

The QUANTREG Procedure (Experimental)

Contents

OVERVIEW	3
Quantile Regression	6
GETTING STARTED	7
Analysis of Fish-Habitat Relationships	7
Growth Charts for Body Mass Index	13
SYNTAX	16
PROC QUANTREG Statement	16
BY Statement	19
CLASS Statement	19
ID Statement	20
MODEL Statement	20
OUTPUT Statement	23
PERFORMANCE Statement	24
TEST Statement	24
WEIGHT Statement	25
DETAILS	25
Quantile Regression as an Optimization Problem	25
Optimization Algorithms	26
Confidence Interval	34
Covariance-Correlation	38
Linear Test	38
Leverage Point and Outlier Detection	39
INEST= Data Set	40
OUTEST= Data Set	40
Computational Resources	41
ODS Table Names	42
ODS Graphics (Experimental)	42
EXAMPLES	44
Example 1. Comparison of Algorithms	44
Example 2. Quantile Regression for Econometric Growth Data	48
Example 3. Quantile Regression Analysis of Birth-Weight Data	54
Example 4. Nonparametric Quantile Regression for Ozone Levels	58

2 ♦ *The QUANTREG Procedure (Experimental)*

Example 5. Quantile Polynomial Regression for Salary Data	61
REFERENCES	64
SUBJECT INDEX	67
SYNTAX INDEX	69

The QUANTREG Procedure (Experimental)

Overview

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 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.

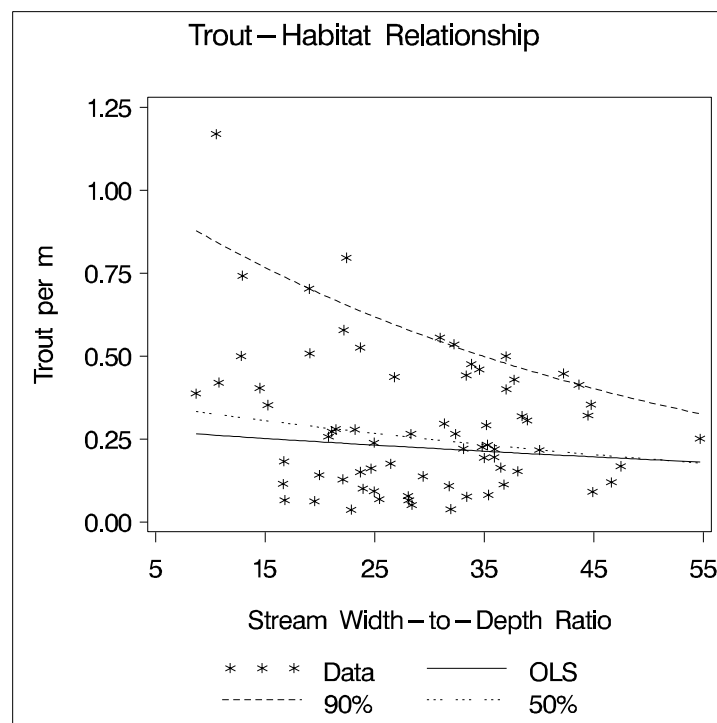


Figure 1. Trout Density in Streams

4 ♦ The QUANTREG Procedure (Experimental)

As analyzed by Dunham, Cade, and Terrell (2002), trout density depends on a number of unmeasured limiting factors in addition to the ratio, which is 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 using a simple linear regression model for the logarithm of density (The curves in Figure 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 7.) 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.

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 2, which plots the body mass index for 8,250 men versus their age.

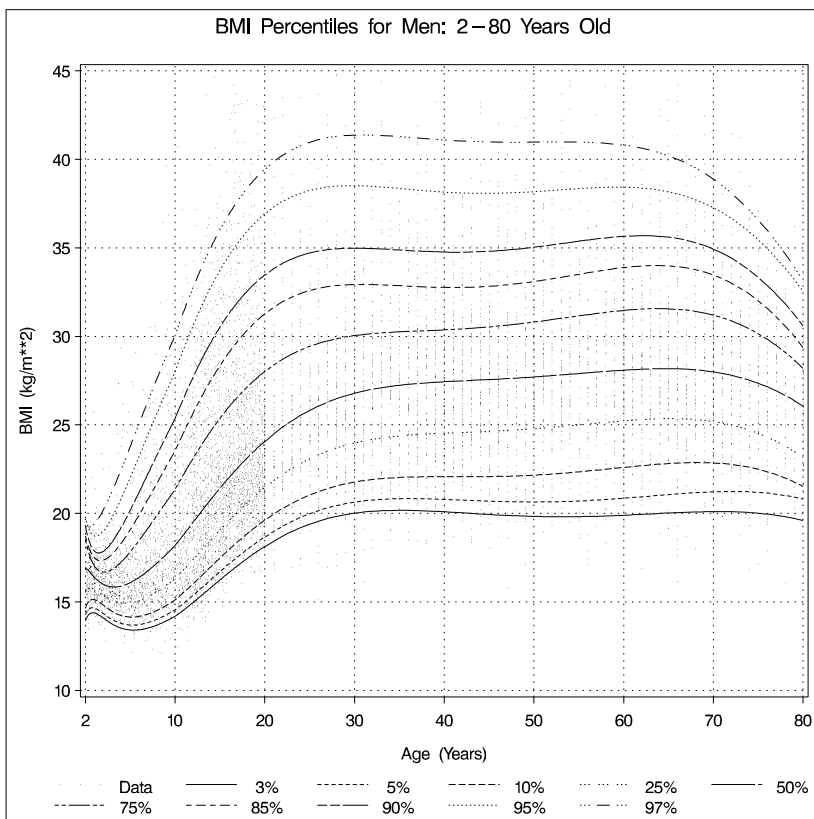


Figure 2. Quantiles for Body Mass Index

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 2](#) were determined with the QUANTREG procedure using polynomial quantile regression; details are provided in the section “[Growth Charts for Body Mass Index](#)” on page 13. Clearly, the rate of change with age (as expressed by the regression coefficients) 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 ability 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.

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.

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

Note: We are eager for your feedback on this experimental procedure. Please send comments to quantreg@sas.com.

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 τ th quantile if his (or her) grade is better than $100\tau\%$ of the students who took the test. The score is also said to be at the 100τ th percentile.

For a random variable Y with probability distribution function

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

the τ th 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 $\{y_1, \dots, y_n\}$ 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 τ th sample quantile $\xi(\tau)$, which is the analogue 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 τ th 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 any quantile $\tau \in (0, 1)$. The quantity $\hat{\beta}(\tau)$ is called the τ th *regression quantile*. The case $\tau = 1/2$, which minimizes the sum of absolute residuals, corresponds to median regression, which is also known as L_1 regression.

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 $\hat{\beta}(\tau)$. The procedure also provides diagnostic plots using ODS Graphics, which is experimental in SAS 9.1. For more information, see the section “[ODS Graphics](#)” on page 42.

Getting Started

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 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 MODEL statement specifies a simple linear regression model with LnDensity as the response variable Y and WDRatio as the covariate X . The option QUANTILE=0.9 requests that the regression quantile function $Q(0.9|X = x) = x'\beta(0.9)$ is to be estimated by solving

$$\hat{\beta}(0.9) = \arg \min_{\beta \in \mathbf{R}^2} \sum_{i=1}^n \rho_{0.9}(y_i - x'_i \beta)$$

By default, the regression coefficients $\hat{\beta}(0.9)$ are estimated with the simplex algorithm, which is explained in the section “[Simplex Algorithm](#)” on page 26. The option ALPHA=0.01 requests 99% 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 using the MCMB resampling method. The SEED= option specifies a seed for the resampling method. The options COVB and CORRB request covariance and correlation 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 3](#) displays model information and summary statistics for the variables in the model. The summary statistics include the median and the median absolute deviation (MAD), which are robust measures of univariate location and scale, respectively.

The QUANTREG Procedure						
Model Information						
Data Set			WORK.TROUT			
Dependent Variable			LnDensity	LOG(Density)		
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 3. Model Fitting Information and Summary Statistics

[Figure 4](#) displays the parameter estimates, standard errors, 99% confidence limits, t values, and p-values computed by the resampling method.

The QUANTREG Procedure							
Quantile and Objective Function							
Quantile							0.9
Objective Function							7.2303
Predicted Value at Mean							-0.5709
Parameter Estimates							
Parameter	DF	Estimate	Standard Error	99% Confidence Limits		t Value	Pr > t
Intercept	1	0.0576	0.2727	-0.6648	0.7801	0.21	0.8333
WDRatio	1	-0.0215	0.0073	-0.0408	-0.0022	-2.96	0.0042

Figure 4. Parameter Estimates

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 1](#). The lower dashed curve for the median can be obtained by changing the option QUANTILE=0.9 to QUANTILE=0.5.

The covariance and correlation matrices for the estimated parameters are shown in [Figure 5](#). The resampling method used for the confidence intervals is used to compute these matrices.

The QUANTREG Procedure			
Estimated Covariance Matrix			
	Intercept	WDRatio	
Intercept	0.074384	-.001934	
WDRatio	-.001934	0.000053	
Estimated Correlation Matrix			
	Intercept	WDRatio	
Intercept	1.00000	-0.97357	
WDRatio	-0.97357	1.00000	

Figure 5. Covariance and Correlation

The tests requested with the TEST statement are shown in [Figure 6](#). Both the Wald test and the Likelihood Ratio test indicate that the coefficient of width-to-depth ratio is significantly different from zero.

The QUANTREG Procedure				
Tests				
Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
Wald	8.7468	1	8.75	0.0031
Likelihood Ratio	9.0528	1	9.05	0.0026

Figure 6. Tests of Significance

In many quantile regression problems it is useful to examine how the estimated regression parameters for each covariate change as a function of the quantile 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.

```
ods html;
ods graphics on;
proc quantreg data=trout alpha=0.1 ci=resampling;
  model LnDensity = WDRatio / quantile=all seed=12345
    plot=quantplot;
run;
ods graphics off;
ods html close;
```

The QUANTILE=ALL option requests an estimate of the quantile process for each regression parameter, which is computed with the default simplex algorithm. 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 7 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.

The QUANTREG Procedure				
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 7. Objective Function

Figure 8 displays a portion of the table of the quantile processes for the estimated parameters and confidence limits.

The QUANTREG Procedure			
Parameter Estimates for Quantile Process			
Label	Quantile	Intercept	WDRatio
.	.	.	.
.	.	.	.
.	.	.	.
t57	0.765705	-0.4220	-0.0133
lower	0.765705	-0.9418	-0.0270
upper	0.765705	0.0977	0.0003
t58	0.786206	-0.3269	-0.0159
lower	0.786206	-0.8336	-0.0298
upper	0.786206	0.1798	-0.0020
.	.	.	.
.	.	.	.
.	.	.	.

Figure 8. Quantile Processes

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 9 displays the process for the intercept, and the right side displays the process for the coefficient of WDRatio.

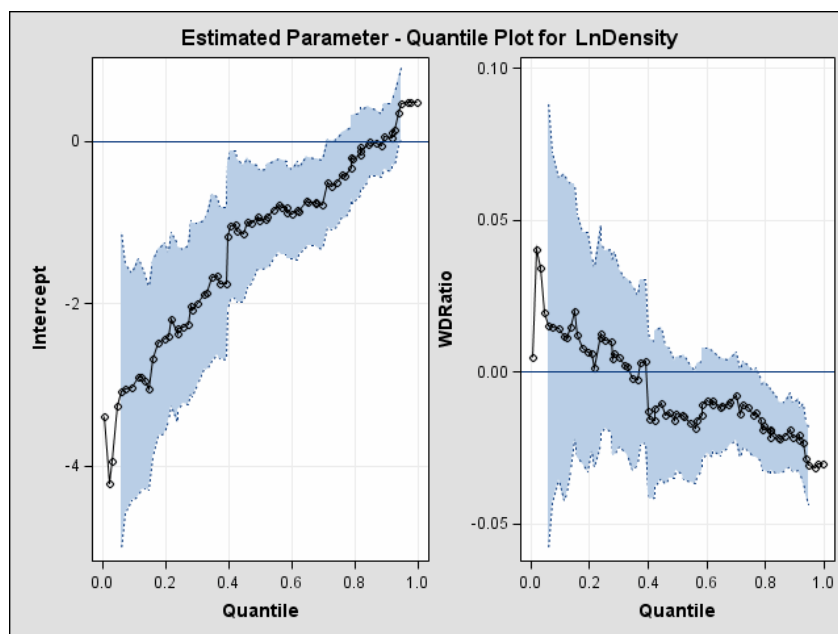


Figure 9. Quantile Processes for Intercept and Slope

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 8](#). 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).

Growth Charts for Body Mass Index

Body mass index is defined as the ratio of weight (kg) to squared height (m²) and is a widely used measure of overweight or underweight. The percentiles of BMI for specified ages are of particular interest. As age increases, these percentiles provide growth patterns of BMI not only for the majority of the population, but also for underweight or overweight extremes of the population. In addition, the percentiles of BMI for a specified age provide a reference for individuals at that age with respect to the population.

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/m²), AGE (year), and SEQN (respondent sequence number) for 8,250 men. More details about the data can be found in [Chen \(2005\)](#).

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.

The following statements invoke the QUANTREG procedure:

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

The MODEL statement provides the model, and the option QUANTILE=0.5 requests median regression, which computes $\hat{\beta}(\frac{1}{2})$ using the interior point algorithm as requested with the ALGORITHM= option. See the section “[Interior Point Algorithm](#)” on page 28 for details about this algorithm.

[Figure 10](#) displays the estimated parameters, standard errors, 95% confidence intervals, t values, and p-values which 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.

The QUANTREG Procedure							
Parameter Estimates							
Parameter	DF	Estimate	Standard		t Value	Pr > t	
			Error	95% Confidence Limits			
Intercept	1	6.41816705	0.59781166	5.24630565	7.59002846	10.74	<.0001
inveage	1	-1.1339904	0.33581906	-1.7922803	-.47570047	-3.38	0.0007
sqrtage	1	-3.7649349	0.50595055	-4.7567254	-2.7731444	-7.44	<.0001
age	1	1.46718520	0.17443918	1.12524047	1.80912992	8.41	<.0001
sqrtage*age	1	-.24610559	0.02837402	-.30172581	-.19048537	-8.67	<.0001
age*age	1	0.01643716	0.00190722	0.01269854	0.02017579	8.62	<.0001
age*age*age	1	-.00003114	0.00000378	-.00003855	-.00002373	-8.23	<.0001

Figure 10. Parameter Estimates with Median Regression: Men

The TEST statement requests Wald and likelihood ratio tests for the significance of the cubic term in AGE. The test results, shown in Figure 11, indicate that this term is significant. Higher-order terms are not significant.

The QUANTREG Procedure				
Tests				
TEST_AGE_CUBIC				
Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
Wald	72.5749	1	72.57	<.0001
Likelihood Ratio	56.2815	1	56.28	<.0001

Figure 11. Test of Significance for Cubic Term

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 12, 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, 22 men are identified as outliers. All of these men have large positive standardized residuals, which indicates that they are overweight for their age. The cutoff value 4.5 is ad hoc; it corresponds to a probability less than $0.5E-5$ if normality is assumed, but the standardized residuals for median regression usually do not meet this assumption.

In order to construct the chart shown in Figure 2, the same model used for median regression is used for other quantiles. Note that the QUANTREG procedure computes fitted values only for a single quantile at a time.

The QUANTREG Procedure							
Diagnostics							
Obs	SEQN	age	weight	height	bmi	Standardized Residual	Outlier
1337	13275	8.916667	73.600000	142.100000	36.450000	4.5506	*
1376	2958	9.166667	67.500000	130.500000	39.640000	5.0178	*
1428	19390	9.416667	70.300000	138.100000	36.860000	4.5122	*
1572	19814	10.250000	72.900000	133.800000	40.720000	4.9485	*
1903	15305	12.000000	143.600000	162.600000	54.310000	6.3591	*
2356	12567	13.500000	114.900000	162.300000	43.620000	4.6933	*
2562	6177	14.333333	123.200000	166.100000	44.660000	4.6809	*
2746	18352	14.916667	117.100000	158.200000	46.790000	4.8641	*
2967	710	15.750000	130.440000	165.300000	47.740000	4.8448	*
3090	2079	16.166667	148.600000	171.700000	50.410000	5.1141	*
3342	1874	17.000000	168.800000	181.800000	51.070000	5.0644	*
3424	17793	17.166667	176.000000	182.100000	53.080000	5.2791	*
3486	7095	17.416667	153.700000	171.300000	52.380000	5.1599	*
3559	903	17.666667	174.600000	172.600000	58.610000	5.8216	*
3686	10568	18.083333	153.700000	175.600000	49.850000	4.7583	*
3858	12027	18.666667	171.500000	180.400000	52.700000	5.0257	*
4347	14686	21.000000	196.800000	193.700000	52.450000	4.7264	*
5273	2304	35.000000	193.300000	178.200000	60.870000	4.9920	*
5669	9031	40.000000	177.200000	174.600000	58.130000	4.6614	*
6209	17923	46.000000	174.100000	174.000000	57.500000	4.5603	*
6282	19911	47.000000	188.300000	172.900000	62.990000	5.1203	*
6366	11309	49.000000	171.300000	163.400000	64.160000	5.2209	*

Diagnostics Summary		
Observation Type	Proportion	Cutoff
Outlier	0.0027	4.5000

Figure 12. Diagnostics with Median Regression

When fitted values are required for multiple quantiles, you can use the following macro.

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

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

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

The 10 output data sets are merged, and the fitted BMI values together with the original BMI values are plotted against AGE to create the display shown in [Figure 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 an effective way to control overweight in a population is to start in childhood.

Compared to the 97th percentile in reference growth charts published by CDC in 2000, the 97th percentile for 10-year-old boys in [Figure 2](#) is 6.4 BMI units higher (an increase of 27%). This can be interpreted as a warning of overweight or obesity. Refer to Chen (2005) for a detailed analysis.

Syntax

```
PROC QUANTREG < options > ;
  BY variables ;
  CLASS variables ;
  ID variables ;
  MODEL response = independents < / options > ;
  OUTPUT < OUT= SAS-data-set > < options > ;
  PERFORMANCE < options > ;
  TEST 'label' 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. 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 on 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 algorithm to estimate the regression parameters. Three algorithms are available: simplex (SIMPLEX), interior point (INTERIOR), and smoothing (SMOOTH). The default algorithm is simplex. The following table summarizes these methods.

Table 1. Options for Estimation Algorithms

ALGORITHM= Value	Algorithm	Suboptions
SIMPLEX	simplex	
INTERIOR	interior point	KAPPA= MAXIT= TOLERANCE=
SMOOTH	smoothing	RRATIO=

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, the step length parameter should not be too large to produce floating point errors for the algorithm. By default, KAPPA = 0.99995. See the section “[Interior Point Algorithm](#)” on page 28 for details.
- MAXIT=*n* sets the maximum number of iterations for the interior point algorithm. By default, *n*=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 28 for details.

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 31 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 2. 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, the QUANTREG procedure computes confidence intervals using the rank method for data sets with less than 5,000 observations and less than 20 variables; otherwise, the default is the resampling method.

By default, confidence intervals are not computed for the quantile process, which is estimated when you specify the QUANTILE=ALL 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 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 i.i.d. By default, the sparsity function is estimated by assuming that the conditional quantile function is locally linear. See the section “[Sparsity](#)” on page 34 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 40 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 sorting order for the levels of the classification variables (specified in the CLASS statement). This ordering determines which parameters in the model correspond to each level in the data. The following table illustrates how the QUANTREG procedure interprets values of the ORDER= option.

Table 3. Options for Order

Value of ORDER=	Levels Sorted By
DATA	order of appearance in the input data set
FORMATTED	formatted 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 on sorting order, refer to the chapter titled “The SORT Procedure” in the *SAS Procedures Guide*.

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 40 for a detailed description of the contents of the OUTEST= data set.

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 only be used for very large data sets, for example, data sets with over 100,000 observations. See Portnoy and Koenker (1997) for details.

BY Statement

BY *variables* ;

You can specify a BY statement with the QUANTREG procedure to obtain separate analysis on observations in groups 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 your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING 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 using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

CLASS Statement

CLASS *variables* ;

Variables that are classification variables rather than quantitative numeric variables must be listed in the CLASS statement. For each explanatory variable listed in the CLASS statement, indicator variables are generated for the levels assumed by the CLASS variable. The parameterization of the CLASS variable is the same as in the GLM procedure. If the CLASS statement is used, it must appear before the MODEL statement.

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 beside 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. Class variables in the MODEL statement must be specified in the CLASS statement.

The optional *label* 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; otherwise it computes the asymptotic correlation based on an estimator of the sparsity function.

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; otherwise it computes the asymptotic covariance based on an estimator of the sparsity function.

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 $\sqrt{\chi_{p,1-\alpha}^2}$, where

α can be specified with the CUTOFFALPHA= option. By default, $\alpha = .025$. 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, $H = \lceil (3n + p + 1)/4 \rceil$, where n is the number of observations and p is the number of independent variables. The LEVERAGE option is ignored if the model includes class 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.

QUANTILE=number list | ALL

specifies the quantiles for the quantile regression. You can specify any number of quantiles in $(0, 1)$. You can also compute the entire quantile process by specifying the ALL option. Only the simplex algorithm is available for computing the quantile process. Statistical inference is available only when a single quantile is specified.

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 39 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. 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.

PLOT=plot option**PLOTS=(plot options)**

You can use the PLOT (or PLOTS) option in the MODEL statement together with the ODS GRAPHICS statement to request various graphical displays. To request these plots you must specify the ODS GRAPHICS statement in addition to these options in the MODEL statement. For more information on the ODS GRAPHICS statement, see Chapter 15, “Statistical Graphics Using ODS” (*SAS/STAT User’s Guide*).

The following plot options are available.

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

creates a plot of Robust Distance against Mahalanobis Distance. See the section “Leverage Point and Outlier Detection” on page 39 for details about the Robust Distance. The LABEL= option specifies how the points on this plot are to be labeled, as summarized by the following table.

Table 4. Options for Label

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

By default, LABEL=ALL.

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.

RESHISTOGRAM

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.

RESQQPLOT

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

QUANTPLOT<(EFFECTS) </ <NOBANDS> <UNPACKPANEL> > >

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 NOBANDS 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. You can use the UNPACKPANEL option to create individual process plots.

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

creates the plot of standardized residual against Robust Distance. See the section “Leverage Point and Outlier Detection” on page 39 for details about the Robust Distance. The LABEL= option specifies a label method for points on this plot. These label methods are described in Table 4 on page 22.

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.

OUTPUT Statement

```
OUTPUT <OUT=SAS-data-set> keyword=name <...keyword=name>
;
```

When you specify a single quantile with the QUANTILE= option in the MODEL statement, the OUTPUT statement creates a SAS data set containing statistics calculated after fitting the model. 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 on permanent SAS data sets).

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 39 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 39 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.

RESIDUAL | RES specifies a variable to contain the residuals

$$y_i - \mathbf{x}'_i \mathbf{b}$$

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.

SRESIDUAL | SR specifies a variable to contain the standardized residuals

$$\frac{y_i - \mathbf{x}_i' \mathbf{b}}{\sigma}$$

See the section “Leverage Point and Outlier Detection” on page 39 for how to compute σ .

STDP specifies a variable to contain the estimates of the standard errors of the estimated response. To request this variable, you need to specify either COVB or CORRB in the MODEL statement.

PERFORMANCE Statement

You use the PERFORMANCE statement to specify options that tune the performance of the QUANTREG procedure. By default these options are chosen to maximize performance.

PERFORMANCE < options > ;

The following option is available:

CPUCOUNT=*n*

specifies the number of threads to use in the computation of the interior point algorithm. By default this number equals to the number of processors on the hardware.

TEST Statement

<label:> **TEST** effects </WALD | LR > ;

The TEST statement provides a means for obtaining a test for the canonical linear hypothesis concerning the model parameters:

$$\beta_j = 0, \quad j = q + 1, \dots, p$$

where p is the total number of parameters in the model, and $p - q$ is the number of parameters for the testing of significance.

You can request Wald tests with the WALD option and likelihood ratio tests with the LR option. By default, only Wald tests are provided. See the “Details” section for more information about these tests.

The optional *label* is used to label 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 “Details” for more information about weighted quantile regression.

Details

Quantile Regression as an Optimization Problem

The model for linear quantile regression is

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

where $y = (y_1, \dots, y_n)'$ is the $(n \times 1)$ vector of responses, $A' = (x_1, \dots, x_n)'$ is the $(n \times p)$ regressor matrix, $\beta = (\beta_1, \dots, \beta_p)'$ is the $(p \times 1)$ vector of unknown parameters, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ 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 τ th regression quantile, $0 < \tau < 1$, as any solution to the minimization problem

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

The solution is denoted as $\hat{\beta}(\tau)$, and the L_1 -norm estimate corresponds to $\hat{\beta}(1/2)$. The τ th regression quantile is an extension of the τ th 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 $w_i, i = 1, \dots, n$ with the WEIGHT statement, weighted quantile regression is carried out by solving

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

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

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

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

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 $\mathbf{0}' = (0 \ 0 \ \dots \ 0)_p$. The reformulation presents a standard LP problem:

$$(P) \quad \min_{\theta} d' \theta$$

$$\text{subject to} \quad B\theta = y$$

$$\theta \geq 0$$

This problem has the dual formulation

$$(D) \quad \max_z y' z$$

$$\text{subject to} \quad B' z \leq d$$

which can be simplified as

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

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

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

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

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

The QUANTREG procedure solves this LP problem using the simplex algorithm of Barrodale and Roberts (1973). This algorithm solves the primary LP problem (P) by two stages, which exploit the special structure of the coefficient matrix B . The first stage 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. 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 1993). It greatly reduces the computing time required by the general simplex algorithm, and it is suitable for data sets with less than 5,000 observations and 50 variables.

Interior Point Algorithm

There are many variations of interior point algorithms. The QUANTREG procedure uses the primal-dual 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

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

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

$$\begin{aligned} Az &= b \\ z + v &= u \\ A't + s - w &= c \\ ZSe &= 0 \\ VWe &= 0 \\ z, s, v, w &\geq 0 \end{aligned}$$

$$\begin{array}{ll} \text{where} & W = \text{diag}(w), \text{ (that is, } W_{i,j} = w_i \text{ if } i = j, W_{i,j} = 0 \text{ otherwise)} \\ & V = \text{diag}(v), Z = \text{diag}(z), S = \text{diag}(s) \end{array}$$

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

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

$$0 = z'_{opt}s_{opt} = s'_{opt}z_{opt} = (c - A't_{opt} + w_{opt})'z_{opt} =$$

$$c'z_{opt} - t'_{opt}(Az_{opt}) + w_{opt}'z_{opt} = c'z_{opt} - b't_{opt} + u'w_{opt}$$

$$\text{therefore} \quad 0 = c'z_{opt} - b't_{opt} + u'w_{opt}$$

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:

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

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

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

$$\Delta s_{aff} = -Se - Z^{-1}S\Delta z_{aff}$$

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

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 below), 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 $(z_{aff}, t_{aff}, s_{aff}, v_{aff}, w_{aff})$:

$$\begin{aligned} \Delta z_c + \Delta v_c &= 0 \\ A\Delta z_c &= 0 \\ A'\Delta t_c + \Delta s_c - \Delta w_c &= 0 \\ S\Delta z_c + Z\Delta s_c &= -Z_{aff}S_{aff}e + \sigma\mu e \\ V\Delta w_c + W\Delta v_c &= -V_{aff}W_{aff}e + \sigma\mu e \end{aligned}$$

where

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

Therefore, the computations involved in solving the centering step are

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

Then

$$\begin{aligned}
 (\Delta z, \Delta t, \Delta s, \Delta v, \Delta w) &= \\
 (\Delta z_{aff}, \Delta t_{aff}, \Delta s_{aff}, \Delta v_{aff}, \Delta w_{aff}) &+ \\
 (\Delta z_c, \Delta t_c, \Delta s_c, \Delta v_c, \Delta w_c) & \\
 \\
 (z^{k+1}, t^{k+1}, s^{k+1}, v^{k+1}, w^{k+1}) &= \\
 (z^k, t^k, s^k, v^k, w^k) &+ \\
 \kappa(\Delta z, \Delta t, \Delta s, \Delta v, \Delta w) &
 \end{aligned}$$

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

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

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 (2003). 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

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

The function $H_{\gamma,\tau}$ is determined by whether $r_i(\beta) \leq (\tau - 1)\gamma$, $r_i(\beta) \geq \tau\gamma$, or $(\tau - 1)\gamma \leq r_i(\beta) \leq \tau\gamma$. 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 $B_{\gamma,\tau}$:

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

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

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

and introduce

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

thus,

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

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

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

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

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

$$D_{\gamma,\tau}^{(2)}(\beta) = \frac{1}{\gamma} AW_{\gamma,\tau}(\beta)A'$$

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 $\beta \in \mathbf{R}^p \setminus B_{\gamma,\tau}$ 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 $s_\gamma = s$,

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

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 is 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 $D_{\rho\tau}$ 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 utilized. Finally, when γ is small enough, a solution can be found by the modified Newton-Ralphon 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= option in the PROC statement. By default,

$$\text{RRATIO} = \begin{cases} 0.1 & \text{if } n \geq 10000 \text{ and } p \leq 20 \\ 0.9 & \text{if } \frac{p}{n} \geq .1 \text{ or } \{n \leq 5000 \text{ and } p \geq 300\} \\ 0.5 & \text{otherwise} \end{cases}$$

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.

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 using a local estimate of the sparsity function. The rank method, which computes confidence intervals by inverting the rank score test, does not suffer from this problem, but it uses the simplex algorithm and is computationally expensive with large data sets. The resampling method, which uses the bootstrap, can overcome all of these problems, but it is unstable for small data sets.

Based on these properties, the QUANTREG uses a combination of the 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 $\{\epsilon_i\}$, $i = 1, \dots, n$, are i.i.d. with a distribution F and a density $f = F'$, where $f(F^{-1}(\tau)) > 0$ in a neighborhood of τ . Under some mild conditions

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

where $\omega^2(\tau, F) = \tau(1 - \tau)/f^2(F^{-1}(\tau))$ and $\Omega = \lim_{n \rightarrow \infty} n^{-1} \sum x_i x_i'$. 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}_n^{-1} can be estimated by the empirical quantile function of the residuals from the quantile regression fit,

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

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

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

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 i.i.d. assumption. Alternately, Koenker and Machado (1999) considered the non i.i.d. case. By assuming local linearity of the conditional quantile function $Q(\tau|x)$ in x , they proposed a local estimator of the density function 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 i.i.d. confidence intervals. You can request i.i.d. 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). By inverting this test, confidence intervals can be computed for the regression quantiles corresponding to β_2 .

The rankscore function $\hat{a}_n(t) = (\hat{a}_{n1}(t), \dots, \hat{a}_{nn}(t))$ 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 $H_0: \beta_2 = \eta$

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

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

Resampling

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

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

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

which is the estimating equation for the τ th regression quantile, is a pivotal quantity for the τ th quantile regression parameter β_τ . In other words, the distribution of $S(\beta)$ may be generated exactly by a random vector U , which is a weighted sum of independent, re-centered Bernoulli variables. They further showed that for large n the distribution of $\hat{\beta}(\tau) - \beta_\tau$ can be approximated by the conditional distribution of $\hat{\beta}_U - \hat{\beta}_n(\tau)$, where $\hat{\beta}_U$ solves an augmented quantile regression problem with $n+1$ 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. Since resampling methods achieve stability only for relatively large data sets, they are not recommended for small data sets ($n < 5000$ and $p < 20$).

Covariance-Correlation

You can specify the COVB and CORR options in the MODEL statement to request covariance and correlation matrices for the estimated parameters. The QUANTREG procedure computes these matrices only for a single quantile at a time.

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. Otherwise, asymptotic covariance and correlation matrices are computed.

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 i.i.d. and non i.i.d. estimates. By default, the QUANTREG procedure computes non i.i.d. estimates. Since the rank method does not provide a covariance-correlation estimate, the asymptotic covariance-correlation is computed when the confidence intervals are computed using this method.

Bootstrap Covariance-Correlation

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

Linear Test

Two tests are available in the QUANTREG procedure for the linear null hypothesis $H_0 : \beta_2 = 0$. Here β_2 denotes a subset of the parameters, where the parameter vector $\beta(\tau)$ is partitioned as $\beta'(\tau) = (\beta'_1(\tau), \beta'_2(\tau))$, and the covariance matrix Ω for the parameter estimates is partitioned correspondingly as Ω_{ij} with $i = 1, 2; j = 1, 2$; and $\Omega^{22} = (\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12})^{-1}$.

The Wald test 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 $D_0(\tau) = \sum \rho_\tau(y_i - x_i\hat{\beta}(\tau))$, $D_1(\tau) = \sum \rho_\tau(y_i - x_{1i}\hat{\beta}_1(\tau))$, 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.

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

Leverage Point and Outlier Detection

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

Mahalanobis distance is defined as

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

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{C} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})'(x_i - \bar{x})$. Here, $x_i = (x_{i1}, \dots, x_{i(p-1)})'$ 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 $C(p) = \sqrt{\chi_{p;1-\alpha}^2}$ be the cutoff value. The variable LEVERAGE is defined as

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

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

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

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

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

<code>_MODEL_</code>	a character variable of length 8 containing the label of the MODEL statement, if present. Otherwise, the variable's value is blank.
<code>_ALGORITHM_</code>	a character variable of length 8 containing the name of the algorithm used for computing the parameter estimates, either SIMPLEX, or INTERIOR, or SMOOTH
<code>_TYPE_</code>	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.
<code>_STATUS_</code>	a character variable of length 12 containing the status of model fitting, either NORMAL, or NOUNIQUE, or NOVALID
INTERCEPT	a numeric variable containing the intercept parameter estimates
<code>_QUANTILE_</code>	a numeric variable containing the specified quantiles

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

Computational Resources

The various different 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 5. ODS Tables Produced in PROC QUANTREG

ODS Table Name	Description	Statement	Option
ClassLevels	Class 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
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
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

* Depends on data.

ODS Graphics (Experimental)

The QUANTREG procedure uses ODS Graphics, which is experimental in SAS 9.1, 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 MODEL statement. For more information on the ODS GRAPHICS statement, see Chapter 15, “Statistical Graphics Using ODS” (*SAS/STAT User’s Guide*).

For a single quantile, two plots are particularly useful for 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 using the PLOT=RDPLOT and PLOT=DDPLOT option in the MODEL statement with the ODS GRAPHICS option turned on.

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

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

The plot options in the MODEL statement are summarized in the following table. See the **PLOT=** option in the MODEL statement for details.

Table 6. Options for Plots

Keyword	Plot
DDPLOT	Robust distance–Mahalanobis distance
RDPLOT	Standardized robust residual–Robust distance
RESHISTOGRAM	Histogram of standardized robust residuals
RESQQPLOT	Q-Q plot of standardized robust residuals
QUANTPLOT	Scatter plot of regression quantile

ODS Graph Names

The QUANTREG procedure assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. The names are listed in [Table 7](#) on page 43.

To request these graphs you must specify the ODS GRAPHICS statement in addition to the plot options in the MODEL statement. For more information on the ODS GRAPHICS statement, see Chapter 15, “Statistical Graphics Using ODS” (*SAS/STAT User’s Guide*).

Table 7. ODS Graphics Produced by PROC QUANTREG

ODS Graph Name	Plot Description	Statement	Option
DDPlot	Robust distance–Mahalanobis distance	MODEL	DDPLOT
QuantPlot	Scatter plot for regression quantiles with confidence bands	MODEL	QUANTPLOT UNPACKPANEL
QuantPlotE	Scatter plot for regression quantiles without confidence bands	MODEL	QUANTPLOT NOBANDS UNPACKPANEL
QuantPanel	Panel including maximum four quantile plots with confidence bands	MODEL	QUANTPLOT
QuantPanelE	Panel including maximum four quantile plots without confidence bands	MODEL	QUANTPLOT NOBANDS
RDPlot	Standardized robust residual–Robust distance	MODEL	RDPLOT
ResidualHistogram	Histogram of standardized robust residuals	MODEL	RESHISTOGRAM
ResidualQQPlot	Q-Q plot of standardized robust residuals	MODEL	RESQQPLOT

Examples

Example 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 less than 5,000 observations and 50 variables. The simplex algorithm can also compute the entire quantile process, which is shown in [Example 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 + .5 * e;
    output;
  end;
run;
```

The following statements invoke the QUANTREG procedure to fit a median regression model with the default simplex algorithm.

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

Output 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 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 1.2**, which also displays the objective function value and the predicted value of the response at the means of the covariates.

Output 1.2. Quantile and Objective Function with Simplex Algorithm

The QUANTREG Procedure		
Quantile and Objective Function		
Quantile		0.5
Objective Function		2441.1927
Predicted Value at Mean		10.0259

Output 1.3 displays parameter estimates and confidence limits. These estimates are reasonable, which indicates that median regression is robust to the 50 outliers.

Output 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 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 1.4 displays model fit information.

Output 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 1.5 displays the iteration history of the interior point algorithm. Note that the duality gap is less than $1\text{E}-6$ 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 1.5. Iteration History for the Interior Point Algorithm

The QUANTREG Procedure						
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	
5.44624E-8	12	12	1.0000	1.0000	2441.1927	

Output 1.6 displays the parameter estimates obtained with the interior point algorithm, which are identical to those obtained with the simplex algorithm.

Output 1.6. Parameter Estimates with Interior Point Algorithm

The QUANTREG Procedure		
Quantile and Objective Function		
Quantile		0.5
Objective Function		2441.1927
Predicted Value at Mean		10.0259
Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	10.0364
x1	1	5.0106
x2	1	3.0294

The following statements refit the model using the smoothing algorithm.

```
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 1.7** displays the model fit information.

Output 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 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 (2003).

Output 1.8. Iteration History for the Smoothing Algorithm

```

The QUANTREG Procedure

Iteration History of Smoothing Algorithm

Threshold  Iter  Refac  FullUpd  PartUpd  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 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 1.9. Parameter Estimates with Smoothing Algorithm

```

The QUANTREG Procedure

Quantile and Objective Function

Quantile                0.5
Objective Function      2441.1927
Predicted Value at Mean 10.0259

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 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.

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

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

Output 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 may have outliers.

Output 2.1. Model Information and Summary Statistics

The QUANTREG Procedure						
Model Information						
Data Set	MYLIB.GROWTH					
Dependent Variable	GDP					
Number of Independent Variables	13					
Number of Observations	161					
Optimization Algorithm	Simplex					
Method for Confidence Limits	Inv_Rank					
Summary Statistics						
Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
lgdp2	6.9893	7.7454	8.6084	7.7905	0.9543	1.1572
mse2	0.3160	0.7230	1.2675	0.9666	0.8574	0.6835
fse2	0.1270	0.4230	0.9835	0.7117	0.8331	0.5011
fhe2	0.0110	0.0350	0.0890	0.0792	0.1216	0.0400
mhe2	0.0400	0.1060	0.2060	0.1584	0.1752	0.1127
lexp2	3.8670	4.0639	4.2428	4.0440	0.2028	0.2734
lintr2	0.00159	0.5604	1.8804	1.4625	2.5492	1.0064
gedy2	0.0247	0.0343	0.0465	0.0359	0.0141	0.0150
Iy2	0.1395	0.1955	0.2671	0.2010	0.0877	0.0982
gcony2	0.0479	0.0767	0.1276	0.0914	0.0617	0.0566
lblakp2	0	0.0696	0.2407	0.1915	0.3071	0.1031
pol2	0	0.0500	0.2429	0.1683	0.2409	0.0741
ttrad2	-0.0241	-0.0101	0.00731	-0.00569	0.0375	0.0241
GDP	0.00293	0.0196	0.0351	0.0191	0.0248	0.0237

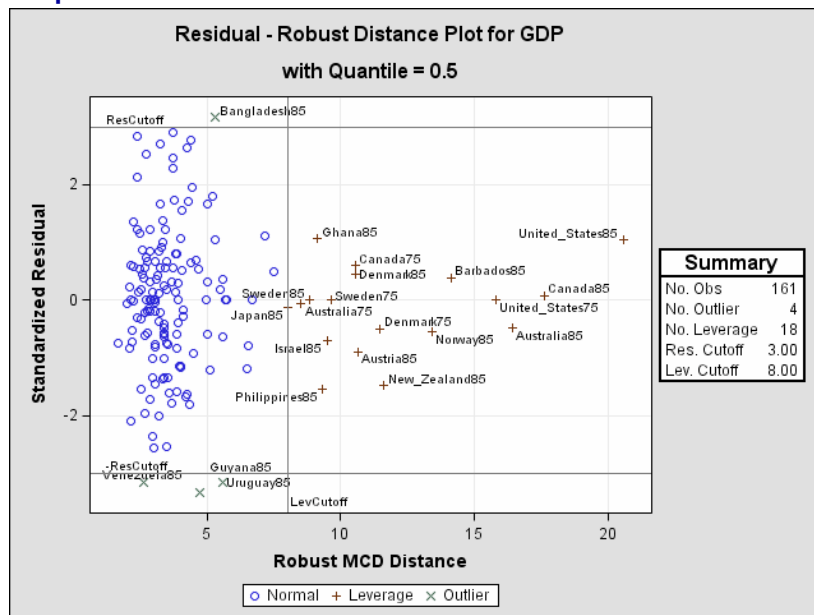
Output 2.2. Parameter Estimates

The QUANTREG Procedure				
Parameter Estimates				
Parameter	DF	Estimate	95% Confidence Limits	
Intercept	1	-0.0433	-0.2453	0.0811
lgdp2	1	-0.0268	-0.0389	-0.0175
mse2	1	0.0109	0.0000	0.0329
fse2	1	-0.0009	-0.0300	0.0116
fhe2	1	0.0120	-0.0830	0.0375
mhe2	1	0.0052	-0.0237	0.0789
lexp2	1	0.0666	0.0276	0.1335
lintr2	1	-0.0022	-0.0052	0.0010
gedy2	1	-0.0503	-0.4308	0.1264
Iy2	1	0.0750	0.0158	0.1148
gcony2	1	-0.0930	-0.2116	0.0042
lblakp2	1	-0.0267	-0.0545	-0.0189
pol2	1	-0.0301	-0.0471	-0.0015
ttrad2	1	0.1640	0.0392	0.2943

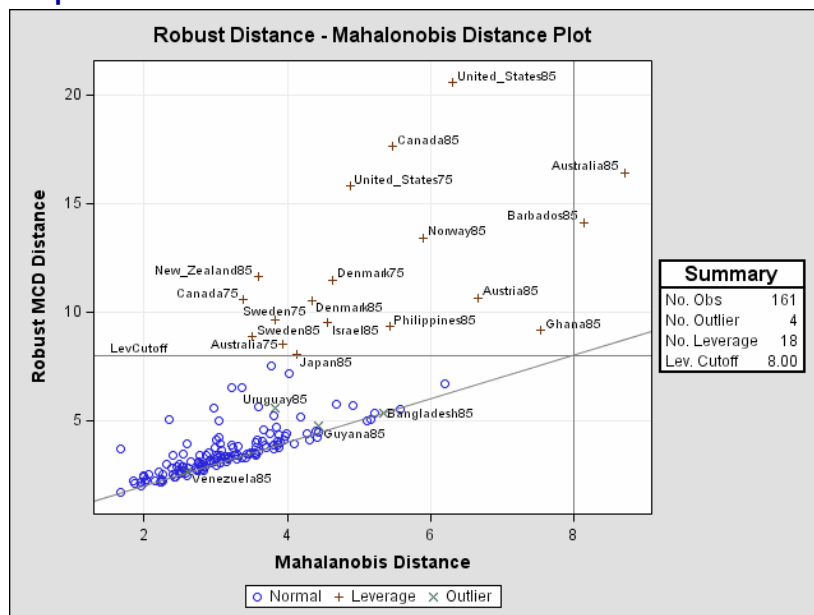
Output 2.2 displays parameter estimates and 95% confidence limits computed with the rank method.

Diagnostics for the median regression fit are displayed in Output 2.3 and Output 2.4, which are requested with the PLOTS= option. Output 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 2.4 plots the robust MCD distance against the Mahalanobis distance. This display is used to diagnose leverage points.

Output 2.3. Residual-Robust Distance Plot



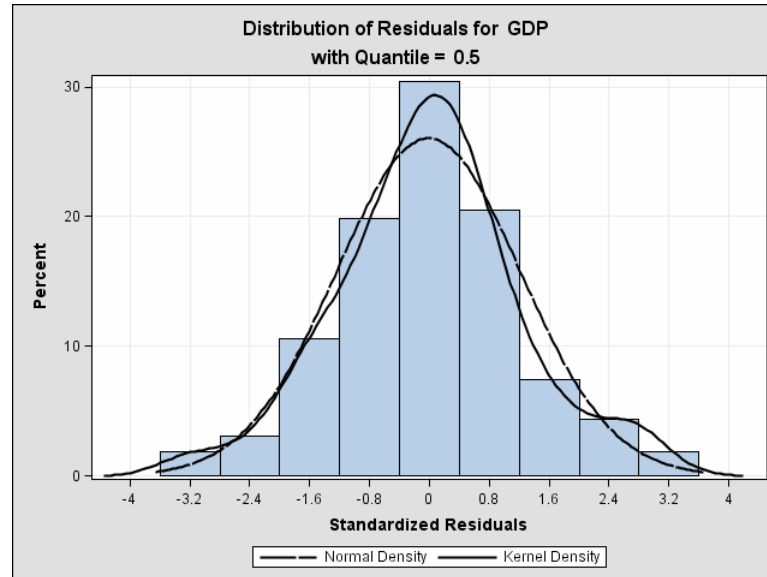
Output 2.4. Robust Distance-Mahalanobis Distance Plot



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

A histogram of the standardized residuals and two fitted density curves are displayed in [Output 2.5](#). This shows that median regression fits the data well.

Output 2.5. Histogram for Residuals



Tests of significance for the initial per-capita GDP (LGDP2) are shown in [Output 2.6](#).

Output 2.6. Tests for Regression Coefficient

The QUANTREG Procedure				
Tests				
Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
Wald	45.3228	1	45.32	<.0001

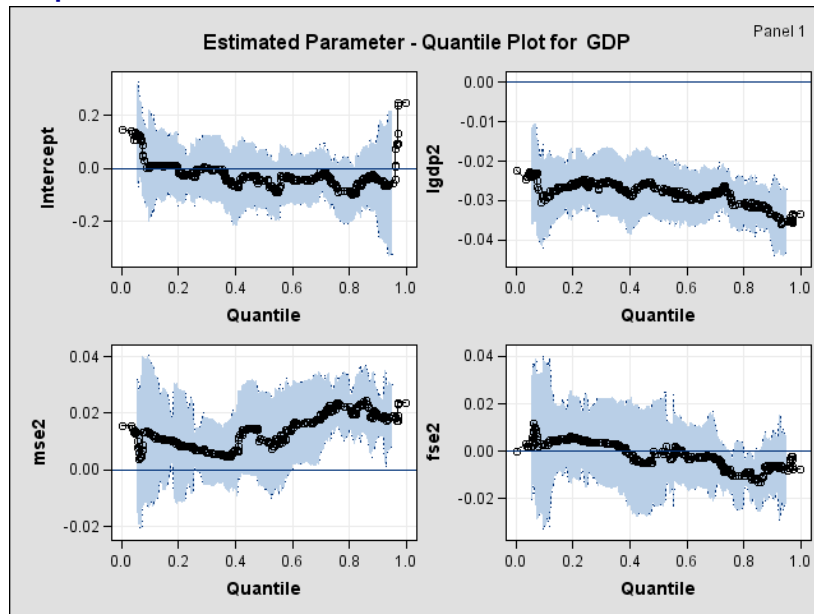
The QUANTREG procedure computes entire quantile processes for covariates when you specify QUANTILE=ALL in the MODEL statement.

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

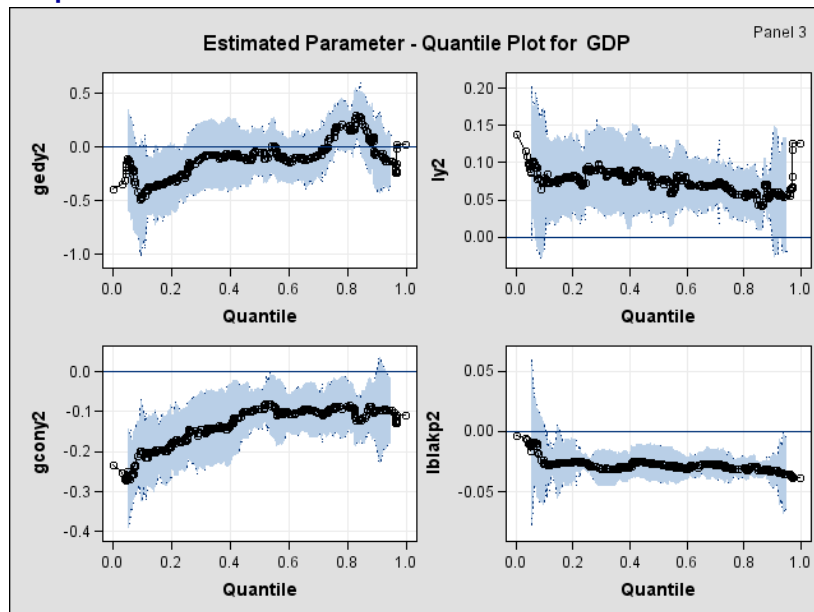
Confidence limits for quantile processes can be computed with the sparsity or resampling methods, but not the rank method because the computation would be prohibitively expensive.

A total of 14 quantile process plots are produced. [Output 2.7](#) and [Output 2.8](#) display two panels of eight selected process plots. The 95% confidence bands are shaded.

Output 2.7. Quantile Processes with 95% Confidence Bands



Output 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:

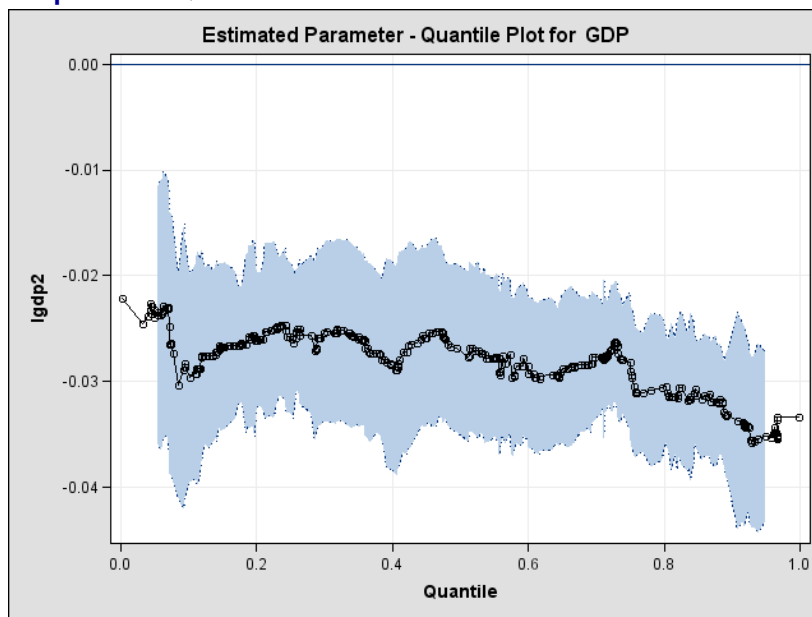
```
ods html;
ods graphics on;
proc quantreg ci=sparsity data=growth;
    model GDP = lgdp2 mse2 fse2 fhe2 mhe2 lexp2 lintr2
              gedy2 Iy2 gcony2 lblakp2 pol2 ttrad2
              / quantile=all plot=quantplot(lgdp2);
run;
ods graphics off;
ods html close;
```

The plot is shown in [Output 2.9](#).

The confidence bands here are computed using the sparsity method with the non i.i.d. assumption, unlike with Koenker and Machado (1999), who used the rank method for a few selected points. [Output 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).

Output 2.9. Quantile Process Plot for LGDP2



Example 3. Quantile Regression Analysis of Birth-Weight Data

This example is patterned after a quantile regression analysis of covariates associated with birthweight 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
novisit	Indicator of No Prenatal Visit
tri2	Indicator of First Visit in Second Trimester
tri3	Indicator of First Visit in Last Trimester
ed_hs	Indicator of Mother with High School Education
ed_smcol	Indicator of Mother with Some College Education
ed_col	Indicator of Mother with College Education
smoke	Indicator of Smoking Mother
cigsper	Number of Cigarettes Smoked Per Day
mom_age	Mother's Age
mom_age2	Square of Mother's Age
m_wtgain	Mother's Weight Gain During Pregnancy
m_wtgain2	Square of Mother's Weight Gain During Pregnancy

There are four indicator variables for the level of education of the mother. Since no indicator variable is provided for less than a high school education, this level serves as a reference level (the regression coefficients of the indicator variables measure the effect relative to this level.) Likewise, there are four indicator variables for the level of prenatal medical care of the mother, and a first visit in the first trimester serves as the reference level.

The following statements fit a regression model for 19 quantiles of birthweight, 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 html;
ods graphics on;
proc quantreg ci=sparsity/iid algorithm=interior
  data=bweight;
  model weight = black married boy novisit tri2 tri3
    ed_hs ed_smcol ed_col smoke cigspers
    mom_age mom_age2 m_wtgain m_wtgain2 /
  quantile= .05 .1 .15 .2 .25 .3 .35 .4
    .45 .5 .55 .6 .65 .7 .75 .8 .85 .9 .95
  plot=quantplot;
run;
ods graphics off;
ods html close;
```

Output 3.1 displays the model information and summary statistics for the variables in the model.

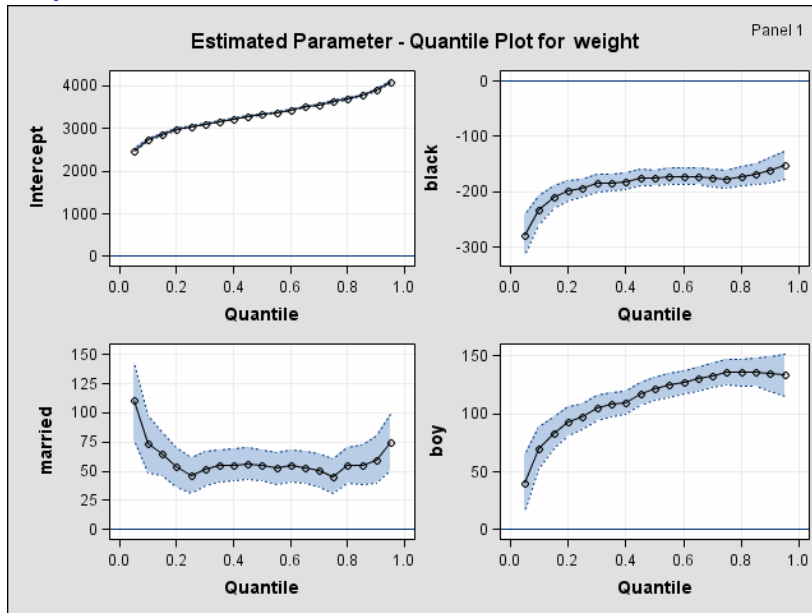
Output 3.1. Model Information and Summary Statistics

The QUANTREG Procedure						
Model Information						
Data Set						MYLIB.BWEIGHT
Dependent Variable						weight
Number of Independent Variables						15
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
novisit	0	0	0	0.00806	0.0894	0
tri2	0	0	0	0.1268	0.3327	0
tri3	0	0	0	0.0223	0.1476	0
ed_hs	0	0	1.0000	0.3490	0.4767	0
ed_smcol	0	0	0	0.2426	0.4286	0
ed_col	0	0	0	0.2490	0.4324	0
smoke	0	0	0	0.1307	0.3370	0
cigspers	0	0	0	1.4766	4.6541	0
mom_age	-4.0000	0	5.0000	0.4161	5.7285	5.9304
mom_age2	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_wtgain2	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

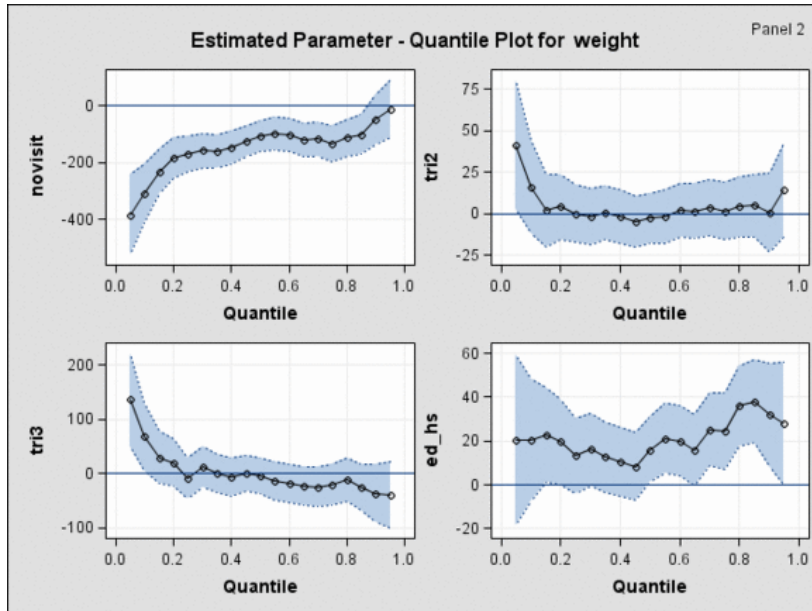
Among the 15 independent variables, the first 11 are categorical variables. For these variables, the mean represents the proportion in each category. The two continuous variables, MOM_AGE and M_WTGAIN, are centered at their medians, which are 27 and 30, respectively.

The quantile plots for the intercept and the 15 independent variables are shown in the following four panels. In each plot, the regression coefficient at a given quantile indicates the effect on birthweight of a unit change in that variable, assuming that the other variables are fixed. The bands represent 95% confidence intervals.

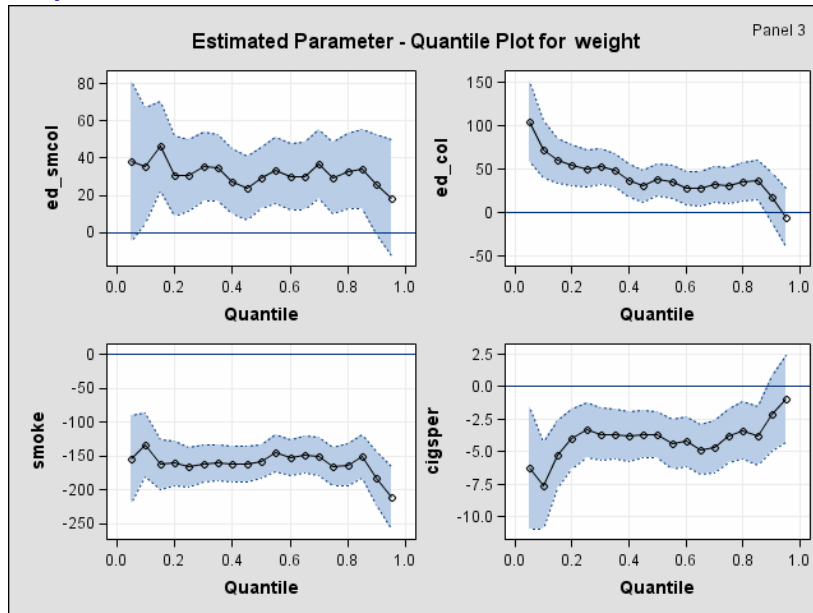
Output 3.2. Quantile Processes with 95% Confidence Bands



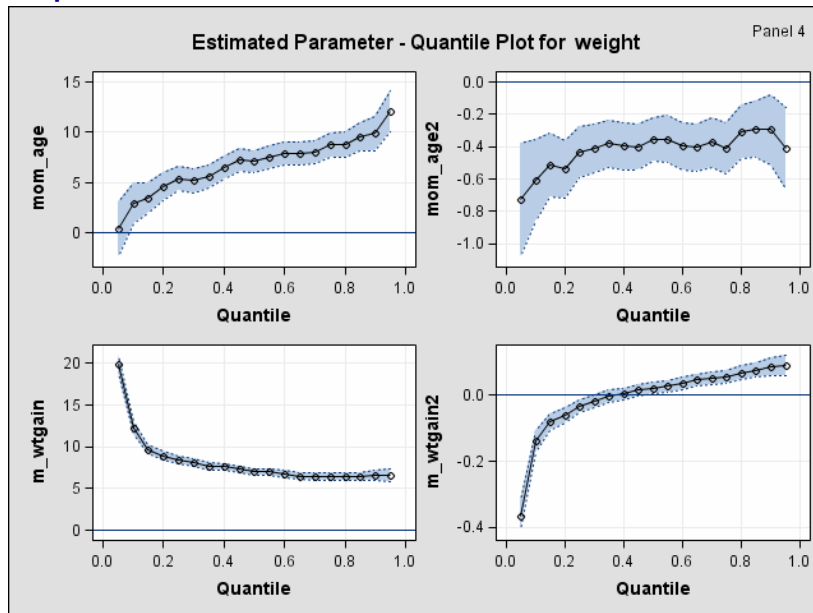
Output 3.3. Quantile Processes with 95% Confidence Bands



Output 3.4. Quantile Processes with 95% Confidence Bands



Output 3.5. Quantile Processes with 95% Confidence Bands



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 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 birthweight 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 birthweight distribution. The third plot shows that marital status has a large positive effect on birthweight, 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 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 birthweight 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 3.3](#) and the first two plots in [Output 3.4](#) are for variables related to education. Education beyond high school is associated with a positive effect on birthweight. The effect of high school education is uniformly around 15 grams across the entire birthweight 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 3.4](#) show that smoking is associated with a large negative effect on birthweight.

The linear and quadratic effects for the two continuous variables are shown in [Output 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.

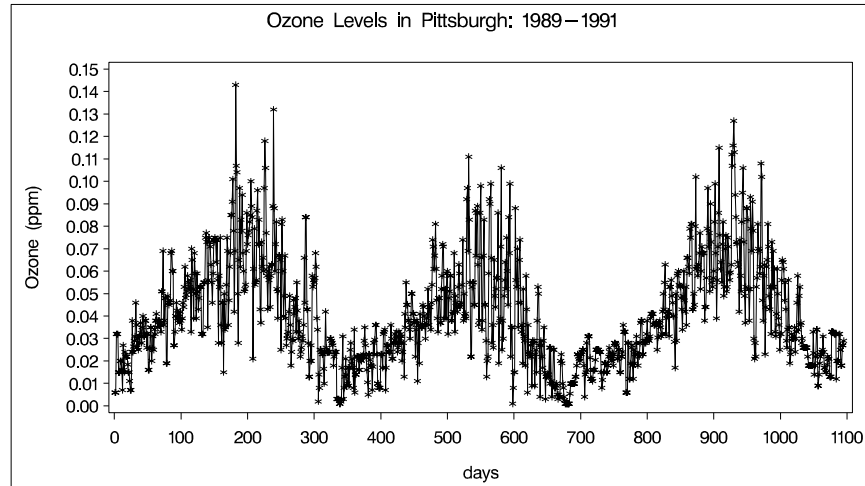
Example 4. Nonparametric Quantile Regression for Ozone Levels

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 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 `DAY` (index of 1,095 days (3 years)).

Output 4.1, which displays the time series plot of ozone concentration for the three years, shows a clear seasonal pattern.

Output 4.1. Time Series of Ozone Levels in Pittsburgh, Pennsylvania



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 construct 15 basis functions for DAY using the TRANSREG procedure.

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

The 15 basis functions (including the implicit intercept) are saved in the output data set `bs`, which is merged with `OZONE` to create a data set named `OZBS`. These basis variables are renamed as `INTERCEPT`, `X1` to `X14`.

The following statements fit the conditional mean using OLS regression.

```
proc reg data=ozbs;
  model ozone = x1-x14;
  output out=outp0 pred=p0;
run;
```

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

You can use the following macro to compute fitted values for multiple quantiles.

```

%macro quantiles(NQuant, Quantiles);
  %do i=1 %to &NQuant;
    proc quantreg data=ozbs algorithm=smooth;
      model ozone = x1-x14 /
        quantile=%scan(&Quantiles,&i,"");
      output out=outp&i pred=p&i;
    run;
  %end;
%mend;

```

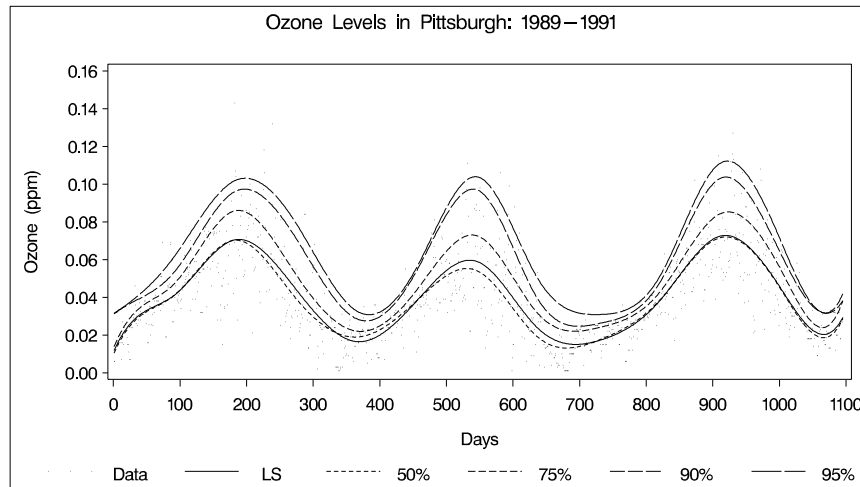
The following statements request fitted values for the median and three upper quantiles.

```

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

```

Output 4.2. Quantiles and Mean Ozone Levels in Pittsburgh, Pennsylvania



Output 4.2 displays the conditional mean curve obtained with the REG procedure and the quantile curves obtained with the QUANTREG procedure.

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

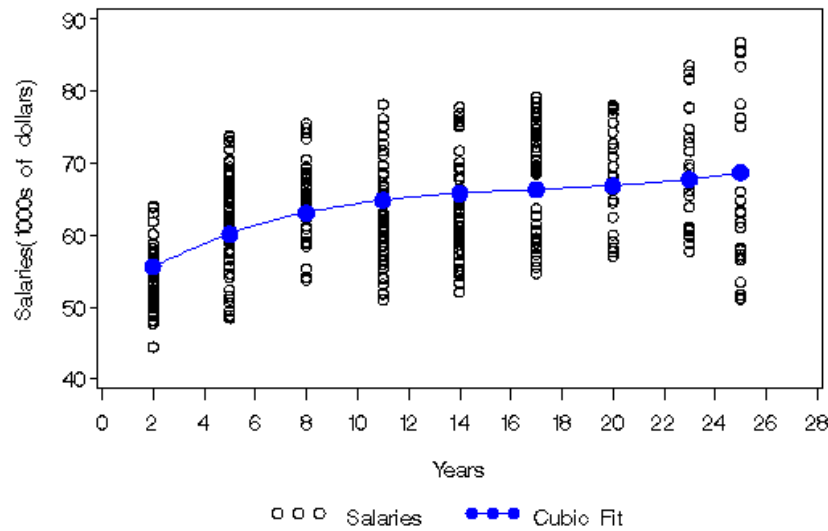
The quantile curves also show that high ozone concentration in 1989 had a longer duration than in 1990 and 1991. This is indicated by the wider spread of the quantile curves in 1989.

Example 5. Quantile Polynomial Regression for Salary Data

In 1991 a university union surveyed the salaries of professors across departments in U.S. colleges and universities that list programs in statistics. The goal was to examine the relationship between faculty salaries and years of service.

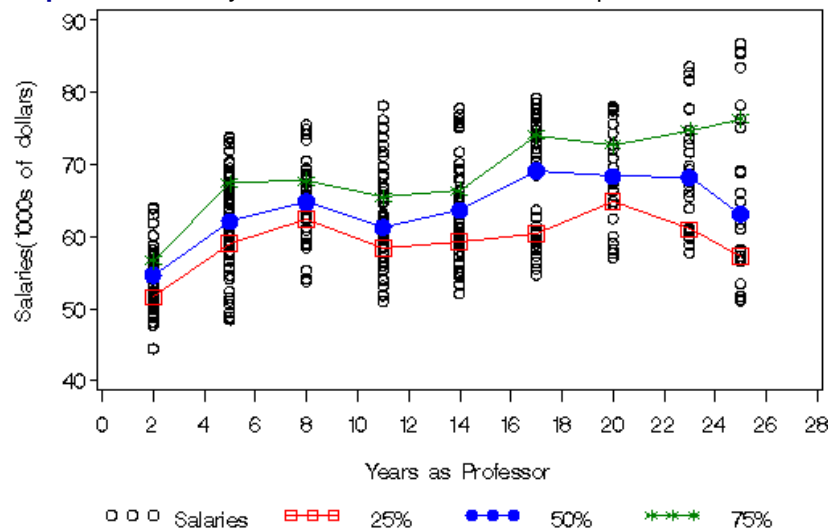
The data include salaries and years of service for 459 professors. The scatter plot in [Output 5.1](#) shows that the relationship is not linear, and a quadratic or cubic regression curve is appropriate. [Output 5.1](#) shows a cubic curve.

Output 5.1. Salary with Years as Professor: Cubic Fit



The curve in [Output 5.1](#) does not adequately describe the conditional salary distributions and how they change with length of service. [Output 5.2](#) shows the 25th, 50th, and 75th percentiles for each number of years, which gives a better picture of the conditional distributions.

Output 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 using polynomial quantile regression.

```
proc quantreg data=salary ci=none;
  model salaries = years years*years years*years*years
    /quantile=.25 .5 .75;
run;
```

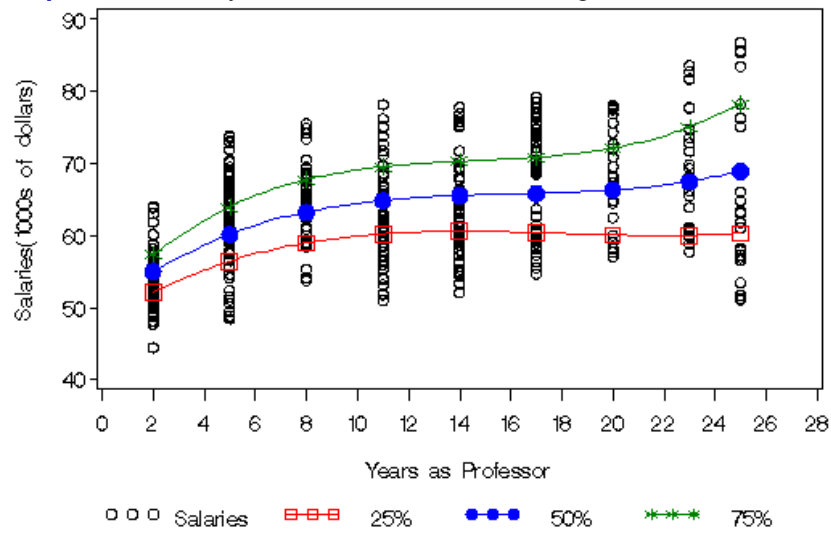
Output 5.3 shows the regression coefficients for the three quantiles.

Output 5.3. Regression Coefficients

The QUANTREG Procedure		
Parameter Estimates		
q = 0.25		
Parameter	DF	Estimate
Intercept	1	48.2509
years	1	2.2234
years*years	1	-0.1292
years*years*years	1	0.0024
Parameter Estimates		
q = 0.5		
Parameter	DF	Estimate
Intercept	1	50.2512
years	1	2.7173
years*years	1	-0.1632
years*years*years	1	0.0034
Parameter Estimates		
q = 0.75		
Parameter	DF	Estimate
Intercept	1	51.0298
years	1	3.6513
years*years	1	-0.2390
years*years*years	1	0.0055

Output 5.4 displays the three cubic percentile curves.

Output 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. (2003), “A Finite Smoothing Algorithm for Quantile Regression,” submitted, preprint available from the author.
- 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, USA.
- 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.
- 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*, P. Mandl and M. Huskova. ed., 349–359, Springer-Verlag, New York.
- Koenker, R. and Bassett, G. W. (1978), “Regression Quantiles,” *Econometrica*, 46, 33–50.

- Koenker, R. and Bassett, G. W. (1982), “Robust Tests for Heteroscedasticity Based on Regression Quantiles,” *Econometrica*, 50, 43–61.
- Koenker, R. and d’Orey, V. (1993), “Computing Regression Quantiles,” *Applied Statistics*, 43, 410–414.
- Koenker, R. and Hallock, K. (2001), “Quantile Regression: An Introduction,” *Journal of Economic Perspectives*, 15, 143–156.
- Koenker, R. and Machado, A. F. (1999), “Goodness of Fit and Related Inference Processes for Quantile Regression,” *Journal of the American Statistical Association*, 94, 1296–1310.
- Koenker, R. and Zhao, Q. (1994), “L-estimation for Linear Heteroscedastic Models,” *Journal of Nonparametric Statistics*, 3, 223–235.
- 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

A

affine step
 QUANTREG procedure, 29

C

centering step
 QUANTREG procedure, 30
complementarity
 QUANTREG procedure, 28
computational resources
 QUANTREG procedure, 41

I

INEST= data sets
 QUANTREG procedure, 40
infeasibility
 QUANTREG procedure, 29

K

Karush-Kuhn-Tucker (KKT) conditions
 QUANTREG procedure, 28

O

ODS graph names
 QUANTREG procedure, 43
OUTEST= data sets
 QUANTREG procedure, 40
output table names
 QUANTREG procedure, 42

P

Primal-Dual with Predictor-Corrector algorithm
 QUANTREG procedure, 29

Q

QUANTREG procedure, 3
 affine step, 29
 centering step, 30
 complementarity, 28
 computational resources, 41
 INEST= data sets, 40
 infeasibility, 29
 Karush-Kuhn-Tucker (KKT) conditions, 28
 ODS graph names, 43
 OUTEST= data sets, 40
 output table names, 42
 Primal-Dual with Predictor-Corrector algorithm,
 29
QUANTTREG procedure
 syntax, 16

S

syntax
 QUANTTREG procedure, 16

Syntax Index

A

- ALGORITHM option
 - PROC QUANTREG statement, 16
- ALPHA= option
 - PROC QUANTREG (QUANTREG), 17

B

- BY statement
 - QUANTREG procedure, 19

C

- CI option
 - PROC QUANTREG statement, 17
- CLASS statement
 - QUANTREG procedure, 19
- CORRB option
 - MODEL statement (QUANTREG), 20
- COVB option
 - MODEL statement (QUANTREG), 20
- CPUCOUNT option
 - PERFORMANCE statement (QUANTREG), 24
- CUTOFF option
 - MODEL statement (QUANTREG), 20

D

- DATA= option
 - PROC QUANTREG statement, 18
- DDPLOT option
 - PROC QUANTREG statement, 22
- DIAGNOSTICS option
 - MODEL statement (QUANTREG), 20

H

- HISTOGRAM option
 - MODEL statement (QUANTREG), 22

I

- ID statement
 - QUANTREG procedure, 20
- INEST= option
 - PROC QUANTREG statement, 18
- ITPRINT option
 - MODEL statement, 20

K

- KAPPA= option
 - PROC QUANTREG statement, 17
- keyword= option
 - OUTPUT statement (QUANTREG), 23

L

- LEVERAGE keyword
 - OUTPUT statement (QUANTREG), 23
- LEVERAGE option
 - MODEL statement, 20

M

- MAHADIST keyword
 - OUTPUT statement (QUANTREG), 23
- MAXIT= option
 - PROC QUANTREG statement, 17
- MODEL statement
 - QUANTREG procedure, 20

N

- NAMELEN= option
 - PROC QUANTREG statement, 18
- NODIAG option
 - MODEL statement (QUANTREG), 21
- NOINT option
 - MODEL statement (QUANTREG), 21
- NOSUMMARY option
 - MODEL statement (QUANTREG), 21

O

- OPTION statement
 - QUANTREG procedure, 20
- ORDER= option
 - PROC QUANTREG statement, 18
- OUT= option
 - OUTPUT statement (QUANTREG), 23
- OUTEST= option
 - PROC QUANTREG statement, 19
- OUTLIER keyword
 - OUTPUT statement (QUANTREG), 23
- OUTPUT statement
 - QUANTREG procedure, 23

P

- PERFORMANCE statement
 - QUANTREG procedure, 24
- PP= option
 - PROC QUANTREG statement, 19
- PREDICTED keyword
 - OUTPUT statement (QUANTREG), 23
- PROC QUANTREG statement,
 - QUANTREG procedure

Q

QQPLOT option
 MODEL statement, 22
 QUANTILES keyword
 OUTPUT statement (QUANTREG), 23
 QUANTILES option
 MODEL statement (QUANTREG), 21
 QUANTPLOT option
 MODEL statement (QUANTREG), 22
 QUANTREG procedure, BY statement, 19
 QUANTREG procedure, CLASS statement, 19
 QUANTREG procedure, ID statement, 20
 QUANTREG procedure, MODEL statement, 20
 CORRB option, 20
 COVB option, 20
 CUTOFF option, 20
 DDPLOT option, 22
 DIAGNOSTICS option, 20
 HISTOGRAM option, 22
 ITPRINT option, 20
 LEVERAGE option, 20
 NODIAG option, 21
 NOINT option, 21
 NOSUMMARY option, 21
 PLOT= plot option, 21
 QQPLOT option, 22
 QUANTILES option, 21
 QUANTPLOT option, 22
 RDPLOT option, 22
 SCALE option, 21
 SINGULAR= option, 21
 QUANTREG procedure, OPTION2 statement, 20
 QUANTREG procedure, OUTPUT statement, 23
 keyword= option, 23
 LEVERAGE keyword, 23
 MAHADIST keyword, 23
 OUT= option, 23
 OUTLIER keyword, 23
 PREDICTED keyword, 23
 QUANTILES keyword, 23
 RESIDUAL keyword, 23
 ROBDIST keyword, 23
 SRESIDUAL keyword, 24
 STD_ERR keyword, 24
 QUANTREG procedure, PERFORMANCE state-
 ment, 24
 CPUCOUNT option, 24
 QUANTREG procedure, PROC QUANTREG state-
 ment, 16
 ALGORITHM option, 16
 ALPHA= option, 17
 CI option, 17
 DATA= option, 18
 INEST= option, 18
 KAPPA= option, 17
 MAXIT= option, 17
 NAMELEN= option, 18
 ORDER= option, 18
 OUTEST= option, 19

PP option, 19

RRATIO= option, 17

TOLERANCE= option, 17

QUANTREG procedure, TEST statement, 24

QUANTREG procedure, WEIGHT statement, 25

QUANTREG procedure, MODEL statement

SEED option, 21

R

RDPLOT option

MODEL statement, 22

RESIDUAL keyword

OUTPUT statement (QUANTREG), 23

ROBDIST keyword

OUTPUT statement (QUANTREG), 23

RRATIO= option

PROC QUANTREG statement, 17

S

SCALE option

MODEL statement (QUANTREG), 21

SEED option

MODEL statement (QUANTREG), 21

SINGULAR= option

MODEL statement (QUANTREG), 21

SRESIDUAL keyword

OUTPUT statement (QUANTREG), 24

STD_ERR keyword

OUTPUT statement (QUANTREG), 24

T

TEST statement

QUANTREG procedure, 24

TOLERANCE= option

PROC QUANTREG statement, 17

W

WEIGHT statement

QUANTREG procedure, 25