed.

PROC QUANTREG Statement

PROC QUANTREG < *options* > ;

The PROC QUANTREG statement invokes the procedure. You can specify the following options in the PROC QUANTREG statement.

ALGORITHM=*algorithm* **< (** *suboptions* **) >**
specifies an algorthm to estimate the regression parameters. Three algorithms are available: simplex (SIMPLEX), interior point (INTERIOR), and smoothing (SMOOTH).

The default algorithm depends on the number of the observations (n) and the number of the covariates (p) in the model estimation. See Table 73.1 for the relevant defaults.

Table 73.1 The Default Estimation Algorithm

	$p \leq 100$	$p > 100$
$n \leq 5000$	Simplex	Smoothing
$n > 5000$	Interior point	Smoothing

Table 73.2 summarizes the options available for each of these methods.

Table 73.2 Options for Estimation Algorithms

ALGORITHM= Value	Algorithm	Suboptions
SIMPLEX	Simplex	MAXSTATIONARY=
INTERIOR	Interior point	KAPPA= MAXIT= TOLERANCE=
SMOOTH	Smoothing	RRATIO=

With ALGORITHM=SIMPLEX you can specify the following *suboption*:

- MAXSTATIONARY=*m* specifies that if the objective funtion has not improved for *m* consecutive iterations, the algorithm termintates. By default, *m*=1000.

With ALGORITHM=INTERIOR you can specify the following *suboptions*:

- KAPPA=*value* specifies the step length parameter for the interior point algorithm. This parameter should be between 0 and 1. The larger the parameter, the faster the algorithm. However, numeric instability can occur as the parameter approaches 1. By default, KAPPA=0.99995. See the section "Interior Point Algorithm" on page 6099 for details.
- MAXIT=*m* sets the maximum number of iterations for the interior point algorithm. By default, *m*=1000.
- TOLERANCE=*value* specifies the tolerance for the convergence criterion of the interior point algorithm. The default *value* is 1E−8. The QUANTREG procedure uses the duality gap as the convergence criterion. See the section "Interior Point Algorithm" on page 6099 for details.

With the interior point algorithm, you can use the PERFORMANCE statement to enable parallel computing when multiple processors are available in the hardware.

With ALGORITHM=SMOOTH you can specify the following *suboption*:

- RRATIO=*value* specifies the reduction ratio for the smoothing algorithm. This ratio is used for reducing the threshold of the smoothing algorithm. The *value* should be between 0 and 1. In theory, the smaller the reduction ration, the faster the smoothing algorithm. However, the optimal ratio is quite data dependent in practice. See the section "Smoothing Algorithm" on page 6102 for details.

ALPHA=*value*

sets the confidence level for the confidence intervals for regression parameters. The *value* must be between 0 and 1. The default is ALPHA=0.05, corresponding to a 0.95 confidence interval.

CI=NONE | RANK | SPARSITY< (BF | HS) >< /IID > | RESAMPLING< (NREP=*n*) >

specifies a method to compute confidence intervals for regression parameters. When you specify CI=SPARSITY or CI=RESAMPLING, the QUANTREG procedure also computes standard errors, t values, and p-values for regression parameters.

The following table summarizes these methods.

Table 73.3 Options for Confidence Intervals

Value of CI=	Method	Additional Options
NONE	No confidence intervals computed	
RANK	By inverting rank-score tests	
SPARSITY	By estimating sparsity function	HS BF IID
RESAMPLING	By resampling	NREP

By default, when there are fewer than 5,000 observations, fewer than 20 variables in the data set, and the algorithm is simplex, the QUANTREG procedure computes confidence intervals by using the inverting rank-score test method; otherwise, the resampling method is used.

By default, confidence intervals are not computed for the quantile process, which is estimated when you specify the QUANTILE=PROCESS option in the MODEL statement. Confidence intervals for the quantile process are computed with the sparsity or resampling methods when you specify CI=SPARSITY or CI=RESAMPLING, respectively. The rank method for confidence intervals is not available with quantile processes because it is computationally prohibitive.

With the SPARSITY option, there are two suboptions for estimating the sparsity function. If you specify the IID suboption, the sparsity function is estimated by assuming that the errors in the linear model are independent and identically distributed (iid). By default, the sparsity function is estimated by assuming that the conditional quantile function is locally linear. See the section "Sparsity" on page 6106 for details. With both methods two bandwidth selection methods are available. You can specify the Bofinger method with the BF suboption or the Hall-Sheather method with the HS suboption. By default, the Hall-Sheather method is used.

With the RESAMPLING option, you can specify the number of repeats with the NREP=*n* suboption. By default, NREP=200. The value of n must be greater than 50.

DATA=*SAS-data-set*

specifies the input SAS data set used by the QUANTREG procedure. By default, the most recently created SAS data set is used.

INEST=*SAS-data-set*

specifies an input SAS data set that contains initial estimates for all the parameters in the model. The interior point algorithm and the smoothing algorithm use these estimates as a start. See the section "INEST= Data Set" on page 6112 for a detailed description of the contents of the INEST= data set.

NAMELEN=*n*

specifies the length of effect names in tables and output data sets to be *n* characters, where *n* is a value between 20 and 200. The default length is 20 characters.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the order in which to sort the levels of the classification variables (which are specified in the CLASS statement). This option applies to the levels for all classification variables, except when you use the (default) ORDER=FORMATTED option with numeric classification variables that have no explicit format. With this option, the levels of such variables are ordered by their internal value.

The ORDER= option can take the following values:

Value of ORDER=	Levels Sorted By
DATA	Order of appearance in the input data set
FORMATTED	External formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
FREQ	Descending frequency count; levels with the most observations come first in the order
INTERNAL	Unformatted value

By default, ORDER=FORMATTED. For FORMATTED and INTERNAL, the sort order is machine-dependent. For more information about sorting order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

OUTEST=*SAS-data-set*

specifies an output SAS data set containing the parameter estimates for all quantiles. See the section "OUTEST= Data Set" on page 6112 for a detailed description of the contents of the OUTEST= data set.

PLOT | PLOTS< (*global-plot-options***) > < =***plot-request* **>**

PLOT | PLOTS< (*global-plot-options***) > < =(***plot-request* **< . . .** *plot-request* **>) >**

specifies options that control details of the plots. These plots fall into two categories, diagnostic plots and fit plots. If you do not specify the PLOTS= option, PROC QUANTREG produces the quantile fit plot by default when a single continuous variable is specified in the model. You can use the PLOTS= option in the PROC statement to request various diagnostic plots. In addition to these two categories of plots, you can use the PLOT= option in the MODEL statement to request the quantile process plot for any effects specified in the model.

To request any plots you must specify the ODS GRAPHICS statement. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

The *global-plot-options* apply to all plots generated by the QUANTREG procedure. The following global plot option is available:

ONLY
suppresses the default quantile fit plot. Only plots specifically requested are displayed.

You can specify more than one plot request within the parentheses after PLOTS=. For a single plot request, you can omit the parentheses. The following plot requests are available.

ALL
creates all appropriate plots.

DDPLOT<(LABEL=ALL | LEVERAGE | NONE | OUTLIER)>
creates a plot of robust distance against Mahalanobis distance. See the section "Leverage Point and Outlier Detection" on page 6111 for details about robust distance. The LABEL= option specifies how the points on this plot are to be labeled, as summarized by the following table.

Table 73.4 Options for Label

Value of LABEL=	Label Method
ALL	Label all points
LEVERAGE	Label leverage points
NONE	No labels
OUTLIERS	Label outliers

By default, the QUANTREG procedure labels both outliers and leverage points.

If you specify ID variables in the ID statement, the values of the first ID variable are used as labels; otherwise, observation numbers are used as labels.

FITPLOT<(NOLIMITS | SHOWLIMITS | NODATA)>
creates a plot of fitted conditional quantiles against the single continuous variable that is specified in the model. This plot is produced only when the response is modeled as a function of a single continuous variable. Multiple lines or curves are drawn on this plot if you specify several quantiles with the QUANTILE= option in the MODEL statement. By default, confidence limits are added to the plot when a single quantile is requested, and the confidence limits are not shown on the plot when multiple quantiles are requested. The NOLIMITS option suppresses these limits. The SHOWLIMITS option adds these limits when multiple quantiles are requested. The NODATA option suppresses the observed data, which are superimposed on the plot by default.

HISTOGRAM
creates a histogram for the standardized residuals based on the quantile regression estimates. The histogram is superimposed with a normal density curve and a kernel density curve.

NONE
suppresses all plots.

QQPLOT
creates the normal quantile-quantile plot for the standardized residuals based on the quantile regression estimates.

RDPLOT<(LABEL=ALL | LEVERAGE | NONE | OUTLIER)>

creates the plot of standardized residual against robust distance. See the section "Leverage Point and Outlier Detection" on page 6111 for details about robust distance. The LABEL= option specifies a label method for points on this plot. These label methods are described in Table 73.4.

By default, the QUANTREG procedure labels both outliers and leverage points.

If you specify ID variables in the ID statement, the values of the first ID variable are used as labels; otherwise, observation numbers are used as labels.

PP

requests preprocessing to speed up the interior point algorithm or the smoothing algorithm. The preprocessing uses a subsampling algorithm to reduce the original problem to a smaller one iteratively. It assumes that the data set is evenly distributed. Preprocessing should be used only for very large data sets, such as data sets with more than 100,000 observations. See Portnoy and Koenker (1997) for details.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC QUANTREG to obtain separate analyses on observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the QUANTREG procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLASS Statement

CLASS *variables* < / **TRUNCATE** > **;**

The CLASS statement names the classification variables to be used in the model. Typical classification variables are Treatment, Sex, Race, Group, and Replication. If you use the CLASS statement, it must appear before the MODEL statement.

Classification variables can be either character or numeric. By default, class levels are determined from the entire set of formatted values of the CLASS variables.

NOTE: Prior to SAS 9, class levels were determined by using no more than the first 16 characters of the formatted values. To revert to this previous behavior, you can use the TRUNCATE option in the CLASS statement.

In any case, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*. You can adjust the order of CLASS variable levels with the ORDER= option in the PROC QUANTREG statement. You can specify the following option in the CLASS statement after a slash (/):

TRUNCATE
specifies that class levels should be determined by using only up to the first 16 characters of the formatted values of CLASS variables. When formatted values are longer than 16 characters, you can use this option to revert to the levels as determined in releases prior to SAS 9.

EFFECT Statement (Experimental)

EFFECT *name = spline(variables < / options >)* **;**

The experimental EFFECT statement names a constructed effect that you can use to specify terms in the MODEL statement. Each constructed effect corresponds to a collection of columns that are referred to using the name you supply. You can specify multiple EFFECT statements, and all EFFECT statements must precede the MODEL statement.

After the keyword EFFECT, the name of the effect is specified. This name must appear in only one EFFECT statement and must not be the name of a variable in the input data set. After an equal sign, the *effect-type* is specified followed by the list of variables used in defining the effect within parentheses. In SAS 9.2, the QUANTREG procedure supports only spline effects. You can also specify options pertaining to the effect definition after a slash following the variable list. The SPLIT option is always turned on for the QUANTREG procedure. For more details about the syntax with spline effects, see the section "EFFECT Statement (Experimental)" on page 418 of Chapter 19, "Shared Concepts and Topics."

The QUANTREG procedure supports the regression spline effect. A spline effect expands variables into spline bases whose form depends on the options that you specify. Design matrix columns are generated separately for each of the numeric variables specified, and these columns are collectively

referred to by the name that you specify. By default the spline basis generated for each variable is a cubic B-spline basis with three equally spaced knots positioned between the minimum and maximum values of that variable. You can find more details about regression splines and spline bases in the section "EFFECT Statement (Experimental)" on page 418 of Chapter 19, "Shared Concepts and Topics."

ID Statement

ID *variables* **;**

When the diagnostics table is requested with the DIAGNOSTICS option in the MODEL statement, the variables listed in the ID statement are displayed in addition to the observation number. These values are useful for identifying observations. If the ID statement is omitted, only the observation number is displayed.

MODEL Statement

<label:> MODEL *response* = <*effects*> </ *options*> **;**

Main effects and interaction terms can be specified in the MODEL statement, as in the GLM procedure (Chapter 39, "The GLM Procedure.") Classification variables in the MODEL statement must be specified in the CLASS statement.

The optional *label*, which must be a valid SAS name, is used to label output from the matching MODEL statement.

Options

You can specify the following options for the model fit.

CORRB

produces the estimated correlation matrix of the parameter estimates. When the resampling method is used to compute the confidence intervals, the QUANTREG procedure computes the bootstrap correlation. When the sparsity method is used to compute the confidence intervals, the procedure computes the asymptotic correlation based on an estimator of the sparsity function. The rank method for confidence intervals does not provide a correlation estimate.

COVB

produces the estimated covariance matrix of the parameter estimates. When the resampling method is used to compute the confidence intervals, the QUANTREG procedure computes the bootstrap covariance. When the sparsity method is used to compute the confidence intervals, the procedure computes the asymptotic covariance based on an estimator of the sparsity function. The rank method for confidence intervals does not provide a covariance estimate.

CUTOFF=*value*
specifies the multiplier of the cutoff value for outlier detection. The default *value* is 3.

DIAGNOSTICS<(ALL)>
requests the outlier diagnostics. By default, only observations identified as outliers or leverage points are displayed. To request that all observations be displayed, specify the ALL option.

ITPRINT
displays the iteration history of the interior point algorithm or the smoothing algorithm.

LEVERAGE<(CUTOFF=*value* | **CUTOFFALPHA=***value* | **H=***n***)>**
requests an analysis of leverage points for the continuous covariates. The results are added to the diagnostics table, which you can request with the DIAGNOSTICS option in the MODEL statement. You can specify the cutoff value for leverage-point detection with the CUTOFF= option. The default cutoff value is $\sqrt{\chi^2_{p;1-\alpha}}$, where α can be specified with the CUTOFFALPHA= option. By default, $\alpha = 0.025$. You can use the H= option to specify the number of points to be minimized for the MCD algorithm used for the leverage-point analysis. By default, H=$[(3n + p + 1)/4]$, where n is the number of observations and p is the number of independent variables. The LEVERAGE option is ignored if the model includes classification variables as covariates.

NODIAG
suppresses the computation for outlier diagnostics. If you specify the NODIAG option, the diagnostics summary table will not be provided.

NOINT
specifies no intercept regression.

NOSUMMARY
suppresses the computation for summary statistics. If you specify the NOSUMMARY option, the summary statistics table will not be provided.

PLOT=*plot-option*

PLOTS=(*plot-options***)**
You can use the PLOTS= option in the MODEL statement together with the ODS GRAPHICS statement to request the quantile process plot in addition to all plots available with the PLOT= option in the PROC statement. The plot options in the PROC statement overwrite the plot options in the MODEL statement if you specify the same options in both statements.

You can specify the following plot option only in the MODEL statement:

QUANTPLOT<(EFFECTS)</ <NOLIMITS> <EXTENDCI> <UNPACK> <OLS> > >
plots the regression quantile process. The estimated coefficient of each specified covariate effect is plotted as a function of the quantile. If you do not specify a covariate effect, quantile processes are plotted for all covariate effects in the MODEL statement. You can use the NOLIMITS option to suppress confidence bands for the quantile processes. By default, confidence bands are plotted, and process plots are displayed in panels, each of which can hold up to four plots. By default, the confidence limits are plotted for quantiles in the range between 0.05 and 0.95. You can use the EXTENDCI option

to plot the confidence limits even for quantiles outside this range. You can use the UNPACK option to create individual process plots. For an individual process plot, you can superimpose the ordinary least squares estimate and its confidence limits by specifying the OLS option. The confidence level of the ordinary least squares estimate is specified with the ALPHA= option in the PROC statement.

Again, to request these plots you must specify the ODS GRAPHICS statement. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

QUANTILE=*number-list* | **PROCESS**

specifies the quantile levels for the quantile regression. You can specify any number of quantile levels in (0, 1). You can also compute the entire quantile process by specifying the PROCESS option. Only the simplex algorithm is available for computing the quantile process.

If you do not specify the QUANTILE= option, the QUANTREG procedure fits a median regression, which corresponds to QUANTILE=0.5.

SCALE=*number*

specifies the scale value used to compute the standardized residuals. By default, the scale is computed as the corrected median of absolute residuals. See the section "Leverage Point and Outlier Detection" on page 6111 for details.

SEED=*number*

specifies the seed for the random number generator used to compute the MCMB confidence intervals. This seed is also used to randomly select the subgroups for preprocessing when you specify the PP option in the PROC statement. If you do not specify a seed, or if you specify a value less than or equal to zero, the seed is generated from reading the time of day from the computer clock.

By default or if you specify zero, the QUANTREG procedure generates a seed between one and one billion.

SINGULAR=*value*

sets the tolerance for testing singularity of the information matrix and the crossproducts matrix for the initial least squares estimates. Roughly, the test requires that a pivot be at least this value times the original diagonal value. By default, SINGULAR=1E−12.

OUTPUT Statement

OUTPUT < *OUT=SAS-data-set* > *keyword=name* < *. . . keyword=name* > < / *COLUMNWISE* > **;**

The OUTPUT statement creates a SAS data set containing statistics calculated after fitting models for all specified quantiles with the QUANTILE= option in the MODEL statement. At least one specification of the form *keyword=name* is required.

All variables in the original data set are included in the new data set, along with the variables created as options to the OUTPUT statement. These new variables contain fitted values and estimated

quantiles. If you want to create a permanent SAS data set, you must specify a two-level name (refer to *SAS Language Reference: Concepts* for more information about permanent SAS data sets).

If you specify multiple quantiles in the MODEL statement, the COLUMNWISE option arranges the created OUTPUT data set in column-wise form. This arrangement repeats the input data for each quantile. By default, the OUTPUT data set is created in row-wise form. For each appropriate keyword specified in the OUTPUT statement, one variable for each specified quantile is generated. These variables appear in the sorted order of the specified quantiles.

The following specifications can appear in the OUTPUT statement:

OUT=*SAS-data-set* specifies the new data set. By default, the procedure uses the DATA*n* convention to name the new data set.

keyword=name specifies the statistics to include in the output data set and gives names to the new variables. Specify a keyword for each desired statistic (see the following list of keywords), an equal sign, and the variable to contain the statistic.

The keywords allowed and the statistics they represent are as follows:

LEVERAGE specifies a variable to indicate leverage points. To include this variable in the OUTPUT data set, you must specify the LEVERAGE option in the MODEL statement. See the section "Leverage Point and Outlier Detection" on page 6111 for how to define LEVERAGE.

MAHADIST | MD specifies a variable to contain the Mahalanobis distance. To include this variable in the OUTPUT data set, you must specify the LEVERAGE option in the MODEL statement.

OUTLIER specifies a variable to indicate outliers. See the section "Leverage Point and Outlier Detection" on page 6111 for how to define OUTLIER.

PREDICTED | P specifies a variable to contain the estimated response.

QUANTILE | Q specifies a variable to contain the quantile for which the quantile regression is fitted. If you specify the COLUMNWISE option, this variable is created by default. If multiple quantiles are specified in the MODEL statement and the COLUMNWISE option is not specified, this variable is not created.

RESIDUAL | RES specifies a variable to contain the residuals (unstandardized)

$$y_i - \mathbf{x}_i^T \hat{\beta}$$

ROBDIST | RD specifies a variable to contain the robust MCD distance. To include this variable in the OUTPUT data set, you must specify the LEVERAGE option in the MODEL statement.

SPLINE | SP specifies a variable to contain the estimated spline effect, which includes all spline effects in the model and their interactions.

SRESIDUAL | SR specifies a variable to contain the standardized residuals

$$\frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}}$$

See the section "Leverage Point and Outlier Detection" on page 6111 for how to compute σ.

STDP
: specifies a variable to contain the estimates of the standard errors of the estimated response.

PERFORMANCE Statement

The PERFORMANCE statement is used to change default options that affect the performance of PROC QUANTREG and to request tables that show the performance options in effect and timing details.

PERFORMANCE < *options* > **;**

The following options are available:

CPUCOUNT=*1-1024*

CPUCOUNT=*ACTUAL*
: specifies the number of processors to use in the computation of the interior point algorithm. CPUCOUNT=ACTUAL sets CPUCOUNT to be the number of physical processors available. Note that this can be less than the physical number of CPUs if the SAS process has been restricted by system administration tools. Setting CPUCOUNT= to a number greater than the actual number of available CPUs might result in reduced performance. This option overrides the SAS system option CPUCOUNT=. If CPUCOUNT=1, then NOTHREADS is in effect, and PROC QUANTREG uses singly threaded code.

DETAILS
: requests the "PerfSettings" table that shows the performance settings in effect and the "Timing" table that provides a broad timing breakdown of the PROC QUANTREG step.

THREADS
: enables multithreaded computation for the interior point algorithm. This option overrides the SAS system option THREADS | NOTHREADS. If you do not specify the ALGORITHM=INTERIOR option, then PROC QUANTREG ignores this option and uses singly threaded code.

NOTHREADS
: disables multithreaded computation for the interior point algorithm. This option overrides the SAS system option THREADS | NOTHREADS.

TEST Statement

< *label:* > TEST *effects* < / *options* > **;**

The TEST statement provides a means of obtaining a test for the canonical linear hypothesis concerning the parameters of the tested effects

$$\beta_j = 0, \quad j = i_1, \ldots, i_q$$

where q is the total number of parameters of the tested effects. The tested *effects* can be any set of effects in the MODEL statement. The Wald, rank, and likelihood ratio tests for the regression parameter estimates are available. See the section "Linear Test" on page 6109 for more information about these tests. You can specify the following options in the statement after a slash (/):

WALD

requests Wald tests.

LR

requests likelihood ratio tests.

RANKSCORE < (NORMAL | WILCOXON | SIGN) >

requests rank score tests. For the rank score tests, the available score functions include normal scores, Wilcoxon scores, and sign scores, which are asymptotically optimal for the Gaussian, logistic, and Laplace location shift models, respectively. By default, the NORMAL option is used. For additional information about the score functions, see Koenker (2005). Currently, the rank tests are implemented for iid error models.

You can submit multiple TEST statements. The optional *label*, which must be a valid SAS name, is used to identify output from the corresponding TEST statement.

WEIGHT Statement

WEIGHT *variable* **;**

The WEIGHT statement specifies a weight variable in the input data set.

To request weighted quantile regression, place the weights in a variable and specify the name in the WEIGHT statement. The values of the WEIGHT variable can be nonintegral and are not truncated. Observations with nonpositive or missing values for the weight variable do not contribute to the fit of the model. See the section "Details: QUANTREG Procedure" on page 6097 for more information about weighted quantile regression.

Details: QUANTREG Procedure

Quantile Regression as an Optimization Problem

The model for linear quantile regression is

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

where $y = (y_1, \ldots, y_n)'$ is the $(n \times 1)$ vector of responses, $A' = (x_1, \ldots, x_n)'$ is the $(n \times p)$ regressor matrix, $\beta = (\beta_1, \ldots, \beta_p)'$ is the $(p \times 1)$ vector of unknown parameters, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is the $(n \times 1)$ vector of unknown errors.

L_1 regression, also known as median regression, is a natural extension of the sample median when the response is conditioned on the covariates. In L_1 regression, the least absolute residuals estimate $\hat{\beta}_{LAR}$, referred to as the L_1-*norm estimate*, is obtained as the solution of the minimization problem

$$\min_{\beta \in \mathbf{R}^p} \sum_{i=1}^{n} |y_i - x_i'\beta|$$

More generally, for quantile regression Koenker and Bassett (1978) defined the τth *regression quantile*, $0 < \tau < 1$, as any solution to the minimization problem

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

The solution is denoted as $\hat{\beta}(\tau)$, and the L_1-norm estimate corresponds to $\hat{\beta}(1/2)$. The τth regression quantile is an extension of the τth sample quantile $\hat{\xi}(\tau)$, which can be formulated as the solution of

$$\min_{\xi \in \mathbf{R}} \left[\sum_{i \in \{i: y_i \geq \xi\}} \tau |y_i - \xi| + \sum_{i \in \{i: y_i < \xi\}} (1-\tau)|y_i - \xi| \right]$$

If you specify weights $w_i, i = 1, \ldots, n$, with the WEIGHT statement, weighted quantile regression is carried out by solving

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

Weighted regression quantiles β_w can be used for L-estimation; refer to Koenker and Zhao (1994).

Optimization Algorithms

The optimization problem for median regression has been formulated and solved as a linear programming (LP) problem since the 1950s. Variations of the simplex algorithm, especially the method of Barrodale and Roberts (1973), have been widely used to solve this problem. The simplex algorithm is computationally demanding in large statistical applications, and in theory the number of iterations can increase exponentially with the sample size. This algorithm is often useful with data containing no more 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.

In addition to the interior point method, various heuristic approaches are available 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-Raphson algorithm can be used iteratively to obtain a solution after a finite number of iterations. The smoothing algorithm extends naturally to general quantile regression.

The QUANTREG procedure implements the simplex, interior point, and smoothing algorithms. The remainder of this section describes these algorithms in more detail.

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 reformulating it as the constrained minimization problem

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

where e denotes an $(n \times 1)$ vector of ones.

Let $B = [A' \; -A' \; I \; -I]$, $\theta = (\phi' \; \varphi' \; \mu' \; \nu')'$, and $d = (\mathbf{0}' \; \mathbf{0}' \; e' \; e')'$, where $\mathbf{0}' = (0 \; 0 \; \ldots \; 0)_p$. 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

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

which can be simplified as

$$\max_z y'z; \;\; \text{subject to } Az = 0, z \in [-1, 1]^n$$

By setting $\eta = \frac{1}{2}z + \frac{1}{2}e, b = \frac{1}{2}Ae$, the problem becomes

$$\max_\eta y'\eta; \;\; \text{subject to } A\eta = b, \eta \in [0, 1]^n$$

For quantile regression, the minimization problem is $\min_\beta \sum \rho_\tau(y_i - x_i'\beta)$, and a similar set of steps leads to the dual formulation

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

The QUANTREG procedure solves this LP problem by 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 picks the columns in A' or $-A'$ as pivotal columns. The second stage interchanges the columns in I or $-I$ as basis or nonbasis columns, respectively. 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.

Although this special version of the simplex algorithm was introduced for median regression, it extends naturally to quantile regression for any given quantile and even to the entire quantile process (Koenker and d'Orey 1994). It greatly reduces the computing time required by the general simplex algorithm, and it is suitable for data sets with fewer 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 predictor-corrector algorithm implemented by Lustig, Marsden, and Shanno (1992). The text by Roos, Terlaky, and Vial (1997) provides more information about this particular algorithm. The following brief introduction of this algorithm uses the notation in the first reference.

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

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

To simplify the computation, this is treated as the *primal* problem. The problem has n variables. The index i denotes a variable number, and k denotes an iteration number. If k is 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 algorithm solves the system of equations to satisfy the Karush-Kuhn-Tucker (KKT) conditions for optimality:

$$Az = b$$
$$z + v = u$$
$$A't + s - w = c$$
$$ZSe = 0$$
$$VWe = 0$$
$$z, s, v, w \geq 0$$

where $W = diag(w)$ (that is, $W_{i,j} = w_i$ if $i = j$, $W_{i,j} = 0$ otherwise)
$V = diag(v), Z = diag(z), S = diag(s)$

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, $c'z_{opt} = b't_{opt} - u'w_{opt}$, as

$$0 = v'_{opt}w_{opt} = (u - z_{opt})'w_{opt} = u'w_{opt} - z'_{opt}w_{opt}$$

$$0 = z'_{opt}s_{opt} = s'_{opt}z_{opt} = (c - A't_{opt} + w_{opt})'z_{opt} =$$
$$c'z_{opt} - t'_{opt}(Az_{opt}) + w'_{opt}z_{opt} = c'z_{opt} - b't_{opt} + u'w_{opt}$$

Therefore

$$0 = c'z_{opt} - b't_{opt} + u'w_{opt}$$

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.

Before the optimum is reached, it is possible for a solution (z, t, s, v, w) to violate the KKT conditions in one of several ways:

- Primal bound constraints can be broken, $\delta_b = u - z - v \neq 0$.
- Primal constraints can be broken, $\delta_c = b - Az \neq 0$.
- Dual constraints can be broken, $\delta_d = c - A't - s + w \neq 0$.
- Complementarity conditions are unsatisfied, $z's \neq 0$ and $v'w \neq 0$.

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:

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

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

The QUANTREG procedure implements a predictor-corrector variant of the primal-dual interior point algorithm. First, Newton's method is used to find a direction $(\Delta z_{aff}^k, \Delta t_{aff}^k, \Delta s_{aff}^k, \Delta v_{aff}^k, \Delta w_{aff}^k)$ in which to move. This is known as the *affine* step.

In iteration k, the *affine* step system that must be solved is

$$\begin{aligned}
&\Delta z_{aff} + \Delta v_{aff} = \delta_b \\
&A\Delta z_{aff} = \delta_c \\
&A'\Delta t_{aff} + \Delta s_{aff} - \Delta w_{aff} = \delta_d \\
&S\Delta z_{aff} + Z\Delta s_{aff} = -ZSe \\
&V\Delta w_{aff} + W\Delta z_{aff} = -VWe
\end{aligned}$$

Therefore, the computations involved in solving the affine step are

$$\begin{aligned}
&\Theta = SZ^{-1} + WV^{-1} \\
&\rho = \Theta^{-1}(\delta_d + (S - W)e - V^{-1}W\delta_b) \\
&\Delta t_{aff} = (A\Theta^{-1}A')^{-1}(\delta_c + A\rho) \\
&\Delta z_{aff} = \Theta^{-1}A'\Delta t_{aff} - \rho \\
&\Delta v_{aff} = \delta_b - \Delta z_{aff} \\
&\Delta w_{aff} = -We - V^{-1}W\Delta z_{aff} \\
&\Delta s_{aff} = -Se - Z^{-1}S\Delta z_{aff}
\end{aligned}$$

$$\begin{aligned}
&(z_{aff}, t_{aff}, s_{aff}, v_{aff}, w_{aff}) = (z, t, s, v, w) + \\
&\kappa(\Delta z_{aff}, \Delta t_{aff}, \Delta s_{aff}, \Delta v_{aff}, \Delta w_{aff})
\end{aligned}$$

κ is the *step length* as before.

The success of the affine step is gauged by calculating the complementarity of $z's$ and $v'w$ at $(z_{aff}^k, t_{aff}^k, s_{aff}^k, v_{aff}^k, w_{aff}^k)$ and comparing it with the complementarity at the starting point $(z^k, t^k, s^k, v^k, w^k)$. If the affine step was successful in reducing the complementarity by a substantial amount, the need for centering is not great, and a value close to zero is assigned to σ in a second linear system (see following), which is used to determine a centering vector. If, however, the affine step was unsuccessful, then centering is deemed beneficial, and a value close to 1.0 is assigned to σ. In other words, the value of σ is adaptively altered depending on progress made toward the optimum.

The following linear system is solved to determine a centering vector $(\Delta z_c, \Delta t_c, \Delta s_c, \Delta v_c, \Delta w_c)$ from $(z_{aff}, t_{aff}, s_{aff}, v_{aff}, w_{aff})$:

$$\begin{aligned}
&\Delta z_c + \Delta v_c = 0 \\
&A\Delta z_c = 0 \\
&A'\Delta t_c + \Delta s_c - \Delta w_c = 0
\end{aligned}$$

$$S\Delta z_c + Z\Delta s_c = -Z_{aff}S_{aff}e + \sigma\mu e$$
$$V\Delta w_c + W\Delta v_c = -V_{aff}W_{aff}e + \sigma\mu e$$

where

$\zeta_{start} = z's + v'w$, complementarity at the start of the iteration
$\zeta_{aff} = z'_{aff}s_{aff} + v'_{aff}w_{aff}$, the affine complementarity
$\mu = \zeta_{aff}/2n$, the average complementarity
$\sigma = (\zeta_{aff}/\zeta_{start})^3$

Therefore, the computations involved in solving the centering step are

$$\rho = \Theta^{-1}(\sigma\mu(Z^{-1} - V^{-1})e - Z^{-1}Z_{aff}S_{aff}e + V^{-1}V_{aff}W_{aff}e)$$
$$\Delta t_c = (A\Theta^{-1}A')^{-1}A\rho$$
$$\Delta z_c = \Theta^{-1}A'\Delta t_c - \rho$$
$$\Delta v_c = -\Delta z_c$$
$$\Delta w_c = \sigma\mu V^{-1}e - V^{-1}V_{aff}W_{aff}e - V^{-1}W_{aff}\Delta v_c$$
$$\Delta s_c = \sigma\mu Z^{-1}e - Z^{-1}Z_{aff}S_{aff}e - Z^{-1}S_{aff}\Delta z_c$$

Then

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

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

where, as before, 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.

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 the second linear system is small because the matrix $(A\Theta^{-1}A')$ has already been factorized in order 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

To minimize the sum of the absolute residuals $D_{LAR}(\beta)$, the smoothing algorithm approximates the nondifferentiable function D_{LAR} by the following smooth function, which is referred to as the

Huber function:

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

where

$$H_\gamma(t) = \begin{cases} t^2/(2\gamma) & \text{if } |t| \le \gamma \\ |t| - \gamma/2 & \text{if } |t| > \gamma \end{cases}$$

Here $r_i(\beta) = y_i - x_i'\beta$, and the *threshold* γ is a positive real number. The function D_γ is continuously differentiable and a minimizer β_γ of D_γ is close to a minimizer $\hat{\beta}_{LAR}$ of $D_{LAR}(\beta)$ when γ is close to zero.

The advantage of the smoothing algorithm as described in Madsen and Nielsen (1993) is that the L_1 solution $\hat{\beta}_{LAR}$ can be detected when $\gamma > 0$ is small. In other words, it is not necessary to let γ converge to zero in order to find a minimizer of $D_{LAR}(\beta)$. The algorithm terminates before going through the entire sequence of values of γ that are generated by the algorithm. Convergence is indicated by no change of the status of residuals $r_i(\beta)$ as γ goes through this sequence.

The smoothing algorithm extends naturally from L_1 regression to general quantile regression; refer to Chen (2007). The 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

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

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 $\mathbf{R}^p$ 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 $B_{\gamma,\tau}$:

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

Define the sign vector $s_\gamma(\beta) = (s_1(\beta), \ldots, s_n(\beta))'$ as

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

and introduce

$$w_i = w_i(\beta) = 1 - s_i^2(\beta)$$

Therefore,

$$H_{\gamma,\tau}(r_i(\beta)) = \frac{1}{2\gamma} w_i r_i^2(\beta)$$
$$+s_i[\frac{1}{2} r_i(\beta) + \frac{1}{4}(1-2\tau)\gamma + s_i(r_i(\beta)(\tau - \frac{1}{2}) - \frac{1}{4}(1-2\tau+2\tau^2)\gamma)]$$

yielding

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

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

The gradient of $D_{\gamma,\tau}$ is given by

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

and for $\beta \in \mathbf{R}^p \backslash B_{\gamma,\tau}$ 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 $\mathbf{R}^p$, whereas the Hessian is piecewise constant.

Following Madsen and Nielsen (1993), the vector s is referred to as a *γ-feasible sign vector* if there exists $\beta \in \mathbf{R}^p \backslash B_{\gamma,\tau}$ with $s_\gamma(\beta) = s$. If s is *γ-feasible*, then Q_s is defined as the quadratic function $Q_s(\alpha)$ that is derived from $D_{\gamma,\tau}(\beta)$ by substituting s for s_γ. Thus, for any β with $s_\gamma = s$,

$$Q_s(\alpha) = \frac{1}{2}(\alpha-\beta)' D^{(2)}_{\gamma,\tau}(\beta)(\alpha-\beta) + D^{(1)}_{\gamma,\tau}(\beta)(\alpha-\beta) + D_{\gamma,\tau}(\beta)$$

In the domain $C_s = \{\alpha | s_\gamma(\alpha) = s\}$

$$D_{\gamma,\tau}(\alpha) = Q_s(\alpha)$$

For each $\gamma > 0$ and $\theta \in \mathbf{R}^p$, there can be one or several corresponding quadratics Q_s. If $\theta \notin B_{\gamma,\tau}$ then Q_s is characterized by θ and γ, but for $\theta \in B_{\gamma,\tau}$ the quadratic is not unique. Therefore, a *reference*

$$(\gamma, \theta, s)$$

determines the quadratic.

Again following Madsen and Nielsen (1993), let

(γ, θ, s) be a *feasible reference* if s is a *γ-feasible sign vector* with $\theta \in C_s$, and

(γ, θ, s) be a *solution reference* if it is feasible and θ minimizes $D_{\gamma,\tau}$.

The smoothing algorithm for minimizing D_{ρ_τ} is based on minimizing $D_{\gamma,\tau}$ for a set of decreasing γ. For each new value of γ, information from the previous solution is used. Finally, when γ is small enough, a solution can be found by the modified Newton-Raphson algorithm as stated by Madsen and Nielsen (1993):

find an initial solution reference $(\gamma, \beta_\gamma, s)$

repeat

decrease γ

find a solution reference $(\gamma, \beta_\gamma, s)$

until $\gamma = 0$

β_0 is the solution.

By default, the initial solution reference is found by letting β_γ be the least squares solution. Alternatively, you can specify the initial solution reference with the INEST= option in the PROC statement. Then γ and s are chosen according to these initial values.

There are several approaches for determining a decreasing sequence of values of γ. The QUANTREG procedure uses a strategy by Madsen and Nielsen (1993). The computation involved is not significant comparing with the Newton-Raphson step. You can control the ratio of consecutive decreasing values of γ with the RRATIO= suboption of the ALGORITHM= option in the PROC statement. By default,

$$\text{RRATIO} = \begin{cases} 0.1 & \text{if } n \geq 10000 \text{ and } p \leq 20 \\ 0.9 & \text{if } \frac{p}{n} \geq 0.1 \text{ or } \{n \leq 5000 \text{ and } p \geq 300\} \\ 0.5 & \text{otherwise} \end{cases}$$

For the L_1 and quantile regression, it turns out that the smoothing algorithm is very efficient and competitive, especially for a *fat* data set—namely, when $\frac{p}{n} > 0.05$ and AA' is dense. Refer to Chen (2007) for a complete smoothing algorithm and details.

Confidence Interval

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 by 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 approach, addresses these problems, but at a computation cost.

Based on these properties, the QUANTREG uses a combination of the resampling and rank methods as the default. For data sets with more than either 5,000 observations or 20 variables, the QUANTREG procedure uses the MCMB resampling method; otherwise it uses the rank method. You can request a particular method by using the CI= option in the PROC statement.

Sparsity

Consider the linear model

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

and assume that $\{\epsilon_i\}$, $i = 1, \ldots, n$, are iid 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 $\Omega = \lim_{n\rightarrow\infty} n^{-1} \sum x_i x_i'$. Refer to 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 first be estimated.

Since

$$s(t) = \frac{d}{dt} F^{-1}(t)$$

$s(t)$ can be estimated by the difference quotient of the empirical quantile function—that is,

$$\hat{s}_n(t) = [\hat{F}_n^{-1}(t + h_n) - \hat{F}_n^{-1}(t - h_n)]/2h_n$$

where $\hat{F}_n$ is an estimate of F^{-1} and h_n is a bandwidth that tends to zero as $n \rightarrow \infty$.

The QUANTREG procedure provides two bandwidth methods. The Bofinger bandwidth

$$h_n = n^{-1/5}\left(\frac{4.5s^2(t)}{(s^{(2)}(t))^2}\right)^{1/5}$$

is an optimizer of mean squared error for standard density estimation, and the Hall-Sheather bandwidth

$$h_n = n^{-1/3} z_\alpha^{2/3}\left(\frac{1.5s(t)}{s^{(2)}(t)}\right)^{1/3}$$

is based on Edgeworth expansions for studentized quantiles, where $s^{(2)}(t)$ is the second derivative of $s(t)$ and z_α satisfies $\Phi(z_\alpha) = 1 - \alpha/2$ for the construction of $1-\alpha$ confidence intervals. The quantity

$$\frac{s(t)}{s^{(2)}(t)} = \frac{f^2}{2(f^{(1)}/f)^2 + [(f^{(1)}/f)^2 - f^{(2)}/f]}$$

is not sensitive to f and can be estimated by assuming f is Gaussian.

$\hat{F}^{-1}$ can be estimated by the empirical quantile function of the residuals from the quantile regression fit,

$$\hat{F}^{-1}(t) = r_{(i)}, \quad \text{for } t \in [(i-1)/n, i/n),$$

or the empirical quantile function of regression proposed by Bassett and Koenker (1982),

$$\hat{F}^{-1}(t) = \overline{x}'\hat{\beta}(t)$$

The QUANTREG procedure interpolates the first empirical quantile function and gets the piecewise linear version

$$\hat{F}^{-1}(t) = \begin{cases} r_{(1)} & \text{if } t \in [0, 1/2n) \\ \lambda r_{(i+1)} + (1-\lambda) r_{(i)} & \text{if } t \in [(2j-1)/2n, (2i+1)/2n) \\ r_{(n)} & \text{if } t \in [(2n-1), 1] \end{cases}$$

$\hat{F}^{-1}$ is set to a constant if $t \pm h_n$ falls outside $[0, 1]$.

This estimator of the sparsity function is sensitive to the iid assumption. Alternately, Koenker and Machado (1999) considered the non-iid case. By assuming local linearity of the conditional quantile function $Q(\tau|x)$ in x, they proposed a local estimator of the density function by using the difference quotient. A Huber sandwich estimate of the covariance and standard error is computed and used to construct the confidence intervals. One difficulty with this method is the selection of the bandwidth when using the difference quotient. With a small sample size, either the Bofinger or the Hall-Sheather bandwidth tends to be too large to assure local linearity of the conditional quantile function. The QUANTREG procedure uses a heuristic bandwidth selection in these cases.

By default, the QUANTREG procedure computes non-iid confidence intervals. You can request iid confidence intervals with the IID option in the PROC statement.

Inversion of Rank Tests

The classical theory of rank tests can be extended to test 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) for more details. By inverting this test, confidence intervals can be computed for the regression quantiles that correspond to β_2.

The rank score function $\hat{a}_n(t) = (\hat{a}_{n1}(t), \ldots, \hat{a}_{nn}(t))$ can be obtained by solving the dual problem

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

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: $\beta_2 = \eta$

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

for large 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. In order to obtain confidence intervals for β_2, a critical value can be specified for T_n. The dual vector $\hat{a}_n(\eta)$ is a piecewise constant in η, and η can 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 can be restored by taking one simplex pivot. The process can 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 (1994).

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 iid 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.

In contrast with these bootstrap methods, Parzen, Wei, and Ying (1994) observed that

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

which is the estimating equation for the τth regression quantile, is a pivotal quantity for the τth quantile regression parameter β_τ. In other words, the distribution of $S(\beta)$ can 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 $n+1$ observations with $x_{n+1} = -n^{-1/2}u/\tau$ and y_{n+1} sufficiently large for a given realization of u. By exploiting the asymptotically pivot role of the quantile regression "gradient condition," this approach also achieves some robustness to certain heteroscedasticity.

Although the bootstrap method by Parzen, Wei, and Ying (1994) is much simpler, it is too time-consuming for relatively large data sets, especially for high-dimensional data sets. The QUANTREG procedure implements a new, general resampling method developed by He and Hu (2002), which is 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.

Covariance-Correlation

You can specify the COVB and CORRB options in the MODEL statement to request covariance and correlation matrices for the estimated parameters.

The QUANTREG procedure provides two methods for computing 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. Asymptotic covariance and correlation matrices are computed when asymptotic confidence intervals are computed. The rank method for confidence intervals does not provide a covariance-correlation estimate.

Asymptotic Covariance-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, the QUANTREG procedure provides both iid and non-iid estimates. By default, the QUANTREG procedure computes non-iid estimates.

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

Three 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 statistic, 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 $D_0(\tau) = \sum \rho_\tau(y_i - x_i\hat{\beta}(\tau))$ and $D_1(\tau) = \sum \rho_\tau(y_i - x_{1i}\hat{\beta}_1(\tau))$, 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.

The rank test statistic under iid error models is given by

$$T_R(\tau) = S_n' M_n S_n / A^2(\varphi)$$

where

$$S_n = n^{-1/2}(X_2 - \hat{X_2})'\hat{b_n}$$

$$\hat{X_2} = X_1(X_1'X_1)^{-1}X_1'X_2$$

$$M_n = (X_2 - \hat{X_2})(X_2 - \hat{X_2})'/n$$

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

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

$$A^2(\varphi) = \int_0^1 (\varphi(t) - \bar{\varphi}(t))^2 dt$$

$$\bar{\varphi}(t) = \int_0^1 \varphi(t)dt$$

and $\varphi(t)$ is a score function. The following three score functions are available in the QUANTREG procedure:

Wilcoxon scores: $\phi(t) = t - 1/2$

normal scores: $\phi(t) = \Phi^{-1}(t)$, where Φ is the normal distribution function

sign scores: $\phi(t) = 1/2\ \text{sign}(t - 1/2)$

An important feature of the rank test statistic $T_R(\tau)$ is that, unlike Wald tests or likelihood ratio tests, no estimation of the sparsity function $s(\tau)$ is required.

Koenker and Machado (1999) prove that the three test statistics ($T_W(\tau)$, $T_{LR}(\tau)$, and $T_R(\tau)$) are asymptotically equivalent and that their distributions 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 $\bar{C}(A) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x})$ are the empirical multivariate location and scale. Here, $x_i = (x_{i1}, \ldots, x_{i(p-1)})'$ does not include the intercept variable. The relationship between the Mahalanobis distance $MD(x_i)$ and the 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

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

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

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

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

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 $\sigma = \text{median}\{|r_i|/\beta_0, i = 1, \ldots, n\}$, where $\beta_0 = \Phi^{-1}(0.75)$ is an adjustment constant for consistency with the normal distribution.

An ODS table called DIAGNOSTICS contains these two variables.

INEST= Data Set

The INEST= data set specifies initial estimates for all the parameters in the model. The INEST= data set must contain the intercept variable (named Intercept) and all independent variables in the MODEL statement.

If BY processing is used, the INEST= data set should also include the BY variables, and there must be at least one observation for each BY group. If there is more than one observation in one BY group, the first one read is used for that BY group.

If the INEST= data set also contains the _TYPE_ variable, only observations with the _TYPE_ value 'PARMS' are used as starting values.

You can specify starting values for the interior point algorithm or the smoothing algorithm in the INEST= data set. The INEST= data set has the same structure as the OUTEST= data set but is not required to have all the variables or observations that appear in the OUTEST= data set. One simple use of the INEST= option is passing the previous OUTEST= data set directly to the next model as an INEST= data set, assuming that the two models have the same parameterization. If you specify more than one quantile in the MODEL statement, the same initial values are used for all quantiles.

OUTEST= Data Set

The OUTEST= data set contains parameter estimates for the specified model with all quantiles. A set of observations is created for each quantile specified. You can also specify a label in the MODEL statement to distinguish between the estimates for different models used by the QUANTREG procedure.

Note that, if the QUANTREG procedure does not produce valid solutions, the parameter estimates are set to missing in the OUTEST data set.

If created, this data set contains all variables specified in the MODEL statement and the BY statement. Each observation consists of parameter values for a specified quantile with the dependent variable having the value -1.

The following variables are also added to the data set:

MODEL	a character variable of length 8 containing the label of the MODEL statement, if present. Otherwise, the variable's value is blank.
ALGORITHM	a character variable of length 8 containing the name of the algorithm used for computing the parameter estimates, either SIMPLEX, INTERIOR, or SMOOTH
TYPE	a character variable of length 8 containing the type of the observation. it is fixed as PARMS to indicate that the observation includes parameter estimates.
STATUS	a character variable of length 12 containing the status of model fitting, either NORMAL, NOUNIQUE, or NOVALID

INTERCEPT	a numeric variable containing the intercept parameter estimates
QUANTILE	a numeric variable containing the specified quantile levels

Any BY variables specified are also added to the OUTEST= data set.

Computational Resources

The various algorithms need different amounts of memory for working space. Let p be the number of parameters estimated and n be the number of observations used in the model estimation.

For the simplex algorithm, the minimum working space (in bytes) needed is

$$2np + 6n + 10p$$

for the interior point algorithm,

$$np + p^2 + 13n + 4p$$

and for the smoothing algorithm,

$$np + p^2 + 6n + 4p$$

For the last two algorithms, if you want to use preprocessing, an extra amount

$$np + 6n + 2p$$

is needed.

If sufficient space is available, the input data set is kept in memory; otherwise, the input data set is reread as necessary, and the execution time of the procedure increases substantially.

ODS Table Names

The QUANTREG procedure assigns a name to each table it creates. You can specify these names when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table.

Table 73.5 ODS Tables Produced in PROC QUANTREG

ODS Table Name	Description	Statement	Option
ClassLevels	Classification variable levels	CLASS	default
CorrB	Parameter estimate correlation matrix	MODEL	CORRB
CovB	Parameter estimate covariance matrix	MODEL	COVB
Diagnostics	Outlier diagnostics	MODEL	DIAGNOSTICS
DiagSummary	Summary of the outlier diagnostics	MODEL	DIAGNOSTICS
IPIterHistory	Iteration history (Interior Point)	MODEL	ITPRINT

Table 73.5 (continued)

ODS Table Name	Description	Statement	Option
ModelInfo	Model information	MODEL	default
NObs	Number of observations	PROC	default
ObjFunction	Objective function	MODEL	default
ParameterEstimates	Parameter estimates	MODEL	default
ParmInfo	Parameter indices	MODEL	default
PerfSettings	Performance settings	PERFORMANCE	DETAILS
ProcessEst	Quantile process estimates	MODEL	QUANTILE=
ProcessObj	Objective function for quantile process	MODEL	QUANTILE=
SMIterHistory	Iteration history (Smoothing)	MODEL	ITPRINT
SummaryStatistics	Summary statistics for model variables	MODEL	default
Tests	Results for tests	TEST	default
ScalableTiming	Timing details	PERFORMANCE	DETAILS

ODS Graphics

The QUANTREG procedure uses ODS Graphics to produce graphical displays for model fitting and model diagnostics. To request these plots you must specify the ODS GRAPHICS statement in addition to specific plot options in the PROC statement or the MODEL statement. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

For a single quantile, two plots are particularly useful in revealing outliers and leverage points. The first is a scatter plot of the standardized residuals for the specified quantile against the robust distances. The second is a scatter plot of the robust distances against the classical Mahalanobis distances. You can request these two plots by using the PLOT=RDPLOT and PLOT=DDPLOT options.

You can also request a normal quantile-quantile plot and a histogram of the standardized residuals for the specified quantile with the PLOT=QQPLOT and PLOT=HISTOGRAM options, respectively.

You can request a plot of fitted conditional quantiles by the single continuous variable specified in the model with the PLOT=FITPLOT option.

All these plots can be requested by specifying corresponding plot options in either the PROC statement or the MODEL statement. If you specify same plot options in both statements, options in the PROC statement override options in the MODEL statement.

You can specify the PLOT=QUANTPLOT option in only the MODEL statement to request a quantile process plot with confidence bands.

The plot options in the PROC statement and the MODEL statement are summarized in Table 73.6. See the PLOT= option in the PROC statement and the PLOT= option in the MODEL statement for details.

Table 73.6 Options for Plots

Keyword	Plot
ALL	All appropriate plots
DDPLOT	Robust distance vs. Mahalanobis distance
FITPLOT	Conditional quantile fit vs. independent variable
HISTOGRAM	Histogram of standardized robust residuals
NONE	No plot
QUANTPLOT	Scatter plot of regression quantile
QQPLOT	Q-Q plot of standardized robust residuals
RDPLOT	Standardized robust residual vs. robust distance

The names of the graphs that PROC QUANTREG generates are listed in Table 73.7, along with the required statements and options. The following subsections provide information about these graphs.

Fit Plot

When the model has a single independent continuous variable (with or without the intercept), the QUANTREG procedure automatically creates a plot of fitted conditional quantiles against this independent variable for one or more quantiles specified in the MODEL statement.

The following example reuses the trout data set in the section "Analysis of Fish-Habitat Relationships" on page 6075 to show the fit plot for one or several quantiles.

```
ods graphics on;

proc quantreg data=trout ci=resampling;
   model LnDensity = WDRatio / quantile=0.9 seed=1268;
run;

proc quantreg data=trout ci=resampling;
   model LnDensity = WDRatio / quantile=0.5 0.75 0.9 seed=1268;
run;

ods graphics off;
```

For a single quantile, the confidence limits for the fitted conditional quantiles are also plotted if you specify the CI=RESAMPLING or CI=SPARSITY option. (See Figure 73.14.) For multiple quantiles, confidence limits are not plotted by default. (See Figure 73.15.) You can add the confidence limits on the plot by specifying the option PLOT=FITPLOT(SHOWLIMITS).

The QUANTREG procedure also provides fit plots for quantile regression splines and polynomials if they are based on a single continuous variable. Refer to Example 73.4 and Example 73.5 for some examples.

Figure 73.14 Fit Plot with Confidence Limits

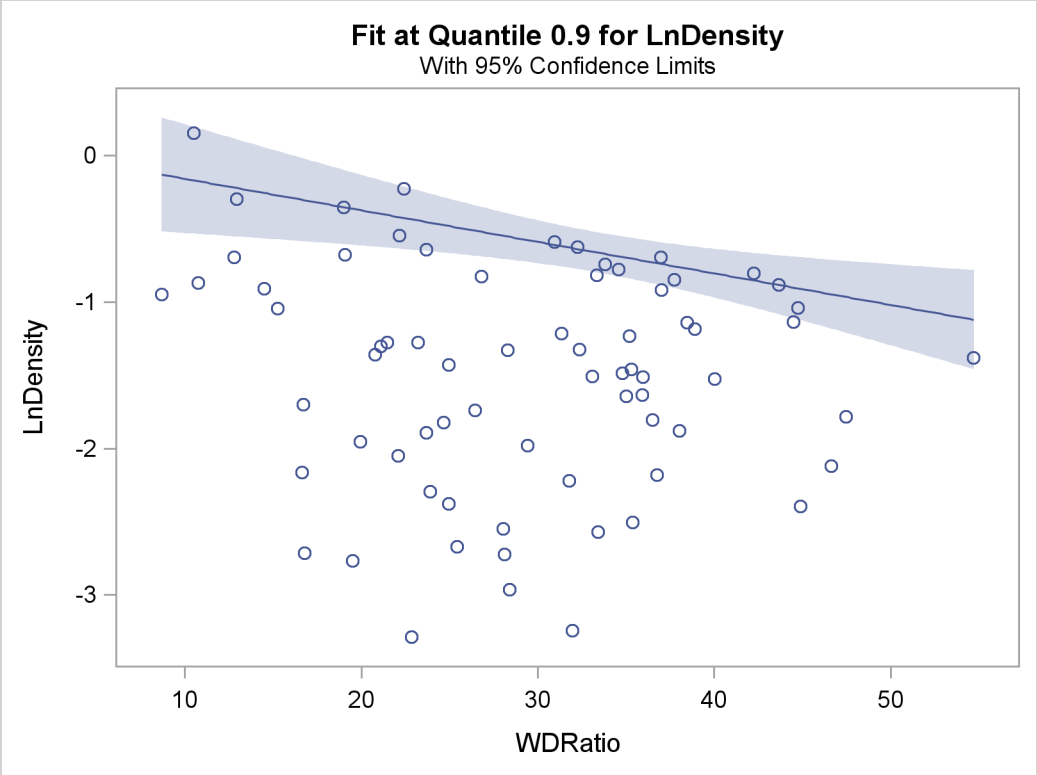
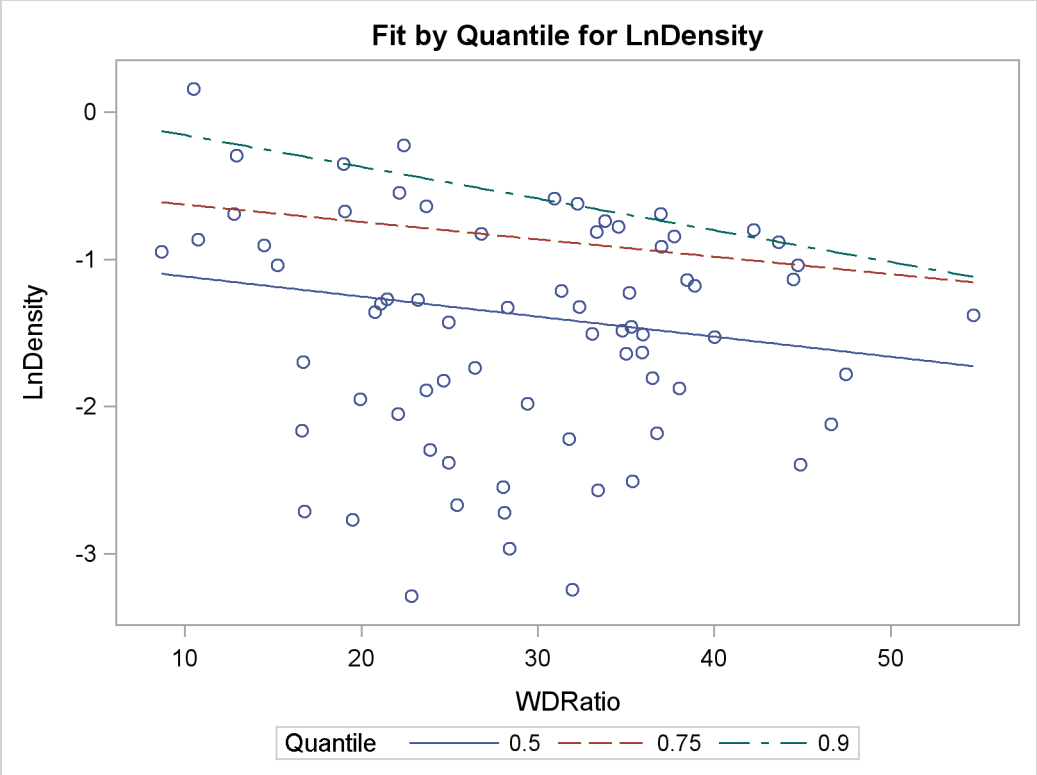


Figure 73.15 Fit Plot for Multiple Quantiles



Quantile Process Plot

A quantile process plot is a scatter plot of an estimated regression parameter against quantile. You can request this plot with the PLOT=QUANTPLOT option in the MODEL statement when multiple regression quantiles or the entire quantile process is computed. Quantile process plots are often used to check model variations at different quantiles, which is usually called model heterogeneity.

By default, panels are used to hold multiple process plots (up to four in each panel). You can use the UNPACK option to request individual process plots. Figure 73.10 in the section "Analysis of Fish-Habitat Relationships" on page 6075 shows a panel with two quantile process plots. Output 73.2.9 in Example 73.2 shows a single quantile process plot. Example 73.3 demonstrates more quantile process plots and their usage.

Distance-Distance Plot

The distance-distance plot (DDPLOT) is mainly used for leverage-point diagnostics. It is a scatter plot of the robust distances against the classical Mahalanobis distances for the continuous independent variables. See the section "Leverage Point and Outlier Detection" on page 6111 for details about the robust distance. If there is a classification variable specified in the model, this plot is not created.

You can use the PLOT=DDPLOT option to request this plot. The following statements use the growth data set in Example 73.2 to create a single plot, shown in Output 73.2.4 in Example 73.2:

```
ods graphics on;

proc quantreg data=growth ci=resampling plot=ddplot;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2
                lintr2 gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
                / quantile=.5 diagnostics leverage(cutoff=8) seed=1268;
    id Country;
run;

ods graphics off;
```

The reference lines represent the cutoff values. The diagonal line is also drawn to show the distribution of the distances. By default, all outliers and leverage points are labeled with observation numbers. To change the default, you can use the LABEL= option as described in Table 73.4.

Residual-Distance Plot

The residual-distance plot (RDPLOT) is used for both outlier and leverage-point diagnostics. It is a scatter plot of the standardized residuals against the robust distances. See the section "Leverage Point and Outlier Detection" on page 6111 for details about the robust distance. If a classification variable is specified in the model, this plot is not created.

You can use the PLOT=RDPLOT option to request this plot. The following statements use the growth data set in Example 73.2 to create a single plot, shown in Output 73.2.3 in Example 73.2:

```
ods graphics on;

proc quantreg data=growth ci=resampling plot=rdplot;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2
                lintr2 gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
                / quantile=.5 diagnostics leverage(cutoff=8) seed=1268;
    id Country;
run;

ods graphics off;
```

The reference lines represent the cutoff values. By default, all outliers and leverage points are labeled with observation numbers. To change the default, you can use the LABEL= option as described in Table 73.4.

If you specify ID variables in the ID statement, instead of observation numbers, the values of the first ID variable are used as labels.

Histogram and Q-Q Plot

PROC QUANTREG produces a histogram and a Q-Q plot for the standardized residuals. The histogram is superimposed with a normal density curve and a kernel density curve. Using the growth data set in Example 73.2, the following statements create the plot shown in Output 73.2.5 in Example 73.2:

```
ods graphics on;

proc quantreg data=growth ci=resampling plot=histogram;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2
                lintr2 gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
                / quantile=.5 diagnostics leverage(cutoff=8) seed=1268;
    id Country;
run;

ods graphics off;
```

ODS Graph Names

The QUANTREG procedure assigns a name to each graph it creates. You can use these names to reference the graphs when using ODS. The names are listed in Table 73.7.

To request these graphs you must specify the ODS GRAPHICS statement in addition to the plot options in the PROC statement or the MODEL statement. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

Table 73.7 ODS Graphics Produced by PROC QUANTREG

ODS Graph Name	Plot Description	Statement	Option
DDPlot	Robust distance versus Mahalanobis distance	PROC MODEL	DDPLOT
FitPlot	Quantile fit versus independent variable	PROC MODEL	FITPLOT
Histogram	Histogram of standardized robust residuals	PROC MODEL	HISTOGRAM
QQPlot	Q-Q plot of standardized robust residuals	PROC MODEL	QQPLOT
QuantPanel	Panel of quantile plots with confidence limits	MODEL	QUANTPLOT
QuantPlot	Scatter plot for regression quantiles with confidence limits	MODEL	QUANTPLOT UNPACK
RDPlot	Standardized robust residual versus robust distance	PROC MODEL	RDPLOT

Examples: QUANTREG Procedure

Example 73.1: Comparison of Algorithms

This example illustrates and compares the three algorithms for regression estimation available in the QUANTREG procedure. The simplex algorithm is the default because of its stability. Although this algorithm is slower than the interior point and smoothing algorithms for large data sets, the difference is not as significant for data sets with fewer than 5,000 observations and 50 variables. The simplex algorithm can also compute the entire quantile process, which is shown in Example 73.2.

The following statements generate 1,000 random observations. The first 950 observations are from a linear model, and the last 50 observations are significantly biased in the y-direction. In other words, 5% of the observations are contaminated with outliers.

```
data a (drop=i);
   do i=1 to 1000;
      x1=rannor(1234);
      x2=rannor(1234);
      e=rannor(1234);
      if i > 950 then y=100 + 10*e;
      else y=10 + 5*x1 + 3*x2 + 0.5 * e;
      output;
   end;
run;
```

The following statements invoke the QUANTREG procedure to fit a median regression model with the default simplex algorithm. They produce the results in Output 73.1.1 through Output 73.1.3.

```
proc quantreg data=a;
   model y = x1 x2;
run;
```

Output 73.1.1 displays model information and summary statistics for variables in the model. It indicates that the simplex algorithm is used to compute the optimal solution and the rank method is used to compute confidence intervals of the parameters.

By default, the QUANTREG procedure fits a median regression model. This is indicated by the quantile value 0.5 in Output 73.1.2, which also displays the objective function value and the predicted value of the response at the means of the covariates.

Output 73.1.3 displays parameter estimates and confidence limits. These estimates are reasonable, which indicates that median regression is robust to the 50 outliers.

Output 73.1.1 Model Fit Information and Summary Statistics with Simplex Algorithm

The QUANTREG Procedure

Model Information

Data Set	WORK.A
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Optimization Algorithm	Simplex
Method for Confidence Limits	Inv_Rank

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
x1	-0.6546	0.0230	0.7099	0.0222	0.9933	1.0085
x2	-0.7891	-0.0747	0.6839	-0.0401	1.0394	1.0857
y	6.1045	10.6936	14.9569	14.4864	20.4087	6.5696

Output 73.1.2 Quantile and Objective Function with Simplex Algorithm

Quantile and Objective Function

Quantile	0.5
Objective Function	2441.1927
Predicted Value at Mean	10.0259

Output 73.1.3 Parameter Estimates with Simplex Algorithm

Parameter Estimates

Parameter	DF	Estimate	95% Confidence Limits	
Intercept	1	10.0364	9.9959	10.0756
x1	1	5.0106	4.9602	5.0388
x2	1	3.0294	2.9944	3.0630

The following statements refit the model by using the interior point algorithm:

```
proc quantreg algorithm=interior(tolerance=1e-6)
              ci=none data=a;
   model y = x1 x2 / itprint nosummary;
run;
```

The TOLERANCE= option specifies the stopping criterion for convergence of the interior point algorithm, which is controlled by the duality gap. Although the default criterion is 1E−8, the value 1E−6 is often sufficient. The ITPRINT option requests the iteration history for the algorithm. The option CI=NONE suppresses the computation of confidence limits, and the option NOSUMMARY suppresses the table of summary statistics.

Output 73.1.4 displays model fit information.

Output 73.1.4 Model Fit Information with Interior Point Algorithm

The QUANTREG Procedure

Model Information

Data Set	WORK.A
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Optimization Algorithm	Interior

Output 73.1.5 displays the iteration history of the interior point algorithm. Note that the duality gap is less than 1E−6 in the final iteration. The table also provides the number of iterations, the number of corrections, the primal step length, the dual step length, and the objective function value at each iteration.

Output 73.1.5 Iteration History for the Interior Point Algorithm

Iteration History of Interior Point Algorithm

Duality Gap	Iter	Correction	Primal Step	Dual Step	Objective Function
2623	1	1	0.3113	0.4910	3303.4688
3215	2	2	0.0427	1.0000	2461.3774
1127	3	3	0.9882	0.3653	2451.1337
760.88658	4	4	0.3381	1.0000	2442.8104
77.10290	5	5	1.0000	0.8916	2441.2627
8.43666	6	6	0.9370	0.8381	2441.2085
1.82868	7	7	0.8375	0.7674	2441.1985
0.40584	8	8	0.6980	0.8636	2441.1948
0.09550	9	9	0.9438	0.5955	2441.1930
0.00665	10	10	0.9818	0.9304	2441.1927
0.0002248	11	11	0.9179	0.9994	2441.1927
5.44651E-8	12	12	1.0000	1.0000	2441.1927

Output 73.1.6 displays the parameter estimates obtained with the interior point algorithm, which are identical to those obtained with the simplex algorithm.

Output 73.1.6 Parameter Estimates with Interior Point Algorithm

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	10.0364
x1	1	5.0106
x2	1	3.0294

The following statements refit the model by using the smoothing algorithm. They produce the results in Output 73.1.7 through Output 73.1.9.

```
proc quantreg algorithm=smooth(rratio=.5) ci=none data=a;
   model y = x1 x2 / itprint nosummary;
run;
```

The RRATIO= option controls the reduction speed of the threshold. Output 73.1.7 displays the model fit information.

Output 73.1.7 Model Fit Information with Smoothing Algorithm

The QUANTREG Procedure

Model Information

Data Set	WORK.A
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Optimization Algorithm	Smooth

Output 73.1.8 displays the iteration history of the smoothing algorithm. The threshold controls the convergence. Note that the thresholds decrease by a factor of at least 0.5, the value specified with the RRATIO= option. The table also provides the number of iterations, the number of factorizations, the number of full updates, the number of partial updates, and the objective function value in each iteration. For details concerning the smoothing algorithm, refer to Chen (2007).

Output 73.1.8 Iteration History for the Smoothing Algorithm

Iteration History of Smoothing Algorithm

Threshold	Iter	Refac	Full Update	Partial Update	Objective Function
227.24557	1	1	1000	0	4267.0988
116.94090	15	4	1480	2420	3631.9653
1.44064	17	4	1480	2583	2441.4719
0.72032	20	5	1980	2598	2441.3315
0.36016	22	6	2248	2607	2441.2369
0.18008	24	7	2376	2608	2441.2056
0.09004	26	8	2446	2613	2441.1997
0.04502	28	9	2481	2617	2441.1971
0.02251	30	10	2497	2618	2441.1956
0.01126	32	11	2505	2620	2441.1946
0.00563	34	12	2510	2621	2441.1933
0.00281	35	13	2514	2621	2441.1930
0.0000846	36	14	2517	2621	2441.1927
1E-12	37	14	2517	2621	2441.1927

Output 73.1.9 displays the parameter estimates obtained with the smoothing algorithm, which are identical to those obtained with the simplex and interior point algorithms. All three algorithms should have the same parameter estimates unless the problem does not have a unique solution.

Output 73.1.9 Parameter Estimates with Smoothing Algorithm

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	10.0364
x1	1	5.0106
x2	1	3.0294

The interior point algorithm and the smoothing algorithm offer better performance than the simplex algorithm for large data sets. Refer to Chen (2004) for more details on choosing an appropriate algorithm on the basis of data set size. All three algorithms should have the same parameter estimates, unless the optimization problem has multiple solutions.

Example 73.2: Quantile Regression for Econometric Growth Data

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. The following statements request median regression for a preliminary exploration. They produce the results in Output 73.2.1 through Output 73.2.6.

```
data growth;
   length Country$ 22;
   input Country GDP lgdp2 mse2 fse2 fhe2 mhe2 lexp2 lintr2 gedy2
         Iy2 gcony2 lblakp2 pol2 ttrad2 @@;
datalines;
Algeria75          .0415 7.330 .1320 .0670 .0050 .0220 3.880 .1138 .0382
                   .1898 .0601 .3823 .0833 .1001
Algeria85          .0244 7.745 .2760 .0740 .0070 .0370 3.978 -.107 .0437
                   .3057 .0850 .9386 .0000 .0657
Argentina75        .0187 8.220 .7850 .6200 .0740 .1660 4.181 .4060 .0221
                   .1505 .0596 .1924 .3575 -.011
Argentina85        -.014 8.407 .9360 .9020 .1320 .2030 4.211 .1914 .0243
                   .1467 .0314 .3085 .7010 -.052


   ... more lines ...

Zambia75           .0120 6.989 .3760 .1190 .0130 .0420 3.757 .4388 .0339
                   .3688 .2513 .3945 .0000 -.032
Zambia85           -.046 7.109 .4200 .2740 .0110 .0270 3.854 .8812 .0477
                   .1632 .2637 .6467 .0000 -.033
Zimbabwe75         .0320 6.860 .1450 .0170 .0080 .0450 3.833 .7156 .0337
                   .2276 .0246 .1997 .0000 -.040
Zimbabwe85         -.011 7.180 .2200 .0650 .0060 .0400 3.944 .9296 .0520
                   .1559 .0518 .7862 .7161 -.024
;

ods graphics on;

proc quantreg data=growth ci=resampling
                plots=(rdplot ddplot reshistogram);
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2
                lintr2 gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
                / quantile=.5 diagnostics leverage(cutoff=8) seed=1268;
    id Country;
    test_lgdp2: test lgdp2 / lr wald;
run;

ods graphics off;
```

The QUANTREG procedure employs the default simplex algorithm to estimate the parameters. The MCMB resampling method is used to compute confidence limits.

Output 73.2.1 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 might have outliers.

Output 73.2.2 displays parameter estimates and 95% confidence limits computed with the rank method.

Output 73.2.1 Model Information and Summary Statistics

The QUANTREG Procedure

Model Information

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

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
lgdp2	6.9890	7.7450	8.6080	7.7905	0.9543	1.1579
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	4.0640	4.2430	4.0440	0.2028	0.2728
lintr2	0.00160	0.5604	1.8805	1.4625	2.5491	1.0058
gedy2	0.0248	0.0343	0.0466	0.0360	0.0141	0.0151
Iy2	0.1396	0.1955	0.2671	0.2010	0.0877	0.0981
gcony2	0.0480	0.0767	0.1276	0.0914	0.0617	0.0566
lblakp2	0	0.0696	0.2407	0.1916	0.3070	0.1032
pol2	0	0.0500	0.2429	0.1683	0.2409	0.0741
ttrad2	−0.0240	−0.0100	0.00730	−0.00570	0.0375	0.0239
GDP	0.00290	0.0196	0.0351	0.0191	0.0248	0.0237

Output 73.2.2 Parameter Estimates

Parameter Estimates

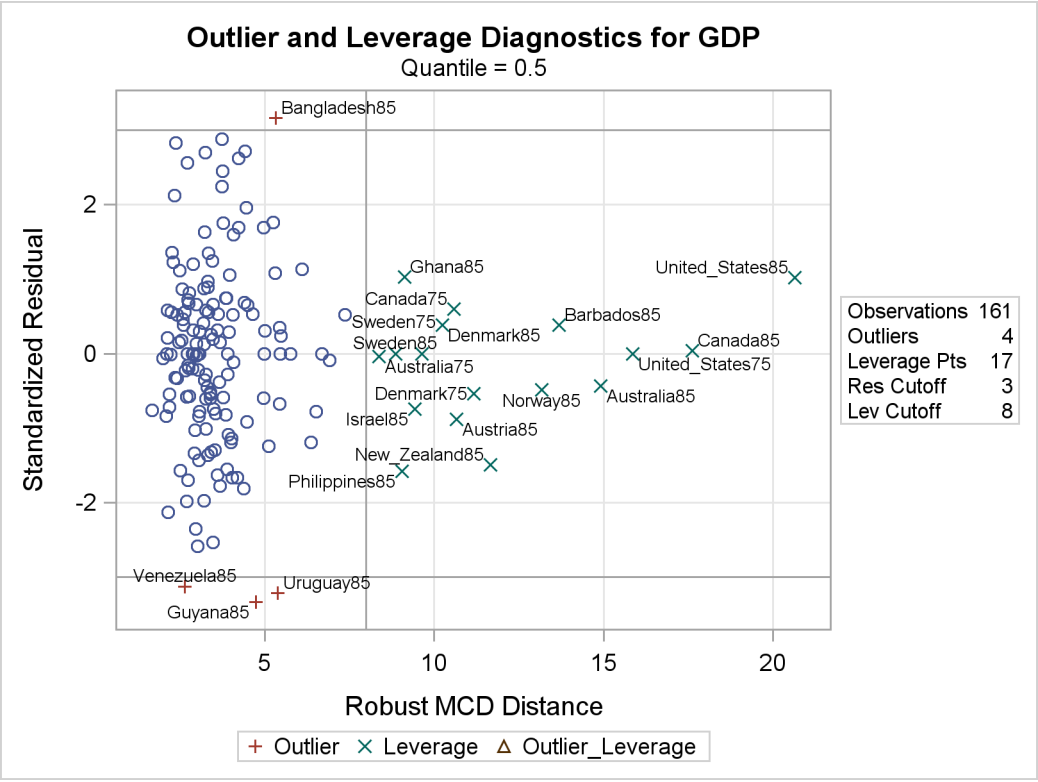
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	−0.0488	0.0733	−0.1937	0.0961	−0.67	0.5065
lgdp2	1	−0.0269	0.0041	−0.0350	−0.0188	−6.58	<.0001
mse2	1	0.0110	0.0080	−0.0048	0.0269	1.38	0.1710
fse2	1	−0.0011	0.0088	−0.0185	0.0162	−0.13	0.8960
fhe2	1	0.0148	0.0321	−0.0485	0.0782	0.46	0.6441
mhe2	1	0.0043	0.0268	−0.0487	0.0573	0.16	0.8735
lexp2	1	0.0683	0.0229	0.0232	0.1135	2.99	0.0033
lintr2	1	−0.0022	0.0015	−0.0052	0.0008	−1.44	0.1513
gedy2	1	−0.0508	0.1654	−0.3777	0.2760	−0.31	0.7589
Iy2	1	0.0723	0.0248	0.0233	0.1213	2.92	0.0041
gcony2	1	−0.0935	0.0382	−0.1690	−0.0181	−2.45	0.0154
lblakp2	1	−0.0269	0.0084	−0.0435	−0.0104	−3.22	0.0016
pol2	1	−0.0301	0.0093	−0.0485	−0.0117	−3.23	0.0015
ttrad2	1	0.1613	0.0740	0.0149	0.3076	2.18	0.0310

Diagnostics for the median regression fit are displayed in Output 73.2.3 and Output 73.2.4, which are requested with the PLOTS= option. Output 73.2.3 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. Output 73.2.4 plots the robust MCD distance against the Mahalanobis distance. This display is used to diagnose leverage points.

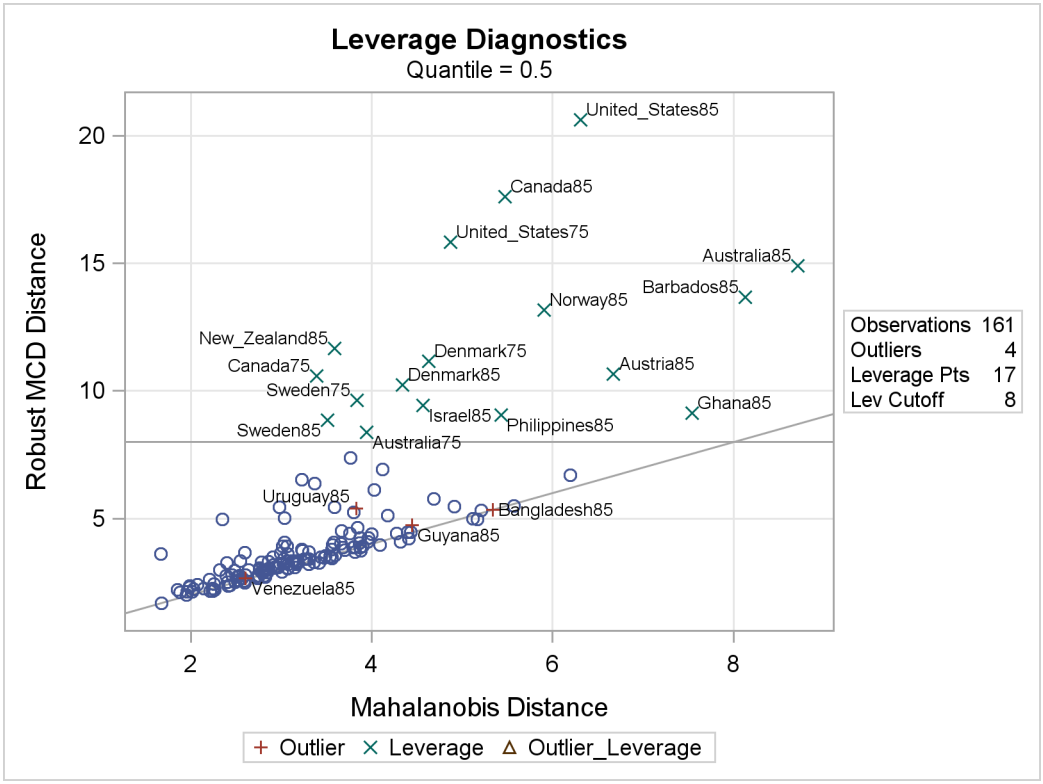
The cutoff value 8 specified with the LEVERAGE option is close to the maximum of the Mahalanobis distance. Eighteen points are diagnosed as high leverage points, and almost all are countries with high human capital, which is the major contributor to the high leverage as observed from the summary statistics.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

A histogram of the standardized residuals and two fitted density curves are displayed in Output 73.2.5. This shows that median regression fits the data well.

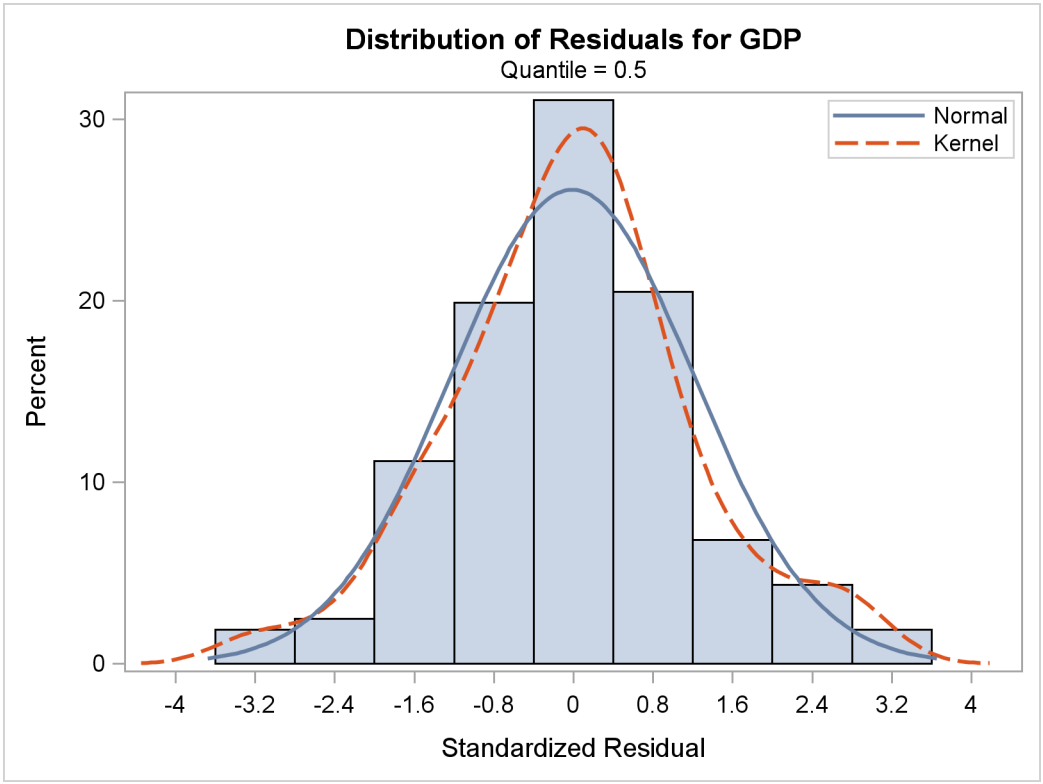
Output 73.2.3 Residual-Robust Distance Plot



Output 73.2.4 Robust Distance-Mahalanobis Distance Plot



Output 73.2.5 Histogram for Residuals



Tests of significance for the initial per-capita GDP (LGDP2) are shown in Output 73.2.6.

Output 73.2.6 Tests for Regression Coefficient

Test test_lgdp2 Results

Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
Wald	43.2684	1	43.27	<.0001
Likelihood Ratio	36.3047	1	36.30	<.0001

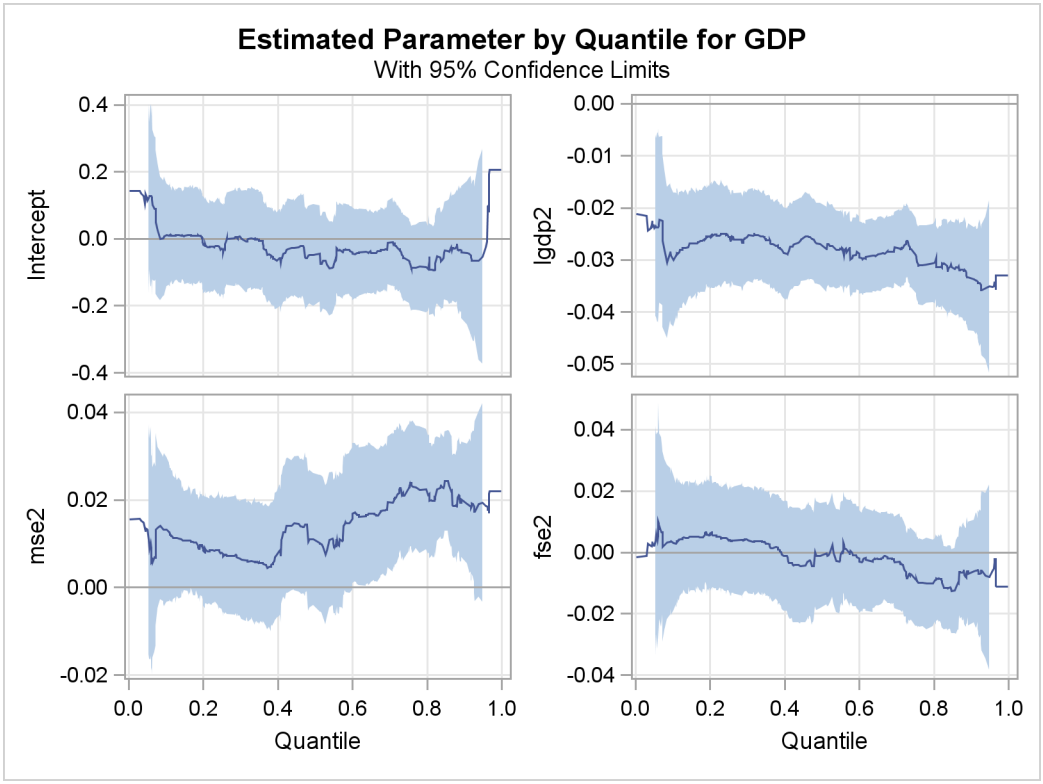
The QUANTREG procedure computes entire quantile processes for covariates when you specify QUANTILE=PROCESS in the MODEL statement, as follows:

```
ods graphics on;

proc quantreg data=growth ci=resampling;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2 lintr2
                gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
                / quantile=process plot=quantplot seed=1268;
run;

ods graphics off;
```

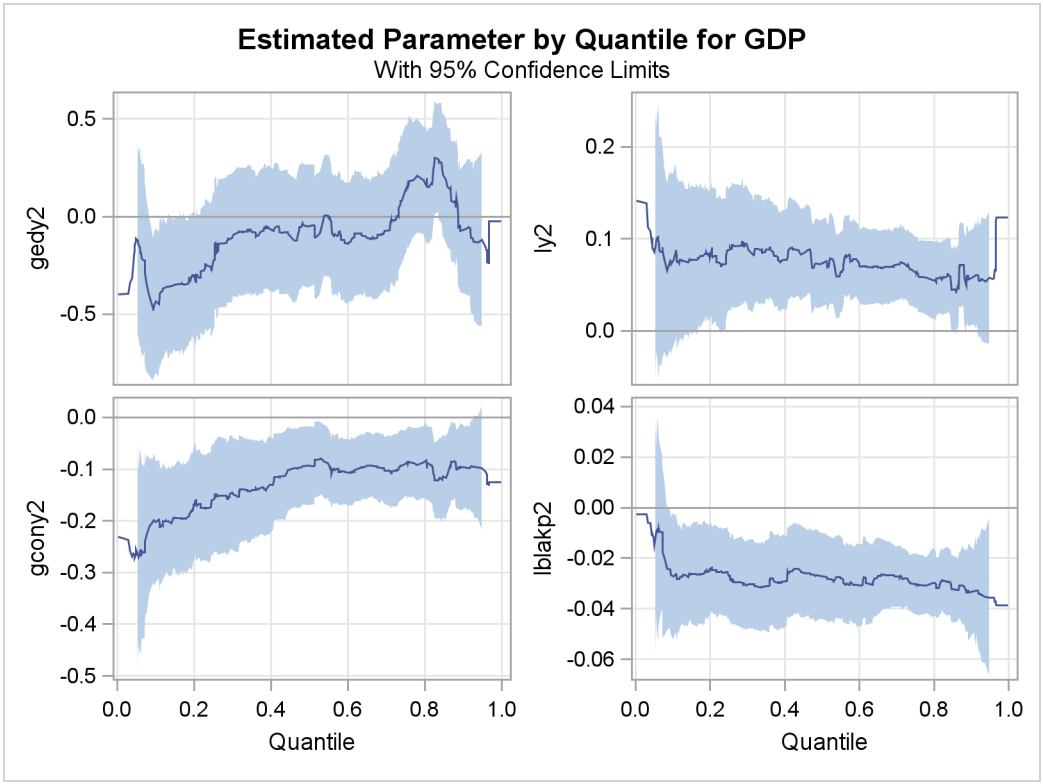
Confidence limits for quantile processes can be computed with the sparsity or resampling methods, but not the rank method, because the computation would be prohibitively expensive.

A total of 14 quantile process plots are produced. Output 73.2.7 and Output 73.2.8 display two panels of eight selected process plots. The 95% confidence bands are shaded.

Output 73.2.7 Quantile Processes with 95% Confidence Bands



Output 73.2.8 Quantile Processes with 95% Confidence Bands

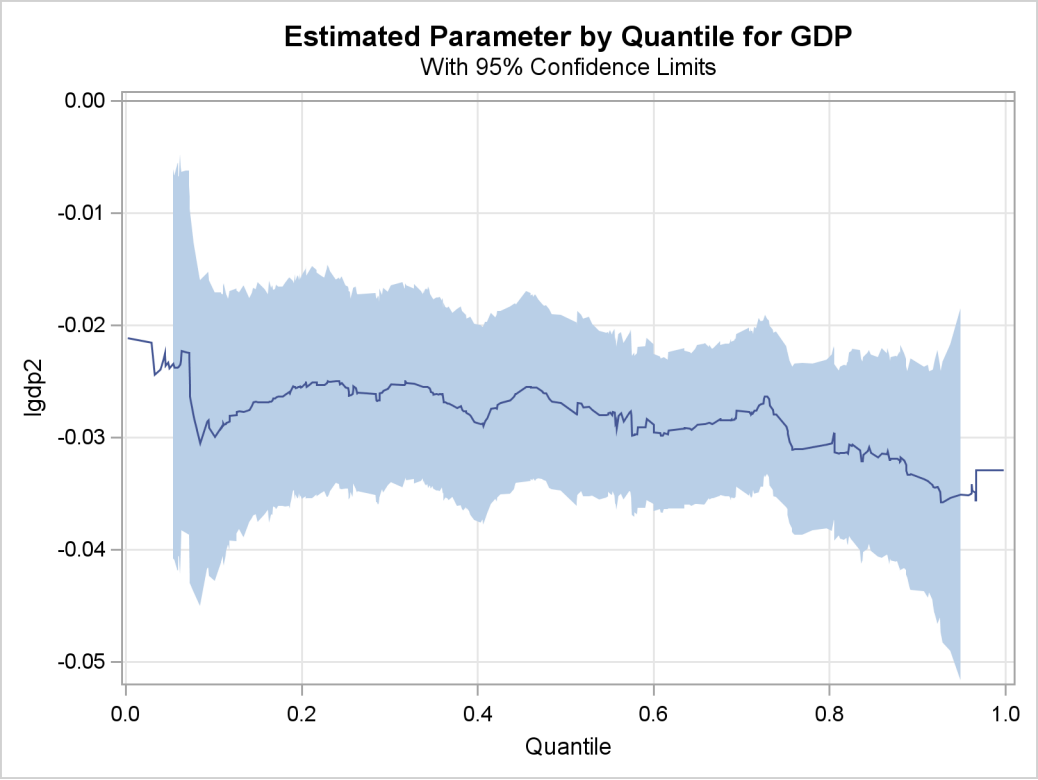


As pointed out by Koenker and Machado (1999), previous studies of the Barro growth data have focused on the effect of the initial per-capita GDP on the growth of this variable (annual change per-capita GDP). A single process plot for this effect can be requested with the following statements:

```
ods graphics on;

proc quantreg data=growth ci=resampling;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2 lintr2
                gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
                / quantile=process plot=quantplot(lgdp2) seed=1268;
run;

ods graphics off;
```

The plot is shown in Output 73.2.9.

Output 73.2.9 Quantile Process Plot for LGDP2



The confidence bands here are computed with the MCMB resampling method, unlike in Koenker and Machado (1999), where the rank method was used to compute confidence limits for a few selected points. Output 73.2.9 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 an analysis of the effects of the other covariates, refer to Koenker and Machado (1999).

Example 73.3: Quantile Regression Analysis of Birth-Weight Data

This example is patterned after a quantile regression analysis of covariates associated with birth weight that was carried out by Koenker and Hallock (2001). Their study used a subset of the June 1997 Detailed Natality Data published by the National Center for Health Statistics and demonstrated that conditional quantile functions provide more complete information about the covariate effects than ordinary least squares regression.

As in Koenker and Hallock (2001) and Abreveya (2001), this example uses data for live, singleton births to mothers in the United States who were recorded as black or white, and who were between the ages of 18 and 45. For convenience, this example uses 50,000 observations, which were randomly selected from the qualified observations. Observations with missing data for any of the variables were deleted.

The following table describes the variables in the data.

Variable	Description
weight	Infant's birth weight
black	Indicator of black mother
married	Indicator of married mother
boy	Indicator of boy
visit	Prenatal visit: 0 = no visit, 1 = visit in second trimester, 2 = visit in last trimester, 3 = visit in first trimester
ed	Mother's education level: 0 = high school, 1 = some college, 2 = college, 3 = less than high school
smoke	Indicator of smoking mother
cigsper	Number of cigarettes smoked per day
mom_age	Mother's age
m_wtgain	Mother's weight gain during pregnancy

There are four levels of education of the mother. By default, the QUANTREG procedure treats the highest level (3 - less than high school) as a reference level. The regression coefficients of other levels measure the effect relative to this level. Likewise, there are four levels of prenatal medical care of the mother, and a first visit in the first trimester serves as the reference level. These two variables are treated as classification variables in the model.

The following statements fit a regression model for 19 quantiles of birth weight, which are evenly spaced in the interval (0, 1). The model includes linear and quadratic effects for the age of the mother and for weight gain during pregnancy.

```
ods graphics on;

proc quantreg ci=sparsity/iid algorithm=interior(tolerance=1.e-4)
              data=sashelp.bweight;
   class visit ed;
   model weight = black married boy visit ed smoke
                  cigsper mom_age mom_age*mom_age
                  m_wtgain m_wtgain*m_wtgain /
                  quantile= 0.05 to 0.95 by 0.05
                  plot=quantplot;
run;

ods graphics off;
```

Output 73.3.1 Model Information and Summary Statistics

The QUANTREG Procedure

Model Information

Data Set	SASHELP.BWEIGHT
Dependent Variable	weight
Number of Independent Variables	9
Number of Continuous Independent Variables	7
Number of Class Independent Variables	2
Number of Observations	50000
Optimization Algorithm	Interior
Method for Confidence Limits	Sparsity

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
black	0	0	0	0.1628	0.3692	0
married	0	1.0000	1.0000	0.7126	0.4525	0
boy	0	1.0000	1.0000	0.5158	0.4998	0
smoke	0	0	0	0.1307	0.3370	0
cigsper	0	0	0	1.4766	4.6541	0
mom_age	-4.0000	0	5.0000	0.4161	5.7285	5.9304
mom_age*mom_age	4.0000	16.0000	49.0000	32.9877	39.2861	22.2390
m_wtgain	-8.0000	0	9.0000	0.7092	12.8761	11.8608
m_wtgain*m_wtgain	16.0000	64.0000	196.0	166.3	298.8	88.9561
weight	3062.0	3402.0	3720.0	3370.8	566.4	504.1

Output 73.3.1 displays the model information and summary statistics for the variables in the model.

Among the 11 independent variables, black, married, boy, and smoke are binary variables. For these variables, the mean represents the proportion in the category. The two continuous variables, mom_age and m_wtgain, are centered at their medians, which are 27 and 30, respectively.

The quantile plots for the intercept and the other 15 factors with nonzero degree of freedom are shown in the following four panels. In each plot, the regression coefficient at a given quantile indicates the

effect on birth weight of a unit change in that factor, assuming that the other factors are fixed. The bands represent 95% confidence intervals.

Although the data set used here is a subset of the Natality data set, the results are quite similar to those of Koenker and Hallock (2001) for the full data set.

In Output 73.3.2, the first plot is for the intercept. As explained by Koenker and Hallock (2001), the intercept "may be interpreted as the estimated conditional quantile function of the birth-weight distribution of a girl born to an unmarried, white mother with less than a high school education, who is 27 years old and had a weight gain of 30 pounds, didn't smoke, and had her first prenatal visit in the first trimester of the pregnancy."

The second plot shows that infants born to black mothers weigh less than infants born to white mothers, especially in the lower tail of the birth-weight distribution. The third plot shows that marital status has a large positive effect on birth weight, especially in the lower tail. The fourth plot shows that boys weigh more than girls for any chosen quantile; this difference is smaller in the lower quantiles of the distribution.

In Output 73.3.3, the first three plots deal with prenatal care. Compared with babies born to mothers who had a prenatal visit in the first trimester, babies born to mothers who received no prenatal care weigh less, especially in the lower quantiles of the birth-weight distributions. As noted by Koenker and Hallock (2001), "babies born to mothers who delayed prenatal visits until the second or third trimester have substantially *higher* birthweights in the lower tail than mothers who had a prenatal visit in the first trimester. This might be interpreted as the self-selection effect of mothers confident about favorable outcomes."

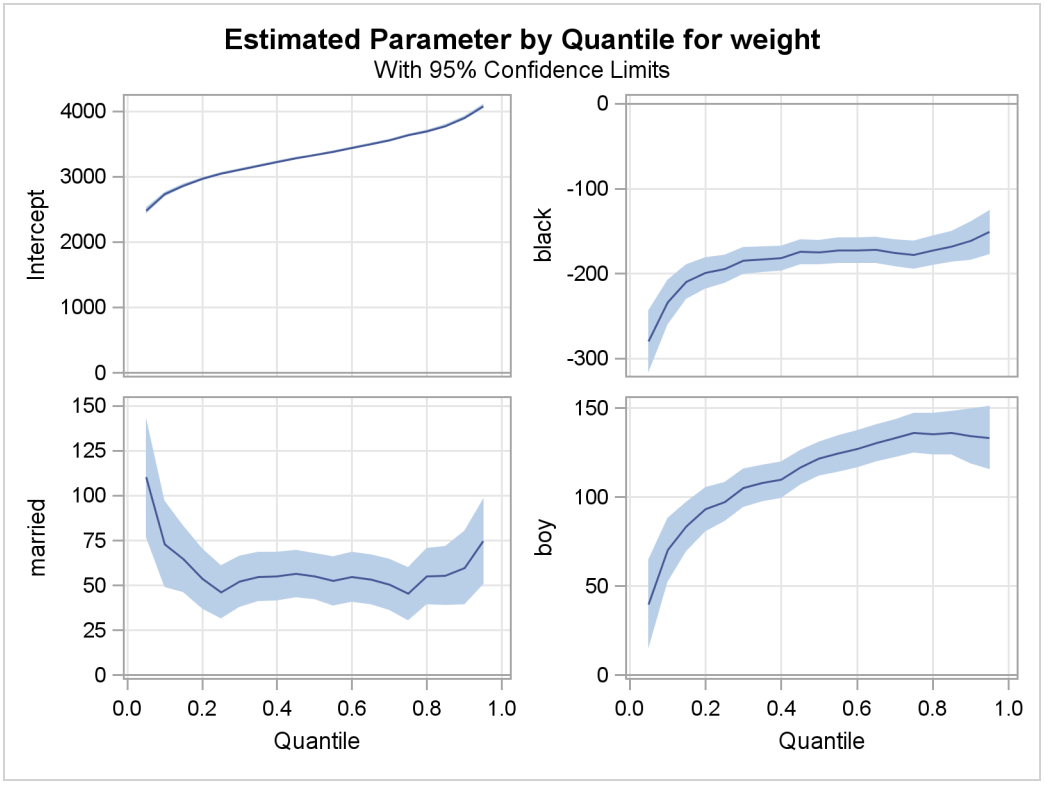
The fourth plot in Output 73.3.3 and the first two plots in Output 73.3.4 are for variables related to education. Education beyond high school is associated with a positive effect on birth weight. The effect of high school education is uniformly around 15 grams across the entire birth-weight distribution (this is a pure location shift effect), while the effect of some college and college education is more positive in the lower quantiles than the upper quantiles.

The remaining two plots in Output 73.3.4 show that smoking is associated with a large negative effect on birth weight.

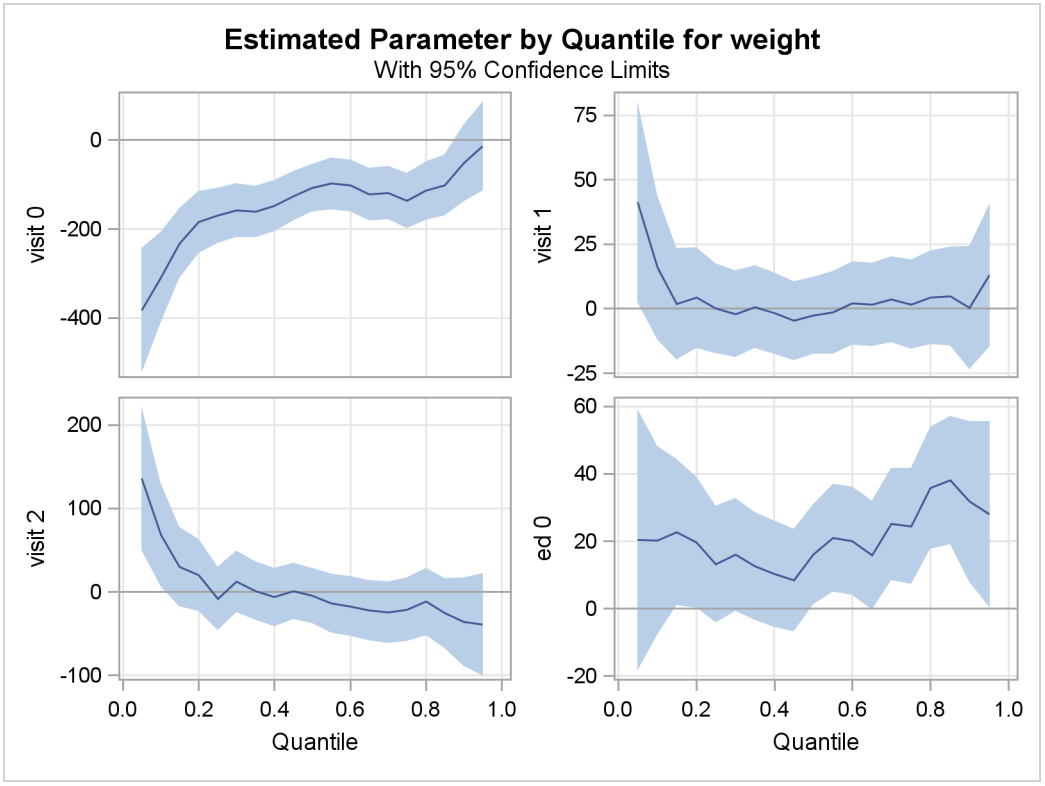
The linear and quadratic effects for the two continuous variables are shown in Output 73.3.5. Both of these variables are centered at their median. At the lower quantiles, the quadratic effect of the mother's age is more concave. The optimal age at the first quantile is about 33, and the optimal age at the third quantile is about 38. The effect of the mother's weight gain is clearly positive, as indicated by the narrow confidence bands for both linear and quadratic coefficients.

Refer to Koenker and Hallock (2001) for more details about the covariate effects discovered with quantile regression.

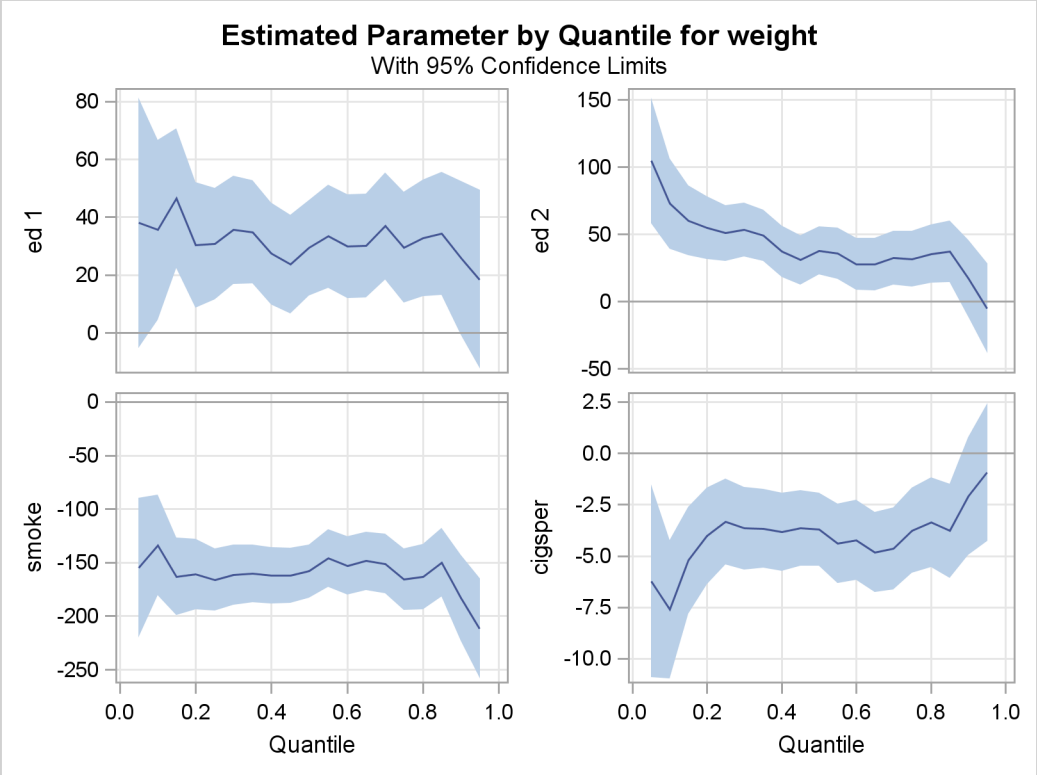
Output 73.3.2 Quantile Processes with 95% Confidence Bands



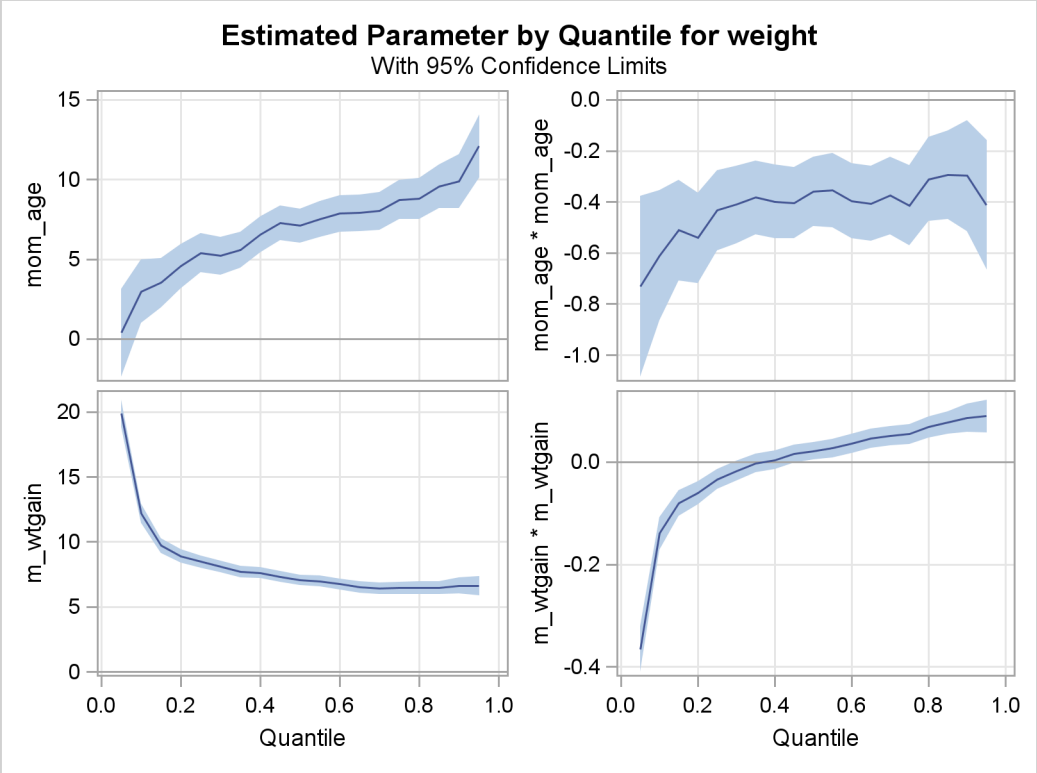
Output 73.3.3 Quantile Processes with 95% Confidence Bands



Output 73.3.4 Quantile Processes with 95% Confidence Bands



Output 73.3.5 Quantile Processes with 95% Confidence Bands



Example 73.4: Nonparametric Quantile Regression for Ozone Levels

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 by using nonparametric quantile regression with cubic B-splines.

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 two variables: ozone (daily-maximum one-hour ozone concentration (ppm)) and days (index of 1,095 days (3 years)).

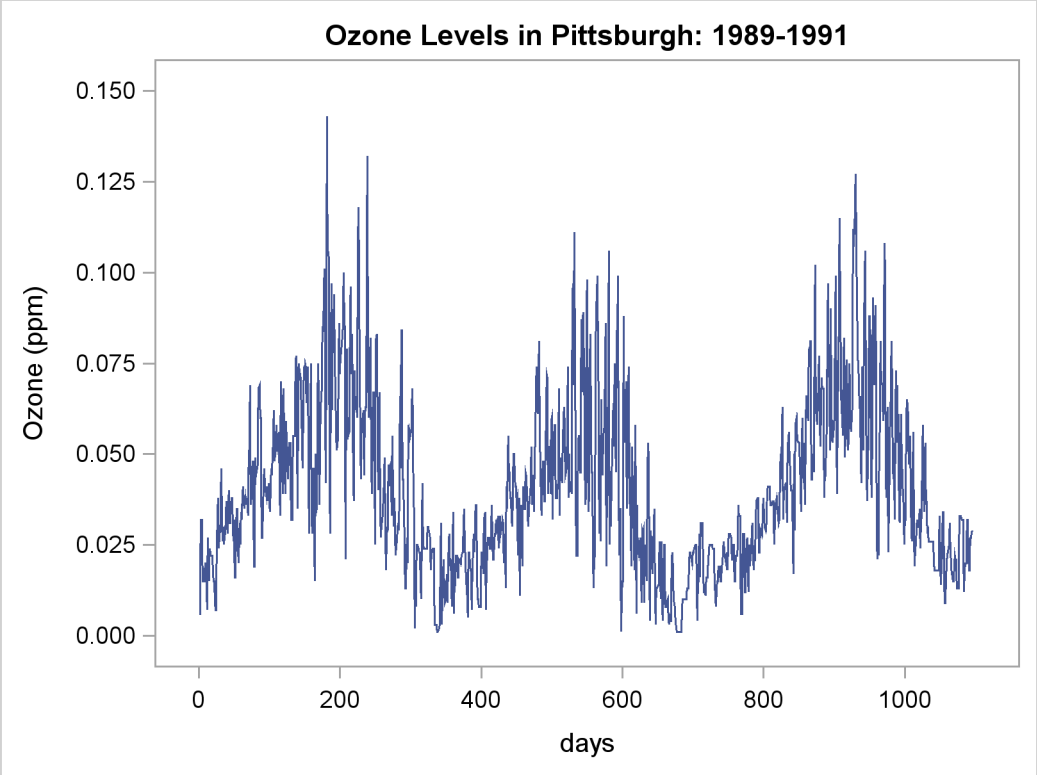
```
data ozone;
  days = _n_;
  input ozone @@;
datalines;
0.0060 0.0060 0.0320 0.0320 0.0320 0.0150 0.0150 0.0150 0.0200 0.0200
0.0160 0.0070 0.0270 0.0160 0.0150 0.0240 0.0220 0.0220 0.0220 0.0185
0.0150 0.0150 0.0110 0.0070 0.0070 0.0240 0.0380 0.0240 0.0265 0.0290

   ... more lines ...

0.0220 0.0210 0.0210 0.0130 0.0130 0.0130 0.0330 0.0330 0.0330 0.0325
0.0320 0.0320 0.0320 0.0120 0.0200 0.0200 0.0200 0.0320 0.0320 0.0250
0.0180 0.0180 0.0270 0.0270 0.0290
;
```

Output 73.4.1, which displays the time series plot of ozone concentration for the three years, shows a clear seasonal pattern.

In this example, 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 create the spline basis and fit multiple quantile regression spline curves:

```
ods graphics on;

proc quantreg data=ozone algorithm=smooth plot=fitplot(nodata);
   effect sp = spline( days / knotmethod = list
          (90 182 272 365 455 547 637 730 820 912 1002) );
   model ozone = sp / quantile = 0.5 0.75 0.90 0.95 seed=1268;
run;

ods graphics off;
```

Output 73.4.1 Time Series of Ozone Levels in Pittsburgh, Pennsylvania

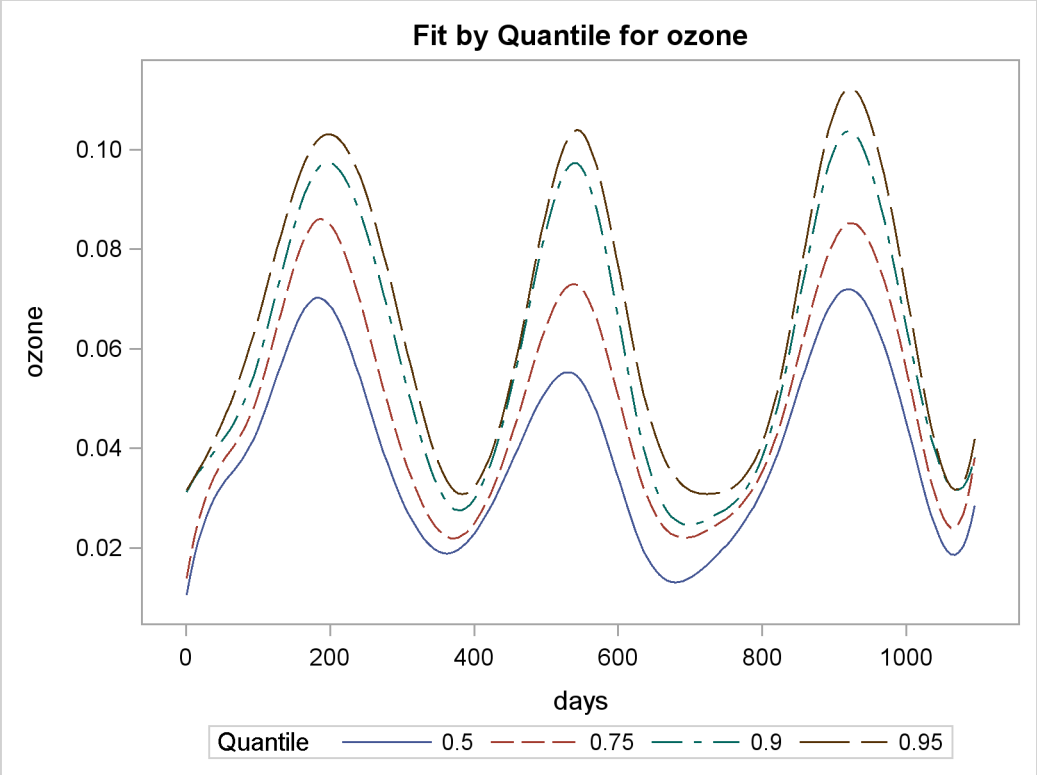


The EFFECT statement creates spline bases for the variable days. The KNOTMETHOD=LIST option provides all internal knots for these bases. Cubic spline bases are generated by default. These bases are treated as components of the spline effect *sp*, which is used in the MODEL statement. Spline fits for four quantiles are requested with the QUANTILE= option.

When you enable ODS Graphics, the QUANTREG procedure automatically generates a fit plot, which includes all fitted curves.

Output 73.4.2 displays these curves obtained with the QUANTREG procedure. The curves show that peak ozone levels occur in the summer. For the three years (1989–1991), the median curve (labeled 50%) does not cross the 0.08 ppm line, which is the 1997 EPA 8-hour standard. The median curve 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 days with low ozone concentration in 1990, but the top 10% and 5% tend to have higher ozone concentration levels.

Output 73.4.2 Quantiles of Ozone Levels in Pittsburgh, Pennsylvania



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.

Example 73.5: Quantile Polynomial Regression for Salary Data

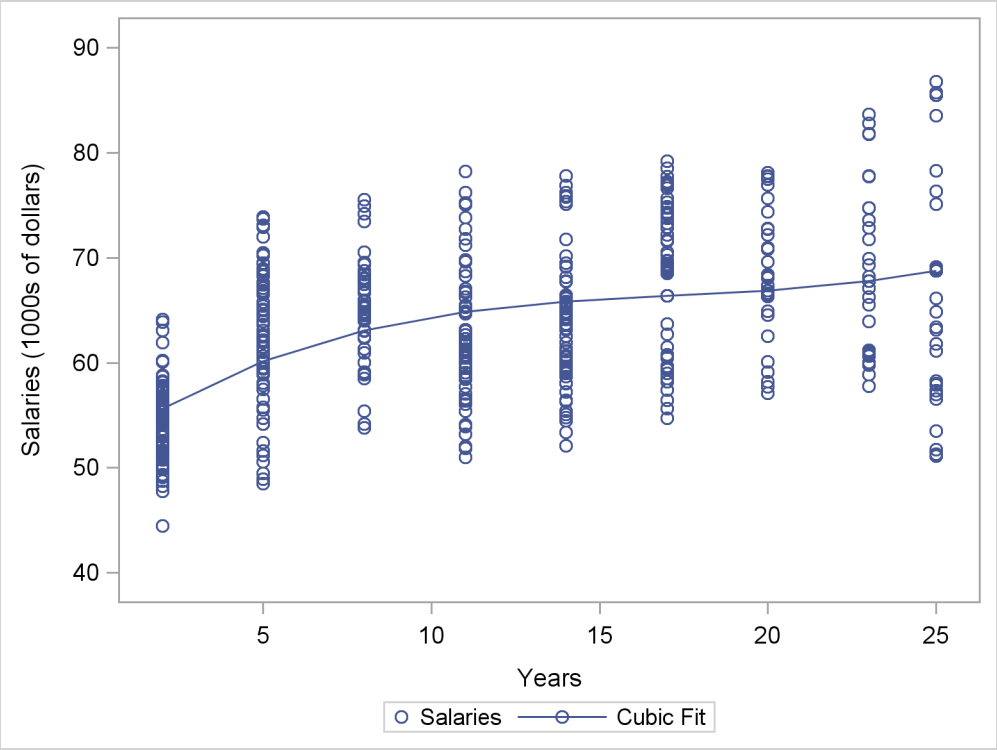
This example uses the data set from a university union survey of salaries of professors in 1991. The survey covered departments in U.S. colleges and universities that list programs in statistics. The goal here is to examine the relationship between faculty salaries and years of service.

The data include salaries and years of service for 459 professors. The scatter plot in Output 73.5.1 shows that the relationship is not linear, and a quadratic or cubic regression curve is appropriate. Output 73.5.1 shows a cubic curve.

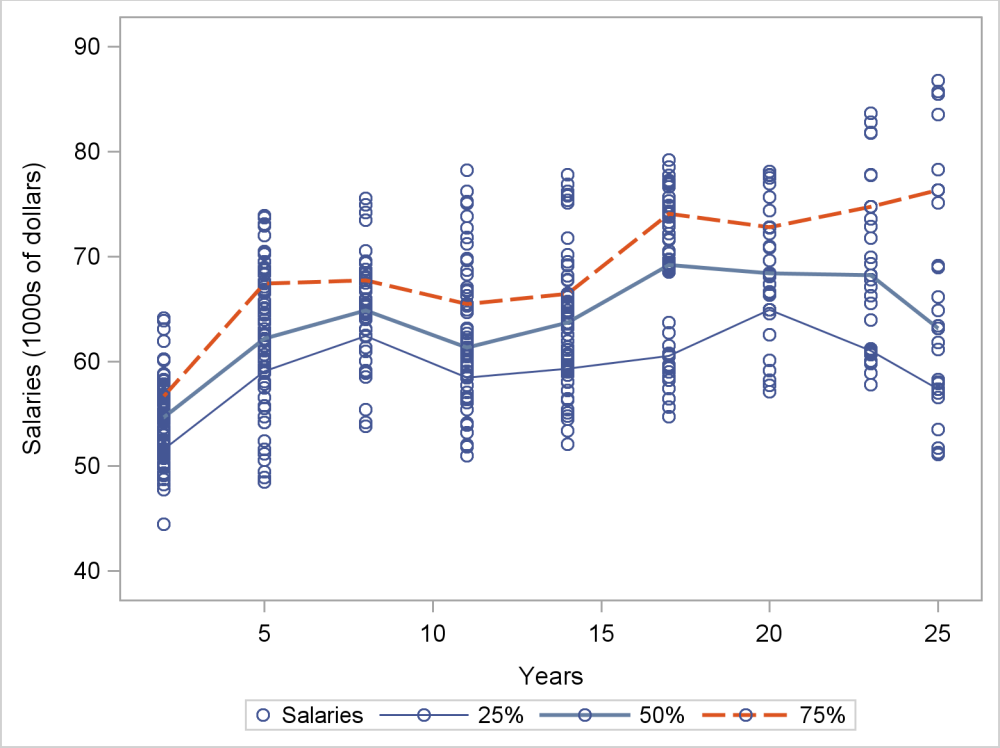
The curve in Output 73.5.1 does not adequately describe the conditional salary distributions and how they change with length of service. Output 73.5.2 shows the 25th, 50th, and 75th percentiles for each number of years, which gives a better picture of the conditional distributions.

```
data salary;
  label salaries='Salaries (1000s of dollars)'
        years   ='Years';
  input salaries years @@;
datalines;
54.94   2  58.24   2  58.11   2  52.23   2  52.98    2  57.62    2
44.48   2  57.22   2  54.24   2  54.79   2  56.42    2  61.90    2
63.90   2  64.10   2  47.77   2  54.86   2  49.31    2  53.37    2

   ... more lines ...

85.72  25  64.87  25  51.76  25  51.11  25  51.31   25  78.28   25
57.91  25  86.78  25  58.27  25  56.56  25  76.33   25  61.83   25
69.13  25  63.15  25  66.13  25
;
```

Output 73.5.1 Salary with Years as Professor: Cubit Fit



Output 73.5.2 Salary with Years as Professor: Sample Quantiles



These descriptive percentiles do not clearly show trends with length of service. The following statements use the QUANTREG procedure to obtain a smooth version by using polynomial quantile regression. The results are shown in Output 73.5.3 and Output 73.5.4.

```
ods graphics on;

proc quantreg data=salary ci=sparsity;
   model salaries = years years*years years*years*years
                    /quantile=0.25 0.5 0.75
                    plot=fitplot(showlimits);
run;

ods graphics off;
```

Output 73.5.3 shows the regression coefficients for the three quantiles.

Output 73.5.3 Regression Coefficients

The QUANTREG Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	48.2509	1.3484	45.6011	50.9007	35.78	<.0001
years	1	2.2234	0.5455	1.1514	3.2953	4.08	<.0001
years*years	1	-0.1292	0.0500	-0.2275	-0.0308	-2.58	0.0101
years*years*years	1	0.0024	0.0013	-0.0001	0.0049	1.86	0.0634

Parameter Estimates

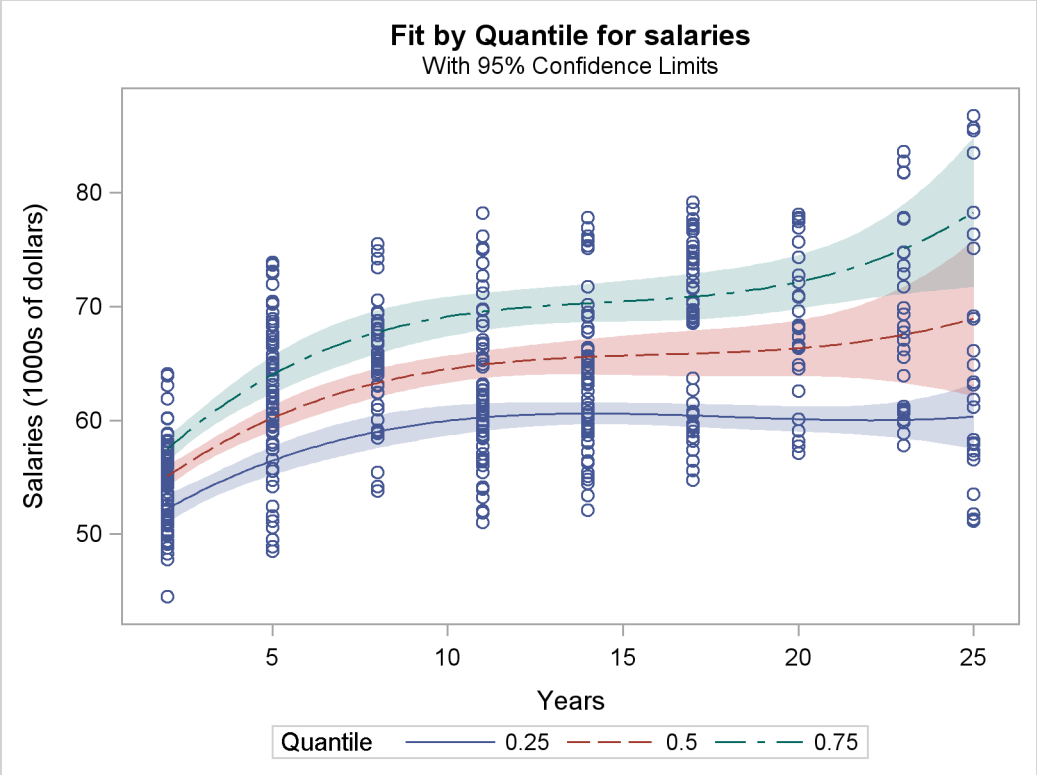
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	50.2512	1.2812	47.7334	52.7690	39.22	<.0001
years	1	2.7173	0.5947	1.5485	3.8860	4.57	<.0001
years*years	1	-0.1632	0.0632	-0.2873	-0.0390	-2.58	0.0101
years*years*years	1	0.0034	0.0018	-0.0002	0.0070	1.85	0.0647

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	51.0298	1.5886	47.9078	54.1517	32.12	<.0001
years	1	3.6513	0.7594	2.1590	5.1436	4.81	<.0001
years*years	1	-0.2390	0.0764	-0.3892	-0.0888	-3.13	0.0019
years*years*years	1	0.0055	0.0021	0.0013	0.0096	2.60	0.0098

Output 73.5.4 displays the three cubic percentile curves with 95% confidence limits.

Output 73.5.4 Salary with Years as Professor: Regression Quantiles



The three curves show that salary dispersion increases gradually with length of service. After 15 years, a salary over $70,000 is relatively high, while a salary less than $60,000 is relatively low. Note that percentile curves of this type are useful in medical science as reference curves; see Yu, Lu, and Stabder (2003).

References

Abreveya, J. (2001), "The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes," *Journal of Economics*, 26, 247–257.

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 Journal of Numerical Analysis,* 10, 839–848.

Bassett, G. W. and Koenker, R. (1982), "An Empirical Quantile Function for Linear Models with iid Errors," *Journal of the American Statistical Association*, 77, 401–415.

Cade, B. S. and Noon B. R. (2003), "A Gentle Introduction to Quantile Regression for Ecologists," *Frontiers in Ecology and the Environment*, 1(8), 412–420.

Chen, C. (2004), "An Adaptive Algorithm for Quantile Regression," *Theory and Applications of Recent Robust Methods*, ed. M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel, 39–48.

Chen, C. (2005), "Growth Charts of Body Mass Index (BMI) with Quantile Regression," *Proceedings of 2005 International Conference on Algorithmic Mathematics and Computer Science*, June 20–23, 2005, Las Vegas, Nevada.

Chen, C. (2007), "A Finite Smoothing Algorithm for Quantile Regression," *Journal of Computational and Graphical Statistics*, 16, 136–164.

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.

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.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning,* New York: Springer-Verlag.

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

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley & Sons.

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

Koenker, R. (1994), "Confidence Intervals for Regression Quantiles," *Asymptotic Statistics*, eds. P. Mandl and M. Huskova, 349–359, New York: Springer-Verlag.

Koenker, R. (2005), *Quantile Regression*, New York: Cambridge University Press.

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. (1994), "Remark AS R92: A Remark on Algorithm AS 229: Computing Dual Regression Quantiles and Regression Rank Scores," *Applied Statistics*, 43, 410–414.

Koenker, R. (1995), "Rank Tests for Linear Models," *The Handbook of Statistics*, 15, edited by C.R. Rao and G.S. Madalla.

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.

Koenker, R. and Zhao, Q. (1994), "L-Estimation for Linear Heteroscedastic Models," *Journal of Nonparametric Statistics*, 3, 223–235.

Kuczmarski, R. J., Ogden, C. L., Guo, S. S., et al. (2002), "2000 CDC Growth Charts for the United States: Methods and Development," *Vital Health Stat.*, 11, 246, 1–190.

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

Madsen, K. and Nielsen, H. B. (1993), "A Finite Smoothing Algorithm for Linear L_1 Estimation," *SIAM Journal on 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.

Portnoy, S. and Koenker, R. (1997), "The Gaussian Hare and the Laplacian Tortoise: Computation of Squared-Error vs. Absolute-Error Estimators," *Statistical Science*, 12, 279–300.

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.

Yu, K., Lu, Z., and Stabder, J. (2003), "Quantile Regression: Application and Current Research Areas," *The Statistician*, 52, 331–350.

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.