Quantile Regression

statgraphics® centurion

Revised: 12/2/2019

Summary

The **Quantile Regression** procedure fits linear models to describe the relationship between selected quantiles of a dependent variable Y and one or more independent variables. The independent variables may be either quantitative or categorical. Unlike standard multiple regression procedures in which the model is used to predict mean response, quantile regression models may be used to predict any percentile. Median regression is a special case where the quantile to be predicted is the 50^{th} percentile.

The models are estimated by the "quantreg" package in R. To run the procedure, R must be installed on your computer together with that package. For information on downloading and installing R, refer to the document titled "R – Installation and Configuration".

Sample StatFolio: *quantilereg.sgp*

Sample Data:

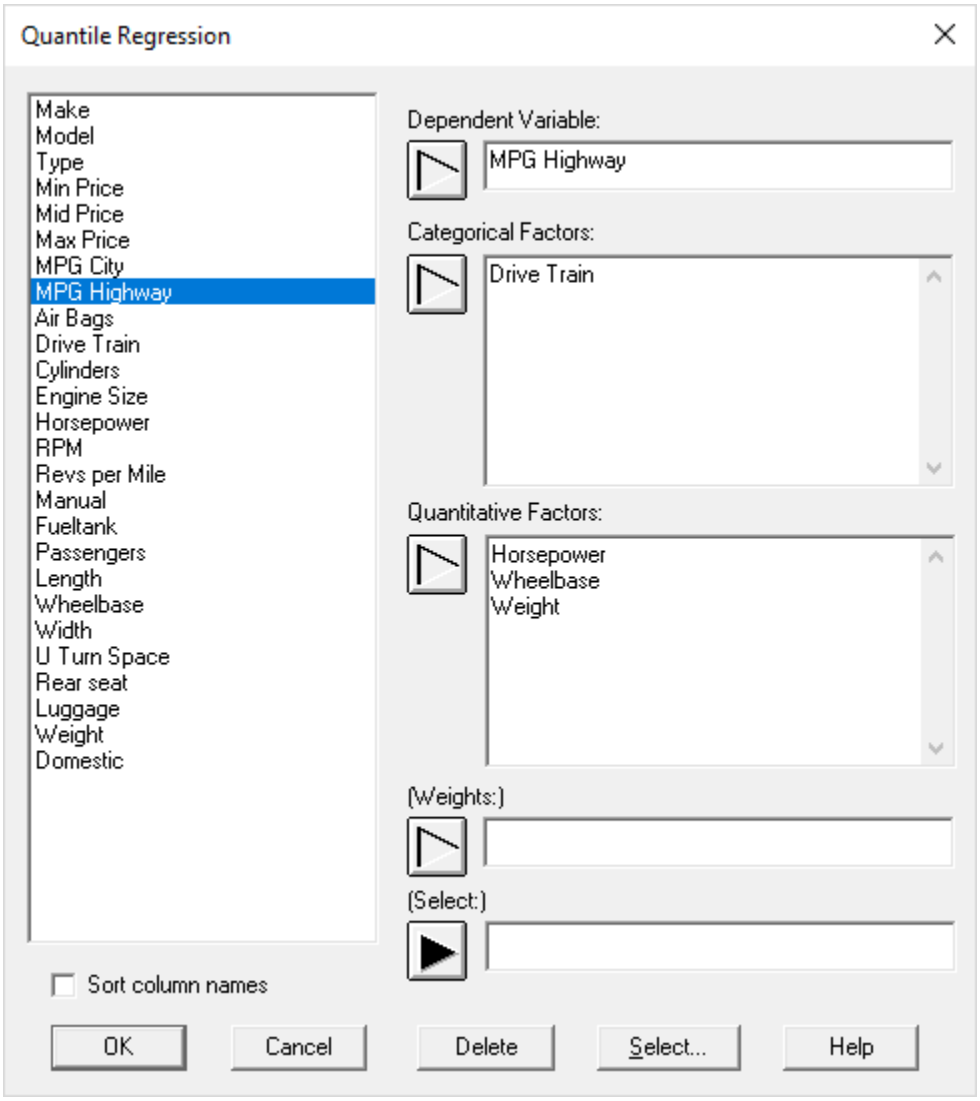
The file *93cars.sgd* contains information on 26 variables for n = 93 models of automobiles, taken from Lock (1993). The table below shows a partial list of 7 columns from that file:

Make	*Model*	*MPG Highway*	*Weight*	*Horsepower*	*Wheelbase*	*Drive Train*
Acura	Integra	31	2705	140	102	front
Acura	Legend	25	3560	200	115	front
Audi	90	26	3375	172	102	front
Audi	100	26	3405	172	106	front
BMW	535i	30	3640	208	109	rear
Buick	Century	31	2880	110	105	front
…	…	…	…	…	…	…

A model is desired that can predict *MPG Highway* from *Weight*, *Horsepower*, *Wheelbase*, and *Drive Train*.

Data Input

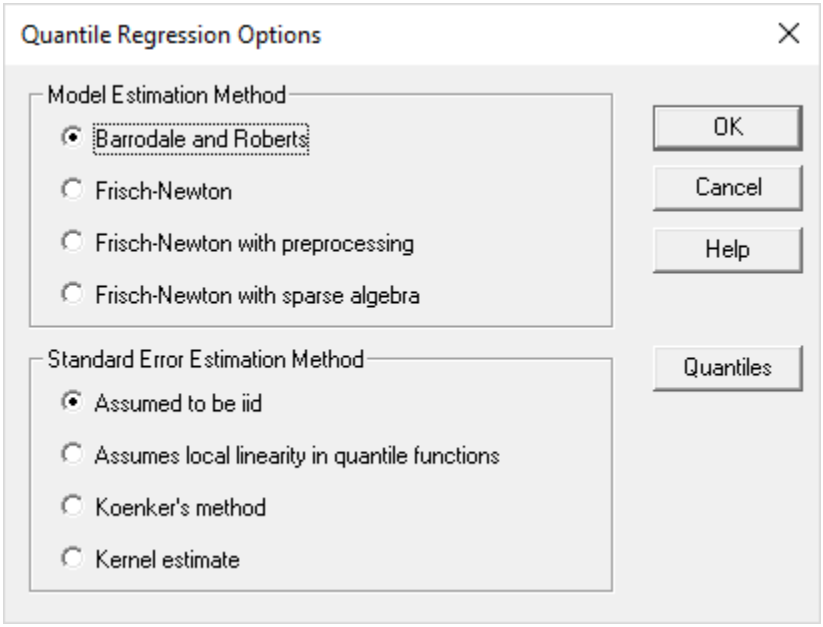
The data input dialog box requests information about the input variables:



- **Dependent Variable**: a numeric variable containing the *n* values of the dependent variable.
- **Categorical Factors**: names of numeric or character variables containing the *n* values of independent variables that should be treated as categorical factors.
- **Quantitative Factors**: names of numeric variables containing the *n* values of independent variables that should be treated as quantitative factors.
- **Weights:** optional weights to be applied to each of the *n* observations.
- **Select**: optional subset selection.

Analysis Options

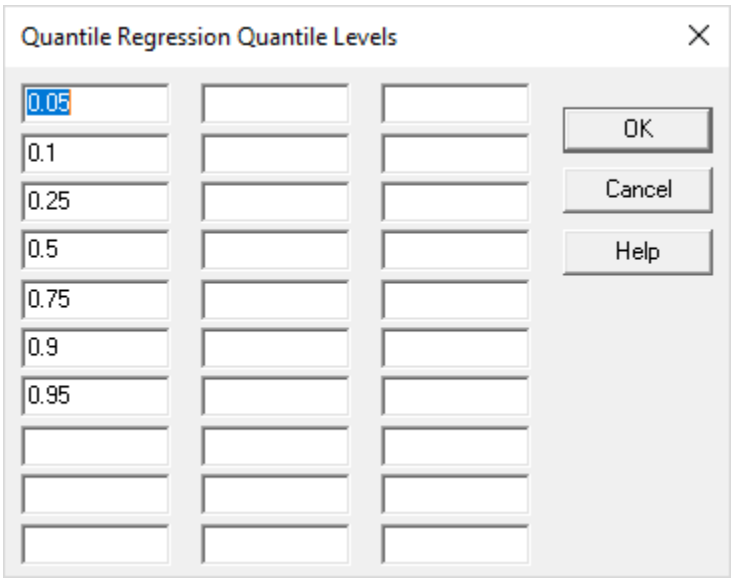
The *Analysis Options* dialog box is used to specify options for fitting the quantile regression model:



- **Model Estimation Method:** the algorithmic method used to calculate the model. The default Barrodale and Roberts method is said to be quite efficient for up to several thousand observations. The other methods may be preferable for very large data sets. For more details, refer to the reference for the R package "quantreg".

- **Standard Error Estimation Method:** method used to estimate the standard errors of the estimated model coefficients. For details, refer to the reference for the R package "quantreg".

- **Quantiles button:** push this button to display a dialog box on which to specify one or more quantiles for which a regression model is desired:



Between 1 and 30 values may be entered. Each value τ must satisfy $0 < \tau < 1$.

Analysis Summary

The *Analysis Summary* displays the output generated by R when the model is fit:

Quantile Regression

```
Sys.setenv("RSTUDIO_PANDOC"="")
d<-read.csv("C:\\Users\\NEIL~1.STA\\AppData\\Local\\Temp\\data.csv",dec=".",sep=",",
stringsAsFactors=FALSE)
setwd("C:\\Users\\NEIL~1.STA\\AppData\\Local\\Temp\\")
library("quantreg")

## Warning: package 'quantreg' was built under R version 3.5.3

## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

fit1=rq(MPG.Highway~as.factor(Drive.Train)+Horsepower+Wheelbase+Weight,data=d,
na.action=na.omit,method="br",tau=c(0.05,0.1,0.25,0.5,0.75,0.9,0.95))

## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique

## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique

summary(fit1,se="iid")

## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique

##
## Call: rq(formula = MPG.Highway ~ as.factor(Drive.Train) + Horsepower +
##     Wheelbase + Weight, tau = c(0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
##     0.95), data = d, na.action = na.omit, method = "br")
##
## tau: [1] 0.05
##
## Coefficients:
##                              Value    Std. Error t value  Pr(>|t|)
## (Intercept)                   9.17474  9.74679    0.94131  0.34931
## as.factor(Drive.Train)front   4.69651  1.35497    3.46614  0.00084
## as.factor(Drive.Train)rear    5.51610  1.61980    3.40541  0.00103
## Horsepower                    0.00355  0.01338    0.26565  0.79117
## Wheelbase                     0.37512  0.13991    2.68109  0.00887
## Weight                       -0.00919  0.00214   -4.30212  0.00005

…
```

The output includes:

- The R statements used to fit the model. The function “rq” does the regression. Note the formula for the model, which indicates that *Drive Train* is a categorical factor.

- The estimated coefficients for each specified quantile. The above table indicates that the fitted model for the 5th percentile is

Q5 = 9.17474 + 4.69651*(DriveTrain=”front”) + 5.51610*(DriveTrain=”rear”)

+ 0.00355*Horsepower + 0.37512*Wheelbase - 0.00919*Weight

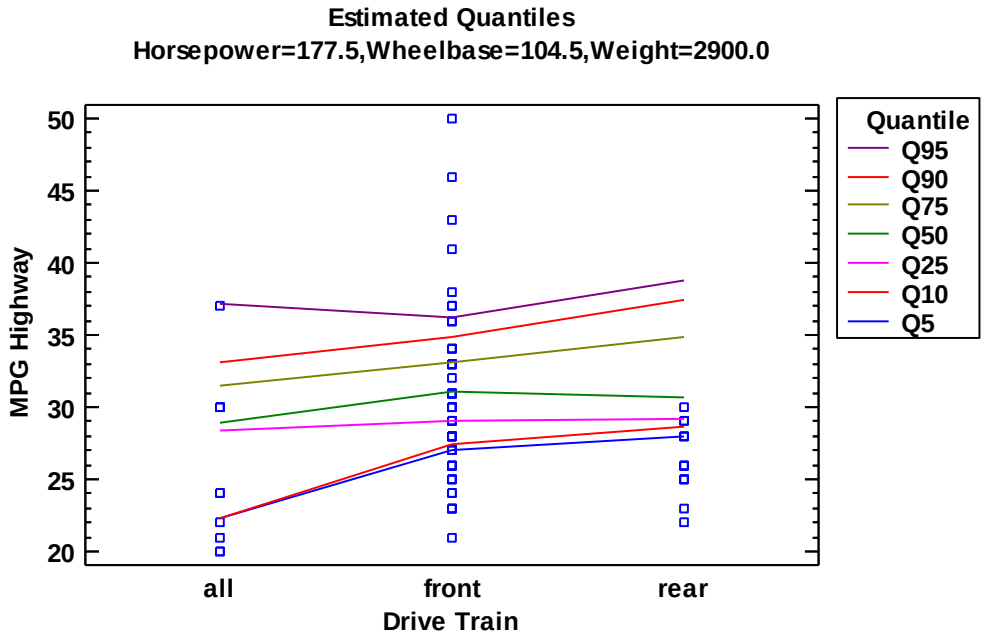
The statement *DriveTrain=”front”* represents an indicator variable that takes the value 1 when the statement is true and 0 when it is false. For a categorical factor with *k* unique levels, there will be *k-1* indicator variables.

- *Pr(>|t|)* – P values that test the statistical significance of each model coefficient. Values less than 0.05 indicate coefficients that are significantly different than 0 at the 5% significance level. In the above output, all coefficients are statistically significant except the one for *Horsepower* and the one corresponding to the intercept.

A separate table is displayed for each quantile specified on the *Analysis Options – Quantiles* dialog box.

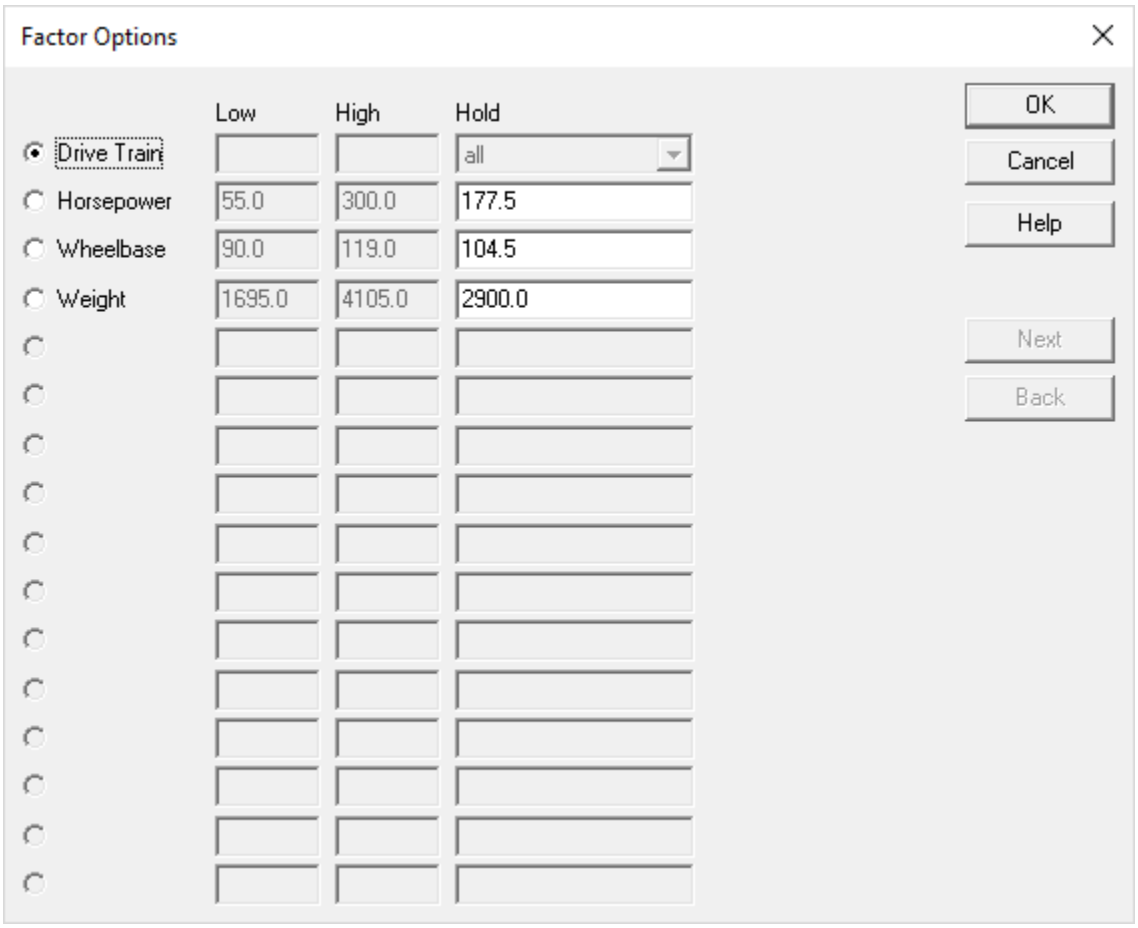
Quantile Plot

This plot shows the fitted regression models for each specified quantile:



One factor is varied along the horizontal axis. The other factors are fixed at values specified on the *Pane Options* dialog box.

Pane Options



- **Radio buttons**: Select one factor to plot along the horizontal axis.
- **Low:** for a selected quantitative factor, the low end of the range for varying the factor.
- **High:** for a selected quantitative factor, the high end of the range for varying the factor.
- **Hold**: for factors not selected, the value at which the factor should be held constant.

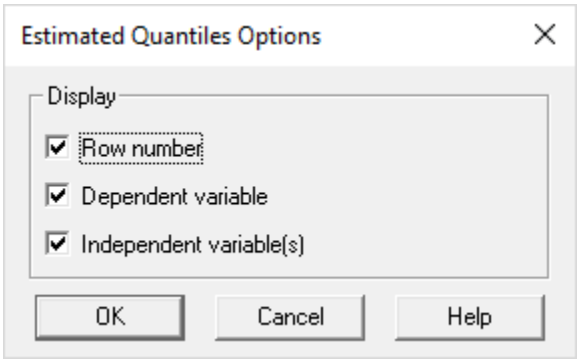
Estimated Quantiles

This table displays the estimated quantiles corresponding to each row of the datasheet that has no missing data for the dependent variable or predictive factors.

Estimated Quantiles

Row	MPG Highway	Drive Train	Horsepower	Wheelbase	Weight	Q5	Q10	Q25	Q50	Q75	Q90	Q95
1	31.0	front	140.0	102.0	2705.0	27.8193	28.6034	30.0455	32.0801	34.2925	36.3458	37.7247
2	25.0	front	200.0	115.0	3560.0	25.0	25.0	25.3687	27.7516	28.2509	28.602	30.0737
3		front	172.0	102.0	3375.0							
4	26.0	front		106.0	3405.0							
5	30.0	rear	208.0	109.0	3640.0	22.9214	23.0405	23.9107	25.1848	28.5187	29.9612	30.0
6	31.0	front	110.0	105.0	2880.0	27.2586	28.2222	28.8553	31.0	32.7113	34.268	35.6796
7	28.0	front	170.0	111.0	3470.0	24.2125	24.4721	25.2548	27.4046	28.3359	28.9233	29.8409
8	25.0	rear	180.0	116.0	4105.0	21.1959	21.3735	20.9392	22.4155	24.6681	25.0	25.0
9	27.0	front	170.0	108.0	3495.0	22.8153	23.0501	24.5361	26.4353	27.7143	28.4145	28.6714
10	25.0	front	200.0	114.0	3620.0	24.0496	24.0192	24.6892	26.9499	27.5458	27.8593	28.9582
...	...	...	...	...	...	...	...	...	...	...	...	...

Pane Options



- **Display:** Select the information to display in the table.

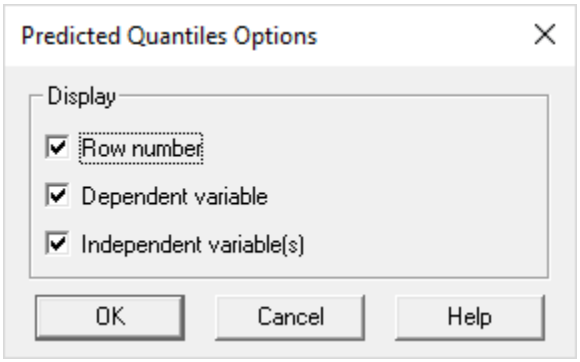
Predicted Quantiles

This table displays predicted quantiles corresponding to each row of the datasheet for which the value of the dependent variable is missing (or has not been selected by the *Select* field on the data input dialog box) but has no missing values for the predictive factors.

Predicted Quantiles

Row	MPG Highway	Drive Train	Horsepower	Wheelbase	Weight	Q5	Q10	Q25	Q50	Q75	Q90	Q95
3		front	172.0	102.0	3375.0	21.6212	21.8629	24.56	26.0844	28.128	29.307	28.7471

Pane Options



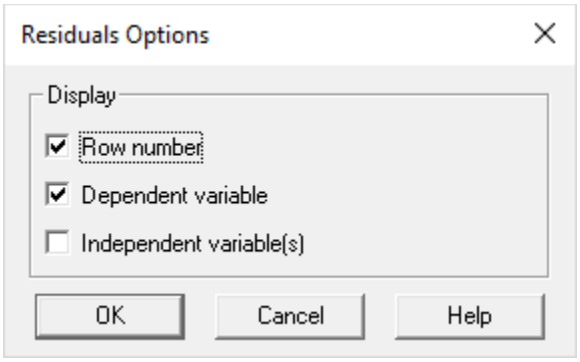
- **Display:** Select the information to display in the table.

Residuals

This table displays the residuals from the fitted quantile regression models:

Row	MPG Highway	Q5	Q10	Q25	Q50	Q75	Q90	Q95
1	31.0	3.18071	2.39659	0.954475	-1.08006	-3.2925	-5.34578	-6.7247
2	25.0	-3.55271E-15	7.10543E-15	-0.368682	-2.75158	-3.25095	-3.60195	-5.07365
3								
4	26.0							
5	30.0	7.07859	6.95945	6.08932	4.81518	1.48135	0.0388284	3.55271E-15
6	31.0	3.74138	2.7778	2.14467	1.42109E-14	-1.71129	-3.268	-4.67958
7	28.0	3.78754	3.5279	2.74522	0.59539	-0.335863	-0.923286	-1.84093
8	25.0	3.80409	3.62654	4.06075	2.5845	0.331933	7.10543E-15	0.0
9	27.0	4.18469	3.94991	2.46389	0.564683	-0.714273	-1.41445	-1.67143
10	25.0	0.950395	0.98076	0.310818	-1.94993	-2.54575	-2.85926	-3.95815
...	...	...	...	...	...	...	...	...

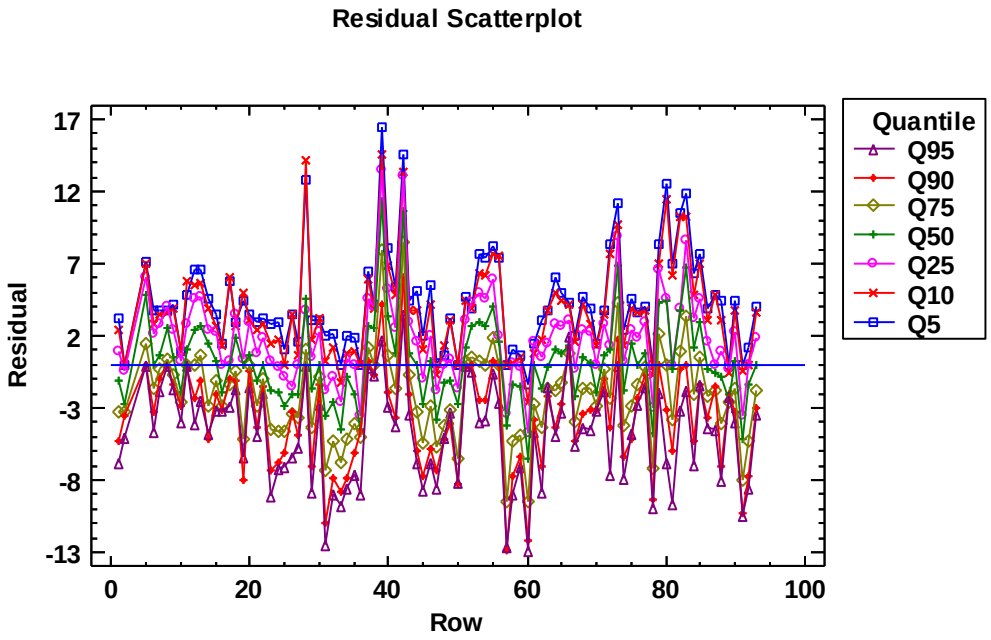
Pane Options



- **Display:** Select the information to display in the table.

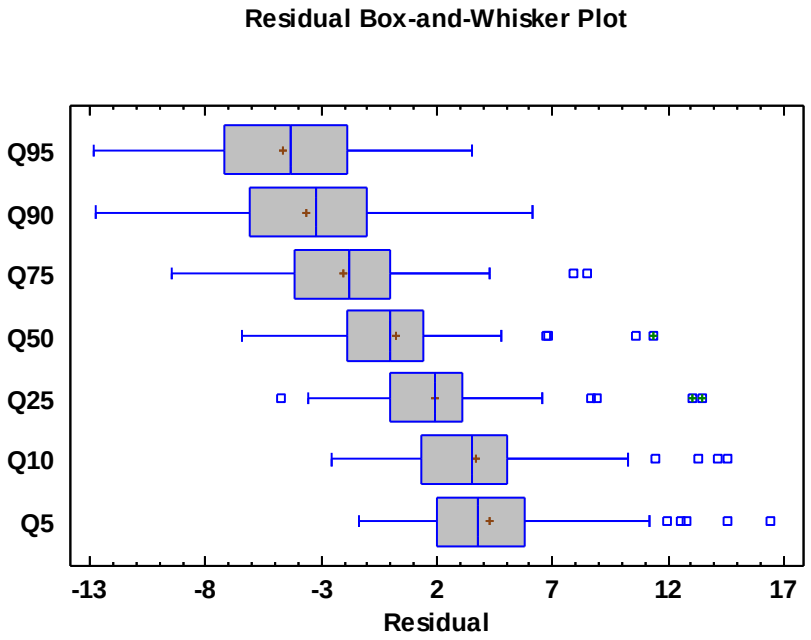
Residual Scatterplot

This graph displays the residuals from the fitted quantile regression models, plotted versus row number in the datasheet.



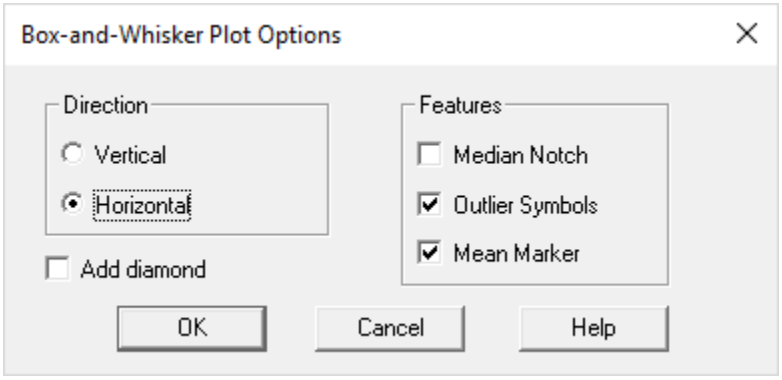
Residual Box-and-Whisker Plot

This graph displays a box-and-whisker plot for the residuals from each quantile regression model.



The boxes cover the center 50% of the residuals from each model. The vertical lines within the boxes indicate the location of the medians, while the plus signs show the location of the means. The whiskers extend out to the largest and smallest residuals, unless residuals are far enough from the box to be declared to be "outside points". For more information on how this plot is constructed, refer to the PDF document titled *Box-and-Whisker Plot*.

Pane Options



- **Direction**: the orientation of the plot, corresponding to the direction of the whiskers.

- **Median Notch**: if selected, a notch will be added to the plot showing an approximate 100(1-α)% confidence interval for the median at the default system confidence level (set on the *General* tab of the *Preferences* dialog box on the *Edit* menu).

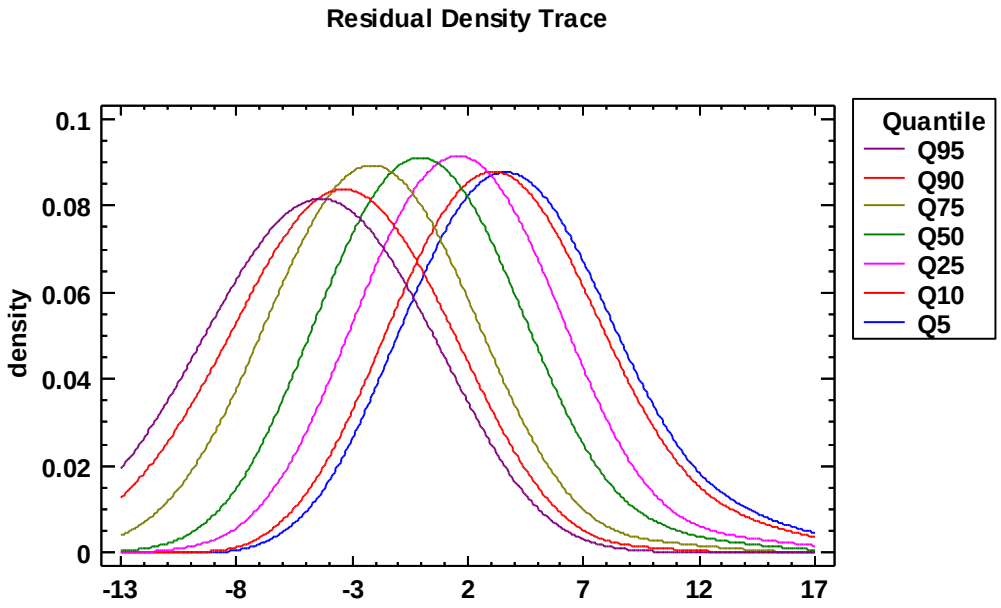
- **Outlier Symbols**: if selected, indicates the location of outside points.

- **Mean Marker**: if selected, shows the location of the sample mean as well as the median.

- **Add diamond:** if selected, adds a diamond to the plot showing a 100(1-α)% confidence interval for the mean at the default system confidence level.

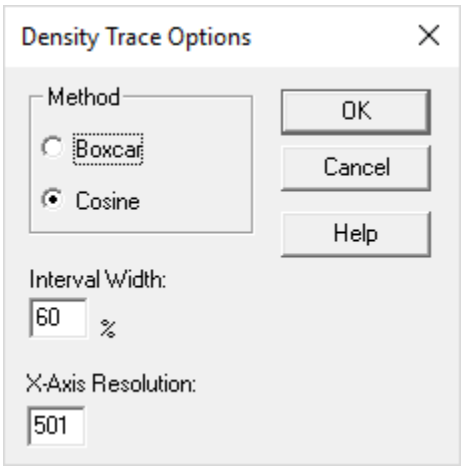
Residual Density Trace

This graph displays a nonparametric density trace for the residuals from each quantile regression model.



The *Density Trace* provides a nonparametric estimate of the probability density function of the population from which the residuals were sampled. It is created by counting the number of observations which fall within a window of fixed width moved across the range of the data. For more information on how this plot is constructed, refer to the PDF document titled *One Variable Analysis*.

Pane Options

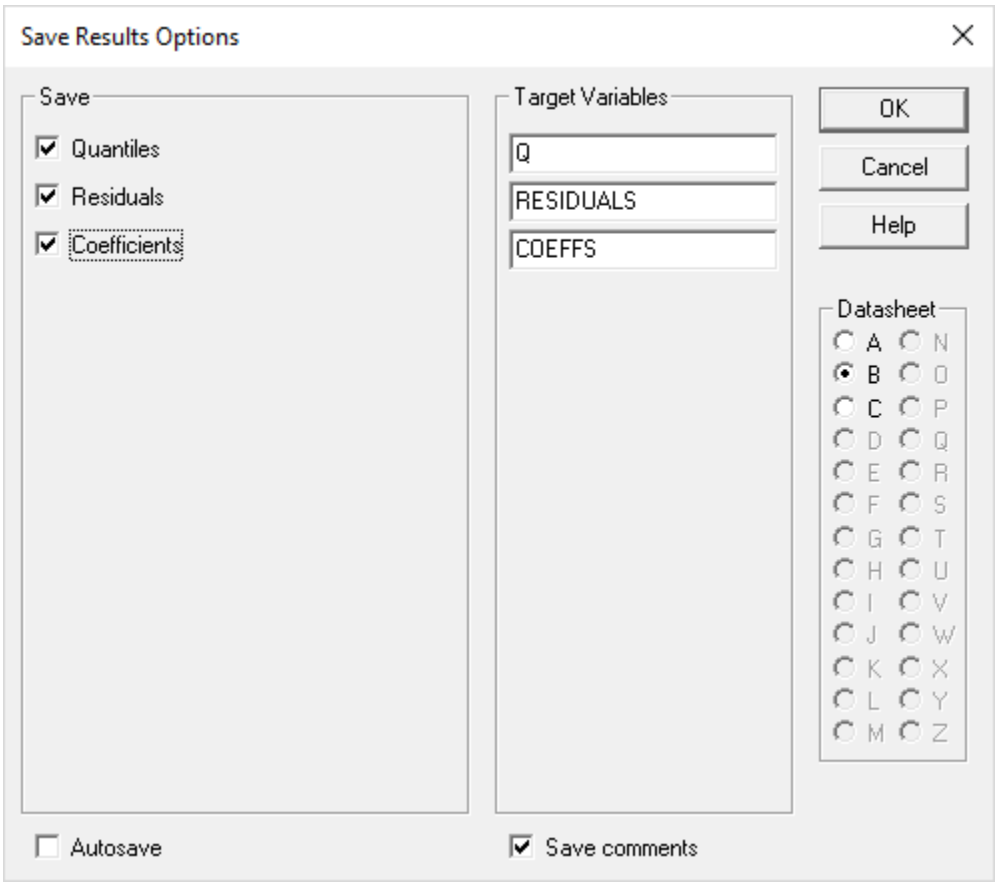


- **Method:** the desired weighting function. The boxcar function weights all values within the window equally. The cosine function gives decreasing weight to observations further from the center of the window. The default selection is determined by the setting on the *EDA* tab of the *Preferences* dialog box accessible from the *Edit* menu.

- **Interval Width:** the width of the window *h* within which observations affect the estimated density, as a percentage of the range covered by the x-axis. *h* = 60% is not unreasonable for a small sample but may not give as much detail as a smaller value in larger samples.

- **X-Axis Resolution**: the number of points at which the density is estimated.

Save Results



The following results may be saved to the datasheet:

1. *Quantiles* – the estimated and predicted quantiles for each row in the datasheet.
2. *Residuals* – the residuals from the fitted quantile regression models.
3. *Coefficients* – the estimated model coefficients.

References

R Package "quantreg" (2019) - https://cran.r-project.org/web/packages/quantreg/quantreg.pdf

Koenker, Roger – Quantile Regression in R – A Vignette. https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf