
# SAS/STAT® 15.1 User's Guide
# The QUANTREG Procedure

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

**SAS/STAT® 15.1 User's Guide**



SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

# Chapter 100
# The QUANTREG Procedure

## Contents

# Overview: QUANTREG Procedure

The QUANTREG procedure uses quantile regression to model the effects of covariates on the conditional quantiles of a response variable.

Quantile regression was introduced by Koenker and Bassett (1978) as an extension of ordinary least squares (OLS) regression, which models the relationship between one or more covariates $X$ and the *conditional mean* of the response variable $Y$ given $X = x$. Quantile regression extends the OLS regression to model the *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 100.1** Trout Density in Streams

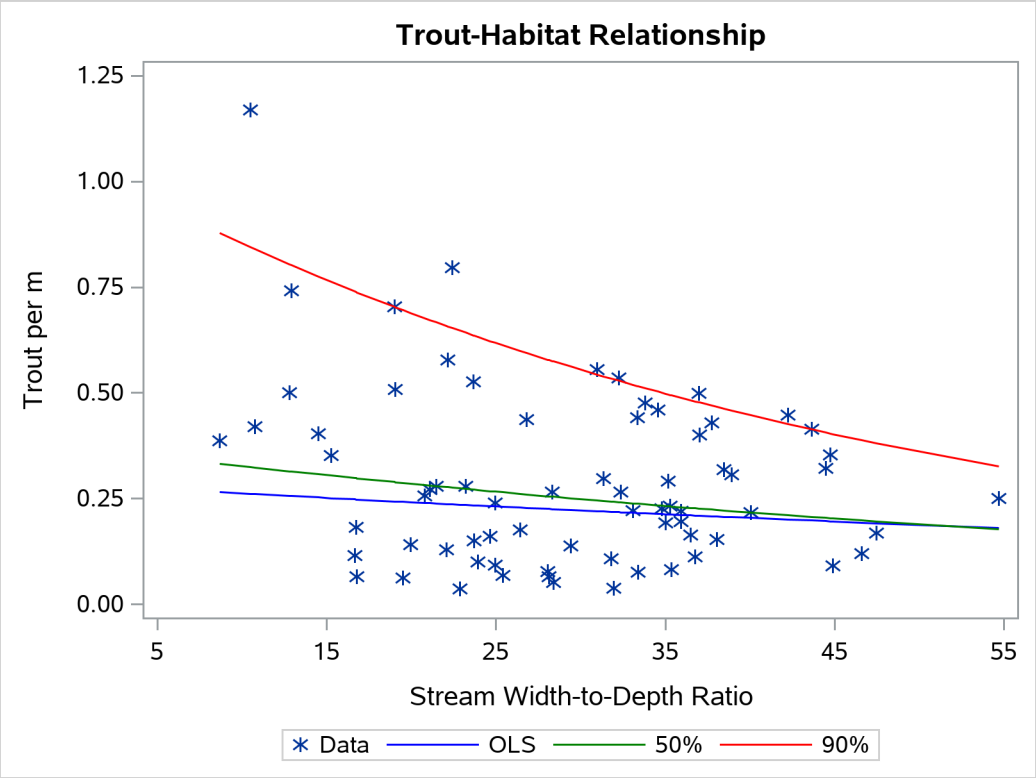
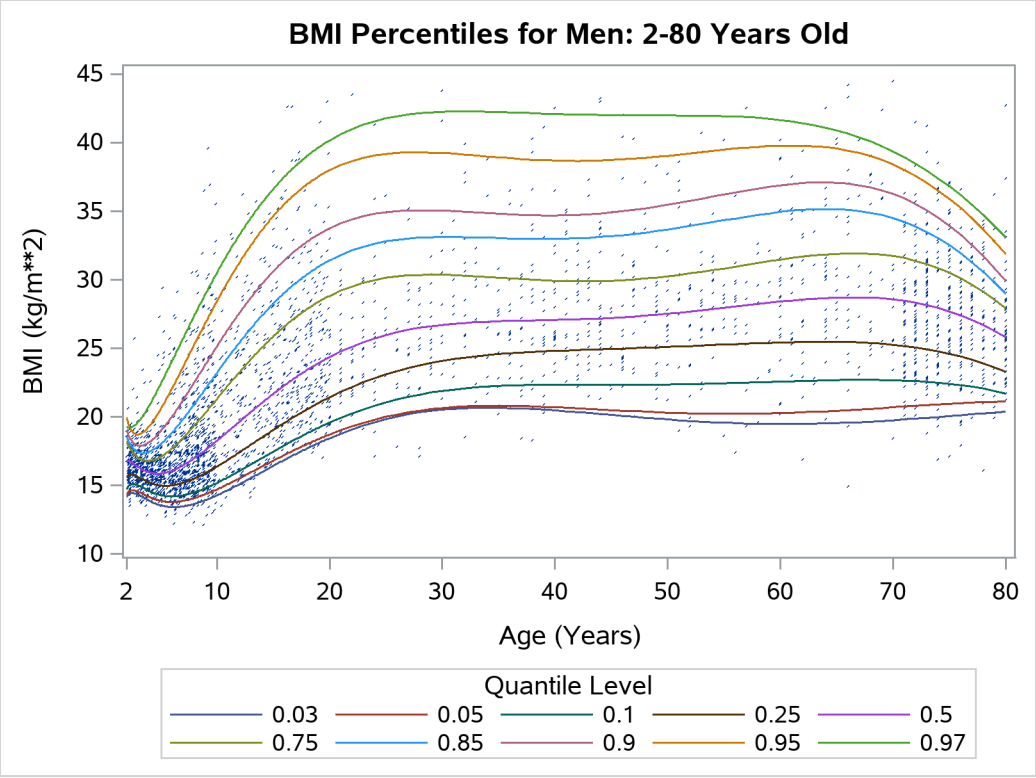



Figure 100.1 illustrates an ecological study in which modeling upper conditional quantiles reveals additional information. The points represent measurements of trout density and stream width-to-depth ratio that were taken at 13 streams over seven years.

As analyzed by Dunham, Cade, and Terrell (2002), both the ratio and the trout density depend on a number of unmeasured limiting factors that are 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 red and green curves represent the conditional 90th and 50th percentiles of density as determined by the QUANTREG procedure. The analysis was done by using a simple linear regression model for the logarithm of density. (The curves in Figure 100.1 were obtained by transforming the fitted lines back to the original scale. For more information, see the section "Analysis of Fish-Habitat Relationships" on page 8252.) 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 0. Similarly, the slope parameter for the mean, which is obtained with OLS regression, is not significantly different from 0.

**Figure 100.2** Percentiles for Body Mass Index




Quantile regression is especially useful when the data 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 100.2, which plots the body mass index of 8,250 men versus their age.

Here, both upper (overweight) and lower (underweight) conditional quantiles are important because they provide the basis for developing growth charts and establishing health standards. The curves in Figure 100.2 were determined by using the QUANTREG procedure to perform polynomial quantile regression. For more information, see the section "Growth Charts for Body Mass Index" on page 8258. Clearly, the rate of change with age (as expressed by the regression coefficients), particularly for ages less than 20, is different for each conditional quantile.

Heterogeneous data occur in many fields, including biomedicine, econometrics, survival analysis, and ecology. Quantile regression, which includes median regression as a special case, provides a complete picture of the covariate effect when a set of percentiles is modeled. So it can 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 because 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 about 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, quantile regression 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, consider using the ROBUSTREG procedure (Chapter 104, "The ROBUSTREG Procedure.")

Unlike OLS regression, quantile regression is equivariant to monotone transformations of the response variable. For example, 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.

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

## Features

The main features of the QUANTREG procedure are as follows:

- offers simplex, interior point, and smoothing algorithms for estimation
- provides sparsity, rank, and resampling methods for confidence intervals
- provides asymptotic and bootstrap methods for covariance and correlation matrices of the estimated parameters
- provides the Wald, likelihood ratio, and rank tests for the regression parameter estimates and the Wald test for heteroscedasticity
- provides outlier and leverage-point diagnostics
- enables parallel computing when multiple processors are available
- provides rowwise or columnwise output data sets with multiple quantiles

- provides regression quantile spline fits
- produces fit plots, diagnostic plots, and quantile process plots by using ODS Graphics

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

## Quantile Regression

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

For a random variable $Y$ with probability distribution function

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

the $\tau$ quantile of $Y$ is defined as the inverse function

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

where the quantile level $\tau$ ranges between 0 and 1. In particular, the median is $Q(1/2)$.

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

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

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

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

where $\rho_\tau(z) = z(\tau - I(z < 0))$, $0 < \tau < 1$, and where $I(\cdot)$ denotes the indicator function. The loss function $\rho_\tau$ assigns a weight of $\tau$ 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 $\tau$ 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 $E(Y|X = x) = \mathbf{x}'\boldsymbol{\beta}$ by solving for

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

The estimated parameter $\hat{\boldsymbol{\beta}}$ 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, $Q_Y(\tau|X = x) = \mathbf{x}'\boldsymbol{\beta}(\tau)$, by solving the following equation for $\tau \in (0, 1)$:

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

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

The following set of regression quantiles is referred to as the *quantile process*:

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

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

# Getting Started: QUANTREG Procedure

The following examples demonstrate how you can use the QUANTREG procedure to fit linear models for selected quantiles or for the entire quantile process. The first example explains the use of the procedure in a 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 100.1 indicates a nonlinear relationship, so it is reasonable to fit regression models for the conditional quantiles of the log of density. Because regression quantiles are equivariant under any monotonic (linear or nonlinear) transformation (Koenker and Hallock 2001), the exponential transformation converts the conditional quantiles to the original density scale.

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

```
data trout;
   input Density WDRatio LnDensity @@;
   datalines;
 0.38732     8.6819    -0.94850    1.16956    10.5102     0.15662
 0.42025    10.7636    -0.86690    0.50059    12.7884    -0.69197
 0.74235    12.9266    -0.29793    0.40385    14.4884    -0.90672
 0.35245    15.2476    -1.04284    0.11499    16.6495    -2.16289
 0.18290    16.7188    -1.69881    0.06619    16.7859    -2.71523
 0.70330    19.0141    -0.35197    0.50845    19.0548    -0.67639

   ... more lines ...

 0.25125    54.6916    -1.38129
;
```

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

```
ods graphics on;

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

The MODEL statement specifies a simple linear regression model with LnDensity as the response variable $Y$ and WDRatio as the covariate $X$. The QUANTILE= option requests that the regression quantile function $Q(\tau|X = x) = \mathbf{x}'\boldsymbol{\beta}(\tau)$ be estimated by solving the following equation, where $\tau = (0.5, 0.9)$:

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

By default, the regression coefficients $\hat{\boldsymbol{\beta}}(\tau)$ are estimated by using the simplex algorithm, which is explained in the section "Simplex Algorithm" on page 8285. The ALPHA= option requests 90% confidence limits for the regression parameters, and the option CI=RESAMPLING specifies that the intervals be computed by using the Markov chain marginal bootstrap (MCMB) resampling method of He and Hu (2002). When you specify the CI=RESAMPLING option, the QUANTREG procedure also computes standard errors, $t$ values, and $p$-values of regression parameters by using the MCMB resampling method. The SEED= option specifies a seed for the resampling method. The COVB option requests covariance matrices for the estimated regression coefficients, and the TEST statement requests tests for the hypothesis that the slope parameter (the coefficient of WDRatio) is 0.

Figure 100.3 displays model information and summary statistics for the variables in the model. The summary statistics include the median and the standardized median absolute deviation (MAD), which are robust measures of univariate location and scale, respectively. For more information about the standardized MAD, see Huber (1981, p. 108).

**Figure 100.3** Model Fitting Information and Summary Statistics

**The QUANTREG Procedure**

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

**Summary Statistics**

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

Figure 100.4 and Figure 100.5 display the parameter estimates, standard errors, 95% confidence limits, $t$ values, and $p$-values that are computed by the resampling method.

**Figure 100.4** Parameter Estimates at QUANTILE=0.5

**Parameter Estimates**

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

**Figure 100.5** Parameter Estimates at QUANTILE=0.9

**Parameter Estimates**

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

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

$$\hat{y}_{90} = \exp(0.0576 - 0.0215x)$$

This is the upper dashed curve that is plotted in Figure 100.1. The lower dashed curve for the median can be obtained in a similar fashion.

The covariance matrices for the estimated parameters are shown in Figure 100.6. The resampling method that is used for the confidence intervals is also used to compute these matrices.

**Figure 100.6** Covariance Matrices of the Estimated Parameters

**The QUANTREG Procedure**
**Quantile Level = 0.5**

| Estimated Covariance Matrix for Quantile Level = 0.5 | | |
|---|---|---|
| | **Intercept** | **WDRatio** |
| **Intercept** | 0.156191 | -.004653 |
| **WDRatio** | -.004653 | 0.000151 |

**The QUANTREG Procedure**
**Quantile Level = 0.9**

| Estimated Covariance Matrix for Quantile Level = 0.9 | | |
|---|---|---|
| | **Intercept** | **WDRatio** |
| **Intercept** | 0.067914 | -.001877 |
| **WDRatio** | -.001877 | 0.000056 |

The tests requested by the TEST statement are shown in Figure 100.7. Both the Wald test and the likelihood ratio test indicate that the coefficient of width-to-depth ratio is significantly different from 0 at the 90th percentile, but the difference is not significant at the median.

**Figure 100.7** Tests of Significance

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

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

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

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

Figure 100.8 displays a portion of the objective function table for the entire quantile process. The objective function is evaluated at 77 values of $\tau$ in the interval $(0, 1)$. The table also provides predicted values of the conditional quantile function $Q(\tau)$ at the mean for WDRatio, which can be used to estimate the conditional density function.

**Figure 100.8** Objective Function

**Objective Function for Quantile Process**

| Label | Quantile Level | 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 100.9 displays a portion of the table of the quantile processes for the estimated parameters and confidence limits.

**Figure 100.9** Objective Function

**Parameter Estimates for Quantile Process**

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

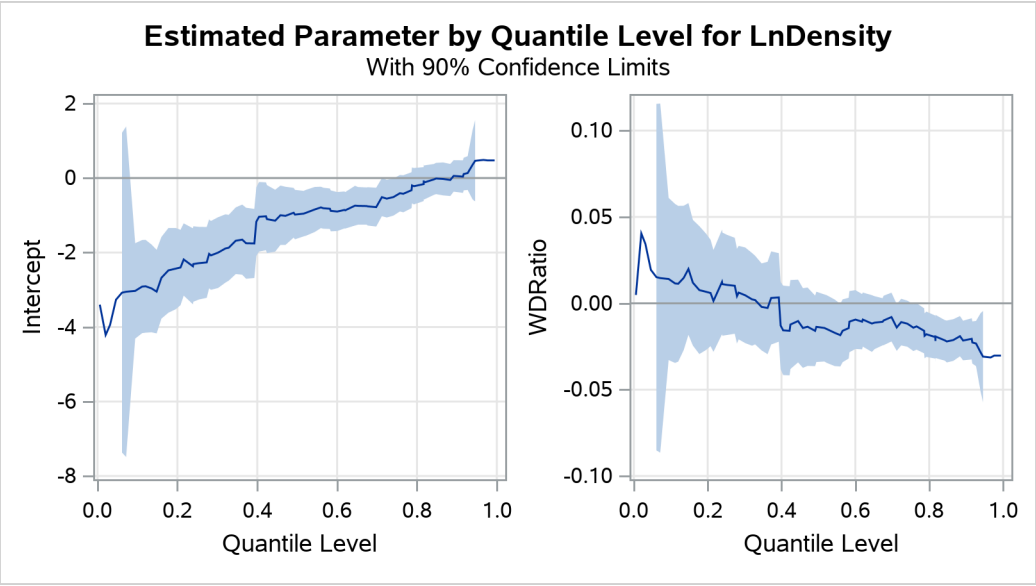
When ODS Graphics is enabled, the PLOT=QUANTPLOT option in the MODEL statement requests a plot of the estimated quantile processes.

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

The left side of Figure 100.10 displays the process for the intercept, and the right side displays the process for the coefficient of WDRatio.

The process plot for WDRatio shows that the slope parameter changes from positive to negative as the quantile increases and that 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 100.9, which shows that 0 falls between the lower and upper confidence limits of the slope parameter for quantiles below 0.78. Since the confidence intervals for the extreme quantiles are not stable because of insufficient data, the confidence band is not displayed outside the interval (0.05, 0.95).

**Figure 100.10** Quantile Processes for Intercept and Slope



## Growth Charts for Body Mass Index

Body mass index (BMI) is defined as the ratio of weight (kg) to squared height ($m^2$) and is a widely used measure for categorizing individuals as overweight or underweight. The percentiles of BMI for specified ages are of particular interest. As age increases, these percentiles provide growth patterns of BMI not only for the majority of the population, but also for underweight or overweight extremes of the population. In addition, the percentiles of BMI for a specified age provide a reference for individuals at that age with respect to the population.

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

A SAS data set named bmimen was created by merging and cleaning the 1999–2000 and 2001–2002 survey results for men that is published by the National Center for Health Statistics. This data set contains the variables Weight (kg), Height (m), BMI ($kg/m^2$), Age (year), and SeQN (respondent sequence number) for 8,250 men (Chen 2005).

The data set that is used in this example is a subset of the original data set of Chen (2005). It contains the two variables BMI and Age with 3,264 observations.

```
data bmimen;
   input BMI Age @@;
   SqrtAge = sqrt(Age);
   InveAge = 1/Age;
   LogBMI  = log(BMI);
   datalines;
18.6  2.0 17.1  2.0 19.0  2.0 16.8  2.0 19.0  2.1  15.5   2.1
16.7  2.1 16.1  2.1 18.0  2.1 17.8  2.1 18.3  2.1  16.9   2.1
15.9  2.1 20.6  2.1 16.7  2.1 15.4  2.1 15.9  2.1  17.7   2.1

   ... more lines ...

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

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

The following statements invoke the QUANTREG procedure:

```
proc quantreg data=bmimen algorithm=interior(tolerance=1e-5) ci=resampling;
   model logbmi = inveage sqrtage age sqrtage*age
                  age*age age*age*age
                  / diagnostics cutoff=4.5 quantile=.5 seed=1268;
   id age bmi;
   test_age_cubic: test age*age*age / wald lr rankscore(tau);
run;
```

The MODEL statement provides the model, and the option QUANTILE=0.5 requests median regression. The ALGORITHM= option requests that the interior point algorithm be used to compute $\hat{\boldsymbol{\beta}}(\frac{1}{2})$. For more information about this algorithm, see the section "Interior Point Algorithm" on page 8286.

Figure 100.11 displays the estimated parameters, standard errors, 95% confidence intervals, $t$ values, and $p$-values that are computed by the resampling method, which is requested by the CI= option. All of the parameters are considered significant because the $p$-values are smaller than 0.001.

**Figure 100.11** Parameter Estimates with Median Regression: Men

**The QUANTREG Procedure**

**Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 7.8909 | 0.8192 | 6.2848 | 9.4971 | 9.63 | <.0001 |
| InveAge | 1 | -1.8354 | 0.4362 | -2.6905 | -0.9802 | -4.21 | <.0001 |
| SqrtAge | 1 | -5.1247 | 0.7157 | -6.5280 | -3.7214 | -7.16 | <.0001 |
| Age | 1 | 1.9759 | 0.2545 | 1.4770 | 2.4748 | 7.77 | <.0001 |
| SqrtAge*Age | 1 | -0.3347 | 0.0426 | -0.4182 | -0.2513 | -7.86 | <.0001 |
| Age*Age | 1 | 0.0227 | 0.0029 | 0.0170 | 0.0285 | 7.75 | <.0001 |
| Age*Age*Age | 1 | -0.0000 | 0.0000 | -0.0001 | -0.0000 | -7.38 | <.0001 |

The TEST statement requests Wald, likelihood ratio, and rank tests for the significance of the cubic term in Age. The test results, shown in Figure 100.12, indicate that this term is significant. Higher-order terms are not significant.

**Figure 100.12** Test of Significance for Cubic Term

**Test test_age_cubic Results**

| Test | Test Statistic | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Wald | 54.4787 | 1 | 54.48 | <.0001 |
| Likelihood Ratio | 56.9473 | 1 | 56.95 | <.0001 |
| Rank_Tau | 42.5730 | 1 | 42.57 | <.0001 |

Median regression and, more generally, quantile regression are robust to extremes of the response variable. The DIAGNOSTICS option in the MODEL statement requests a diagnostic table of outliers, shown in Figure 100.13, which uses a cutoff value that is specified in the CUTOFF= option. The variables that are specified in the ID statement are included in the table.

With CUTOFF=4.5, 14 men are identified as outliers. All of these men have large positive standardized residuals, which indicates that they are overweight for their age. The cutoff value 4.5 is ad hoc. It corresponds to a probability less than 0.5E–5 if normality is assumed, but the standardized residuals for median regression usually do not meet this assumption.

In order to construct the chart shown in Figure 100.2, the same model that is used for median regression is used for other quantiles. The QUANTREG procedure can compute fitted values for multiple quantiles.

**Figure 100.13** Diagnostics with Median Regression

**Diagnostics**

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

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

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

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

proc sgplot data=outbmi;
   title 'BMI Percentiles for Men: 2-80 Years Old';
   yaxis label='BMI (kg/m**2)' min=10 max=45 values=(10 15 20 25 30 35 40 45);
   xaxis label='Age (Years)' min=2 max=80 values=(2 10 20 30 40 50 60 70 80);

   scatter x=age y=bmi /markerattrs=(size=1);
   series  x=age y=pbmi/group=QUANTILE;
run;
```

The fitted values are stored in the OUTPUT data set outp. The COLUMNWISE option arranges these fitted values for all quantiles in the single variable p by groups of the quantiles. After the exponential transformation, both the fitted BMI values and the original BMI values are plotted against age to create the display shown in Figure 100.2.

The fitted quantile curves reveal important information. During the quick growth period (ages 2 to 20), the dispersion of BMI increases dramatically. It becomes stable during middle age, and then it contracts after age 60. This pattern suggests that effective population weight control should start in childhood.

Compared to the 97th percentile in reference growth charts that were published by the Centers for Disease Control and Prevention (CDC) in 2000 (Kuczmarski, Ogden, and Guo 2002), the 97th percentile for 10-year-old boys in Figure 100.2 is 6.4 BMI units higher (an increase of 27%). This can be interpreted as a warning of overweight or obesity. See Chen (2005) for a detailed analysis.

# Syntax: QUANTREG Procedure

The following statements are available in the QUANTREG procedure:

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

The PROC QUANTREG statement invokes the QUANTREG procedure. The CLASS statement specifies which explanatory variables are treated as categorical. The ID statement names variables to identify observations in the outlier diagnostics tables. The MODEL statement is required and specifies the variables used in the regression. Main effects and interaction terms can be specified in the MODEL statement, as in the GLM procedure (Chapter 50, "The GLM Procedure.") The OUTPUT statement creates an output data set that contains predicted values, residuals, and estimated standard errors. The PERFORMANCE statement tunes the performance of PROC QUANTREG by using single or multiple processors available in the hardware. The TEST statement requests linear tests for the model parameters. The WEIGHT statement identifies a variable in the input data set whose values are used to weight the observations. Multiple OUTPUT and TEST statements are allowed in one invocation of PROC QUANTREG.

The EFFECT and ESTIMATE statements are also available in other procedures. Summary descriptions of functionality and syntax for these statements are provided in this chapter, and you can find full documentation about them in Chapter 19, "Shared Concepts and Topics."

## PROC QUANTREG Statement

**PROC QUANTREG** < *options* > **;**

The PROC QUANTREG statement invokes the QUANTREG procedure. Table 100.1 summarizes the *options* available in the PROC QUANTREG statement.

**Table 100.1** PROC QUANTREG Statement Options

| Option | Description |
|---|---|
| ALGORITHM= | Specifies an algorithm to estimate the regression parameters |
| ALPHA= | Specifies the level of significance |
| CI= | Specifies a method to compute confidence intervals |
| DATA= | Specifies the input SAS data set |
| INEST= | Specifies an input SAS data set that contains initial estimates |

**Table 100.1** *continued*

| Option | Description |
|---|---|
| NAMELEN= | Specifies the length of effect names |
| ORDER= | Specifies the order in which to sort classification variables |
| OUTEST= | Specifies an output SAS data set containing the parameter estimates |
| PLOT | Specifies options that control details of the plots |

You can specify the following *options* in the PROC QUANTREG statement.

**ALGORITHM=***algorithm* **<(** *suboptions* **)>**
specifies an algorithm for estimating the regression parameters.

You can specify one of the following four *algorithms*:

**SIMPLEX <** *suboption* **>**
uses the simplex algorithm to estimate the regression parameters. You can specify the following *suboption*:

**MAXSTATIONARY**=*m* requests that the algorithm terminate if the objective function has not improved for *m* consecutive iterations. By default, *m* = 1000.

**INTERIOR <(** *suboptions* **)>**
uses the interior point algorithm to estimate the regression parameters. You can specify the following *suboptions*:

**KAPPA**=*value* specifies the step-length parameter for the interior point algorithm. The *value* should be between 0 and 1. The larger the *value*, the faster the algorithm. However, numeric instability can occur as the *value* approaches 1. By default, KAPPA=0.99995. For more information, see the section "Interior Point Algorithm" on page 8286.

**MAXIT**=*m* sets the maximum number of iterations for the interior point algorithm. By default, MAXIT=1000.

**TOLERANCE**=*value* specifies the tolerance for the convergence criterion of the interior point algorithm. By default, TOLERANCE=1E–8. The QUANTREG procedure uses the duality gap as the convergence criterion. For more information, see the section "Interior Point Algorithm" on page 8286.

You can also use the PERFORMANCE statement to enable parallel computing when multiple processors are available in the hardware.

**IPM <(** *suboptions* **)>**
uses the efficient interior point algorithm to estimate the regression parameters. You can specify the following *suboptions*:

**MAXIT**=*m* sets the maximum number of iterations for the efficient interior point algorithm. By default, MAXIT=1000.

**TOLERANCE**=*value* specifies the tolerance for the convergence criterion of the efficient interior point algorithm. The QUANTREG procedure uses the complementarity value as the convergence criterion. By default, TOLERANCE=1E–8.

**SMOOTH** < *suboption* >
uses the smoothing algorithm to estimate the regression parameters. You can specify the following *suboption*:

**RRATIO**=*value* specifies the reduction ratio for the smoothing algorithm. This ratio is used to reduce the threshold of the smoothing algorithm. The *value* should be between 0 and 1. In theory, the smaller the *value*, the faster the smoothing algorithm. However, in practice, the optimal ratio is quite dependent on the data. For more information, see the section "Smoothing Algorithm" on page 8289.

The default algorithm depends on the number of observations ($n$) and the number of covariates ($p$) in the model estimation. See Table 100.2 for the relevant defaults.

**Table 100.2** The Default Estimation Algorithm

| | $p \leq 100$ | $p > 100$ |
|---|---|---|
| $n \leq 5000$ | SIMPLEX | SMOOTH |
| $n > 5000$ | INTERIOR | SMOOTH |

**ALPHA**=*value*
specifies the level of significance $\alpha$ for $100(1-\alpha)\%$ confidence intervals for regression parameters. The *value* must be between 0 and 1. The default is ALPHA=0.05, which corresponds to a 0.95 confidence interval.

**CI=NONE | RANK | SPARSITY**< **(BF | HS)** >< **/IID** > **| RESAMPLING**< **(NREP=***n***)** >
specifies a method for computing 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.

Table 100.3 summarizes these methods.

**Table 100.3** Options for Confidence Intervals

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

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

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

When you specify the SPARSITY option, you have two suboptions for estimating the sparsity function. If you specify the IID suboption, the sparsity function is estimated by assuming that the errors in the linear model are independent and identically distributed (iid). By default, the sparsity function is estimated by assuming that the conditional quantile function is locally linear. For more information, see the section "Sparsity" on page 8293. For both methods, two bandwidth selection methods are available: You can specify the BF suboption for the Bofinger method or the HS suboption for the Hall-Sheather method. By default, the Hall-Sheather method is used.

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

**DATA=***SAS-data-set*
specifies the input SAS data set that contains the training observations to be 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. For a detailed description of the contents of the INEST= data set, see the section "INEST= Data Set" on page 8304.

**NAMELEN=***n*
restricts the length of effect names in tables and output data sets to *n* characters, where *n* is a value between 20 and 200. By default, NAMELEN=20.

**ORDER=DATA | FORMATTED | FREQ | INTERNAL**
specifies the sort order for the levels of the classification variables (which are specified in the CLASS statement).

This option applies to the levels for all classification variables, except when you use the (default) ORDER=FORMATTED option with numeric classification variables that have no explicit format. In that case, the levels of such variables are ordered by their internal value.

The ORDER= option can take the following values:

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

By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine-dependent.

For more information about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Language Reference: Concepts*.

**OUTEST=***SAS-data-set*
specifies an output SAS data set to contain the parameter estimates for all quantiles. See the section "OUTEST= Data Set" on page 8304 for a detailed description of the contents of the OUTEST= data set.

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

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

When you specify only one *plot-request*, you can omit the parentheses around the plot request.

Here are some examples:

```
plots=ddplot
plots=(ddplot rdplot)
```

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;

proc quantreg plots=fitplot;
   model y=x1;
run;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

You can specify the following *global-plot-options*, which apply to all plots that PROC QUANTREG generates:

**MAXPOINTS=NONE | *number***
suppresses plots that have elements that require processing more than *number* points. The default is MAXPOINTS=5000. This cutoff is ignored if you specify MAXPOINTS=NONE.

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

You can specify the following *plot-requests*:

**ALL**
creates all appropriate plots.

**DDPLOT< (LABEL=ALL | LEVERAGE | NONE | OUTLIER) >**
creates a plot of robust distance against Mahalanobis distance. For more information about robust distance, see the section "Leverage Point and Outlier Detection" on page 8303. The LABEL= option specifies how the points on this plot are to be labeled, as summarized by Table 100.4.

**Table 100.4** Options for Label

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

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

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

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

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

**NONE**
suppresses all plots.

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

**RDPLOT< (LABEL=ALL | LEVERAGE | NONE | OUTLIER) >**
creates the plot of standardized residual against robust distance. For more information about robust distance, see the section "Leverage Point and Outlier Detection" on page 8303.

The LABEL= option specifies a label method for points on this plot. These label methods are described in Table 100.4.

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

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

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

## BY Statement

**BY** *variables* **;**

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

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

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

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

## CLASS Statement

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

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

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

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

In any case, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*. You can adjust the order of CLASS variable levels with the ORDER= option in the PROC QUANTREG statement.

You can specify the following *option* in the CLASS statement after a slash (/):

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

## CONDDIST Statement

**<** *label***:>** **CONDDIST** *options* **;**

The CONDDIST statement estimates and displays the conditional and marginal probability functions for the response random variable. These functions include the cumulative distribution functions (CDFs) and the probability density functions (PDFs). For more information about the probability functions, see the section "Estimating Probability Functions by Using the CONDDIST Statement" on page 8298.

You can specify multiple CONDDIST statements. You can use the optional *label*, which must be a valid SAS name, to identify the output group for its corresponding CONDDIST statement.

Table 100.5 summarizes the *options* available in the CONDDIST statement.

**Table 100.5** CONDDIST Statement Options

| Option | Description |
|---|---|
| HIDERAW= | Hides the observed marginal distribution of the training response for the analysis |
| MWU= | Performs the Mann-Whitney U test against the observed marginal CDF sample of the training response variable |

**Table 100.5** *continued*

| Option | Description |
|---|---|
| OBS= | Specifies a subset of the training observations for the analysis by using the indices of these observations in the DATA= data set of the PROC QUANTREG statement |
| PDF=KDE | Specifies the kernel density estimator for estimating the probability density functions |
| PLOTS= | Specifies options for graphically displaying the probability functions |
| SHOWAVG= | Requests the distribution of the training response that is conditional on the average explanatory covariates for the analysis |
| TESTDATA= | Specifies an input SAS data set that contains test observations for estimating the probability functions |

You can specify the following *options*:

**HIDERAW**

**HR**

hides the observed marginal distribution of the training response for the analysis. This distribution is estimated from the sample of all the response values $y_i$ in the training data without using the quantile regression model. The CONDDIST statement assigns the type "Observed" and the label "TrainObs" to this distribution.

**MWU**

**WILCOXON**

**RANKSUM**

performs the Mann-Whitney U test (also called the Wilcoxon rank-sum test) for each CDF sample of all specified observations against the CDF sample of the observed training response for the DATA= data set in the PROC QUANTREG statement.

**OBS=***number-list*

specifies the observation indices to use for the conditional distribution analysis. Each observation index identifies a training observation in the DATA= data set that you specify in the PROC QUANTREG statement. The CONDDIST statement types the distributions of these observations as "Fit for Obs" and labels the distributions by using the observation ID values (if available) or the observation indices.

**PDF=KDE(***kde-options***)**

specifies the kernel density estimator for estimating the probability density functions. For more information about the kernel density estimates, see the section "Probability Density Functions" on page 8301.

You can specify the following *kde-options*:

**C=***value*

specifies the standardized bandwidth parameter $c$ for the kernel density estimator.

You can specify one of the following *values*:

*number* specifies a positive number.

**MISE** minimizes the approximate mean integrated square error (MISE).

**SJPI** computes the bandwidth parameter by using the Sheather-Jones plug-in method.

By default, C=MISE.

**K=NORMAL | TRIANGULAR | QUADRATIC**

specifies the type of kernel function to use for the kernel density estimator. NORMAL species the normal kernel. TRIANGULAR species the triangular kernel. QUADRATIC species the quadratic kernel. By default, K=NORMAL.

**LOWER=***value*

**L=***value*

specifies the lower bound value for the kernel density curves that suppresses the lower tails of the PDF estimates. If *value* is smaller than the minimum quantile-grid value of all the PDF estimates, the CONDDIST statement ignores this suboption and outputs the minimum quantile-grid value as the lower bound value in the density estimation table.

**LUADJUST=TRIM | SCALE | REFLECT**

**LUA=TRIM | SCALE | REFLECT**

specifies the adjustment type for kernel density estimation when you specify the LOWER=*value* or UPPER=*value* option (or both).

Let $\{\hat{f}_1, \ldots, \hat{f}_l\}$ denote the lower tail of the PDF estimates, $\{\hat{f}_{l+1}, \ldots, \hat{f}_{l+m}\}$ denote the remaining PDF estimates, and $\{\hat{f}_{l+m+1}, \ldots, \hat{f}_{l+m+u}\}$ denote the upper tail of the PDF estimates. For simplicity of notation, assume that $\hat{f}_j = 0$ for all $j \leq 0$ and $j > (l + m + u)$. You can specify one of the following types:

**TRIM** trims the tails of the PDF estimates without adjusting the remaining PDF estimates, so that $\hat{f}_j = 0$ for all $j \leq l$ and $j > (l + m)$.

**SCALE** suppresses the tails of the PDF estimates and scales the remaining PDF estimates by $P/P_r$, where $P = \sum_{j=1}^{l+m+u} \hat{f}_j$ is the sum of the PDF estimates and $P_r = \sum_{j=l+1}^{l+m} \hat{f}_j$ is the sum of the remaining PDF estimates.

**REFLECT** suppresses the tails of the PDF estimates and adjusts the remaining PDF estimates by using the reflections of the suppressed tails. The adjusted $\hat{f}_{l+j}$ equals

$$\hat{f}_{l+j} + \sum_{k=1}^{K} \left( \hat{f}_{l+(1-k)m+1-j} + \hat{f}_{l+(1+k)m+1-j} \right) \text{ for } j = 1, \ldots, m$$

where $K$ is the smallest integer that satisfies $Km \geq l$ and $Km \geq u$.

By default, LUA=TRIM.

**UPPER=***value*

**U=***value*

specifies the upper bound value for the kernel density curves that suppresses the upper tails of the PDF estimates. If *value* is larger than the maximum quantile-grid value of all the PDF estimates, the CONDDIST statement ignores this suboption and outputs the maximum quantile-grid value as the upper bound value in the density estimation table.

When you specify the PLOT= option to create the CDF plot and the PDF plot, the LOWER= and UPPER= suboptions of the PDF=KDE option also set limits for the horizontal range of the plots.

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

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

specifies graphical options for displaying the probability functions.

You can specify the following *global-plot-options*, which apply to all plots that the CONDDIST statement generates:

**HIDEDROPLINES**

**HDL**

suppresses the drop lines for the responses.

**HIDEDROPNUMBERS**

**HDN**

suppresses the drop numbers for the responses.

**HIDEOBSDOTS**

**HOD**

suppresses the response dots.

**HIDEOBSLABELS**

**HOL**

suppresses the response labels.

**SHOWGRIDS**

**SG**

displays the grid lines.

You can specify the following *plot-requests*:

**ALL**

creates all appropriate plots.

**CDFPLOT**< (*plot-options*) >

plots the cumulative distribution functions (CDFs).

You can use any of the *global-plot-options* for the PLOTS option as the *plot-options* for the CDFPLOT option.

**PDFPLOT< (*plot-options*) >**
plots the probability density functions (PDFs).

You can use any of the *global-plot-options* for the PLOTS option as the *plot-options* for the PDFPLOT option.

**PPPLOT< (*plot-options*) >**
creates the scatter plot of the regression quantile levels versus the sample quantile levels for the relevant response values. This plot is referred to as the probability-probability plot.

You can use any of the *global-plot-options* of the PLOTS option, except the HOD suboption, as the *plot-options* for the PPPLOT option.

When you specify the PLOTS option, the following options in the CONDDIST statement control the visualization of their relevant probability functions:

- the HIDERAW and SHOWAVG options
- the HIDEFIT, HIDERAW, SHOWAVG, and SHOWOBS suboptions of the TESTDATA option

**SHOWAVG**

**SA**
requests the conditional distribution of the training response at average, $\mathbf{Y}|\bar{\mathbf{x}}$, for the analysis, where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the average explanatory covariates vector for all the training observations. The CONDDIST statement assigns the type "Fit at Average" and the label "TrainAvg" to this distribution.

**TESTDATA(*options*)=*SAS-data-set***
specifies the test data set for the conditional distribution analysis. The TESTDATA= data set must contain all the explanatory variables that you specify in the MODEL statement.

You can specify the following *options*:

**HIDEFIT**

**HF**
hides the fitted counterfactual marginal distribution of the test response for the analysis. This marginal distribution for $\mathbf{Y}$ integrates out the quantile regression model by pooling together all the quantile process predictions of all the test observations in the TESTDATA= data set. The CONDDIST statement assigns the type "Fit and Pooled" and the label "TestFit" to this distribution.

**HIDERAW**

**HR**
hides the observed marginal distribution of the test response for the analysis. This distribution is estimated from the sample of all the response values $y_i$ in the TESTDATA= data set without using the quantile regression model. The CONDDIST statement assigns the type "Observed" and the label "TestObs" to this distribution.

**MWU**

**WILCOXON**

**RANKSUM**

performs the Mann-Whitney U test for each CDF sample of all specified observations against the observed CDF sample for the test response of the TESTDATA= data set.

**SHOWAVG**

**SA**

requests the conditional distribution of the test response at average, $\mathbf{Y}|\bar{\mathbf{x}}$, for the analysis, where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the average explanatory covariates vector for the TESTDATA= data set. The CONDDIST statement assigns the type "Fit at Average" and the label "TestAvg" to this distribution.

**SHOWOBS**

**SO**

requests the conditional distributions of all the test observations in the analysis. The CONDDIST statement assigns the type "Fit for Obs" to the distributions of these observations and labels these distributions by using the observation ID values (if available) or the observation indices. By default, the TESTDATA= option ignores these distributions.

By default, if the TESTDATA= data set contains fewer than 16 observations, the CONDDIST statement ignores the observed marginal distribution of $\mathbf{Y}$. Otherwise, the CONDDIST statement estimates the observed marginal distributions of $\mathbf{Y}$.

## EFFECT Statement

**EFFECT** *name*=*effect-type* **(***variables* < / *options* >**) ;**

The EFFECT statement enables you to construct special collections of columns for design matrices. These collections are referred to as *constructed effects* to distinguish them from the usual model effects that are formed from continuous or classification variables, as discussed in the section "GLM Parameterization of Classification Variables and Effects" on page 393 in Chapter 19, "Shared Concepts and Topics."

You can specify the following *effect-types*:

**COLLECTION** specifies a collection effect that defines one or more variables as a single effect with multiple degrees of freedom. The variables in a collection are considered as a unit for estimation and inference.

**LAG** specifies a classification effect in which the level that is used for a particular period corresponds to the level in the preceding period.

**MULTIMEMBER | MM** specifies a multimember classification effect whose levels are determined by one or more variables that appear in a CLASS statement.

**POLYNOMIAL | POLY** specifies a multivariate polynomial effect in the specified numeric variables.

**SPLINE** specifies a regression spline effect whose columns are univariate spline expansions of one or more variables. A spline expansion replaces the original variable with an expanded or larger set of new variables.

Table 100.6 summarizes the *options* available in the EFFECT statement.

**Table 100.6** EFFECT Statement Options

| Option | Description |
|---|---|
| **Collection Effects Options** | |
| DETAILS | Displays the constituents of the collection effect |
| **Lag Effects Options** | |
| DESIGNROLE= | Names a variable that controls to which lag design an observation is assigned |
| DETAILS | Displays the lag design of the lag effect |
| NLAG= | Specifies the number of periods in the lag |
| PERIOD= | Names the variable that defines the period. This option is required. |
| WITHIN= | Names the variable or variables that define the group within which each period is defined. This option is required. |
| **Multimember Effects Options** | |
| NOEFFECT | Specifies that observations with all missing levels for the multimember variables should have zero values in the corresponding design matrix columns |
| WEIGHT= | Specifies the weight variable for the contributions of each of the classification effects |
| **Polynomial Effects Options** | |
| DEGREE= | Specifies the degree of the polynomial |
| MDEGREE= | Specifies the maximum degree of any variable in a term of the polynomial |
| STANDARDIZE= | Specifies centering and scaling suboptions for the variables that define the polynomial |
| **Spline Effects Options** | |
| BASIS= | Specifies the type of basis (B-spline basis or truncated power function basis) for the spline effect |
| DEGREE= | Specifies the degree of the spline effect |
| KNOTMETHOD= | Specifies how to construct the knots for the spline effect |

For more information about the syntax of these *effect-types* and how columns of constructed effects are computed, see the section "EFFECT Statement" on page 403 in Chapter 19, "Shared Concepts and Topics."

## ESTIMATE Statement

**ESTIMATE** < '*label*' > *estimate-specification* < (*divisor*=*n*) >
< , . . . < '*label*' > *estimate-specification* < (*divisor*=*n*) > >
< / *options* > ;

The ESTIMATE statement provides a mechanism for obtaining custom hypothesis tests. Estimates are formed as linear estimable functions of the form $\mathbf{L}\boldsymbol{\beta}$. You can perform hypothesis tests for the estimable functions, construct confidence limits, and obtain specific nonlinear transformations.

Table 100.7 summarizes the *options* available in the ESTIMATE statement.

**Table 100.7** ESTIMATE Statement Options

| Option | Description |
|---|---|
| **Construction and Computation of Estimable Functions** | |
| DIVISOR= | Specifies a list of values to divide the coefficients |
| NOFILL | Suppresses the automatic fill-in of coefficients for higher-order effects |
| SINGULAR= | Tunes the estimability checking difference |
| **Degrees of Freedom and *p*-Values** | |
| ADJUST= | Determines the method of multiple comparison adjustment of estimates |
| ALPHA=$\alpha$ | Determines the confidence level $(1-\alpha)$ |
| LOWER | Performs one-sided, lower-tailed inference |
| STEPDOWN | Adjusts multiplicity-corrected *p*-values further in a step-down fashion |
| TESTVALUE= | Specifies values under the null hypothesis for tests |
| UPPER | Performs one-sided, upper-tailed inference |
| **Statistical Output** | |
| CL | Constructs confidence limits |
| CORR | Displays the correlation matrix of estimates |
| COV | Displays the covariance matrix of estimates |
| E | Prints the **L** matrix |
| JOINT | Produces a joint *F* or chi-square test for the estimable functions |
| PLOTS= | Produces ODS statistical graphics if the analysis is sampling-based |
| SEED= | Specifies the seed for computations that depend on random numbers |
| **Generalized Linear Modeling** | |
| CATEGORY= | Specifies how to construct estimable functions for multinomial data |
| EXP | Exponentiates and displays estimates |
| ILINK | Computes and displays estimates and standard errors on the inverse linked scale |

For more information about the syntax of the ESTIMATE statement, see the section "ESTIMATE Statement" on page 451 in Chapter 19, "Shared Concepts and Topics."

## ID Statement

**ID** *variables* **;**

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

## MODEL Statement

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

You can specify main effects and interaction terms in the MODEL statement, as you can in the GLM procedure (Chapter 50, "The GLM Procedure.") Classification variables in the MODEL statement must also be specified in the CLASS statement.

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

### Options

Table 100.8 summarizes the *options* available in the MODEL statement.

**Table 100.8** MODEL Statement Options

| Option | Description |
|---|---|
| CORRB | Produces the estimated correlation matrix |
| COVB | Produces the estimated covariance matrix |
| CUTOFF= | Specifies the multiplier of the cutoff value for outlier detection |
| DIAGNOSTICS | Requests the outlier diagnostics |
| ITPRINT | Displays the iteration history |
| LEVERAGE | Requests an analysis of leverage points |
| NODIAG | Suppresses the computation for outlier diagnostics |
| NOINT | Specifies no-intercept regression |
| NOSUMMARY | Suppresses the computation for summary statistics |
| PLOT= | Requests plots |
| QUANTILE= | Specifies the quantile levels |
| SCALE= | Specifies the scale value used to compute the standardized residuals |
| SEED= | Specifies the seed for the random number generator |
| SINGULAR= | Specifies the tolerance for testing singularity |

You can specify the following *options* for the model fit.

**CORRB**

produces the estimated correlation matrix of the parameter estimates. When the resampling method is used to compute the confidence intervals, the QUANTREG procedure computes the bootstrap correlation. When the sparsity method is used to compute the confidence intervals, PROC QUANTREG

bases its computation of the asymptotic correlation on an estimator of the sparsity function. The rank method for confidence intervals does not provide a correlation estimate.

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

**CUTOFF=***value*
specifies the multiplier of the cutoff value for outlier detection. By default, CUTOFF=3.

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

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

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

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

**NOINT**
specifies no intercept regression.

**NOSUMMARY**
suppresses the computation of summary statistics. If you specify the NOSUMMARY option, the summary statistics table is not provided.

**PLOT=***plot-option*

**PLOTS=(***plot-option***)**
You can use the PLOTS= option in the MODEL statement together with ODS Graphics to request the quantile process plot in addition to all that plots that you request in the PLOT= option in the PROC QUANTREG statement.

You can specify the following *plot-option* in the MODEL statement:

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

ODS Graphics must be enabled before you request plots.

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

**QUANTILE=***number-list* | **PROCESS** | **FQPR< (***suboption***) >**

**QUANTLEV=***number-list* | **PROCESS** | **FQPR< (***suboption***) >**
specifies the quantile levels for the quantile regression. A valid quantile level must be a number in the range of (0,1). You can specify the following values for the QUANTILE= option:

*number-list*
computes quantile regression for quantile levels that are specified in the *number-list*.

**PROCESS**
computes the entire quantile process regression. If you specify QUANTILE=PROCESS, the value of the ALGORITHM= option in the PROC QUANTREG statement must be SIMPLEX either by default or by specifying it.

The QUANTILE=PROCESS option produces the quantile process estimates table and the quantile process objective function table. The size of these two tables are proportional to the number of the training observations and can be large for a large training data set. You can suppress displaying these two tables by specifying the following ODS EXCLUDE statement:

```
ods exclude ProcessEst ProcessObj;
```

In addition, you can output these tables to data sets for further processing by specifying the following ODS OUTPUT statement:

```
ods output ProcessEst=PE ProcessObj=PO;
```

**FQPR< (***suboption***) >**
uses a fast quantile process regression method to approximate quantile process regression on a grid of $n$ equally spaced quantile levels. If you specify QUANTILE=FQPR, the value of the ALGORITHM= option in the PROC QUANTREG statement must be IPM. For more details about the fast quantile process regression method, see the section "Fast Quantile Process Regression" on page 8292. You can specify the following suboptions for the FQPR option:

**N=***n*

specifies the number *n* of equally spaced quantile levels at which to fit the quantile process regression.

**OBSRATIO=***value*

**OR=***value*

specifies the number of equally spaced quantile levels as its ratio to the total number of training observations. For example, if the number of training observations is 1,000 and you specify the OR=0.2 suboption, a quantile process regression model is fit for $n = 0.2 \times 1{,}000 = 200$ equally spaced quantile levels. The FQPR option ignores the OR= suboption if a valid N= suboption is specified.

**L=***value*

specifies the starting quantile level of the quantile-level grid. By default, **L**=$1/2n$ if **U** is not specified, otherwise **L**=**U**$/(2n-1)$.

**U=***value*

specifies the ending quantile level of the quantile-level grid. **U**=$(2n-1)/2n$ if **L** is not specified, otherwise **U**=$(\mathbf{L}+2n-2)/(2n-1)$.

If you specify neither the N=*n* nor the OR=*value* suboption, the FQPR option determines the number of quantile levels as the lesser of 100 and half the number of the training observations.

Unlike the QUANTILE=PROCESS option, the QUANTILE=FQPR option does not display the quantile process estimates table and the quantile process objective function table. Instead the QUANTILE=FQPR option produces the average parameter estimates table and the average objective function table for the specified quantile-level grid. However, you can output the two quantile process tables to data sets by specifying the following ODS OUTPUT statement:

```
ods output ProcessEst=PE ProcessObj=PO;
```

By default, QUANTILE=0.5, which fits a median regression.

**SCALE=***number*

specifies the scale value to use 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 8303 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 QUANTREG statement. If you do not specify a seed, or if you specify a value less than or equal to 0, the seed is generated from reading the time of day from the computer clock.

By default or if you specify SEED=0, 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. Approximately, the test requires that a pivot be at least this value times the original diagonal value. By default, SINGULAR=1E–12.

## OUTPUT Statement

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

The OUTPUT statement creates a SAS data set to contain statistics that are calculated after PROC QUANTREG fits models for all specified quantiles that are specified in the QUANTILE= option in the MODEL statement. At least one specification of the form *keyword=name* is required.

All variables in the original data set are included in the new data set, along with the variables that are created from options in the OUTPUT statement. These new variables contain fitted values and estimated quantiles. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see *SAS Language Reference: Concepts*.

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

The following specifications can appear in the OUTPUT statement:

**OUT**=*SAS-data-set* specifies the new data set. By default, PROC QUANTREG 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. For each desired statistic, specify a *keyword* from the following list of *keywords*, an equal sign, and the name of a variable to contain the statistic.

You can specify the following *keywords*:

**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 8303 for how to define LEVERAGE.

**MAHADIST | MD** names 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 8303 for how to define OUTLIER.

**PREDICTED | P** names a variable to contain the estimated response.

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

**RESIDUAL | RES** names a variable to contain the residuals (unstandardized):

$$y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$$

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

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

**SRESIDUAL | SR** names a variable to contain the standardized residuals:

$$\frac{y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}}{\hat{\sigma}}$$

See the section "Leverage Point and Outlier Detection" on page 8303 for how to compute $\sigma$.

**STDP** names a variable to contain the estimates of the standard errors of the estimated response.

## PERFORMANCE Statement

You can use the PERFORMANCE statement to change default options that affect the performance of PROC QUANTREG and to request tables that show the performance options in effect and the timing details.

**PERFORMANCE** < *options* > **;**

You can specify the following *options*:

**CPUCOUNT=***number* **| ACTUAL**
specifies the number of processors to use in the computation of the interior point algorithm. CPUCOUNT=ACTUAL sets CPUCOUNT to be the number of physical processors available, which this can be less than the physical number of CPUs if the SAS process has been restricted by system administration tools. You can specify any integer from 1 to 1024 for *number*. Setting CPUCOUNT= to a *number* greater than the actual number of available CPUs might result in reduced performance. If CPUCOUNT=1, then NOTHREADS is in effect, and PROC QUANTREG uses singly threaded code. This option overrides the SAS system option CPUCOUNT=.

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

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

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

## TEST Statement

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

In quantile regression analysis, you might be interested in testing whether a covariate effect is statistically significant for a given quantile. In other situations, you might be interested in testing whether the coefficients of a covariate are the same across a set of quantiles. You can use the TEST Statement to perform these tests.

You can submit multiple TEST statements, provided that they appear after the MODEL statement. The optional *label*, which must be a valid SAS name, is used to identify output from the corresponding TEST statement. For more information about these tests, see the section "Linear Test" on page 8296.

### Testing Effects of Covariates

You can use TEST statement to obtain a test for the canonical linear hypothesis concerning the parameters of the tested effects,

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

where $q$ is the total number of parameters of the tested effects. The tested *effects* can be any set of effects in the MODEL statement. You can specify three types of tests (Wald, likelihood ratio, and rank methods) for testing effects of covariates by using the following *options* in the TEST statement after a slash (/):

**WALD**
requests Wald tests.

**LR**
requests likelihood ratio tests.

**RANKSCORE < (NORMAL | WILCOXON | SIGN | TAU) >**
requests rank tests. The NORMAL, WILCOXON, and SIGN functions are implemented and suitable for iid error models, and the TAU score function is implemented and appropriate for non-iid error models. By default, the TAU score function is used. See Koenker (2005) for more information about the score functions.

### Testing for Heteroscedasticity

You can test whether there is any difference among the estimated coefficients across quantiles if several quantiles are specified in the MODEL statement. The test for such heteroscedasticity can be requested by the option QINTERACT after a slash (/) in the TEST statement. See Example 100.5.

## 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*. 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. For more information about weighted quantile regression, see the section "Details: QUANTREG Procedure" on page 8284.

# Details: QUANTREG Procedure

## Quantile Regression as an Optimization Problem

The generic model for linear quantile regression is

$$Q_{Y|\mathbf{x}}(\tau) = \mathbf{x}'\boldsymbol{\beta}(\tau)$$

where $Y$ is the response random variable, $\mathbf{x}$ is the explanatory covariates vector, $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \ldots, \beta_p(\tau))'$ is the $(p \times 1)$ vector of the functional model parameters at the quantile level $\tau$, and $Q_{Y|\mathbf{x}}(\cdot)$ is the quantile function for $Y$ conditional on $\boldsymbol{X} = \mathbf{x}$.

This generic model is compatible with the following1 linear model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}(\tau) + \epsilon_i(\tau) \text{ for } i = 1, \ldots, n$$

where $y_i$ is the response value, $\mathbf{x}_i$ is the explanatory covariates vector, and $\epsilon_i(\tau) = y_i - Q_{Y|\mathbf{x}_i}(\tau)$ is an unknown error.

$L_1$ 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 $\hat{\boldsymbol{\beta}}_{LAR}$, referred to as the $L_1$*-norm estimate*, is obtained as the solution of the following minimization problem:

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

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

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

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

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

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

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

Weighted regression quantiles $\boldsymbol{\beta}_w$ can be used for L-estimation (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 that contain 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 the worst-case performance of the simplex algorithm. More important, experience has shown that the interior point algorithm is advantageous for larger problems.

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

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

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

### Simplex Algorithm

Let $\boldsymbol{\mu} = [\mathbf{y} - \mathbf{A}'\boldsymbol{\beta}]_+$, $\boldsymbol{\nu} = [\mathbf{A}'\boldsymbol{\beta} - \mathbf{y}]_+$, $\boldsymbol{\phi} = [\boldsymbol{\beta}]_+$, and $\boldsymbol{\varphi} = [-\boldsymbol{\beta}]_+$, where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the response vector, $\mathbf{A}' = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is the $(n \times p)$ regressor matrix, and $[z]_+$ is the nonnegative part of $z$.

Let $D_{LAR}(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|$. For the $L_1$ problem, the simplex approach solves $\min_{\boldsymbol{\beta}} D_{LAR}(\boldsymbol{\beta})$ by reformulating it as the constrained minimization problem

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

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

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

$$(P) \qquad \min_{\boldsymbol{\theta}} \mathbf{d}'\boldsymbol{\theta}; \text{ subject to } \mathbf{B}\boldsymbol{\theta} = \mathbf{y}, \boldsymbol{\theta} \geq \mathbf{0}$$

This problem has the following dual formulation:

$$(D) \qquad \max_{z} \mathbf{y}'\mathbf{z}; \text{ subject to } \mathbf{B}'\mathbf{z} \leq \mathbf{d}$$

This formulation can be simplified as

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

By setting $\boldsymbol{\eta} = \frac{1}{2} + \frac{1}{2}\mathbf{e}, \mathbf{b} = \frac{1}{2}\mathbf{A}\mathbf{e}$, the problem becomes

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

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

$$\max_{z} \mathbf{y}'\mathbf{z}; \quad \text{subject to } \mathbf{A}\mathbf{z} = (1-\tau)\mathbf{A}\mathbf{e}, \mathbf{z} \in [0,1]^n$$

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

Although this special version of the simplex algorithm was introduced for median regression, it extends naturally to quantile regression for any given quantile and even to the entire quantile process (Koenker and d'Orey 1994). It greatly reduces the computing time that is required by the general simplex algorithm, and it is suitable for data sets with fewer than 5,000 observations and 50 variables.

## Interior Point Algorithm

The ALGORITHM=INTERIOR option implements an interior point algorithm. This algorithm uses the primal-dual predictor-corrector method that is proposed by Lustig, Marsten, and Shanno (1992). Roos, Terlaky, and Vial (1997) provide 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 linear programming setting, let $\mathbf{c} = -\mathbf{y}$, let $\mathbf{b} = (1-\tau)\mathbf{A}\mathbf{e}$, and let $u$ be the general upper bound. The dual form of quantile regression solves the following linear programming primal problem:

$$\min\{\mathbf{c}'\mathbf{z}\}; \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{b}, \mathbf{0} \le \mathbf{z} \le \mathbf{u}$$

This primal 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 $\mathbf{v}$ be the primal slack so that $\mathbf{z} + \mathbf{v} = \mathbf{u}$. Associate dual variables $w$ with these constraints. The interior point algorithm solves the system of equations to satisfy the Karush-Kuhn-Tucker (KKT) conditions for optimality:

$$\mathbf{b} = \mathbf{A}\mathbf{z}$$
$$\mathbf{u} = \mathbf{z} + \mathbf{v}$$
$$\mathbf{c} = \mathbf{A}'\mathbf{t} + \mathbf{s} - \mathbf{w}$$
$$\mathbf{0} = \mathbf{Z}\mathbf{S}\mathbf{e}$$
$$\mathbf{0} = \mathbf{V}\mathbf{W}\mathbf{e}$$
$$\mathbf{z}, \mathbf{s}, \mathbf{v}, \mathbf{w} \ge \mathbf{0}$$

where $\mathbf{W} = \mathrm{diag}(\mathbf{w})$ $\left(\text{that is, } W_{i,j} = \begin{cases} w_i & \text{for } i = j \\ 0 & \text{otherwise} \end{cases}\right)$, $\mathbf{V} = \mathrm{diag}(\mathbf{v})$, $\mathbf{Z} = \mathrm{diag}(\mathbf{z})$, $\mathbf{S} = \mathrm{diag}(\mathbf{s})$.

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

$$
\begin{aligned}
0 &= \mathbf{v}'_{opt}\mathbf{w}_{opt} = (\mathbf{u} - \mathbf{z}_{opt})'\mathbf{w}_{opt} = \mathbf{u}'\mathbf{w}_{opt} - \mathbf{z}'_{opt}\mathbf{w}_{opt} \\
0 &= \mathbf{z}'_{opt}\mathbf{s}_{opt} = \mathbf{s}'_{opt}\mathbf{z}_{opt} = (\mathbf{c} - \mathbf{A}'\mathbf{t}_{opt} + \mathbf{w}_{opt})'\mathbf{z}_{opt} \\
&= \mathbf{c}'\mathbf{z}_{opt} - \mathbf{t}'_{opt}(\mathbf{A}\mathbf{z}_{opt}) + \mathbf{w}'_{opt}\mathbf{z}_{opt} = \mathbf{c}'\mathbf{z}_{opt} - \mathbf{b}'\mathbf{t}_{opt} + \mathbf{u}'\mathbf{w}_{opt}
\end{aligned}
$$

Therefore

$$0 = \mathbf{c}'\mathbf{z}_{opt} - \mathbf{b}'\mathbf{t}_{opt} + \mathbf{u}'\mathbf{w}_{opt}$$

The *duality gap*, $\mathbf{c}'\mathbf{z} - \mathbf{b}'\mathbf{t} + \mathbf{u}'\mathbf{w}$, measures the convergence of the algorithm. You can specify a tolerance for this convergence criterion in the TOLERANCE= option in the PROC QUANTREG statement.

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

- Primal bound constraints can be broken: $\boldsymbol{\delta}_b = \mathbf{u} - \mathbf{z} - \mathbf{v} \neq \mathbf{0}$.
- Primal constraints can be broken: $\boldsymbol{\delta}_c = \mathbf{b} - \mathbf{A}\mathbf{z} \neq \mathbf{0}$.
- Dual constraints can be broken: $\boldsymbol{\delta}_d s = \mathbf{c} - \mathbf{A}'\mathbf{t} - \mathbf{s} + \mathbf{w} \neq \mathbf{0}$.
- Complementarity conditions are unsatisfied: $\mathbf{z}'\mathbf{s} \neq 0$ and $\mathbf{v}'\mathbf{w} \neq 0$.

The interior point algorithm works by using Newton's method to find a direction $(\boldsymbol{\Delta} z^k, \boldsymbol{\Delta} t^k, \boldsymbol{\Delta} s^k, \boldsymbol{\Delta} v^k, \boldsymbol{\Delta} w^k)$ to move from the current solution $(\mathbf{z}^k, \mathbf{t}^k, \mathbf{s}^k, \mathbf{v}^k, \mathbf{w}^k)$ toward a better solution:

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

$\kappa$ is the *step length* and is assigned a value as large as possible, but not so large that a $\mathbf{z}_i^{k+1}$ or $\mathbf{s}_i^{k+1}$ is "too close" to 0. You can control the step length in the KAPPA= option in the PROC QUANTREG 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 $(\boldsymbol{\Delta} z^k_{aff}, \boldsymbol{\Delta} t^k_{aff}, \boldsymbol{\Delta} s^k_{aff}, \boldsymbol{\Delta} v^k_{aff}, \boldsymbol{\Delta} w^k_{aff})$ in which to move. This is known as the *affine* step.

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

$$
\begin{aligned}
\boldsymbol{\delta}_b &= \boldsymbol{\Delta} z_{aff} + \boldsymbol{\Delta} v_{aff} \\
\boldsymbol{\delta}_c &= \mathbf{A}\boldsymbol{\Delta} z_{aff} \\
\boldsymbol{\delta}_d &= \mathbf{A}'\boldsymbol{\Delta} t_{aff} + \boldsymbol{\Delta} s_{aff} - \boldsymbol{\Delta} w_{aff} = \boldsymbol{\delta}_d \\
-\mathbf{ZSe} &= \mathbf{S}\boldsymbol{\Delta} z_{aff} + \mathbf{Z}\boldsymbol{\Delta} s_{aff} \\
-\mathbf{VWe} &= \mathbf{V}\boldsymbol{\Delta} w_{aff} + \mathbf{W}\boldsymbol{\Delta} z_{aff}
\end{aligned}
$$

Therefore, the following computations are involved in solving the affine step, where $\kappa$ is the *step length* as before:

$$
\begin{aligned}
\boldsymbol{\Theta} &= \mathbf{S}\mathbf{Z}^{-1} + \mathbf{W}\mathbf{V}^{-1} \\
\boldsymbol{\rho} &= \boldsymbol{\Theta}^{-1}(\boldsymbol{\delta}_d + (\mathbf{S}-\mathbf{W})\mathbf{e} - \mathbf{V}^{-1}\mathbf{W}\boldsymbol{\delta}_b) \\
\boldsymbol{\Delta} t_{aff} &= (\mathbf{A}\boldsymbol{\Theta}^{-1}\mathbf{A}')^{-1}(\boldsymbol{\delta}_c + \mathbf{A}\boldsymbol{\rho}) \\
\boldsymbol{\Delta} z_{aff} &= \boldsymbol{\Theta}^{-1}\mathbf{A}'\boldsymbol{\Delta} t_{aff} - \boldsymbol{\rho} \\
\boldsymbol{\Delta} v_{aff} &= \boldsymbol{\delta}_b - \boldsymbol{\Delta} z_{aff} \\
\boldsymbol{\Delta} w_{aff} &= -\mathbf{W}\mathbf{e} - \mathbf{V}^{-1}\mathbf{W}\boldsymbol{\Delta} z_{aff} \\
\boldsymbol{\Delta} s_{aff} &= -\mathbf{S}\mathbf{e} - \mathbf{Z}^{-1}\mathbf{S}\boldsymbol{\Delta} z_{aff} \\
(\mathbf{z}_{aff}, \mathbf{t}_{aff}, \mathbf{s}_{aff}, \mathbf{v}_{aff}, \mathbf{w}_{aff}) &= (\mathbf{z}, \mathbf{t}, \mathbf{s}, \mathbf{v}, \mathbf{w}) + \kappa(\boldsymbol{\Delta} z_{aff}, \boldsymbol{\Delta} t_{aff}, \boldsymbol{\Delta} s_{aff}, \boldsymbol{\Delta} v_{aff}, \boldsymbol{\Delta} w_{aff})
\end{aligned}
$$

The success of the affine step is gauged by calculating the complementarity of $\mathbf{z}'\mathbf{s}$ and $\mathbf{v}'\mathbf{w}$ at $(\mathbf{z}^k_{aff}, \mathbf{t}^k_{aff}, \mathbf{s}^k_{aff}, \mathbf{v}^k_{aff}, \mathbf{w}^k_{aff})$ and comparing it with the complementarity at the starting point $(\mathbf{z}^k, \mathbf{t}^k, \mathbf{s}^k, \mathbf{v}^k, \mathbf{w}^k)$. If the affine step was successful in reducing the complementarity by a substantial amount, the need for centering is not great. Therefore, a value close to 0 is assigned to $\sigma$ in the following second linear system, which is used to determine a centering vector.

The following linear system is solved to determine a centering vector $(\boldsymbol{\Delta} z_c, \boldsymbol{\Delta} t_c, \boldsymbol{\Delta} s_c, \boldsymbol{\Delta} v_c, \boldsymbol{\Delta} w_c)$ from $(\mathbf{z}_{aff}, \mathbf{t}_{aff}, \mathbf{s}_{aff}, \mathbf{v}_{aff}, \mathbf{w}_{aff})$:

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

where $\zeta_{start} = \mathbf{z}'\mathbf{s} + \mathbf{v}'\mathbf{w}$, complementarity at the start of the iteration

$\zeta_{aff} = \mathbf{z}'_{aff}\mathbf{s}_{aff} + \mathbf{v}'_{aff}\mathbf{w}_{aff}$, the affine complementarity

$\mu = \zeta_{aff}/2n$, the average complementarity

$\sigma = (\zeta_{aff}/\zeta_{start})^3$

However, if the affine step was unsuccessful, then centering is deemed beneficial, and a value close to 1.0 is assigned to $\sigma$. In other words, the value of $\sigma$ is adaptively altered depending on the progress made toward the optimum.

Therefore, the following computations are involved in solving the centering step:

$$\begin{aligned}
\boldsymbol{\rho} &= \boldsymbol{\Theta}^{-1}(\sigma\mu(\mathbf{Z}^{-1} - \mathbf{V}^{-1})\mathbf{e} - \mathbf{Z}^{-1}\mathbf{Z}_{aff}\mathbf{S}_{aff}\mathbf{e} + \mathbf{V}^{-1}\mathbf{V}_{aff}\mathbf{W}_{aff}\mathbf{e}) \\
\boldsymbol{\Delta} t_c &= (\mathbf{A}\boldsymbol{\Theta}^{-1}\mathbf{A}')^{-1}\mathbf{A}\boldsymbol{\rho} \\
\boldsymbol{\Delta} z_c &= \boldsymbol{\Theta}^{-1}\mathbf{A}'\boldsymbol{\Delta} t_c - \boldsymbol{\rho} \\
\boldsymbol{\Delta} v_c &= -\boldsymbol{\Delta} z_c \\
\boldsymbol{\Delta} w_c &= \sigma\mu\mathbf{V}^{-1}\mathbf{e} - \mathbf{V}^{-1}\mathbf{V}_{aff}\mathbf{W}_{aff}\mathbf{e} - \mathbf{V}^{-1}\mathbf{W}_{aff}\boldsymbol{\Delta} v_c \\
\boldsymbol{\Delta} s_c &= \sigma\mu\mathbf{Z}^{-1}\mathbf{e} - \mathbf{Z}^{-1}\mathbf{Z}_{aff}\mathbf{S}_{aff}\mathbf{e} - \mathbf{Z}^{-1}\mathbf{S}_{aff}\boldsymbol{\Delta} z_c
\end{aligned}$$

Then

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

where, as before, $\kappa$ is the step length, which is assigned a value as large as possible but not so large that a $\mathbf{z}_i^{k+1}$, $\mathbf{s}_i^{k+1}$, $\mathbf{v}_i^{k+1}$, or $\mathbf{w}_i^{k+1}$ is "too close" to 0.

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 $(\mathbf{A}\boldsymbol{\Theta}^{-1}\mathbf{A}')$ has already been factorized in order to solve the first linear system.

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

### Efficient Interior Point Algorithm

The ALGORITHM=IPM option implements a more efficient interior point algorithm than the one that is used when ALGORITHM=INTERIOR. The computing strategy of the ALGORITHM=IPM option is the same as the strategy for the ALGORITHM=INTERIOR option, but the ALGORITHM=IPM option implements the algorithm by using more efficient matrix functions. The ALGORITHM=IPM option uses the complementarity value to measure the convergence of the algorithm, which is different from the dual gap value that is used when ALGORITHM=INTERIOR. The complementarity value is defined as $(\mathbf{z}'\mathbf{s} + \mathbf{v}'\mathbf{w})$. You can specify a tolerance for this complementarity convergence criterion by using the TOLERANCE= option in the PROC QUANTREG statement. Unlike the ALGORITHM=INTERIOR option, the ALGORITHM=IPM option does not support the KAPPA= option.

### Smoothing Algorithm

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

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

where

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

Here $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, and the threshold $\gamma$ is a positive real number. The function $D_\gamma$ is continuously differentiable, and a minimizer $\beta_\gamma$ of $D_\gamma$ is close to a minimizer $\hat{\boldsymbol{\beta}}_{LAR}$ of $D_{LAR}(\boldsymbol{\beta})$ when $\gamma$ is close to 0.

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

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

$$D_{\rho_\tau}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta})$$

can be approximated by the smooth function

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

where

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

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

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

Define the sign vector $\mathbf{s}_\gamma(\boldsymbol{\beta}) = (s_1(\boldsymbol{\beta}), \ldots, s_n(\boldsymbol{\beta}))'$ as

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

and introduce

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

Therefore,

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

This equation yields

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

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

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

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

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

$$D^{(2)}_{\gamma,\tau}(\boldsymbol{\beta}) = \frac{1}{\gamma}\mathbf{A}\mathbf{W}_{\gamma,\tau}(\boldsymbol{\beta})\mathbf{A}'$$

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 *$\gamma$-feasible sign vector* if there exists $\boldsymbol{\beta} \in \mathbf{R}^p \backslash B_{\gamma,\tau}$ with $\mathrm{s}_\gamma(\boldsymbol{\beta}) = \mathrm{s}$. If s is $\gamma$-feasible, then $Q_s$ is defined as the quadratic function $Q_s(\boldsymbol{\alpha})$ that is derived from $D_{\gamma,\tau}(\boldsymbol{\beta})$ by substituting s for $\mathrm{s}_\gamma$. Thus, for any $\boldsymbol{\beta}$ with $\mathrm{s}_\gamma = \mathrm{s}$,

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

In the domain $C_s = \{\alpha | \mathrm{s}_\gamma(\alpha) = \mathrm{s}\}$,

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

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

$$(\gamma, \theta, \mathrm{s})$$

Again following Madsen and Nielsen (1993), let $(\gamma, \theta, \mathrm{s})$ be a *feasible reference* if s is a $\gamma$-feasible sign vector, where $\theta \in C_s$, and let $(\gamma, \theta, \mathrm{s})$ be a *solution reference* if s is feasible and $\theta$ minimizes $D_{\gamma,\tau}$.

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

1. Find an initial solution reference $(\gamma, \boldsymbol{\beta}_\gamma, \mathrm{s})$.
2. Repeat the following substeps until $\gamma = 0$.
   a) Decrease $\gamma$.
   b) Find a solution reference $(\gamma, \boldsymbol{\beta}_\gamma, \mathrm{s})$.

   $\boldsymbol{\beta}_0$ is the solution.

By default, the initial solution reference is found by letting $\boldsymbol{\beta}_\gamma$ be the least squares solution. Alternatively, you can specify the initial solution reference with the INEST= option in the PROC QUANTREG statement. Then $\gamma$ and s are chosen according to these initial values.

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

$$\text{RRATIO} = \begin{cases} 0.1 & \text{if } n \geq 10{,}000 \text{ and } p \leq 20 \\ 0.9 & \text{if } \frac{p}{n} \geq 0.1 \text{ or } \{n \leq 5{,}000 \text{ and } p \geq 300\} \\ 0.5 & \text{otherwise} \end{cases}$$

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 $\mathbf{A}\mathbf{A}'$ is dense. See Chen (2007) for a complete smoothing algorithm and details.

### Fast Quantile Process Regression

The QUANTILE=FQPR option in the MODEL statement implements a fast quantile process regression (FQPR) method. This method can efficiently fit multiple quantile regression models by using the divide-and-conquer strategy proposed by Yao (2017).

The FQPR method begins by fitting a quantile regression model for a selected quantile level in a specified quantile-level grid of $q$-nodes . The quantile level is selected as the closest to 0.5 among all the quantile levels in the grid. Using this fit, FQPR defines two subsets of the data based on whether observed values $y$ are above or below their linear predictors $x\beta$ in this regression fit. Then FQPR proceeds to recursively perform separate quantile process regressions on the two subsets.

The successive quantile regression steps of FQPR are thus fit to smaller and smaller data sets. It is this sequence of reductions in problem size that provides the very significant reduction in computational cost that FQPR can achieve. In particular, FQPR can fit a quantile process regression model for $q$ equally spaced quantiles in the time that it would approximately take to fit just $\log(q)$ quantile regression models to all the data.

The QUANTILE=FQPR option uses the efficient interior point algorithm to fit single-level quantile regression models as described in the section "Efficient Interior Point Algorithm" on page 8289.

## Confidence Interval

The QUANTREG procedure provides three methods to compute confidence intervals for the regression quantile parameter $\boldsymbol{\beta}(\tau)$: 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 uses a local estimate of the sparsity function to compute a Huber sandwich estimate. The rank method, which computes confidence intervals by inverting the rank score test, does not suffer from this problem. However, the rank method uses the simplex algorithm and is computationally expensive with large data sets. The resampling method, which uses the bootstrap approach, addresses these problems, but at a computation cost.

Based on these properties, the QUANTREG uses a combination of the resampling and rank methods as the default. For data sets that have more than either 5,000 observations or more than 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 QUANTREG statement.

## Sparsity

Consider the linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$$

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

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

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

This asymptotic distribution for the regression quantile $\hat{\boldsymbol{\beta}}(\tau)$ 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.

Because

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

$s(t)$ 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 0 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. 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 $z_\alpha$ satisfies $\Phi(z_\alpha) = 1 - \alpha/2$ for the construction of $1-\alpha$ confidence intervals. The following quantity is not sensitive to $f$ and can be estimated by assuming $f$ is Gaussian:

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

$F^{-1}$ can be estimated in either of the following ways:

- by the empirical quantile function of the residuals from the quantile regression fit,

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

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

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

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

$$\hat{F}^{-1}(t) = \begin{cases} r_{(1)} & \text{if } t \in [0, 0.5/n) \\ \lambda r_{(i+1)} + (1-\lambda) r_{(i)} & \text{if } t \in [(i-0.5)/n, (i+0.5)/n) \\ r_{(n)} & \text{if } t \in [(n-0.5)/n, 1] \end{cases}$$

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

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

By default, the QUANTREG procedure computes non-iid confidence intervals. You can request iid confidence intervals by specifying the IID option in the PROC QUANTREG 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 $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\epsilon}$. Here, $(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{A}'$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is the unknown error vector.

See Gutenbrunner and Jureckova (1992) for more details. By inverting this test, confidence intervals can be computed for the regression quantiles that correspond to $\beta_2$.

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

$$\max_a \{(\mathbf{y} - \mathbf{X}_2\boldsymbol{\eta})'a | \mathbf{X}_1'a = (1-t)\mathbf{X}_1'\mathbf{e},\ a \in [0,1]^n\}$$

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

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

yields the $\tau$-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 $H_0$: $\beta_2 = \eta$,

$$S_n(\eta) = n^{-1/2}\mathbf{X}_2'\hat{b}_n(\eta) \rightarrow N(0, \tau(1-\tau)\boldsymbol{\Omega}_n)$$

for large $n$, where $\boldsymbol{\Omega}_n = n^{-1}\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2$.

Let

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

Then $T_n(\hat{\beta}_2(\tau)) = 0$ from the constraint $\mathbf{A}\hat{a} = (1-\tau)\mathbf{Ae}$ in the full model. In order to obtain confidence intervals for $\beta_2$, a critical value can be specified for $T_n$. The dual vector $\hat{a}_n(\eta)$ is a piecewise constant in $\eta$, and $\eta$ can be altered without compromising the optimality of $\hat{a}_n(\eta)$ as long as the signs of the residuals in the primal quantile regression problem do not change. When $\eta$ gets to such a boundary, the solution does change. But it can be restored by taking one simplex pivot. The process can continue in this way until $T_n(\eta)$ exceeds the specified critical value. Because $T_n(\eta)$ is piecewise constant, interpolation can be used to obtain the desired level of confidence interval (Koenker and d'Orey 1994).

### Resampling

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

In contrast with these bootstrap methods, Parzen, Wei, and Ying (1994) observed that the following estimating equation for the $\tau$ regression quantile is a pivotal quantity for the $\tau$ quantile regression parameter $\beta_\tau$:

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

In other words, the distribution of $\mathbf{S}(\boldsymbol{\beta})$ can be generated exactly by a random vector $\mathbf{U}$, which is a weighted sum of independent, re-centered Bernoulli variables. They further showed that for large $n$, the distribution of $\hat{\boldsymbol{\beta}}(\tau) - \beta_\tau$ can be approximated by the conditional distribution of $\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_n(\tau)$, where $\hat{\boldsymbol{\beta}}_U$ solves an augmented quantile regression problem by using $n+1$ observations that have $\mathbf{x}_{n+1} = -n^{-1/2}u/\tau$ and $y_{n+1}$ sufficiently large for a given realization of $u$. By exploiting the asymptotically pivotal 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 called the Markov chain marginal bootstrap (MCMB). For quantile regression, the MCMB method has the advantage that it solves $p$ one-dimensional equations instead of solving $p$-dimensional equations, as the previous bootstrap methods do. This greatly improves the feasibility of the resampling method in computing confidence intervals for regression quantiles.

## Covariance-Correlation

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

The QUANTREG procedure provides two methods for computing the covariance and correlation matrices of the estimated parameters: an asymptotic method and a bootstrap method. Bootstrap covariance and correlation

matrices are computed when resampling confidence intervals are computed. Asymptotic covariance and correlation matrices are computed when asymptotic confidence intervals are computed. The rank method for confidence intervals does not provide a covariance-correlation estimate.

### Asymptotic Covariance-Correlation

This method corresponds to the sparsity method for the confidence intervals. For the sparsity function in the computation of the asymptotic covariance and correlation, the QUANTREG procedure provides both iid and non-iid estimates. By default, the QUANTREG procedure computes non-iid estimates.

### Bootstrap Covariance-Correlation

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

## Linear Test

Consider the linear model

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \epsilon_i$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $p$- and $q$-dimensional unknown parameters and $\{\epsilon_i\}$, $i = 1, \ldots, n$, are errors with unknown density function $f_i$. Let $\mathbf{x}_i' = (\mathbf{x}_{1i}', \mathbf{x}_{2i}')$, and let $\hat{\boldsymbol{\beta}}_1(\tau)$ and $\hat{\boldsymbol{\beta}}_2(\tau)$ be the parameter estimates for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, respectively at the $\tau$ quantile. The covariance matrix $\boldsymbol{\Omega}$ for the parameter estimates is partitioned correspondingly as $\boldsymbol{\Omega}_{ij}$ with $i = 1, 2$; $j = 1, 2$; and $\boldsymbol{\Omega}^{22} = (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}$.

### Testing Effects of Covariates

Three tests are available in the QUANTREG procedure for the linear null hypothesis $H_0 : \beta_2 = 0$ at the $\tau$ quantile:

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

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

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

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

  where $\hat{\omega}(\tau) = \sqrt{\tau(1-\tau)}\hat{s}(\tau)$ and $\hat{s}(\tau)$ is the estimated sparsity function. The estimator that is 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 $D_0(\tau) = \sum \rho_\tau(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}(\tau))$, and let $D_1(\tau) = \sum \rho_\tau(y_i - \mathbf{x}_{1i}'\hat{\boldsymbol{\beta}}_1(\tau))$. Set

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

where $\hat{s}(\tau)$ is the estimated sparsity function.

- The rank test statistic is given by

  $$T_R(\tau) = \mathbf{S}_n' \mathbf{M}_n^{-1} \mathbf{S}_n / A^2(\varphi)$$

  where

  $$\mathbf{S}_n = n^{-1/2}(\mathbf{X}_2 - \hat{\mathbf{X}}_2)'\hat{\mathbf{b}}_n$$

  $$\boldsymbol{\Psi} = \mathrm{diag}(f_i(Q_{y_i}(\tau|\mathbf{x}_{1i}, \mathbf{x}_{2i})))$$

  $$\hat{\mathbf{X}}_2 = \mathbf{X}_1(\mathbf{X}_1'\boldsymbol{\Psi}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$$

  $$\mathbf{M}_n = (\mathbf{X}_2 - \hat{\mathbf{X}}_2)(\mathbf{X}_2 - \hat{\mathbf{X}}_2)'/n$$

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

  $$\hat{\mathbf{a}}(t) = \max_a\{\mathbf{y}'\mathbf{a}|\mathbf{X}_1'\mathbf{a} = (1-t)\mathbf{X}_1'\mathbf{e},\ \mathbf{a} \in [0,1]^n\}$$

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

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

  and $\varphi(t)$ is one of the following score functions:

  - Wilcoxon scores: $\phi(t) = t - 1/2$
  - normal scores: $\phi(t) = \Phi^{-1}(t)$, where $\Phi$ is the normal distribution function
  - sign scores: $\phi(t) = 1/2\mathrm{sign}(t - 1/2)$
  - tau scores: $\phi_\tau(t) = \tau - I(t < \tau)$.

  The rank test statistic $T_R(\tau)$, unlike Wald tests or likelihood ratio tests, requires no estimation of the nuisance parameter $f_i$ under iid error models (Gutenbrunner et al. 1993).

Koenker and Machado (1999) prove that the three test statistics ($T_W(\tau)$, $T_{\mathrm{LR}}(\tau)$, and $T_R(\tau)$) are asymptotically equivalent and that their distributions converge to $\chi_q^2$ under the null hypothesis, where $q$ is the dimension of $\beta_2$.

### Testing for Heteroscedasticity

After you obtain the parameter estimates for several quantiles specified in the MODEL statement, you can test whether there are significant differences for the estimates for the same covariates across the quantiles. For example, if you want to test whether the parameters $\beta_2$ are the same across quantiles, the null hypothesis $H_0$ can be written as $\beta_2(\tau_1) = ... = \beta_2(\tau_k)$, where $\tau_j$, $j = 1, ..., k$, are the quantiles specified in the MODEL statement. See Koenker and Bassett (1982a) for details.

## Estimating Probability Functions by Using the CONDDIST Statement

Because the cumulative distribution function (CDF) is the inverse of the quantile function, you can estimate the CDF and other CDF-based statistics by inverting the relevant quantile function estimates. The CONDDIST statement in the QUANTREG procedure performs this type of analysis by estimating the conditional and marginal probability functions for the response random variable $\mathbf{Y}$. These probability functions include the following:

- $F_{\mathbf{Y}|\mathbf{x}}(t)$ for the conditional CDF and $P_{\mathbf{Y}|\mathbf{x}}(t)$ for the conditional probability density function (PDF) that are computed using the model
- $F_{\mathbf{Y}}(t)$ for the observed marginal CDF and $P_{\mathbf{Y}}(t)$ for the observed marginal PDF that are estimated from the observed response values without using the model
- $F_{\mathbf{P}}(t)$ for the fitted marginal CDF and $P_{\mathbf{P}}(t)$ for the fitted marginal PDF that are estimated from the quantile process predictions of the response variable and with the model integrated out

The QUANTREG procedure estimates the conditional quantile function, $Q_{\mathbf{Y}|\mathbf{x}}(\tau) = \mathbf{x}'\boldsymbol{\beta}(\tau)$, by using the fast quantile process regression (FQPR) algorithm. If you specify the QUANTILE= FQPR option in the MODEL statement, the CONDDIST statement uses the FQPR parameter estimates $\hat{\boldsymbol{\beta}}(\tau)$ to estimate $\boldsymbol{\beta}(\tau)$. For more information about the FQPR option, see the section "Fast Quantile Process Regression" on page 8292. If the QUANTILE=FQPR option is not used in the MODEL statement, the first CONDDIST statement uses the FQPR algorithm to compute $\hat{\boldsymbol{\beta}}(\tau)$ for all the CONDDIST statements. This $\hat{\boldsymbol{\beta}}(\tau)$ is equal to the $\hat{\boldsymbol{\beta}}(\tau)$ that is output from the QUANTILE=FQPR option when you use the default FQPR suboption values.

For purposes of conceptual and computational simplicity, the CONDDIST statement estimates the quantile functions $Q(\tau)$ on a quantile-level grid $\{\tau_j = \tau_1 + (j-1)\tau_s : j = 1, \ldots, q\}$, where $\tau_1$ is the lower end of the grid and $\tau_s$ is the step length of the grid. You can specify $\tau_1$ and $q$ by respectively using the L=*value* and N=*n* suboptions of the QUANTILE= FQPR option in the MODEL statement. You can also specify the upper end of the grid $\tau_q$ by using the U=*value* suboption of the QUANTILE=FQPR option in the MODEL statement, such that

$$\tau_s = \frac{\tau_q - \tau_1}{q - 1}$$

By default, the size of the grid $q$ is the smaller number between 100 and half the number of training observations in the DATA= data set. And by default, $\tau_1 = 0.5/q$, $\tau_q = 1 - 0.5/q$, and $\tau_s = 1/q$.

The estimated quantile function $\left\{\widehat{Q}(\tau_1), \ldots, \widehat{Q}(\tau_q)\right\}$ is also called a CDF sample for $F(t)$.

### Counterfactual Distributions for the TESTDATA= Data Set

The QUANTREG procedure computes the quantile function estimates $\widehat{Q}_{\mathbf{Y}|\mathbf{x}}(\tau) = \mathbf{x}'\hat{\boldsymbol{\beta}}(\tau)$ on the DATA= data set in the PROC QUANTREG statement. To test this $\widehat{Q}_{\mathbf{Y}|\mathbf{x}}(\tau)$ for more observations, you can specify a separate data set by using the TESTDATA= data set option in the CONDDIST statement. Without assuming that the TESTDATA= data set and the DATA= data set share the same conditional distribution for $\mathbf{Y}|\mathbf{X}$, the CONDDIST statement estimates counterfactual probability functions for the response variable $\mathbf{Y}$ of the TESTDATA= data set that impose the quantile function $Q_{\mathbf{Y}|\mathbf{x}}(\tau)$ of the DATA= data set on the TESTDATA= data set.

### Conditional Cumulative Distribution Functions

The CONDDIST statement estimates $F_{\mathbf{Y}|\mathbf{x}}(t)$ by inverting the estimated quantile function on the quantile-level grid $\left\{\widehat{Q}_{\mathbf{Y}|\mathbf{x}}(\tau_1), \ldots, \widehat{Q}_{\mathbf{Y}|\mathbf{x}}(\tau_q)\right\}$, such that the estimated $F_{\mathbf{Y}|\mathbf{x}}(t)$ satisfies

$$\widehat{F}_{\mathbf{Y}|\mathbf{x}}\left(\widehat{Q}_{\mathbf{Y}|\mathbf{x}}(\tau_j)\right) = \tau_i \text{ for } j = 1, \ldots, q$$

By definition, a quantile function $Q(\tau)$ must be nondecreasing, so that $Q(\tau_1) \leq Q(\tau_2)$ for any $0 \leq \tau_1 < \tau_2 \leq 1$. However, quantile regression estimates could result in crossed quantile predictions. In other words, it is possible that $\mathbf{x}'\hat{\boldsymbol{\beta}}(\tau_1) > \mathbf{x}'\hat{\boldsymbol{\beta}}(\tau_2)$ for some $\mathbf{x}$ and some $0 \leq \tau_1 < \tau_2 \leq 1$. This predicament is called crossing. To avoid crossing, the CONDDIST statement defines $\widehat{Q}_{\mathbf{Y}|\mathbf{x}}(\tau_j)$ as the $j$th-smallest prediction among $\left\{\mathbf{x}'\hat{\boldsymbol{\beta}}(\tau_1), \ldots, \mathbf{x}'\hat{\boldsymbol{\beta}}(\tau_q)\right\}$. Therefore, it is possible that $\widehat{Q}_{\mathbf{Y}|\mathbf{x}}(\tau_j) \neq \mathbf{x}'\hat{\boldsymbol{\beta}}(\tau_j)$ for some $\mathbf{x}$ and $\tau_j$.

For an individual observation, the CONDDIST statement assigns the type "Fit for Obs" to this conditional CDF sample and labels this sample by using the observation ID value (if available) or the observation index.

### Conditional Cumulative Distribution Functions at Average

You can also request the conditional CDF sample at average $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ for $F_{\mathbf{Y}|\bar{\mathbf{x}}}(t)$ by using the SHOWAVG option for the training average or the TESTDATA(SHOWAVG)= option for the test average (or both).

The CONDDIST statement assigns the type "Fit at Average" to these conditional CDF samples. For training data, the CONDDIST statement labels this CDF sample as "TrainAvg" and assigns the average of the training response values $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as its response value. For test data, the CONDDIST statement labels this CDF sample as "TestAvg" and assigns the average of the test response values as its response value.

### Observed Marginal Cumulative Distribution Functions

The CONDDIST statement estimates the observed marginal $F_{\mathbf{Y}}(t)$ by inverting the observed quantile function $Q_{\mathbf{Y}}(\tau)$ of the response variable $\widehat{Q}_{\mathbf{Y}}(\tau)$.

Let $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ denote the sorted response values for either the training data that you can specify by using the DATA= data set in the PROC QUANTREG statement or the test data that you can specify by using the TESTDATA= data set in the CONDDIST statement. If you assign the quantile level $(i - 0.5)/n$ to the response value $y_{(i)}$, then $\widehat{Q}_{\mathbf{Y}}(\tau)$ is defined as

$$\widehat{Q}_{\mathbf{Y}}(\tau) = \begin{cases} y_{(1)} & \text{if } \tau \in [0, 0.5/n) \\ \lambda y_{(i+1)} + (1-\lambda) y_{(i)} & \text{if } \tau \in [(i-0.5)/n, (i+0.5)/n) \text{ for } \lambda = n\tau - (i - 0.5) \\ y_{(n)} & \text{if } \tau \in [(n-0.5)/n, 1] \end{cases}$$

To be consistent with the quantile-level grid that is used for estimating the conditional probability functions, the CONDDIST statement estimates $Q_{\mathbf{Y}}(\tau)$ by using the CDF sample $\left\{\widehat{Q}_{\mathbf{Y}}(\tau_1), \ldots, \widehat{Q}_{\mathbf{Y}}(\tau_q)\right\}$.

The CONDDIST statement assigns the type "Observed" to these observed marginal CDF samples. For training data, the CONDDIST statement labels this CDF sample as "TrainObs" and assigns the average of the training response values as its response value. For test data, the CONDDIST statement labels this CDF sample as "testObs" and assigns the average of the test response values as its response value.

## Fitted Marginal Cumulative Distribution Functions

Let $F_{\mathbf{Y}|\mathbf{X}}$ denote the distribution of $\mathbf{Y}|\mathbf{X}$ for the DATA= data set, and let $F^*_{\mathbf{X}}$ denote the marginal distribution of the explanatory covariates vector $\mathbf{X}$ for the TESTDATA= data set. Then the counterfactual marginal distribution of $\mathbf{Y}$ for the TESTDATA= data set is defined as

$$F^*_{\mathbf{P}}(t) = \int F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(t)\, dF^*_{\mathbf{X}}(\mathbf{x})$$

The CONDDIST statement estimates $F^*_{\mathbf{P}}(t)$ by inverting the quantile function of the quantile process predictions for the response variable.

Let $\left\{ p_{ij} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}(\tau_j) : i = 1, \ldots, n \text{ and } j = 1, \ldots, q \right\}$ denote the quantile predictions for all the observations of the TESTDATA= data set. And let $p_{(k)}$ denote the $k$th-smallest value in $\{p_{ij}\}$. The CONDDIST statement estimates $Q^*_{\mathbf{P}}(\tau)$ by defining

$$\widehat{Q}^*_{\mathbf{P}}(\tau_j) = \begin{cases} p_{(jn-0.5n+0.5)} & \text{if } n \text{ is an odd number} \\ 0.5p_{(jn-0.5n)} + 0.5p_{(jn-0.5n+1)} & \text{if } n \text{ is an even number} \end{cases}$$

where $\tau_j = \tau_1 + (j-1)\tau_s$ for $j = 1, \ldots, q$.

The CONDDIST statement assigns the type "Fit and Pooled" and the label "TestFit" to this fitted marginal CDF sample, and assigns the average of the test response values as its response value.

Recall that the observed marginal CDF is also available for the TESTDATA= data set. For clarity, let $F^*_{\mathbf{Y}}(t)$ denote the observed marginal CDF for the TESTDATA= data set. Then, by comparing the fitted marginal CDF $\widehat{F}^*_{\mathbf{P}}(t)$ with the observed marginal CDF $\widehat{F}^*_{\mathbf{Y}}(t)$, you can test whether the TESTDATA= data set follows the same model that is built on the DATA= data set. The CONDDIST statement supports the Mann-Whitney U test for this purpose when you use the MWU option.

## Mann-Whitney U Test Under Resampling

The Mann-Whitney U test (also called the Wilcoxon rank-sum test) is a nonparametric two-sample test. Let $n_1$ denote the size of the first sample and $n_2$ denote the size of the second sample. By merging the two samples into one ordered set $\{a_j : j = 1, \ldots, n = n_1 + n_2\}$, the statistic of the Mann-Whitney U test is defined as

$$U = \sum_{j=1}^{n} c_j R_j \text{ with } c_j = \begin{cases} 1 & \text{if } a_j \text{ belongs to the first sample} \\ 0 & \text{if } a_j \text{ belongs to the second sample} \end{cases}$$

where $R_j$ is the rank of $a_j$. Under some regularity conditions, $U$ asymptotically follows a normal distribution whose expectation equals

$$\mathrm{E}(U) = \frac{n_1}{n} \sum_{j=1}^{n} R_j = \frac{n_1(n+1)}{2}$$

and whose variance equals

$$\mathrm{Var}(U) = \frac{n_1 n_2 (n+1)}{12}$$

You can perform the Mann-Whitney U test against the observed marginal CDF sample of the response variable for the following data sets:

- the DATA= data set in the PROC QUANTREG statement (by using the MWU option in the CONDDIST statement)
- the TESTDATA= data set in the CONDDIST statement (by using the TESTDATA(MWU)= data set option in the CONDDIST statement)

The null and alternative hypotheses of the Mann-Whitney U test for the CONDDIST statement are, respectively,

$$\begin{aligned} H_0 &: \quad F_1 = F_2 \\ H_1 &: \quad F_1 > F_2 \text{ or } F_1 < F_2 \end{aligned}$$

where $F_1$ denotes a CDF under test such as $F_{\mathbf{Y}}$ or $F_{\mathbf{Y}|\mathbf{x}}$, and $F_2$ denotes the observed marginal CDF for either the DATA= data set or the TESTDATA= data set.

Note that the size of the observed marginal CDF sample should be smaller than the number of used observations in its associated data set. Otherwise, the Mann-Whitney U test could output a smaller $p$-value and incorrectly reject $H_0$.

## Regression Quantile Level and Sample Quantile Level

Given a response value $y$ and a CDF sample for $F(t)$: $\left\{\widehat{Q}(\tau_1), \ldots, \widehat{Q}(\tau_q)\right\}$, the CONDDIST statement estimates the quantile level of $y$ on $F(t)$ by using the following linear interpolation method:

$$\hat{\tau}_y \begin{cases} < \tau_1 & \text{if } y < \widehat{Q}(\tau_1) \\ = \lambda\tau_{i+1} + (1-\lambda)\tau_i & \text{if } \widehat{Q}(\tau_i) \le y \le \widehat{Q}(\tau_{i+1}) \text{ for } \lambda = \left(y - \widehat{Q}(\tau_i)\right) \Big/ \left(\widehat{Q}(\tau_{i+1}) - \widehat{Q}(\tau_i)\right) \\ > \tau_q & \text{if } y > \widehat{Q}(\tau_q) \end{cases}$$

Here $\hat{\tau}_y$ is defined as a regression quantile level if the CDF sample is conditional for $F_{\mathbf{Y}|\mathbf{x}}(t)$, or a sample quantile level if the CDF sample is observed and marginal for $F_{\mathbf{Y}}(t)$.

The PLOT=PPPLOT option in the CONDDIST statement creates the scatter plot for the regression quantile levels versus the sample quantile levels.

## Probability Density Functions

The CONDDIST statement estimates the PDF by applying the kernel density estimator to the estimated CDF in the quantile-level grid. An estimated CDF for the CONDDIST statement is in the form of $\left\{\widehat{F}\left(\widehat{Q}(\tau_j)\right) = \tau_j : \tau_j = \tau_1 + (j-1)\tau_s, j = 1, \ldots, q\right\}$. The lower end of the estimated PDF is limited by the quantile level $\max(0, \tau_1 - 0.5\tau_s)$, and the upper end of the estimated PDF is limited by the quantile level $\min(\tau_q + 0.5\tau_s, 1)$.

The general form of the kernel density estimator is

$$\hat{f}_\lambda(t) = \frac{1}{q\lambda l} \sum_{j=1}^{q} K_0\left(\frac{t - t_j}{\lambda}\right)$$

where

$K_0(\cdot)$ is the kernel function

$\lambda$ is the bandwidth

$q$ is the sample size

$t_j$ is the $i$th value $\widehat{F}\left(\widehat{Q}(\tau_j)\right)$

$l = q\tau_s$ is the length of the range of the quantile-level grid

The KDE option provides three kernel functions ($K_0$): normal, quadratic, and triangular. You can specify the function by using the K= *kernel-option* in parentheses after the KDE option. The values of the K= option are NORMAL, QUADRATIC, and TRIANGULAR. By default, a normal kernel is used. The formulas for the kernel functions are as follows:

| | | |
|---|---|---|
| Normal | $K_0(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right)$ | for $-\infty < t < \infty$ |
| Quadratic | $K_0(t) = \frac{3}{4}\left(1 - t^2\right)$ | for $\lvert t\rvert \leq 1$ |
| Triangular | $K_0(t) = 1 - \lvert t\rvert$ | for $\lvert t\rvert \leq 1$ |

The value of $\lambda$, referred to as the bandwidth parameter, determines the degree of smoothness in the estimated probability density function. You specify $\lambda$ indirectly when you specify a standardized bandwidth $c$ by using the C=*value* suboption. Let $Q$ denote the interquartile range and $q$ denote the sample size; then $c$ is related to $\lambda$ by the formula

$$\lambda = cQq^{-\frac{1}{5}}$$

For a specific kernel function, the discrepancy between the density estimator $\hat{f}_\lambda(x)$ and the true density $f(x)$ is measured by the mean integrated square error (MISE):

$$\text{MISE}(\lambda) = \int_x \left\{\text{E}\left(\hat{f}_\lambda(x)\right) - f(x)\right\}^2 dx + \int_x \text{Var}\left(\hat{f}_\lambda(x)\right) dx$$

MISE is the sum of the integrated squared bias and the variance. An approximate mean integrated square error (AMISE) is defined as

$$\text{AMISE}(\lambda) = \frac{1}{4}\lambda^4 \left(\int_t t^2 K(t) dt\right)^2 \int_x \left(f''(x)\right)^2 dx + \frac{1}{q\lambda}\int_t K(t)^2 dt$$

You can derive a bandwidth that minimizes AMISE by treating $f(x)$ as the normal density whose parameters $\mu$ and $\sigma$ are estimated by the sample mean and standard deviation, respectively. If you do not specify a bandwidth parameter or if you specify C=MISE, the bandwidth that minimizes AMISE is used. The value of AMISE can be used to compare different density estimates. You can also specify C=SJPI to select the bandwidth by using the Sheather-Jones plug-in method (Jones, Marron, and Sheather 1996).

The general kernel density estimates assume that the domain of the density to be estimated can take all values on a real line. However, sometimes the domain of a density is an interval that is bounded on one or both sides. For example, if a variable Y is a measurement of only positive values, then the kernel density curve should be bounded so that it is zero for negative Y values. You can use the LOWER= and UPPER= *kde-options* in the PDF= option in the CONDDIST statement to specify the bounds.

## Leverage Point and Outlier Detection

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

Mahalanobis distance is defined as

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

where $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ and $\bar{\mathbf{C}}(\mathbf{A}) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})$ are the empirical multivariate location and scale, respectively. Here, $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(p-1)})'$ does not include the intercept variable. The relationship between the Mahalanobis distance $\mathrm{MD}(\mathbf{x}_i)$ and the matrix $\mathbf{H} = (h_{ij}) = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{A}$ is

$$h_{ii} = \frac{1}{n-1} \mathrm{MD}_i^2 + \frac{1}{n}$$

Robust distance is defined as

$$\mathrm{RD}(\mathbf{x}_i) = [(\mathbf{x}_i - \mathbf{T}(\mathbf{A}))' \mathbf{C}(\mathbf{A})^{-1} (\mathbf{x}_i - \mathbf{T}(\mathbf{A}))]^{1/2}$$

where $\mathbf{T}(\mathbf{A})$ and $\mathbf{C}(\mathbf{A})$ are robust multivariate location and scale estimates that are computed according to the minimum covariance determinant (MCD) method of Rousseeuw and Van Driessen (1999).

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

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

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

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

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

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

You can specify the multiplier $k$ of the cutoff value in the CUTOFF= option in the MODEL statement. You can specify the scale $\sigma$ in the SCALE= option in the MODEL statement. By default, $k = 3$ and the scale $\sigma$ is computed as the corrected median of the absolute residuals:

$\sigma = \text{median}\{|r_i|/\beta_0, i = 1, \ldots, n\}$

where $\beta_0 = \Phi^{-1}(0.75)$ is an adjustment constant for consistency when the normal distribution is used.

An ODS table called DIAGNOSTICS contains the Leverage and Outlier variables.

## INEST= Data Set

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

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

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

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

## OUTEST= Data Set

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

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

If this data set is created, it contains all the variables that are specified in the MODEL statement and the BY statement. Each observation consists of parameter values for a specified quantile, and the dependent variable has the value –1.

The following variables are also added to the data set:

| | |
|---|---|
| _MODEL_ | a character variable of length 8 that contains the label of the MODEL statement, if present. Otherwise, the variable's value is blank. |
| _ALGORITHM_ | a character variable of length 8 that contains the name of the algorithm that is used for computing the parameter estimates, either SIMPLEX, INTERIOR, or SMOOTH. |
| _TYPE_ | a character variable of length 8 that contains the type of the observation. This variable is fixed as PARMS to indicate that the observation includes parameter estimates. |
| _STATUS_ | a character variable of length 12 that contains the status of model fitting (either NORMAL, NOUNIQUE, or NOVALID). |
| Intercept | a numeric variable that contains the intercept parameter estimates. |
| _QUANTILE_ | a numeric variable that contains the specified quantile levels. |

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

## Computational Resources

The various algorithms need different amounts of memory for working space. Let $p$ be the number of parameters that are estimated, $n$ be the number of observations that are used in the model estimation, and $s$ be the size (in bytes) of the double data type.

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

$$(2np + 6n + 10p)s$$

For the interior point algorithm, the minimum working space (in bytes) that is needed is

$$(np + p^2 + 13n + 4p)s$$

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

$$(np + p^2 + 6n + 4p)s$$

For the last two algorithms, if you want to use preprocessing, the following additional amount of working space (in bytes) is needed:

$$(np + 6n + 2p)s$$

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 that it creates. You can specify these names when you use the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the Table 100.9.

**Table 100.9** ODS Tables Produced in PROC QUANTREG

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| AvgParameterEst | Average parameter estimates | MODEL | FQPR |
| AvgObjFunction | Average objective function | MODEL | FQPR |
| ClassLevels | Classification variable levels | CLASS | Default |
| CmpTestObs | Distribution comparisons to the observed marginal distribution of the test response | CONDDIST | TESTDATA(MWU) |
| CmpTrainObs | Distribution comparisons to the observed marginal distribution of the training response | CONDDIST | MWU |
| CondDistEst | Conditional distribution estimates | CONDDIST | Default |
| CorrB | Parameter estimate correlation matrix | MODEL | CORRB |
| CovB | Parameter estimate covariance matrix | MODEL | COVB |

**Table 100.9** (continued)

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| DensityEstInfo | Density estimation information | CONDDIST | PDF=KDE |
| 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 QUANTREG | Default |
| ObjFunction | Objective function | MODEL | Default |
| ParameterEstimates | Parameter estimates | MODEL | Default |
| ParmInfo | Parameter indices | MODEL | Default |
| PerfSettings | Performance settings | PERFORMANCE | DETAILS |
| ProcessEst | Quantile process estimates | MODEL | QUANTILE= |
| ProcessObj | Objective function for quantile process | MODEL | QUANTILE= |
| SMIterHistory | Iteration history (smoothing) | MODEL | ITPRINT |
| SummaryStatistics | Summary statistics for model variables | MODEL | Default |
| TestAcrossQuantiles | Results for across-quantiles test | TEST | QINTERACT |
| Tests | Results for tests | TEST | Default |
| ScalableTiming | Timing details | PERFORMANCE | DETAILS |

## ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 622 in Chapter 21, "Statistical Graphics Using ODS."

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

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

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

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

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

For more information about these plots, see the PLOT= option in the PROC QUANTREG statement and the PLOT= option in the MODEL statement.

Besides the PROC QUANTREG and the MODEL statements, the CONDDIST statement can also output plots for conditional distribution analysis.

You can request the probability-probability plot by using the PLOT=PPPLOT option in the CONDDIST statement. This plot illustrates the power of the estimated conditional distributions on the response values by plotting the points for the regression quantile levels versus the sample quantile levels.

You can request the CDF plot by using the PLOT=CDFPLOT option in the CONDDIST statement. This plot illustrates the CDF estimates for the specified observations.

You can request the PDF plot by using the PLOT=PDFPLOT option in the CONDDIST statement. This plot illustrates the PDF estimates for the specified observations.

For more information about the plots for the CONDDIST statement, see the PLOT= option in the CONDDIST statement.

All the plot options are summarized in Table 100.10.

**Table 100.10** Options for Plots

| Keyword | Plot |
|---|---|
| ALL | All appropriate plots |
| CDFPLOT | Series plot for CDF samples |
| DDPLOT | Robust distance versus Mahalanobis distance |
| FITPLOT | Conditional quantile fit versus independent variable |
| HISTOGRAM | Histogram of standardized robust residuals |
| NONE | No plot |
| PDFPLOT | Series plot for PDF estimates |
| PPPLOT | Regression quantile level versus sample quantile level |
| QUANTPLOT | Scatter plot of regression quantile |
| QQPLOT | Q-Q plot of standardized robust residuals |
| RDPLOT | Standardized robust residual versus robust distance |

The following subsections provide information about these graphs.

## ODS Graph Names

The QUANTREG procedure assigns a name to each graph that it creates. You can use these names to refer to the graphs when you use ODS. The names along with the required statements and options are listed in Table 100.11.

**Table 100.11** Graphs Produced by PROC QUANTREG

| ODS Graph Name | Plot Description | Statement | Option |
|---|---|---|---|
| CDFPlot | Cumulative distribution functions | CONDDIST | CDFPLOT |
| DDPlot | Robust distance versus Mahalanobis distance | PROC QUANTREG, MODEL | DDPLOT |
| FitPlot | Quantile fit versus independent variable | PROC QUANTREG, MODEL | FITPLOT |
| Histogram | Histogram of standardized robust residuals | PROC QUANTREG, MODEL | HISTOGRAM |
| PDFPlot | Probability density functions | CONDDIST | PDFPLOT |
| PPPlot | Regression quantile level versus sample quantile level | CONDDIST | PPPLOT |
| QQPlot | Q-Q plot of standardized robust residuals | PROC QUANTREG, MODEL | QQPLOT |
| QuantPanel | Panel of quantile plots with confidence limits | MODEL | QUANTPLOT |
| QuantPlot | Scatter plot for regression quantiles with confidence limits | MODEL | QUANTPLOT UNPACK |
| RDPlot | Standardized robust residual versus robust distance | PROC QUANTREG, MODEL | RDPLOT |

## Fit Plot

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

The following example reuses the trout data set in the section "Analysis of Fish-Habitat Relationships" on page 8252 to show the fit plot for one or several quantiles:

```
ods graphics on;

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

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

For a single quantile, the confidence limits for the fitted conditional quantiles are also plotted if you specify the CI=RESAMPLING or CI=SPARSITY option. (See Figure 100.14.) For multiple quantiles, confidence limits are not plotted by default. (See Figure 100.15.) You can add the confidence limits on the plot by specifying the option PLOT=FITPLOT(SHOWLIMITS).

The QUANTREG procedure also provides fit plots for quantile regression splines and polynomials if they are based on a single continuous variable. (See Example 100.4 and Example 100.5 for some examples.)

**Figure 100.14** Fit Plot with Confidence Limits

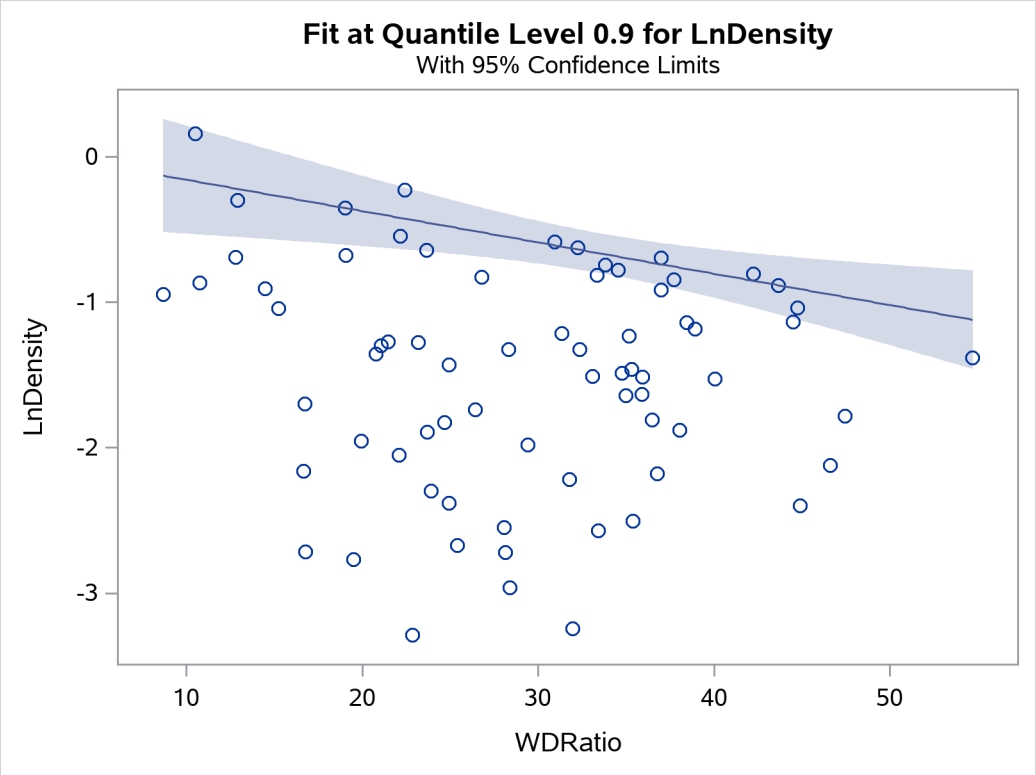
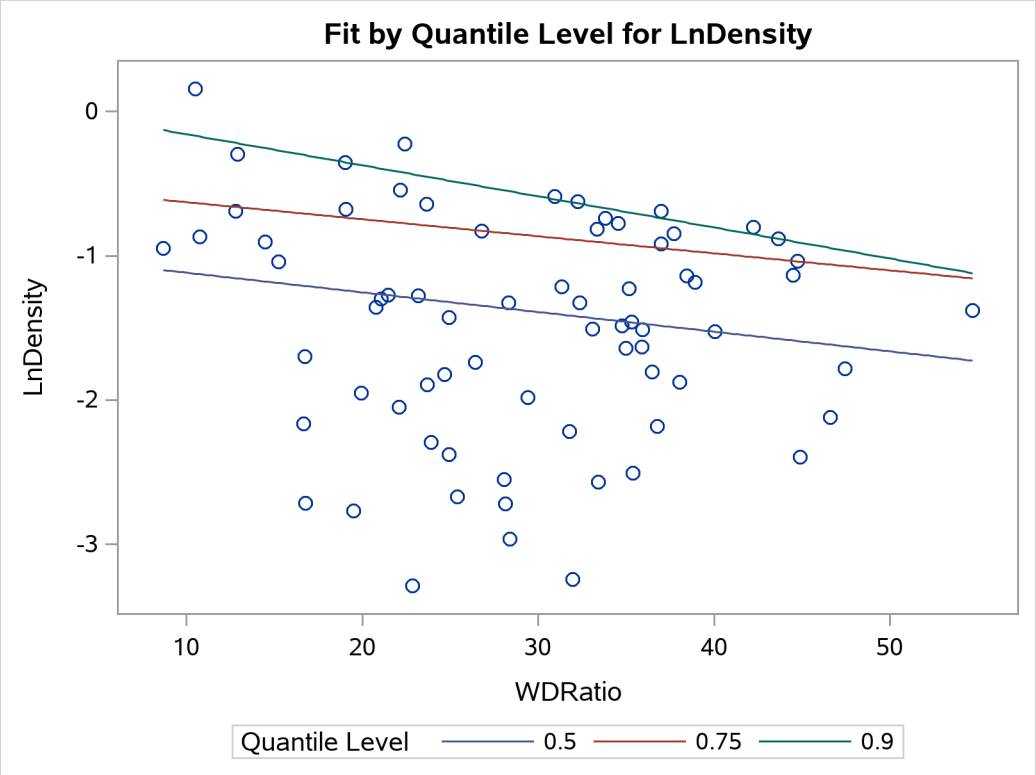


**Figure 100.15** Fit Plot for Multiple Quantiles




## Quantile Process Plot

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

By default, panels are used to hold multiple process plots (up to four in each panel). You can use the UNPACK option to request individual process plots. Figure 100.10 in the section "Analysis of Fish-Habitat Relationships" on page 8252 shows a panel that includes two quantile process plots. Output 100.2.9 in Example 100.2 shows a single quantile process plot. Example 100.3 demonstrates more quantile process plots and their usage.

## Distance-Distance Plot

The distance-distance plot (DDPLOT) is mainly used for leverage-point diagnostics. It is a scatter plot of the robust distances against the classical Mahalanobis distances for the continuous independent variables. For more information about the robust distance, see the section "Leverage Point and Outlier Detection" on page 8303. If is a classification variable is specified in the model, this plot is not created.

You can use the PLOT=DDPLOT option to request this plot. The following statements use the growth data set in Example 100.2 to create a single plot, which is shown in Output 100.2.4 in Example 100.2:

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

The reference lines represent the cutoff values. The diagonal line is also drawn to show the distribution of the distances. By default, all outliers and leverage points are labeled with observation numbers. To change the default, you can use the LABEL= option as described in Table 100.4.

### Residual-Distance Plot

The residual-distance plot (RDPLOT) is used for both outlier and leverage-point diagnostics. It is a scatter plot of the standardized residuals against the robust distances. For more information about the robust distance, see the section "Leverage Point and Outlier Detection" on page 8303. If a classification variable is specified in the model, this plot is not created.

You can use the PLOT=RDPLOT option to request this plot. The following statements use the growth data set in Example 100.2 to create a single plot, which is shown in Output 100.2.3 in Example 100.2:

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

The reference lines represent the cutoff values. By default, all outliers and leverage points are labeled with observation numbers. To change the default, you can use the LABEL= option as described in Table 100.4.

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

### Histogram and Q-Q Plot

PROC QUANTREG produces a histogram and a Q-Q plot for the standardized residuals. The histogram is superimposed with a normal density curve and a kernel density curve. Using the growth data set in Example 100.2, the following statements create the plot that is shown in Output 100.2.5 in Example 100.2:

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

# Examples: QUANTREG Procedure

## Example 100.1: Comparison of Algorithms

This example illustrates and compares the three algorithms for regression estimation available in the QUANTREG procedure. The simplex algorithm is the default because of its stability. Although this algorithm is slower than the interior point and smoothing algorithms for large data sets, the difference is not as significant for data sets with fewer than 5,000 observations and 50 variables. The simplex algorithm can also compute the entire quantile process, which is shown in Example 100.2.

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

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

The following statements invoke the QUANTREG procedure to fit a median regression model with the default simplex algorithm. They produce the results that are shown in Output 100.1.1 through Output 100.1.3.

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

Output 100.1.1 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 that 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 100.1.2, which also displays the objective function value and the predicted value of the response at the means of the covariates.

Output 100.1.3 displays parameter estimates and confidence limits. These estimates are reasonable, which indicates that median regression is robust to the 50 outliers.

**Output 100.1.1** Model Fit Information and Summary Statistics from the 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 100.1.2** Quantile and Objective Function from the Simplex Algorithm

| Quantile Level and Objective Function | |
|---|---|
| **Quantile Level** | 0.5 |
| **Objective Function** | 2441.1927 |
| **Predicted Value at Mean** | 10.0259 |

**Output 100.1.3** Parameter Estimates from the Simplex Algorithm

**Parameter Estimates**

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

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

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

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

Output 100.1.4 displays model fit information.

**Output 100.1.4** Model Fit Information from the 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 100.1.5 displays the iteration history of the interior point algorithm. Note that the duality gap is less than 1E–6 in the final iteration. The table also provides the number of iterations, the number of corrections, the primal step length, the dual step length, and the objective function value at each iteration.

**Output 100.1.5** Iteration History for the Interior Point Algorithm

**The QUANTREG Procedure**
**Quantile Level = 0.5**

| Iteration History of Interior Point Algorithm | | | | |
|---|---|---|---|---|
| **Iter** | **Duality Gap** | **Primal Step** | **Dual Step** | **Objective Function** |
| **1** | 2622.675 | 0.3113 | 0.4910 | 3303.469 |
| **2** | 3214.640 | 0.0427 | 1.0000 | 2461.377 |
| **3** | 1126.899 | 0.9882 | 0.3653 | 2451.134 |
| **4** | 760.887 | 0.3381 | 1.0000 | 2442.810 |
| **5** | 77.102902 | 1.0000 | 0.8916 | 2441.263 |
| **6** | 8.436664 | 0.9370 | 0.8381 | 2441.208 |
| **7** | 1.828685 | 0.8375 | 0.7674 | 2441.199 |
| **8** | 0.405843 | 0.6980 | 0.8636 | 2441.195 |
| **9** | 0.095499 | 0.9438 | 0.5955 | 2441.193 |
| **10** | 0.0066528 | 0.9818 | 0.9304 | 2441.193 |
| **11** | 0.00022482 | 0.9179 | 0.9994 | 2441.193 |
| **12** | 5.44642E-8 | 1.0000 | 1.0000 | 2441.193 |

Output 100.1.6 displays the parameter estimates that are obtained by using the interior point algorithm. These estimates are identical to those obtained by using the simplex algorithm.

**Output 100.1.6** Parameter Estimates from the Interior Point Algorithm

| Parameter Estimates | | |
|---|---|---|
| **Parameter** | **DF** | **Estimate** |
| **Intercept** | 1 | 10.0364 |
| **x1** | 1 | 5.0106 |
| **x2** | 1 | 3.0294 |

The following statements refit the model by using the smoothing algorithm. They produce the results that are shown in Output 100.1.7 through Output 100.1.9.

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

The RRATIO= option controls the reduction speed of the threshold. Output 100.1.7 displays the model fit information.

**Output 100.1.7** Model Fit Information from the 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 100.1.8 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, which is the value specified in 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, see Chen (2007).

**Output 100.1.8** Iteration History for the Smoothing Algorithm

**The QUANTREG Procedure**
**Quantile Level = 0.5**

**Iteration History of Smoothing Algorithm**

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

Output 100.1.9 displays the parameter estimates that are obtained by using the smoothing algorithm. These estimates are identical to those obtained by using 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 100.1.9** Parameter Estimates from the Smoothing Algorithm

**Parameter Estimates**

| Parameter | DF | Estimate |
|---|---|---|
| **Intercept** | 1 | 10.0364 |
| **x1** | 1 | 5.0106 |
| **x2** | 1 | 3.0294 |

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

## Example 100.2: 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 in gross domestic product (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 |
| ly2 | Investment/GDP |
| gcony2 | Public consumption/GDP |
| lblakp2 | Black market premium |
| pol2 | Political instability |
| ttrad2 | Growth rate terms trade |

The goal is to study the effect of the covariates on GDP. The following statements request median regression for a preliminary exploration. They produce the results that are in Output 100.2.1 through Output 100.2.6.

```
data growth;
   length Country$ 22;
   input Country GDP lgdp2 mse2 fse2 fhe2 mhe2 lexp2 lintr2 gedy2
         Iy2 gcony2 lblakp2 pol2 ttrad2 @@;
   datalines;
Algeria75              .0415 7.330 .1320 .0670 .0050 .0220 3.880 .1138 .0382
                       .1898 .0601 .3823 .0833 .1001
Algeria85              .0244 7.745 .2760 .0740 .0070 .0370 3.978 -.107 .0437
                       .3057 .0850 .9386 .0000 .0657
Argentina75            .0187 8.220 .7850 .6200 .0740 .1660 4.181 .4060 .0221
                       .1505 .0596 .1924 .3575 -.011
Argentina85            -.014 8.407 .9360 .9020 .1320 .2030 4.211 .1914 .0243
                       .1467 .0314 .3085 .7010 -.052
Australia75            .0259 9.101 2.541 2.353 .0880 .2070 4.263 6.937 .0348
                       .3272 .0257 .0000 .0080 -.016

   ... more lines ...

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

ods graphics on;

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

The QUANTREG procedure uses the default simplex algorithm to estimate the parameters and uses the MCMB resampling method to compute confidence limits.

Output 100.2.1 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), indicating that this variable might have outliers.

**Output 100.2.1** Model Information and Summary Statistics

**The QUANTREG Procedure**

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

**Summary Statistics**

| Variable | Q1 | Median | Q3 | Mean | Standard Deviation | MAD |
|---|---|---|---|---|---|---|
| **lgdp2** | 6.9890 | 7.7450 | 8.6080 | 7.7905 | 0.9543 | 1.1579 |
| **mse2** | 0.3160 | 0.7230 | 1.2675 | 0.9666 | 0.8574 | 0.6835 |
| **fse2** | 0.1270 | 0.4230 | 0.9835 | 0.7117 | 0.8331 | 0.5011 |
| **fhe2** | 0.0110 | 0.0350 | 0.0890 | 0.0792 | 0.1216 | 0.0400 |
| **mhe2** | 0.0400 | 0.1060 | 0.2060 | 0.1584 | 0.1752 | 0.1127 |
| **lexp2** | 3.8670 | 4.0640 | 4.2430 | 4.0440 | 0.2028 | 0.2728 |
| **lintr2** | 0.00160 | 0.5604 | 1.8805 | 1.4625 | 2.5491 | 1.0058 |
| **gedy2** | 0.0248 | 0.0343 | 0.0466 | 0.0360 | 0.0141 | 0.0151 |
| **ly2** | 0.1396 | 0.1955 | 0.2671 | 0.2010 | 0.0877 | 0.0981 |
| **gcony2** | 0.0480 | 0.0767 | 0.1276 | 0.0914 | 0.0617 | 0.0566 |
| **lblakp2** | 0 | 0.0696 | 0.2407 | 0.1916 | 0.3070 | 0.1032 |
| **pol2** | 0 | 0.0500 | 0.2429 | 0.1683 | 0.2409 | 0.0741 |
| **ttrad2** | -0.0240 | -0.0100 | 0.00730 | -0.00570 | 0.0375 | 0.0239 |
| **GDP** | 0.00290 | 0.0196 | 0.0351 | 0.0191 | 0.0248 | 0.0237 |

Output 100.2.2 displays the parameter estimates and 95% confidence limits that are computed with the rank method.

**Output 100.2.2** Parameter Estimates

**Parameter Estimates**

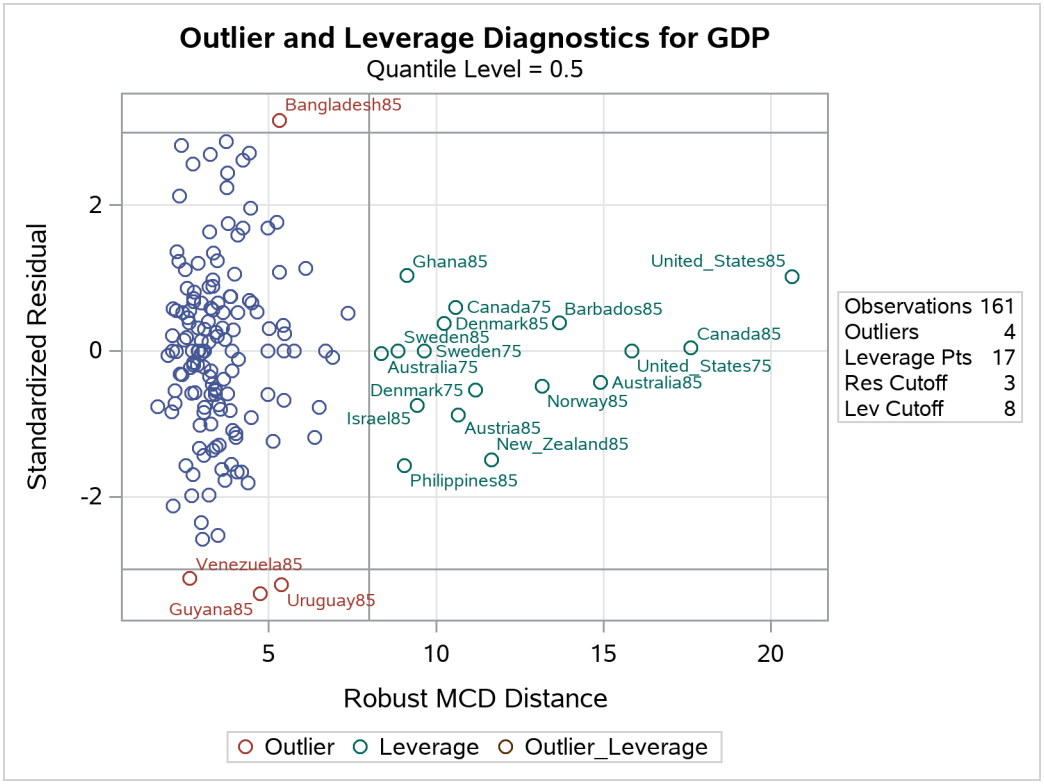
| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|
| **Intercept** | 1 | -0.0488 | 0.0738 | -0.1946 | 0.0970 | -0.66 | 0.5092 |
| **lgdp2** | 1 | -0.0269 | 0.0041 | -0.0351 | -0.0187 | -6.49 | <.0001 |
| **mse2** | 1 | 0.0110 | 0.0077 | -0.0043 | 0.0263 | 1.42 | 0.1563 |
| **fse2** | 1 | -0.0011 | 0.0086 | -0.0182 | 0.0159 | -0.13 | 0.8945 |
| **fhe2** | 1 | 0.0148 | 0.0330 | -0.0504 | 0.0801 | 0.45 | 0.6540 |
| **mhe2** | 1 | 0.0043 | 0.0270 | -0.0492 | 0.0577 | 0.16 | 0.8745 |
| **lexp2** | 1 | 0.0683 | 0.0231 | 0.0227 | 0.1139 | 2.96 | 0.0036 |
| **lintr2** | 1 | -0.0022 | 0.0015 | -0.0053 | 0.0009 | -1.42 | 0.1569 |
| **gedy2** | 1 | -0.0508 | 0.1650 | -0.3770 | 0.2753 | -0.31 | 0.7585 |
| **ly2** | 1 | 0.0723 | 0.0247 | 0.0234 | 0.1212 | 2.92 | 0.0040 |
| **gcony2** | 1 | -0.0935 | 0.0377 | -0.1680 | -0.0191 | -2.48 | 0.0142 |
| **lblakp2** | 1 | -0.0269 | 0.0085 | -0.0438 | -0.0101 | -3.17 | 0.0019 |
| **pol2** | 1 | -0.0301 | 0.0093 | -0.0485 | -0.0117 | -3.23 | 0.0015 |
| **ttrad2** | 1 | 0.1613 | 0.0769 | 0.0093 | 0.3132 | 2.10 | 0.0377 |

Diagnostics for the median regression fit, which are requested in the PLOTS= option, are displayed in Output 100.2.3 and Output 100.2.4. Output 100.2.3 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 100.2.4 plots the robust MCD distance against the Mahalanobis distance. This display is used to diagnose leverage points.

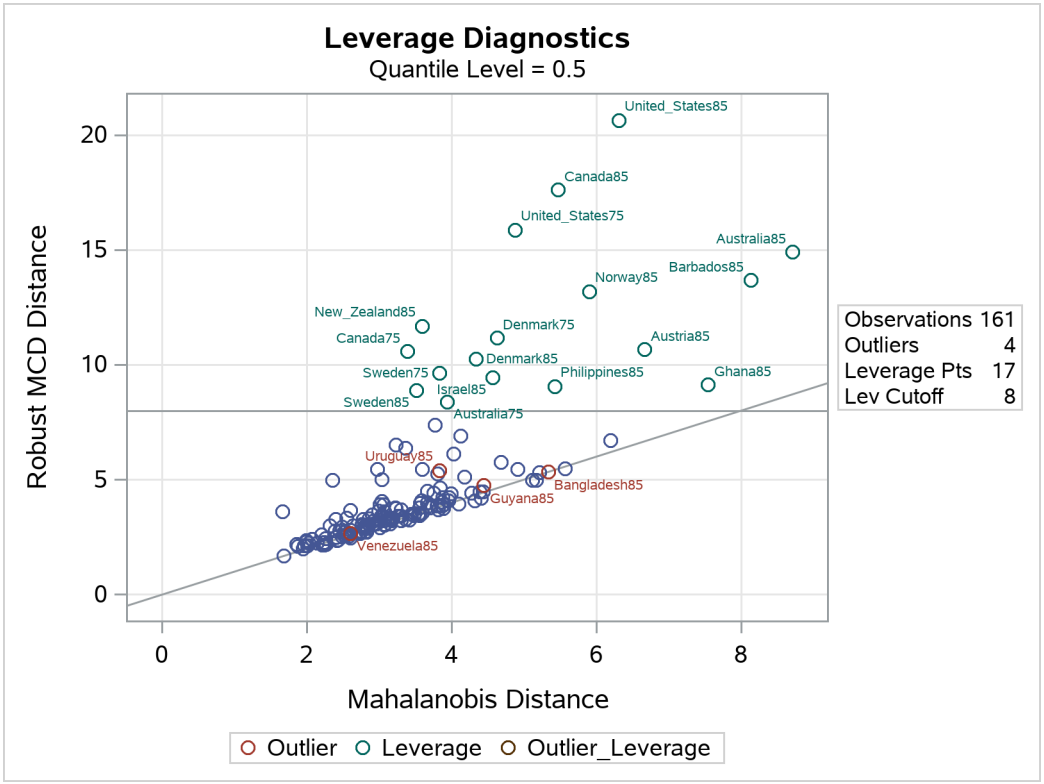
The cutoff value 8, which is specified in the LEVERAGE option, is close to the maximum of the Mahalanobis distance. Eighteen points are diagnosed as high leverage points, and almost all are countries with high human capital, which is the major contributor to the high leverage as observed from the summary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

A histogram of the standardized residuals and two fitted density curves are displayed in Output 100.2.5. This output shows that median regression fits the data well.

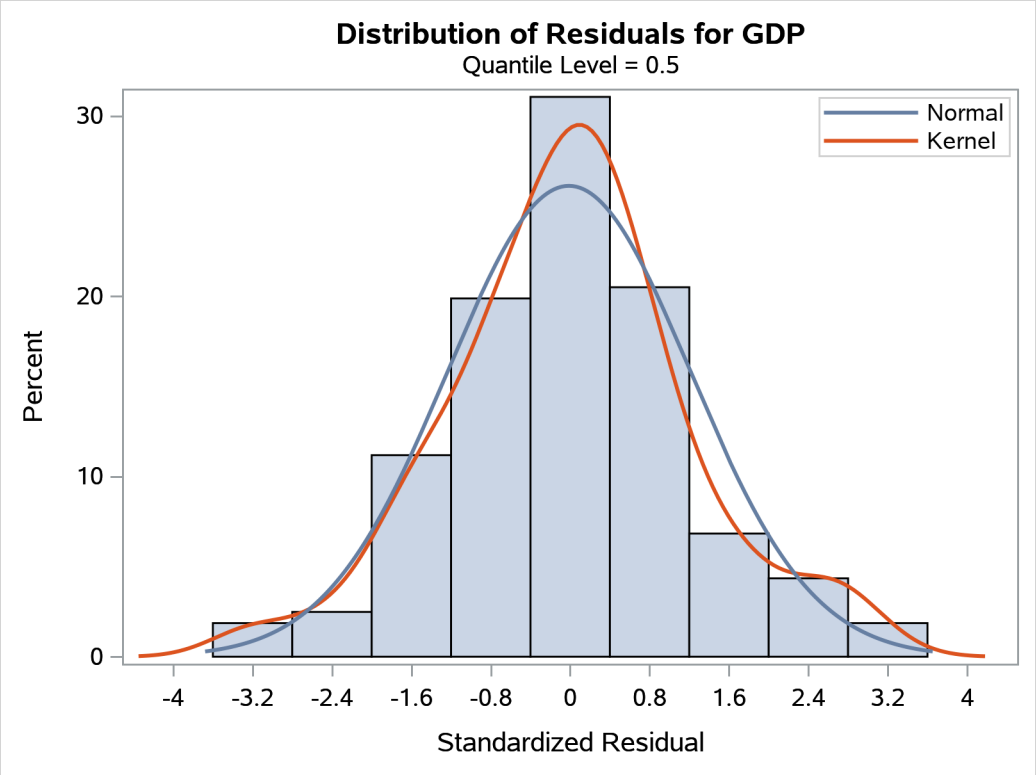
**Output 100.2.3** Plot of Residual versus Robust Distance




**Output 100.2.4** Plot of Robust Distance versus Mahalanobis Distance



**Output 100.2.5** Histogram for Residuals




Tests of significance for the initial per-capita GDP (LGDP2) are shown in Output 100.2.6.

**Output 100.2.6** Tests for Regression Coefficient

**Test test_lgdp2 Results**

| Test | Test Statistic | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| **Wald** | 42.1656 | 1 | 42.17 | <.0001 |
| **Likelihood Ratio** | 36.3047 | 1 | 36.30 | <.0001 |

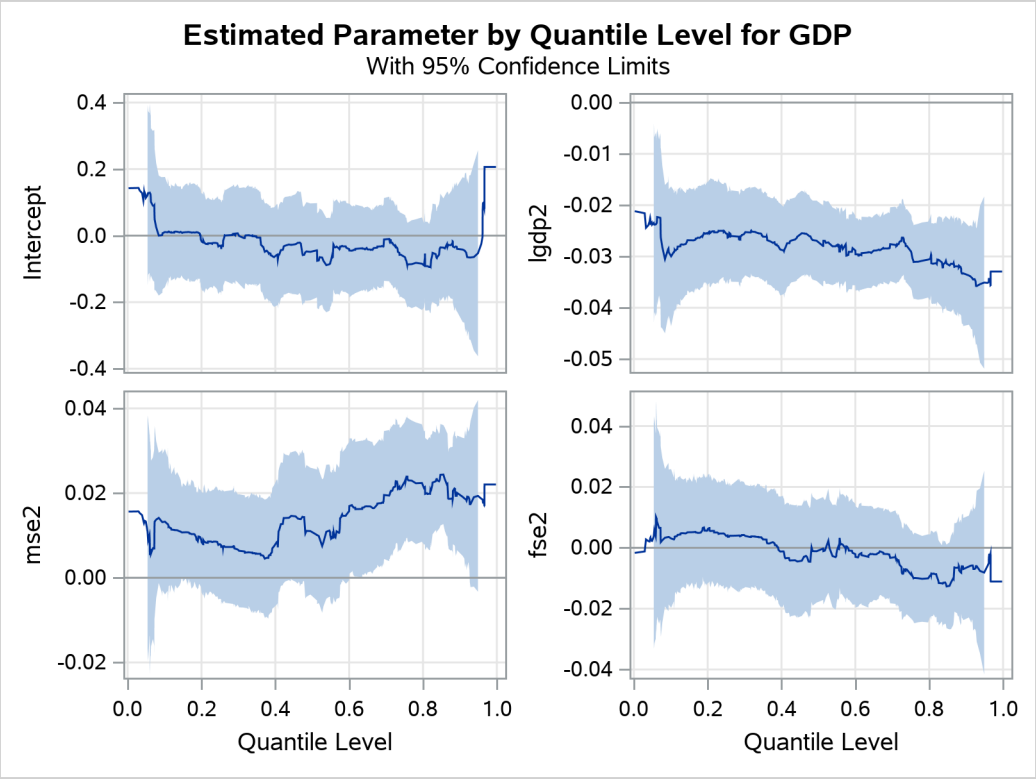
The QUANTREG procedure computes entire quantile processes for covariates when you specify QUANTILE=PROCESS in the MODEL statement, as follows:

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

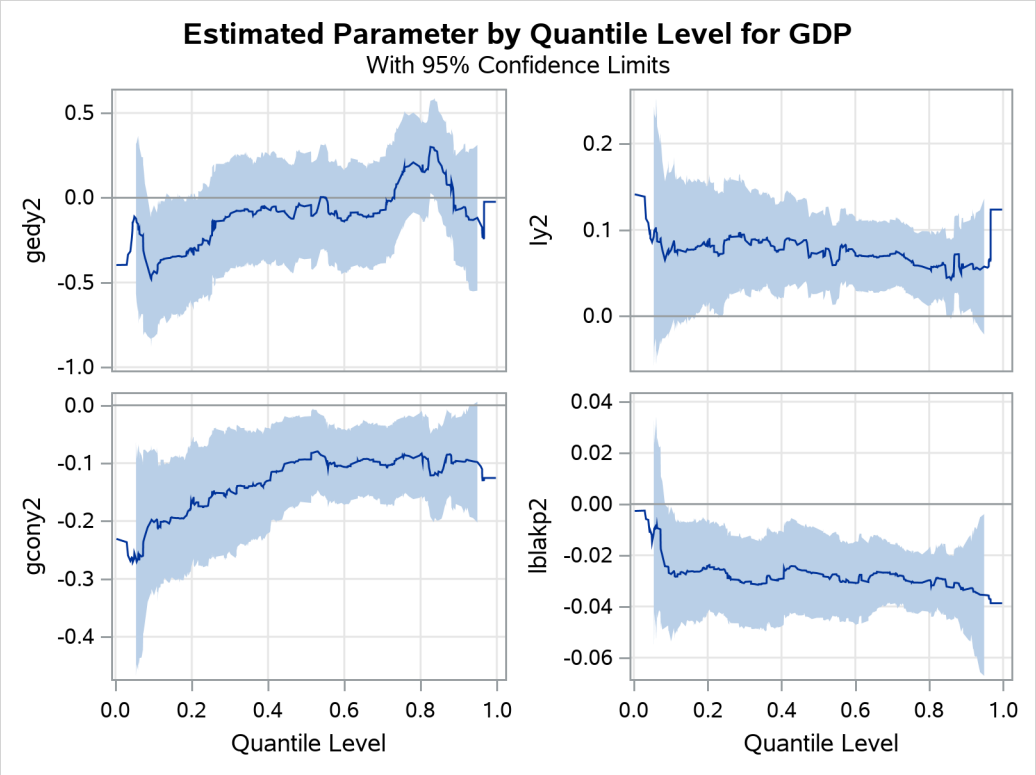
Confidence limits for quantile processes can be computed by using the sparsity or resampling methods. But they cannot be computed by using the rank method, because the computation would be prohibitively expensive.

A total of 14 quantile process plots are produced. Output 100.2.7 and Output 100.2.8 display two panels of eight selected process plots. The 95% confidence bands are shaded.

**Output 100.2.7** Quantile Processes with 95% Confidence Bands



**Output 100.2.8** Quantile Processes with 95% Confidence Bands

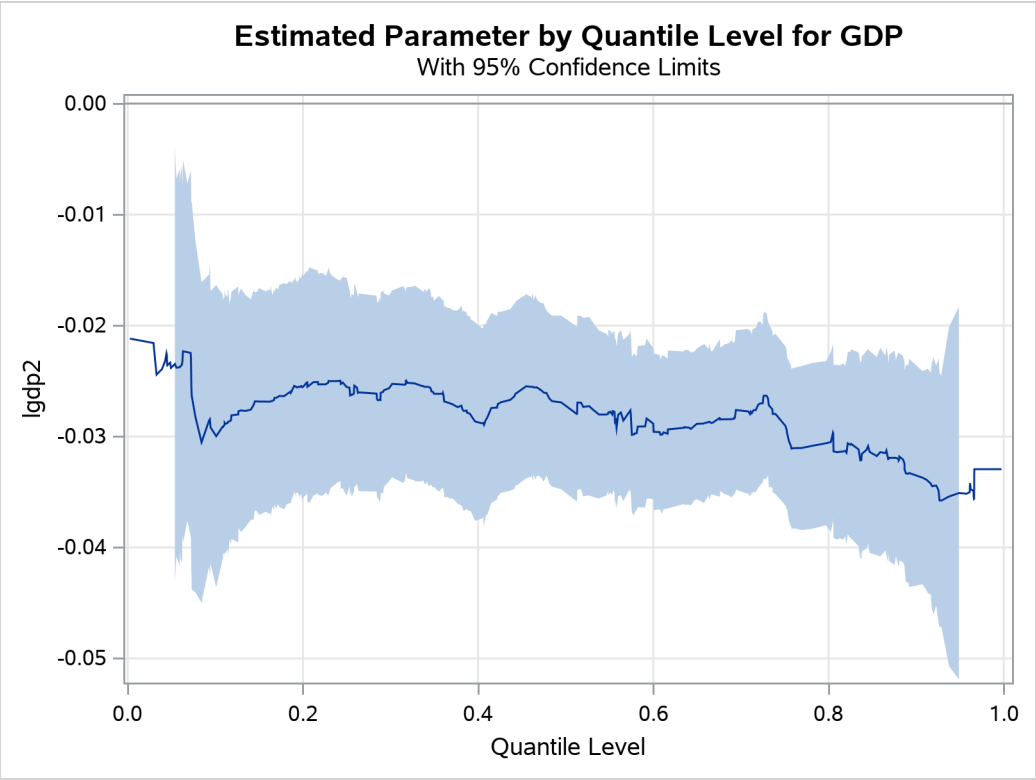



As pointed out by Koenker and Machado (1999), previous studies of the Barro growth data have focused on the effect of the initial per-capita GDP on the growth of this variable (annual change in per-capita GDP). The following statements request a single process plot for this effect:

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

The plot is shown in Output 100.2.9.

**Output 100.2.9** Quantile Process Plot for LGDP2




The confidence bands here are computed by using the MCMB resampling method. In contrast, Koenker and Machado (1999) used the rank method to compute confidence limits for a few selected points. Output 100.2.9 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 divided by 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, see Koenker and Machado (1999).

## Example 100.3: Quantile Regression Analysis of Birth-Weight Data

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

This example is based on Koenker and Hallock (2001); Abreveya (2001); it 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 are randomly selected from the

qualified observations. Observations that have missing data for any of the variables are deleted. The data are available in the data set Sashelp.BWeight. The following step displays in Output 100.3.1 the variables in the data set:

```
proc contents varnum data=sashelp.bweight;
   ods select position;
run;
```

**Output 100.3.1** Sashelp.BWeight Data Set

**The CONTENTS Procedure**

**Variables in Creation Order**

| # | Variable | Type | Len | Label |
|---|---|---|---|---|
| 1 | Weight | Num | 8 | Infant Birth Weight |
| 2 | Black | Num | 8 | Black Mother |
| 3 | Married | Num | 8 | Married Mother |
| 4 | Boy | Num | 8 | Baby Boy |
| 5 | MomAge | Num | 8 | Mother's Age |
| 6 | MomSmoke | Num | 8 | Smoking Mother |
| 7 | CigsPerDay | Num | 8 | Cigarettes Per Day |
| 8 | MomWtGain | Num | 8 | Mother's Pregnancy Weight Gain |
| 9 | Visit | Num | 8 | Prenatal Visit |
| 10 | MomEdLevel | Num | 8 | Mother's Education Level |

The following step creates descriptive labels for the values of the classification variables Visit and MomEdLevel:

```
proc format;
   value vfmt 0 = 'No Visit'       1 = 'Second Trimester'
              2 = 'Last Trimester' 3 = 'First Trimester';
   value efmt 0 = 'High School'    1 = 'Some College'
              2 = 'College'        3 = 'Less Than High School';
run;
```

There are four levels of maternal education. When you specify the ORDER=INTERNAL option, PROC QUANTREG treats the highest unformatted value (3, which represents that the mother's education level is less than high school) as a reference level. The regression coefficients of other levels measure the effect relative to this level. Likewise, there are four levels of prenatal medical care of the mother, and a first visit in the first trimester serves as the reference level.

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

```
ods graphics on;

proc quantreg ci=sparsity/iid algorithm=interior(tolerance=5.e-4)
   data=sashelp.bweight order=internal;
   class Visit MomEdLevel;
   model Weight = Black Married Boy Visit MomEdLevel MomSmoke
   CigsPerDay MomAge MomAge*MomAge
   MomWtGain MomWtGain*MomWtGain /
```

```
   quantile= 0.05 to 0.95 by 0.05
   plot=quantplot;
   format Visit vfmt. MomEdLevel efmt.;
run;
```

Output 100.3.2 displays the model information and summary statistics for the variables in the model.

**Output 100.3.2** Model Information and Summary Statistics

**The QUANTREG Procedure**

| Model Information | | |
|---|---|---|
| **Data Set** | SASHELP.BWEIGHT | Infant Birth Weight |
| **Dependent Variable** | Weight | Infant Birth Weight |
| **Number of Independent Variables** | 9 | |
| **Number of Continuous Independent Variables** | 7 | |
| **Number of Class Independent Variables** | 2 | |
| **Number of Observations** | 50000 | |
| **Optimization Algorithm** | Interior | |
| **Method for Confidence Limits** | Sparsity | |

**Summary Statistics**

| Variable | Q1 | Median | Q3 | Mean | Standard Deviation | MAD |
|---|---|---|---|---|---|---|
| **Black** | 0 | 0 | 0 | 0.1628 | 0.3692 | 0 |
| **Married** | 0 | 1.0000 | 1.0000 | 0.7126 | 0.4525 | 0 |
| **Boy** | 0 | 1.0000 | 1.0000 | 0.5158 | 0.4998 | 0 |
| **MomSmoke** | 0 | 0 | 0 | 0.1307 | 0.3370 | 0 |
| **CigsPerDay** | 0 | 0 | 0 | 1.4766 | 4.6541 | 0 |
| **MomAge** | -4.0000 | 0 | 5.0000 | 0.4161 | 5.7285 | 5.9304 |
| **MomAge*MomAge** | 4.0000 | 16.0000 | 49.0000 | 32.9877 | 39.2861 | 22.2390 |
| **MomWtGain** | -8.0000 | 0 | 9.0000 | 0.7092 | 12.8761 | 11.8608 |
| **MomWtGain*MomWtGain** | 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 11 independent variables, Black, Married, Boy, and MomSmoke are binary variables. For these variables, the mean represents the proportion in the category. The two continuous variables, MomAge and MomWtGain, are centered at their medians, which are 27 and 30, respectively.

The quantile plots for the intercept and the other 15 factors with nonzero degrees of freedom are shown in the following four panels. In each plot, the regression coefficient at a given quantile indicates the effect on birth weight of a unit change in that factor, assuming that the other factors are fixed. The bands represent 95% confidence intervals.

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

In Output 100.3.3, the first plot is for the intercept. As explained by Koenker and Hallock (2001), the intercept "may be interpreted as the estimated conditional quantile function of the birth-weight distribution of a girl born to an unmarried, white mother with less than a high school education, who is 27 years old and had a weight gain of 30 pounds, didn't smoke, and had her first prenatal visit in the first trimester of the pregnancy."

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

In Output 100.3.4, the first three plots deal with prenatal care. Compared with babies born to mothers who had a prenatal visit in the first trimester, babies born to mothers who received no prenatal care weigh less, especially in the lower quantiles of the birth-weight distributions. As noted by Koenker and Hallock (2001), "babies born to mothers who delayed prenatal visits until the second or third trimester have substantially *higher* birthweights in the lower tail than mothers who had a prenatal visit in the first trimester. This might be interpreted as the self-selection effect of mothers confident about favorable outcomes."

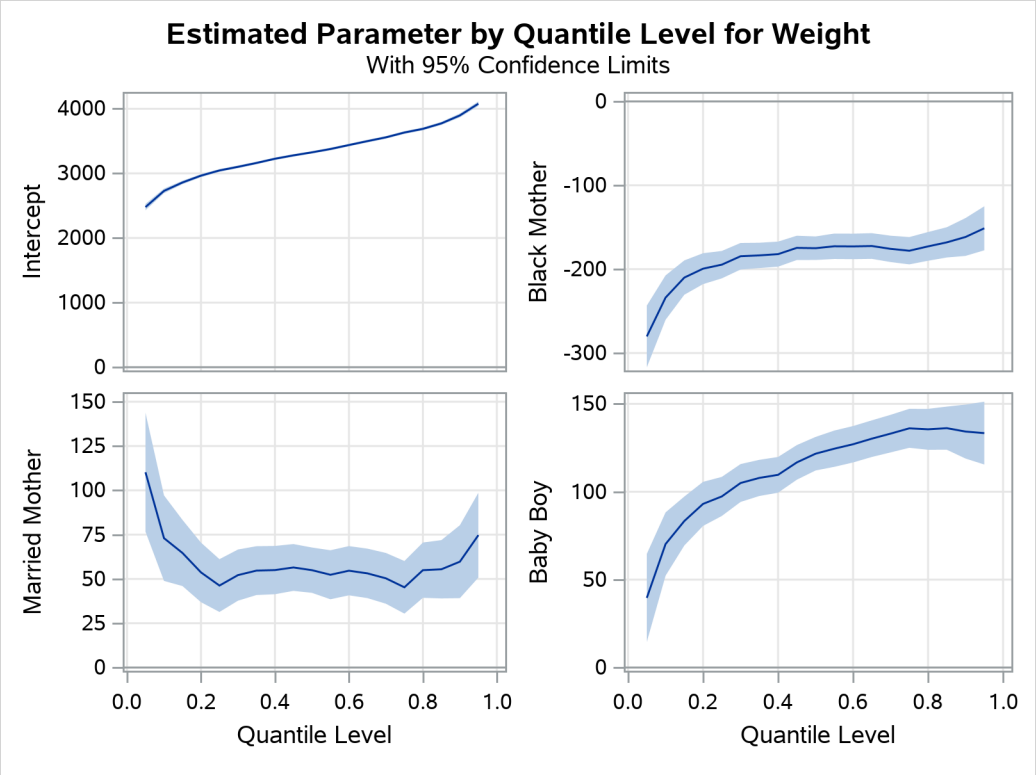
The fourth plot in Output 100.3.4 and the first two plots in Output 100.3.5 are for variables that are related to education. Education beyond high school is associated with a positive effect on birth weight. The effect of high school education is uniformly around 15 grams across the entire birth-weight distribution (this is a pure location shift effect), whereas 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 100.3.5 show that smoking is associated with a large negative effect on birth weight.

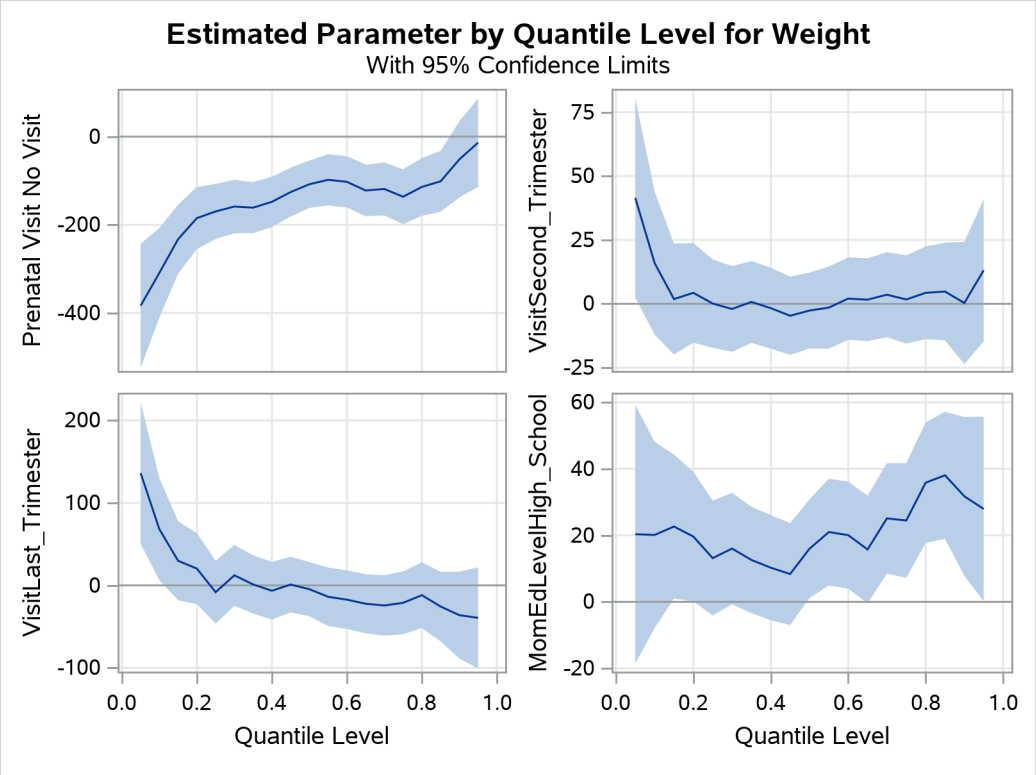
The linear and quadratic effects for the two continuous variables are shown in Output 100.3.6. Both of these variables are centered at their median. At the lower quantiles, the quadratic effect of the mother's age is more concave. The optimal age at the first quantile is about 33, and the optimal age 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.

For more information about the covariate effects that are discovered by using quantile regression, see Koenker and Hallock (2001).

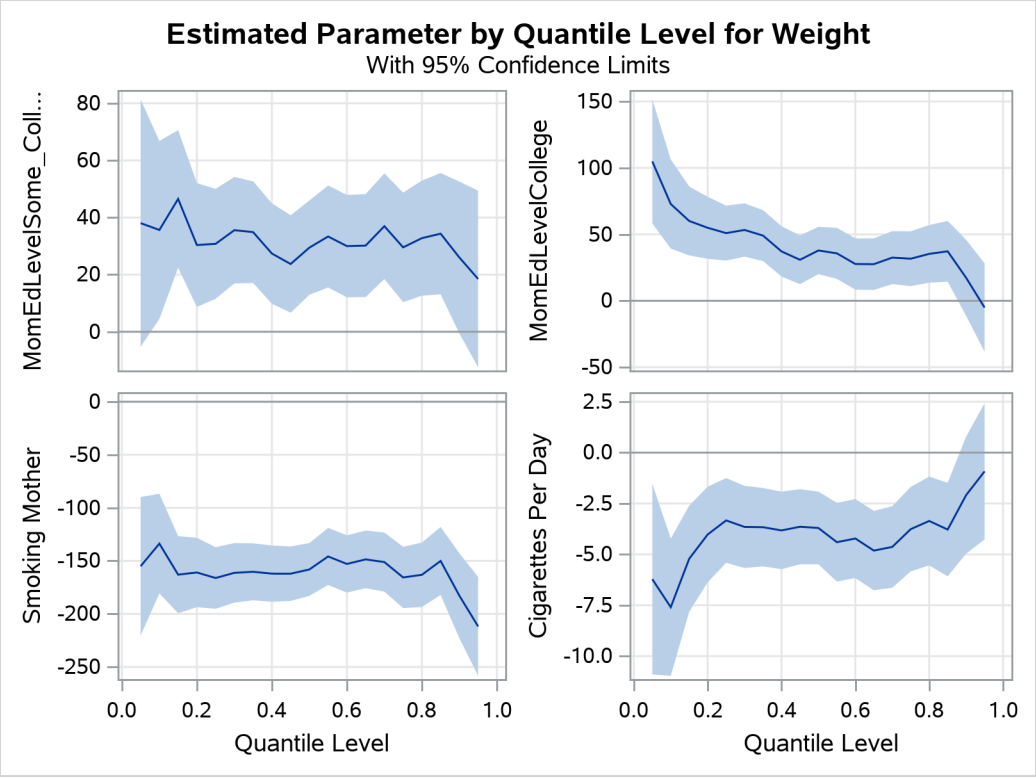
**Output 100.3.3** Quantile Processes with 95% Confidence Bands



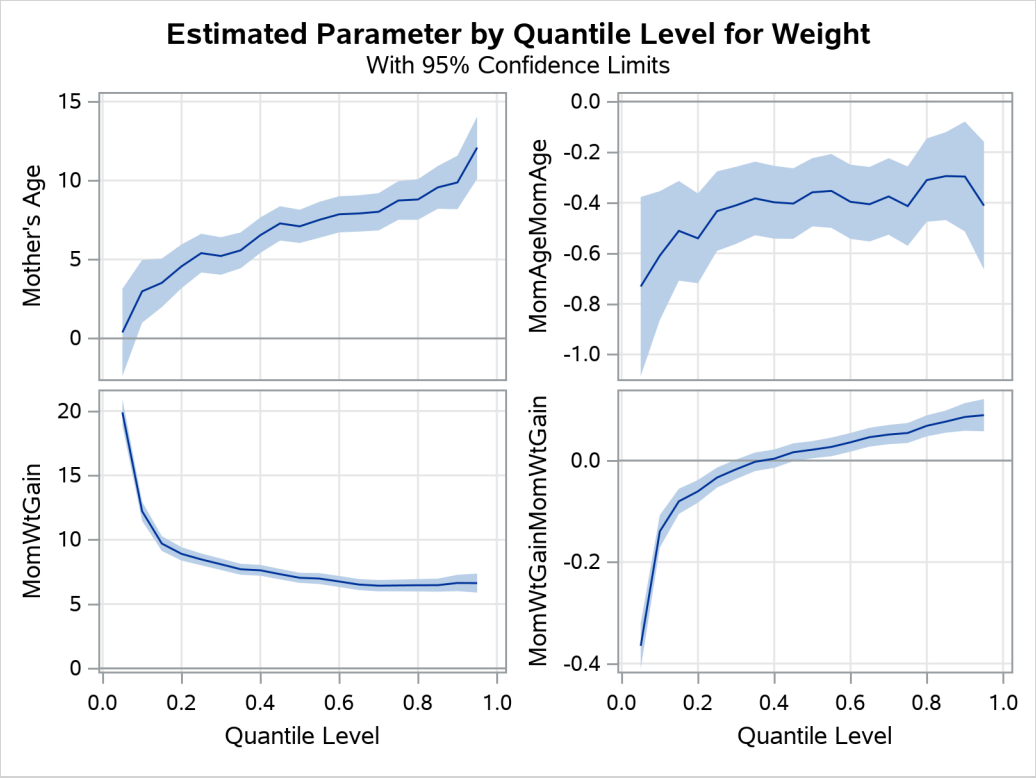

**Output 100.3.4** Quantile Processes with 95% Confidence Bands



**Output 100.3.5** Quantile Processes with 95% Confidence Bands




**Output 100.3.6** Quantile Processes with 95% Confidence Bands



## Example 100.4: Nonparametric Quantile Regression for Ozone Levels

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

The data used here are from Chock, Winkler, and Chen (2000), who studied the association between daily mortality and ambient air pollutant concentrations in Pittsburgh, Pennsylvania. The data set ozone contains the following two variables: Ozone, which represents the daily maximum one-hour ozone concentration (ppm) and Days, which is an index of 1,095 days (3 years).

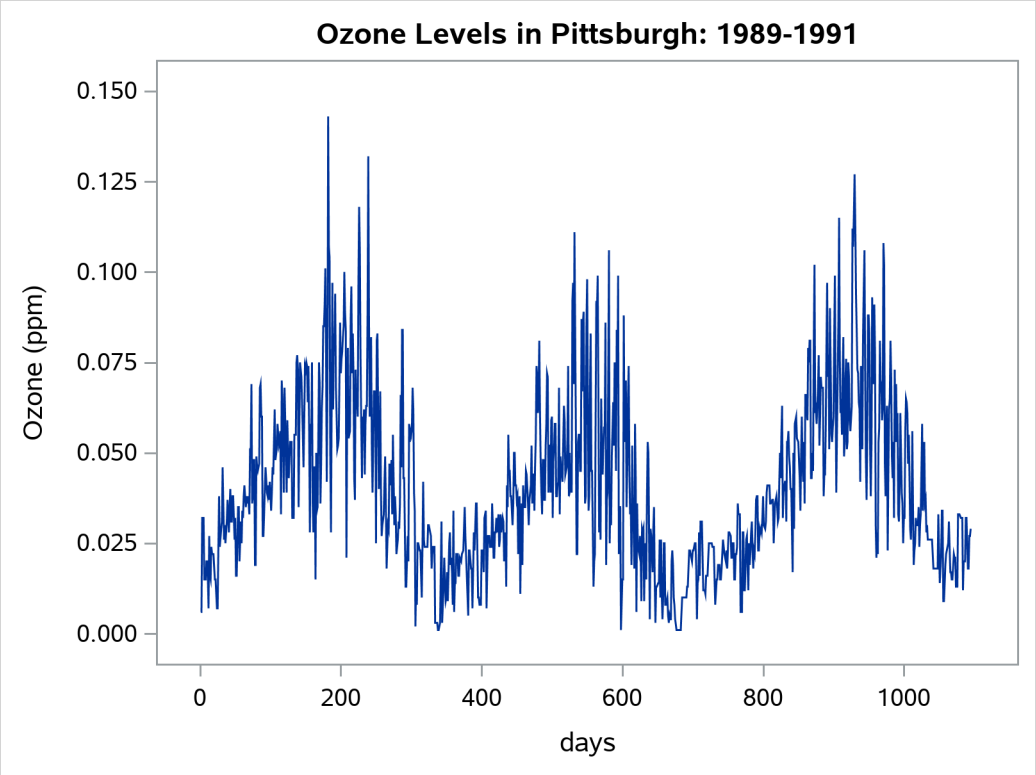
```
data ozone;
   days = _n_;
   input ozone @@;
   datalines;
0.0060 0.0060 0.0320 0.0320 0.0320 0.0150 0.0150 0.0150 0.0200 0.0200
0.0160 0.0070 0.0270 0.0160 0.0150 0.0240 0.0220 0.0220 0.0220 0.0185
0.0150 0.0150 0.0110 0.0070 0.0070 0.0240 0.0380 0.0240 0.0265 0.0290
0.0310 0.0460 0.0360 0.0260 0.0300 0.0250 0.0280 0.0310 0.0370 0.0325

   ... more lines ...

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

Output 100.4.1, which displays the time series plot of ozone concentration for the three years, shows a clear seasonal pattern.

In this example, cubic B-splines are used to fit the seasonal effect. These splines are generated with 11 knots, which split the 3 years into 12 seasons. The following statements create the spline basis and fit multiple quantile regression spline curves:

```
ods graphics on;

proc quantreg data=ozone algorithm=smooth ci=none plot=fitplot(nodata);
   effect sp = spline( days / knotmethod = list
          (90 182 272 365 455 547 637 730 820 912 1002) );
   model ozone = sp / quantile = 0.5 0.75 0.90 0.95 seed=1268;
run;
```

**Output 100.4.1** Time Series of Ozone Levels in Pittsburgh, Pennsylvania

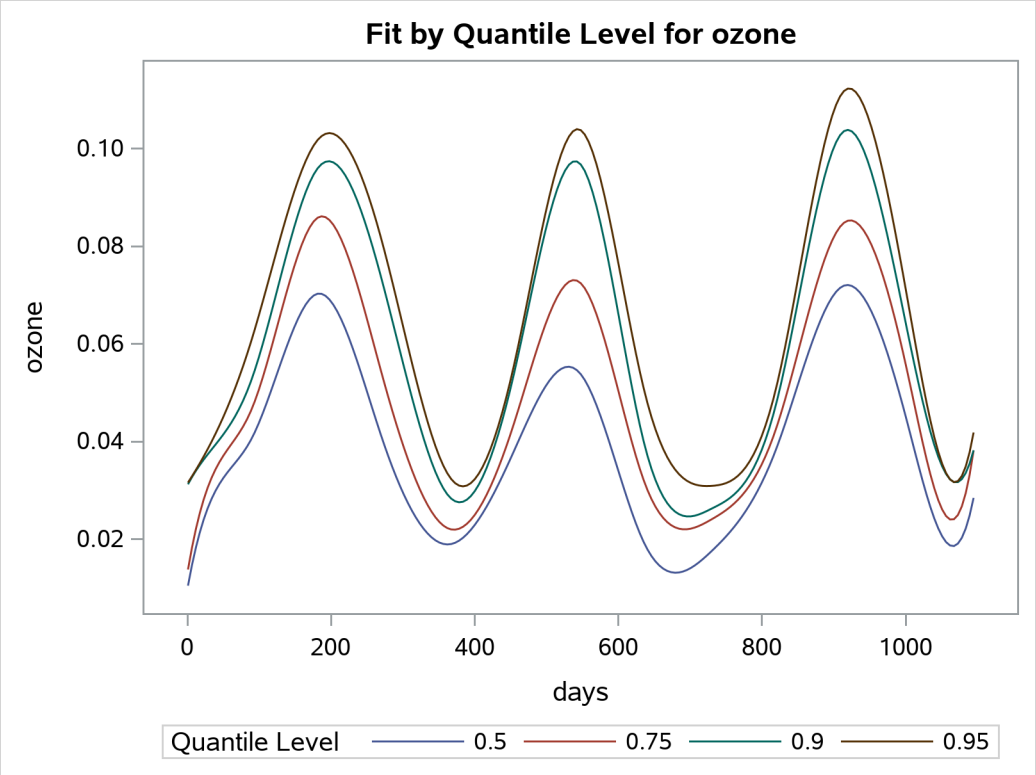



The EFFECT statement creates spline bases for the variable Days. The KNOTMETHOD=LIST option provides all internal knots for these bases. Cubic spline bases are generated by default. These bases are treated as components of the spline effect *sp*, which is specified in the MODEL statement. Spline fits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

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

Output 100.4.2 displays these curves. The curves show that peak ozone levels occur in the summer. For the three years 1989–1991, the median curve (labeled 50%) does not cross the 0.08 ppm line, which is the 1997 EPA eight-hour standard. The median curve and the 75% curve show a drop for the ozone concentration levels in 1990. However, for the 90% and 95% curves, peak ozone levels tend to increase. This indicates that there might have been more days with low ozone concentration in 1990, but the top 10% and 5% tend to have higher ozone concentration levels.

**Output 100.4.2** Quantiles of Ozone Levels in Pittsburgh, Pennsylvania




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

## Example 100.5: Quantile Polynomial Regression for Salary Data

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

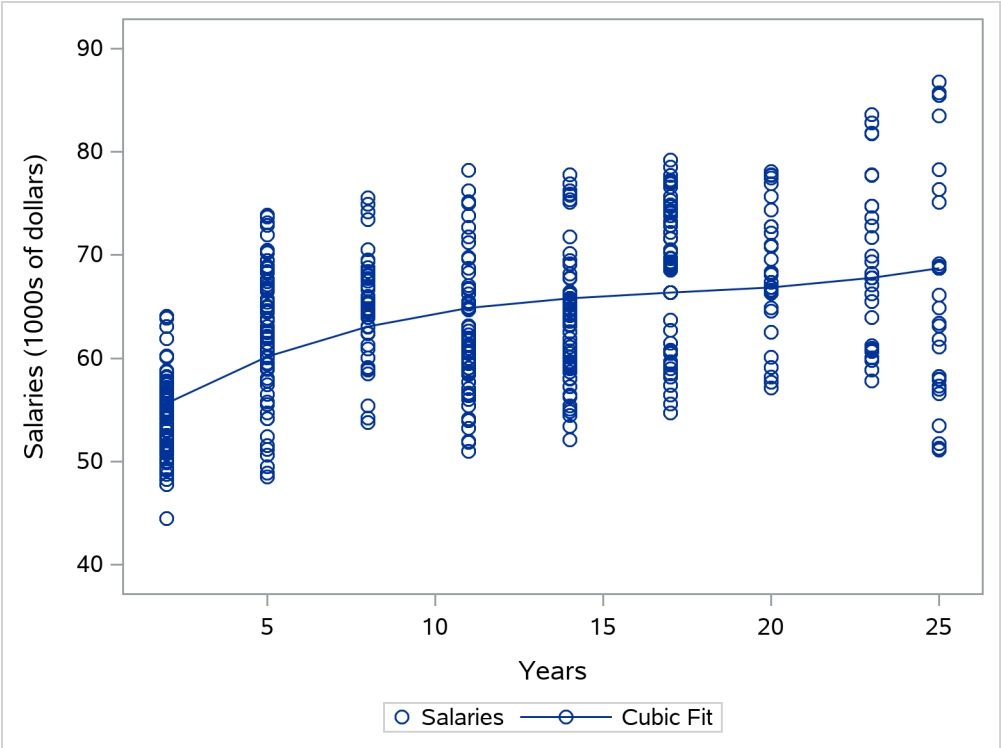
The data include salaries and years of service for 459 professors. The scatter plot in Output 100.5.1 shows that the relationship is not linear and that a quadratic or cubic regression curve is appropriate. Output 100.5.1 shows a cubic curve.

The curve in Output 100.5.1 does not adequately describe the conditional salary distributions and how they change with length of service. Output 100.5.2 shows the 25th, 50th, and 75th percentiles for each number of years, which gives a better picture of the conditional distributions.

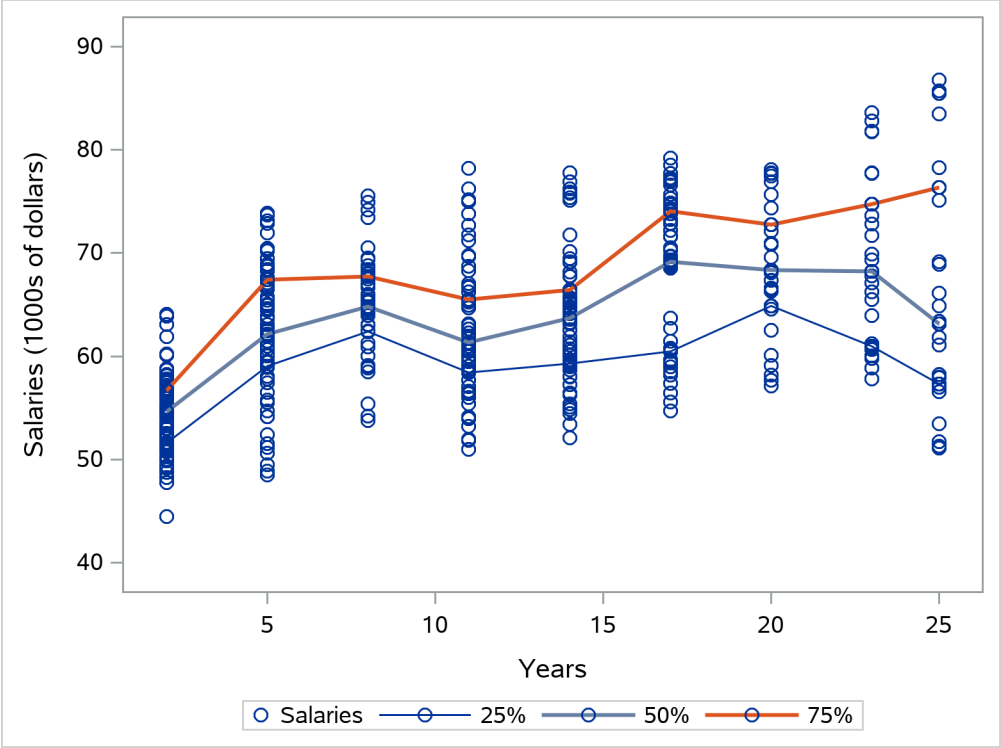
```
data salary;
   input Salaries Years @@;
   label Salaries='Salaries (1000s of dollars)';
   datalines;
54.94   2  58.24   2  58.11   2  52.23   2  52.98    2  57.62    2
44.48   2  57.22   2  54.24   2  54.79   2  56.42    2  61.90    2
63.90   2  64.10   2  47.77   2  54.86   2  49.31    2  53.37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

```
51.69   2  53.66   2  58.77   2  56.77   2  53.06    2  54.86    2
50.96   2  56.46   2  51.67   2  49.37   2  56.86    2  49.85    2

   ... more lines ...

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

**Output 100.5.1** Salary and Years as Professor: Cubic Fit



**Output 100.5.2** Salary and Years as Professor: Sample Quantiles




These descriptive percentiles do not clearly show trends with length of service. The following statements use polynomial quantile regression to obtain a smooth version.

```
ods graphics on;

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

   test  years/QINTERACT;

run;
```

The results are shown in Output 100.5.3 and Output 100.5.5. Output 100.5.3 displays the regression coefficients for the three quantiles, from which you can see a difference among the estimated parameters of the variable *years* across the three quantiles. To test whether the difference is significant, you can specify the option QINTERACT in the TEST statement. Output 100.5.4 indicates that the difference is not significant (the *p*-value is greater than 0.05).

**Output 100.5.3** Regression Coefficients

**The QUANTREG Procedure**
**Quantile Level = 0.25**

**Parameter Estimates**

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

**Parameter Estimates**

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

**Parameter Estimates**

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

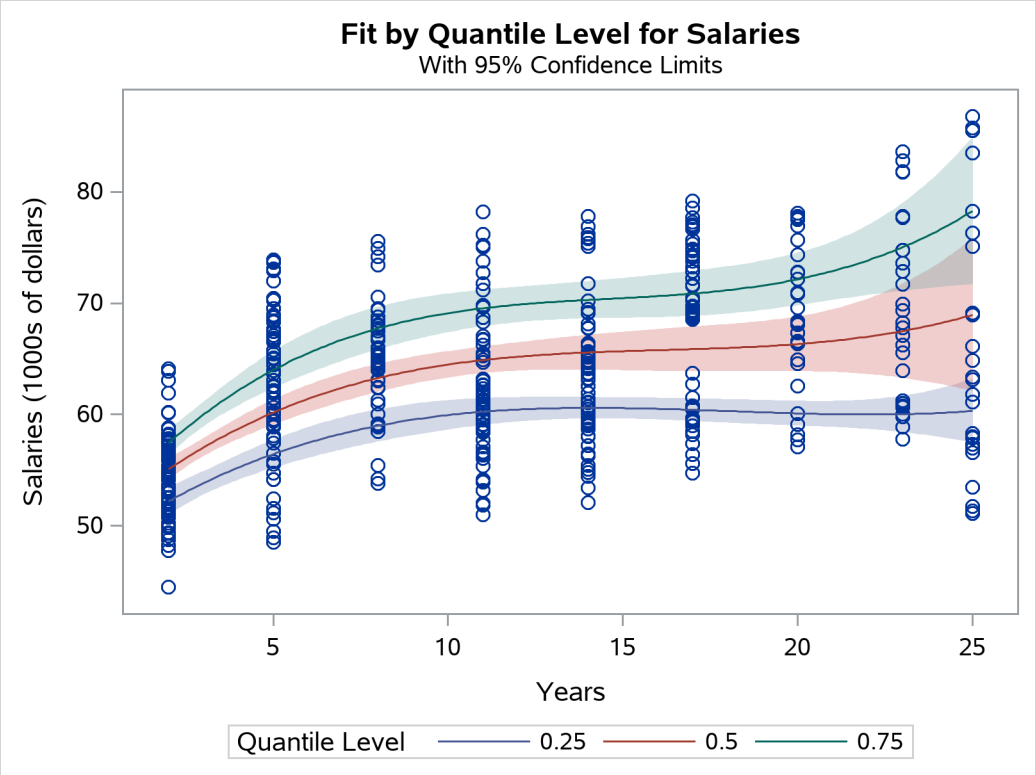
**Output 100.5.4** Tests for Heteroscedasticity

**Test Results Equal Coefficients Across Quantiles**

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 3.4026 | 2 | 0.1825 |

The three fitted quantile curves and their 95% confidence limits in the Output 100.5.5 clearly show that salary dispersion increases gradually with length of service. After 15 years, a salary more than $70,000 is relatively high, whereas a salary less than $60,000 is relatively low. Percentile curves of this type are useful in medical science as reference curves (Yu, Lu, and Stander 2003).

**Output 100.5.5** Salary and Years as Professor: Regression Quantiles




## Example 100.6: Counterfactual Analysis of Smoking-Weight Data

This example demonstrates how you can use the CONDDIST statement in the QUANTREG procedure to perform counterfactual analysis. The data are from the NHANES I Epidemiologic Follow-Up Study (NHEFS) in Hernán and Robins (2018). NHEFS collected medical and behavioral information during an initial physical examination, and again at follow-up interviews approximately one decade later. Chapter 36, "The CAUSALTRT Procedure," uses these same data to analyze the average treatment effect.

The following DATA step creates the input data set SmokingWeight:

```
data SmokingWeight;
   input
   Sex Age Race Education Exercise BaseWeight Weight Change Activity
   YearsSmoke PerDay Quit;
   label
      Activity   = 'Level of daily activity with values in {0,1,2}'
      Age        = 'Age in 1971'
      BaseWeight = 'Weight in kilograms at the 1971 interview'
      Change     = 'Body Weight Change'
      Education  = 'Level of education with values in {0,1,2,3,4}'
      Exercise   = 'Level of recreational exercise with values in {0,1,2}'
      Perday     = 'Number of cigarettes smoked per day in 1971'
      Quit       = '1 for quitting smoking; 0 otherwise'
      Race       = '0 for white; 1 otherwise'
      Sex        = '0 for male; 1 for female'
      Weight     = 'Weight in kilograms at the follow-up interview'
      YearsSmoke = 'Number of years an individual has smoked';
```

```
   datalines;
   0 42 1 1 2 79.04 68.95 -10.09 0 29 30 0
   0 36 0 2 0 58.63 61.23 2.60 0 24 20 0
   1 56 1 2 2 56.81 66.22 9.41 0 26 20 0

   ... more lines ...

   1 51 0 3 0 62.71 . . 0 30 40 0
   0 68 0 1 1 52.39 57.15 4.76 1 46 15 0
   0 26 0 . 0 86.75 87.54 0.79 0 9 20 0
   0 29 0 2 1 90.83 106.59 15.76 1 14 30 1
;
```

The following statements split the SmokingWeight data set into two data subsets:

```
data QuitSmoking KeepSmoking;
   set SmokingWeight;
   if Quit=0 then output KeepSmoking;
   else if
      Quit=1 then output QuitSmoking;
run;
```

The data set KeepSmoking contains all the keep-smoking observations, and the data set QuitSmoking contains all the quit-smoking observations.

The goal of this study is to estimate the effect of quitting smoking on people's changes in body weight. Because NHEFS did not assign subjects to the quit-smoking group or the keep-smoking group, these data are observational but not experimental. To estimate the effect of quitting smoking, this study needs to estimate the counterfactual distribution of body weight changes for the quit-smoking group by assuming that the members of the quit-smoking group were continuing to smoke between their two interviews. With the further assumption that this counterfactual distribution follows the same conditional distribution model of body weight changes for the observed keep-smoking group in the KeepSmoking data set, quantile regression can estimate this counterfactual distribution by taking the following steps:

1. Build a quantile process regression model by using only the keep-smoking data.
2. Predict the quantile process for each observation in the quit-smoking data, which is a sequence of quantile predictions on a prespecified quantile-level grid.
3. Create a univariate sample by pooling together all the quantile predictions for all the observations in the quit-smoking data.
4. Estimate the counterfactual distribution of body weight changes on the univariate sample.

Because these steps do not rely on the observed body weight changes of the quit-smoking data, given the explanatory covariates of the quit-smoking data, the fitted counterfactual marginal distribution is also independent of the observed marginal distribution of the body weight changes for the quit-smoking group. For more information about this estimation method, see the section "Fitted Marginal Cumulative Distribution Functions" on page 8300.

By using the fitted counterfactual distribution, this study can evaluate the effect of quitting smoking on the body weight changes by comparing the following three distributions:

TrainObs: the observed marginal distribution for the keep-smoking group

TestObs: the observed marginal distribution for the quit-smoking group

TestFit: the fitted counterfactual marginal distribution for the quit-smoking group that is fitted by using the model built on the keep-smoking observations

For more information about the observed and fitted marginal distributions, see the section "Estimating Probability Functions by Using the CONDDIST Statement" on page 8298.

The following statements invoke the PROC QUANTREG and CONDDIST statements:

```
ods graphics on;
ods select CondDistEst CmpTrainObs CmpTestObs
           DensityEstInfo cdfplot pdfplot;

proc quantreg data=KeepSmoking ci=none;
   class sex Race Education Activity Exercise;
   model Change = Sex Age Education Activity Exercise YearsSmoke
                  PerDay / quantlev=fqpr(n=255);
   conddist mwu pdf=kde(l=-20 u=30) plot(sg)=(cdfplot pdfplot)
            testdata(mwu)=QuitSmoking;
run;
```

The DATA=KeepSmoking option in the PROC QUANTREG statement specifies the data set KeepSmoking as the training data set. The MODEL statement specifies the variable Change as the response variable and specifies a set of other variables as the explanatory variables. The QUANTLEV=FQPR(N=255) option in the MODEL statement specifies the fast quantile process regression method to build the quantile process regression model on a grid of 255 quantile levels, $\{0.5/255, 1.5/255, \ldots, 254.5/255\}$. For more information about the fast quantile process regression method, see the section "Fast Quantile Process Regression" on page 8292.

The MWU option in the CONDDIST statement performs the Mann-Whitney U test on the CDF sample of the TrainObs distribution. The PDF=KDE(L=–20 U=30) option specifies the kernel density estimation method to estimate probability density functions for the three distributions within the quantile range $(-20, 30)$. The PLOT(SG)=(CDFPLOT PDFPLOT) option produces the CDF plot and the PDF plot, and the SG suboption adds the light gray grid lines to the plots. The TESTDATA(MWU)=QuitSmoking option specifies the data set QuitSmoking as the test data set and performs the Mann-Whitney U test on the CDF sample of the TestObs distribution.

Output 100.6.1 shows the conditional distribution estimates.

**Output 100.6.1** Conditional Distribution Estimates

**The QUANTREG Procedure**
**Conditional Distribution Analysis 1**

**Conditional Distribution Estimates**

| | | | | Quantile Level | | |
|---|---|---|---|---|---|---|
| Data | Type | Label | Response Value | Regression | Sample | Regression Density |
| Training | Observed | TrainObs | 1.984359 | 0.4902 | 0.4892 | 0.065704 |
| Test | Observed | TestObs | 4.524913 | 0.5270 | 0.6573 | 0.053558 |
| Test | Fit and Pooled | TestFit | 4.524913 | 0.6935 | 0.6573 | 0.056206 |

Each row of this table corresponds to the pair of a predicted CDF sample and an assigned or observed response value. The Label column displays the labels of the CDF samples. The Data column indicates whether the CDF samples are associated with the keep-smoking data valued as "Training" or with the quit-smoking data valued as "Test." The Type column shows the types of CDF samples, where the "Observed" type is for the observed marginal distributions and the "Fit and Pooled" type is for the fitted counterfactual marginal distribution. The response values for these three CDF samples are all assigned averages. The Response Value column shows that the average weight change for the keep-smoking data is about 1.984 kilograms and the average weight change for the quit-smoking data is about 4.525 kilograms. If you specify the OBS= or TESTDATA(SHOWOBS)= option (or both) in the CONDDIST statement, the Response Value column shows the observed body weight changes for the relevant observations. The two Quantile Level columns, Regression and Sample, show the regression quantile levels and sample quantile levels for the response values. For more information about these estimated quantile levels, see the section "Regression Quantile Level and Sample Quantile Level" on page 8301.

Output 100.6.2 shows distribution comparisons to the TrainObs distribution from the TestObs distribution in the first row and from the TestFit distribution in the second row by using the Mann-Whitney U test.

**Output 100.6.2** Distribution Comparisons to Observed Marginal Distribution of the Training Data Response

| Distribution Comparisons to TrainObs | | | | |
|---|---|---|---|---|
| Data | Type | Label | z Value | Pr > \|z\| |
| Test | Observed | TestObs | 3.570865 | 0.0004 |
| Test | Fit and Pooled | TestFit | -1.05708 | 0.2905 |

In the first row, the $p$-value 0.0004 rejects the hypothesis that the observed marginal distribution of the body weight changes for the quit-smoking group (labeled TestObs) is the same as that for the keep-smoking group (labeled TrainObs); and the $z$-value $3.571 > 0$ implies that the quit-smoking group gains significantly more body weight than the keep-smoking group. In the second row, the $p$-value 0.2905 does not have enough evidence to reject the hypothesis that the fitted counterfactual marginal distribution of the body weight changes for the quit-smoking group (labeled TestFit) is the same as that for the keep-smoking group (or the TrainObs distribution). The $z$-value $-1.0571 < 0$ implies that the quit-smoking group might even gain less body weight than the keep-smoking group if the quit-smoking group were not quitting smoking, although not enough evidence is available for this implication because the $p$-value 0.2905 is large.

Output 100.6.3 shows distribution comparisons to the TestObs distribution from the TestFit distribution by using the Mann-Whitney U test.

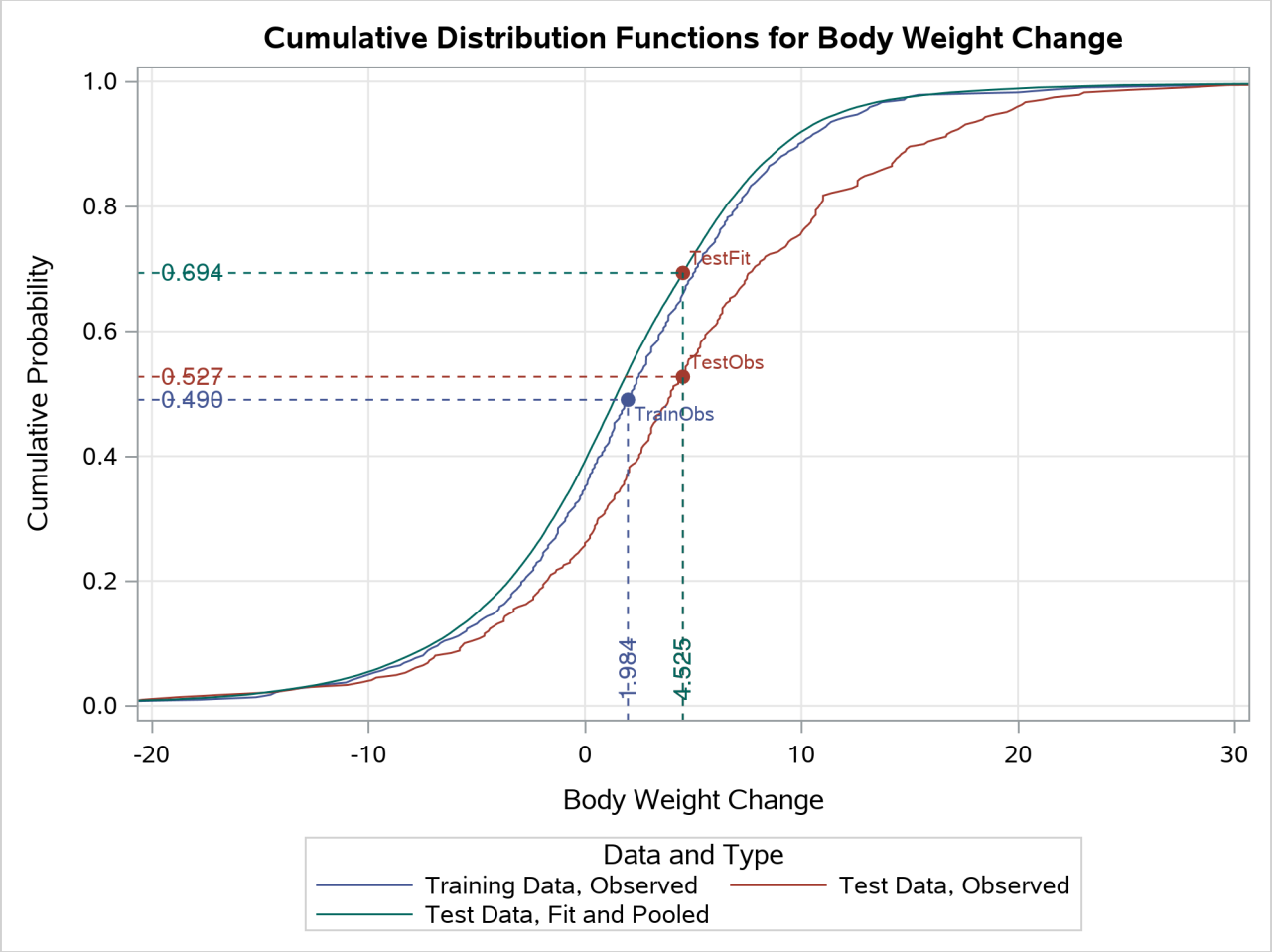
**Output 100.6.3** Distribution Comparisons to Observed Marginal Distribution of the Test Data Response

| Distribution Comparisons to TestObs | | | | |
|---|---|---|---|---|
| Data | Type | Label | z Value | Pr > \|z\| |
| Test | Fit and Pooled | TestFit | -4.48130 | 0.0000 |

The $p$-value 0 strongly rejects the hypothesis that quitting smoking has no impact on the distribution of body weight changes for the quit-smoking group. The $z$-value $-4.4813 < 0$ implies that the quit-smoking group gains significantly less body weight, if they were not actually quitting smoking, in comparison to this group in reality.

Output 100.6.4 shows the CDFs for the TrainObs, TestObs, and TestFit distributions.

**Output 100.6.4** Cumulative Distribution Functions




You can see that the CDF for the counterfactual quit-smoking group (green TestFit curve) is stochastically smaller than the observed marginal distributions for both the keep-smoking group (blue TrainObs curve) and the real quit-smoking group (red TestObs curve). The average weight change for the real quit-smoking group 4.525 has a percentage value of 52.7% according to the TestObs distribution but 69.4% according to the counterfactual TestFit distribution.

Output 100.6.5 shows that the CONDDIST statement uses the kernel density estimation (KDE) method to estimate PDFs.

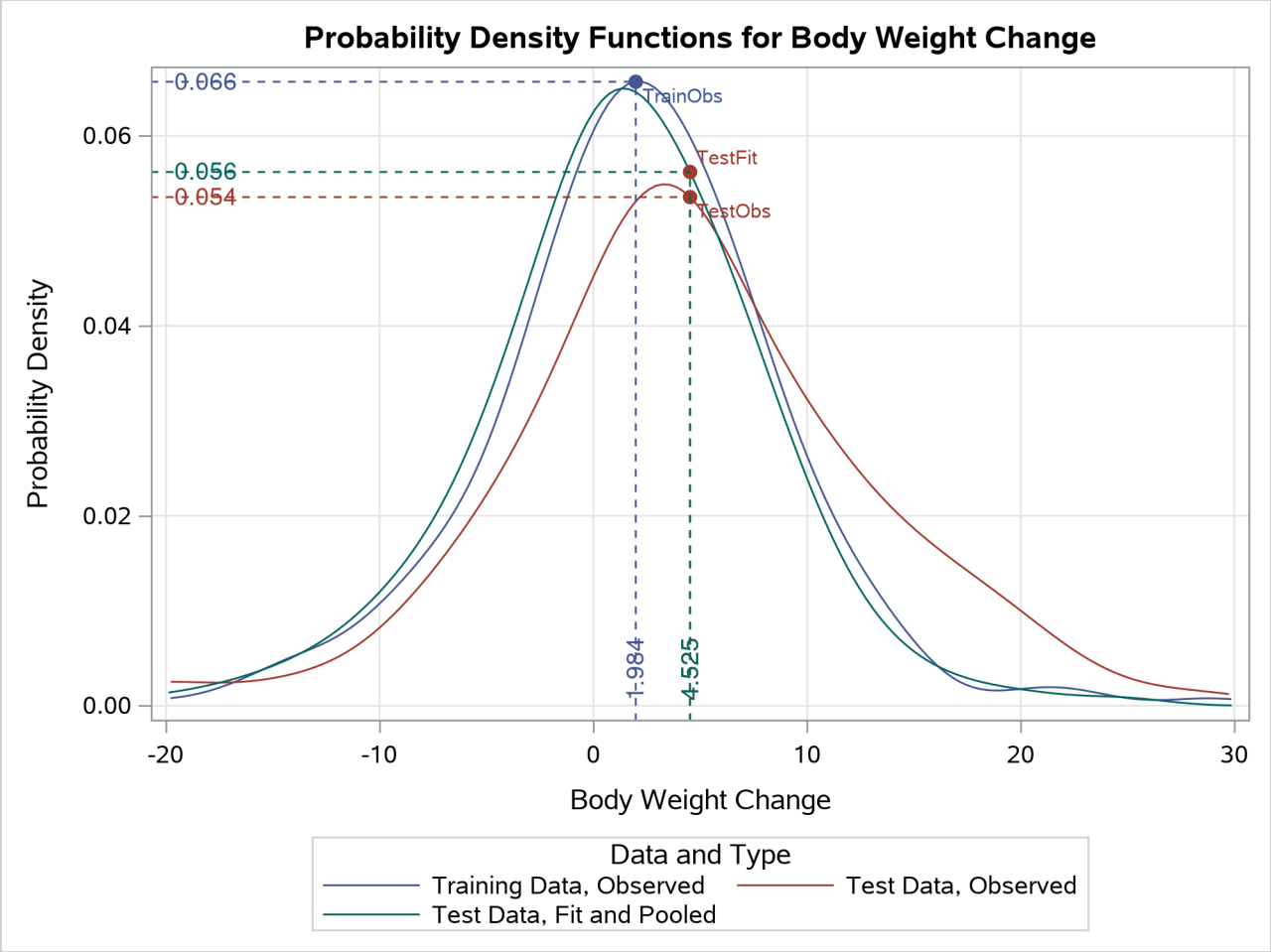
**Output 100.6.5** Density Estimation Information

| Density Estimates Information | |
|---|---|
| **Estimation Method** | Kernel |
| **Kernel Type** | Normal |
| **Grid Size** | 256 |
| **Bandwidth Parameter** | 0.7852 |
| **Bounds Adjustment Type** | Trim |
| **Lower Bound** | -20.0000 |
| **Upper Bound** | 30.0000 |

Each PDF is built on a quantile grid of 255 equally spaced points. Because of the PDF=KDE(L=-20 U=30) option in the SAS program, the KDE method sets the PDF values to zero for the quantile points that are either smaller than L=–20 or greater than U=30. For more information about the KDE method, see the section "Probability Density Functions" on page 8301.

Output 100.6.6 shows the PDFs for the TrainObs, TestObs, and TestFit distributions.

**Output 100.6.6** Probability Density Functions




You can see that the TrainObs and TestFit PDFs are likely to be following the normal distribution. However, the TestObs PDF is right-skewed. The skewness might occur because the quit-smoking time, which is not available in the data set, introduces extra deviance to the right arm of the TestObs distribution.

# 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://admin.nber.org/pub/barro.lee/readme.txt.

Barrodale, I., and Roberts, F. D. K. (1973). "An Improved Algorithm for Discrete $l_1$ Linear Approximation." *SIAM Journal on 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* 1:412–420.

Chen, C. (2004). "An Adaptive Algorithm for Quantile Regression." In *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf, and S. V. Aels, 39–48. Basel: Birkhäuser.

Chen, C. (2005). "Growth Charts of Body Mass Index (BMI) with Quantile Regression." In *Proceedings of 2005 International Conference on Algorithmic Mathematics and Computer Science*, edited by H. R. Arabnia and I. A. Ajwa, 114–120. Bogart, GA: CSREA Press.

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

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

Dunham, J. B., Cade, B. S., and Terrell, J. W. (2002). "Influences of Spatial and Temporal Variation on Fish-Habitat Relationships Defined by Regression Quantiles." *Transactions of the American Fisheries Society* 131:86–98.

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

Gutenbrunner, C., Jureckova, J., Koenker, R., and Portnoy, S. (1993). "Tests of Linear Hypotheses Based on Regression Rank Scores." *Journal of Nonparametric Statistics* 2:307–331.

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

Hernán, M. A., and Robins, J. M. (2018). *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC. Forthcoming.

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

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). "A Brief Survey of Bandwidth Selection for Density Estimation." *Journal of the American Statistical Association* 91:401–407.

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

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

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

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

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

Koenker, R., and Bassett, G. W. (1982b). "Tests of Linear Hypotheses and $l_1$ Estimation." *Econometrica* 50:1577–1583.

Koenker, R., and d'Orey, V. (1994). "Remark AS R92: A Remark on Algorithm AS 229: Computing Dual Regression Quantiles and Regression Rank Scores." *Journal of the Royal Statistical Society, Series C* 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.

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

Lustig, I. J., Marsten, 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. (1997). *Theory and Algorithms for Linear Optimization*. Chichester, UK: John Wiley & Sons.

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

Yao, Y. (2017). "Fast Quantile Process Regression." Paper presented at the 2017 International Conference on Robust Statistics, Wollongong, NSW, Australia.

Yu, K., Lu, Z., and Stander, J. (2003). "Quantile Regression: Application and Current Research Areas." *Journal of the Royal Statistical Society, Series D* 52:331–350.

# Subject Index

# Syntax Index