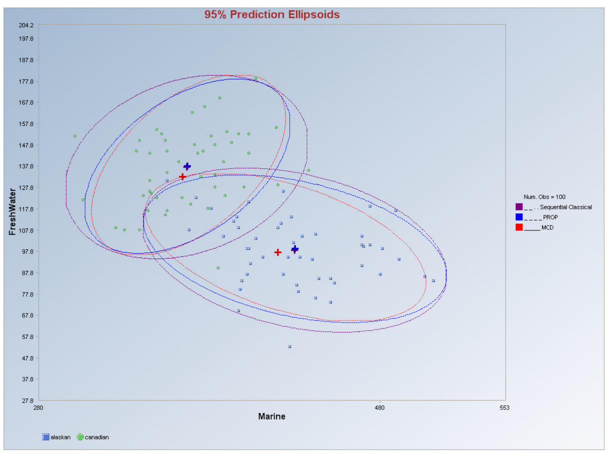
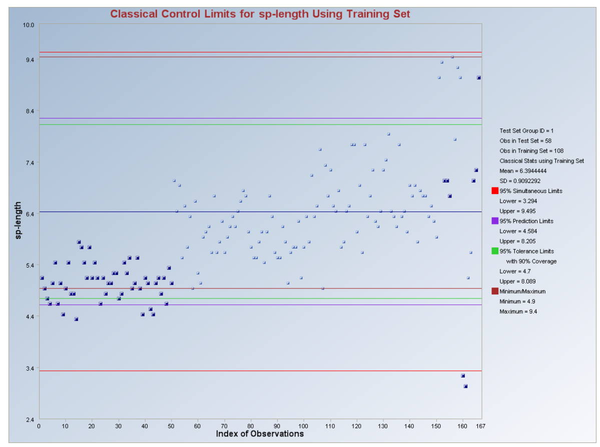
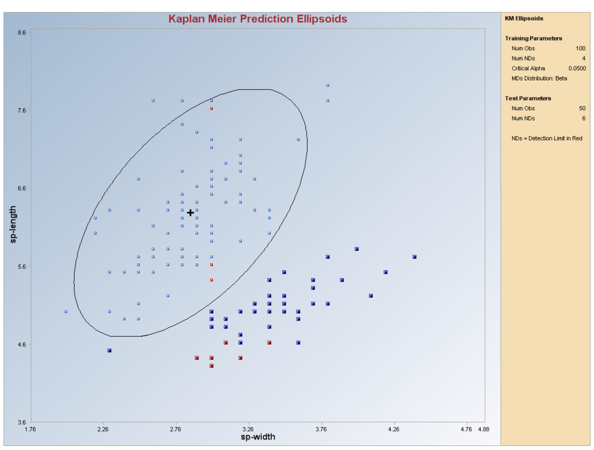
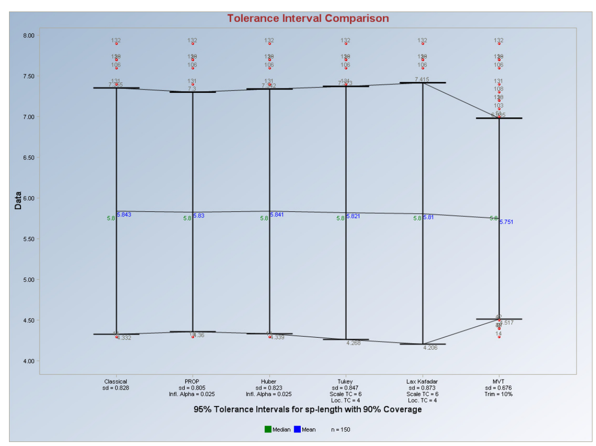
United States
Environmental Protection
Agency

Scout 2008 Version 1.0 User Guide Part II









RESEARCH AND DEVELOPMENT

EPA/600/R-08/038
February 2009
www.epa.gov

Scout 2008 Version 1.0 User Guide

(Second Edition, December 2008)

John Nocerino

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Environmental Sciences Division
Technology Support Center
Characterization and Monitoring Branch
944 E. Harmon Ave.
Las Vegas, NV 89119

Anita Singh, Ph.D.[1]
Robert Maichle[1]
Narain Armbya[1]
Ashok K. Singh, Ph.D.[2]

[1]Lockheed Martin Environmental Services
1050 E. Flamingo Road, Suite N240
Las Vegas, NV 89119

[2]Department of Hotel Management
University of Nevada, Las Vegas
Las Vegas, NV 89154

Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy. Mention of trade names and commercial products does not constitute endorsement or recommendation for use.

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC 20460

7663cmb09

Notice

The United States Environmental Protection Agency (EPA) through its Office of Research and Development (ORD) funded and managed the research described here. It has been peer reviewed by the EPA and approved for publication. Mention of trade names and commercial products does not constitute endorsement or recommendation by the EPA for use.

The Scout 2008 software was developed by Lockheed-Martin under a contract with the USEPA. Use of any portion of Scout 2008 that does not comply with the Scout 2008 User Guide is not recommended.

Scout 2008 contains embedded licensed software. Any modification of the Scout 2008 source code may violate the embedded licensed software agreements and is expressly forbidden.

The Scout 2008 software provided by the USEPA was scanned with McAfee VirusScan and is certified free of viruses.

With respect to the Scout 2008 distributed software and documentation, neither the USEPA, nor any of their employees, assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed. Furthermore, the Scout 2008 software and documentation are supplied "as-is" without guarantee or warranty, expressed or implied, including without limitation, any warranty of merchantability or fitness for a specific purpose.

Acronyms and Abbreviations

% NDs	Percentage of Non-detect observations
ACL	alternative concentration limit
A-D, AD	Anderson-Darling test
AM	arithmetic mean
ANOVA	Analysis of Variance
AOC	area(s) of concern
B*	Between groups matrix
BC	Box-Cox-type transformation
BCA	bias-corrected accelerated bootstrap method
BD	break down point
BDL	below detection limit
BTV	background threshold value
BW	Black and White (for printing)
CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
CL	compliance limit, confidence limits, control limits
CLT	central limit theorem
CMLE	Cohen's maximum likelihood estimate
COPC	contaminant(s) of potential concern
CV	Coefficient of Variation, cross validation
D-D	distance-distance
DA	discriminant analysis
DL	detection limit
DL/2 (t)	UCL based upon DL/2 method using Student's t-distribution cutoff value
DL/2 Estimates	estimates based upon data set with non-detects replaced by half of the respective detection limits
DQO	data quality objective
DS	discriminant scores
EA	exposure area
EDF	empirical distribution function
EM	expectation maximization
EPA	Environmental Protection Agency
EPC	exposure point concentration
FP-ROS (Land)	UCL based upon fully parametric ROS method using Land's H-statistic

Gamma ROS (Approx.)	UCL based upon Gamma ROS method using the bias-corrected accelerated bootstrap method
Gamma ROS (BCA)	UCL based upon Gamma ROS method using the gamma approximate-UCL method
GOF, G.O.F.	goodness-of-fit
H-UCL	UCL based upon Land's H-statistic
HBK	Hawkins Bradu Kaas
HUBER	Huber estimation method
ID	identification code
IQR	interquartile range
K	Next K, Other K, Future K
KG	Kettenring Gnanadesikan
KM (%)	UCL based upon Kaplan-Meier estimates using the percentile bootstrap method
KM (Chebyshev)	UCL based upon Kaplan-Meier estimates using the Chebyshev inequality
KM (t)	UCL based upon Kaplan-Meier estimates using the Student's t-distribution cutoff value
KM (z)	UCL based upon Kaplan-Meier estimates using standard normal distribution cutoff value
K-M, KM	Kaplan-Meier
K-S, KS	Kolmogorov-Smirnov
LMS	least median squares
LN	lognormal distribution
Log-ROS Estimates	estimates based upon data set with extrapolated non-detect values obtained using robust ROS method
LPS	least percentile squares
MAD	Median Absolute Deviation
Maximum	Maximum value
MC	minimization criterion
MCD	minimum covariance determinant
MCL	maximum concentration limit
MD	Mahalanobis distance
Mean	classical average value
Median	Median value
Minimum	Minimum value
MLE	maximum likelihood estimate
MLE (t)	UCL based upon maximum likelihood estimates using Student's t-distribution cutoff value

MLE (Tiku)	UCL based upon maximum likelihood estimates using the Tiku's method
Multi Q-Q	multiple quantile-quantile plot
MVT	multivariate trimming
MVUE	minimum variance unbiased estimate
ND	non-detect or non-detects
NERL	National Exposure Research Laboratory
NumNDs	Number of Non-detects
NumObs	Number of Observations
OKG	Orthogonalized Kettenring Gnanadesikan
OLS	ordinary least squares
ORD	Office of Research and Development
PCA	principal component analysis
PCs	principal components
PCS	principal component scores
PLs	prediction limits
PRG	preliminary remediation goals
PROP	proposed estimation method
Q-Q	quantile-quantile
RBC	risk-based cleanup
RCRA	Resource Conservation and Recovery Act
ROS	regression on order statistics
RU	remediation unit
S	substantial difference
SD, *Sd, sd*	standard deviation
SLs	simultaneous limits
SSL	soil screening levels
S-W, SW	Shapiro-Wilk
TLs	tolerance limits
UCL	upper confidence limit
UCL95, 95% UCL	95% upper confidence limit
UPL	upper prediction limit
UPL95, 95% UPL	95% upper prediction limit
USEPA	United States Environmental Protection Agency
UTL	upper tolerance limit
Variance	classical variance
W*	Within groups matrix

WiB matrix	Inverse of W* cross-product B* matrix
WMW	Wilcoxon-Mann-Whitney
WRS	Wilcoxon Rank Sum
WSR	Wilcoxon Signed Rank
Wsum	Sum of weights
Wsum2	Sum of squared weights

Table of Contents

Chapter 7

Outliers and Estimates

Outliers are inevitable in data sets originating from various applications. There are many graphical (Q-Q plots, Box plots), classical (Dixon, Rosner, Welch, Max MD), sequential classical (Max MD, Kurtosis), and robust estimation and outlier identification methods (Biweight, Huber, MCD, MVE, MVT, OKG, PROP) available in the literature. Classical outlier tests suffer from masking (e.g., extreme outliers may mask intermediate outliers) effects. The use of robust outlier identification procedures is recommended to identify multiple outliers, especially when dealing with multivariate (having multiple contaminants) data sets. Several univariate and multivariate (both classical and robust outlier identification methods (e.g., based upon Biweight, Huber, and PROP influence functions)) are available in this Scout software package.

7.1 Univariate Outliers and Estimates

For historical reasons and also for the sake of comparison, some simple classical outlier tests are also included in the Scout software package. Specifically, the classical outlier tests (often cited in environmental literature), Dixon and Rosner, are available in Scout. For details, refer to ProUCL 4.00.04 Technical Guide. Those classical tests may be used on data sets with and without non-detect observations. For data sets with non-detects, two options are available in Scout to deal with data sets with outliers: 1) exclude non-detects, and 2) replace NDs by DL/2 values. Those options are used only to identify outliers and not to compute any estimates and limits used in the decision-making process.

It is suggested that the classical (and also the robust procedures to be described later) outlier identification procedures be supplemented with graphical displays such as Q-Q plots, box-and-whisker plots (also called box plots), and interquartile range (IQR) plots (upper quartile, Q3, and lower quartile, Q1). Those graphical displays are available in Scout. Box plots with whiskers are sometimes used to identify univariate outliers (e.g., EPA 2006). Typically, a box plot gives a good indication of extreme (outliers) observations that may present in a data set. The statistics (lower quartile, median, upper quartile, and IQR) used in the construction of a box plot do not get distorted by outliers. On a box plot, observations beyond the two whiskers may be considered to be candidates for potential outliers.

On a normal Q-Q plot, observations that are well separated from the bulk (central part) of the data typically represent potential outliers needing further investigation. Moreover, significant and obvious jumps and breaks in a Q-Q plot (for any distribution) are indications of the presence of more than one population. Data sets exhibiting such behavior on Q-Q plots should be partitioned out into component sub-populations before estimating various statistics of interest (e.g., prediction intervals, confidence intervals).

Dixon's Test (Extreme Value Test).

- Used to identify statistical outliers when the sample size is less than or equal to 25.

- Used to identify outliers or extreme values in both the left tail (Case 1) and the right tail (Case 2) of a data distribution. In environmental data sets, extremes found in the right tail may represent potentially contaminated site areas needing further investigation or remediation. The extremes in the left tail may represent ND values.

- Assumes that the data without the suspected outliers are normally distributed; therefore, it is necessary to perform a test for normality on the data without the suspected outliers before applying this test.

- May suffer from masking in the presence of multiple outliers. This means that if more than one outlier is suspected, this test may fail to identify all of the outliers. Therefore, if you decide to use the Dixon's test for multiple outliers, apply the test to the least extreme value first.

Rosner's Test.

- Can be used to identify and detect up to 10 outliers in data sets of sizes 25 and higher.

- Assumes that the data are normally distributed; therefore, it is necessary to perform a test for normality before applying this test.

Depending upon the selected variables and the number of observations associated with them, either the Dixon's Test or the Rosner's Test will be performed.

Biweight Estimates.

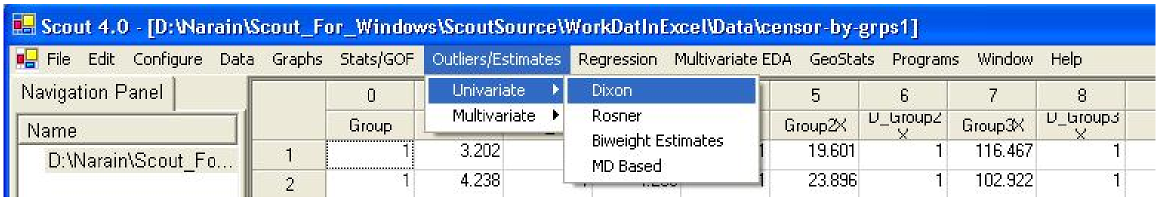
- Based on the estimation methods of Mosteller and Tukey (1977), Kafadar (1981) and LAX (1985).

MD-based test (Grubb's Test)

- This is the multivariate extension of the univariate test known as the Grubbs test. It is based on the assumption of normality. The generalized distances of the multivariate data are calculated and the observation with the distance greater than the critical value is expunged from the data set. The test is iterated until no outliers are detected.

7.1.1 Dixon Test for Univariate Data

1. Click **Outliers/Estimates ► Univariate ►Dixon.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.

- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- The default the number for suspected outliers is 2. In order to use this test, the user has to obtain an initial guess about the number of outliers that may be present in the data set. This can be done by using graphical displays such as a Q-Q plot. On this graphical Q-Q plot, higher observations that are well separated from the rest of the data may be considered to be potential or suspected outliers.

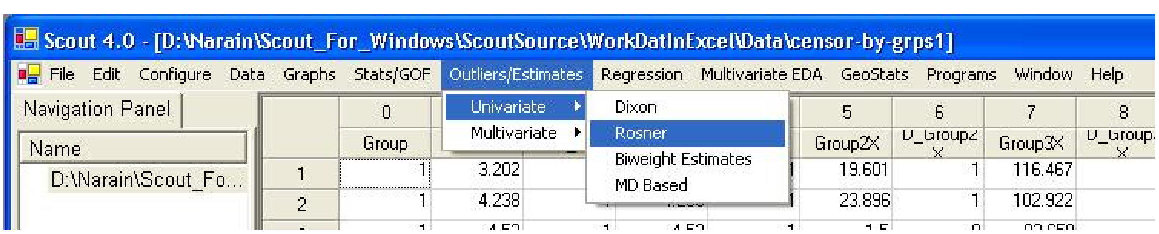
- Click on the "**OK**" button to continue or on the "**Cance**l" button to cancel the Outliers tests.

Output for Dixon Test.

	Dixon Outlier Test for Selected Variables
User Selected Options	
Date/Time of Computation	7/10/2007 3:53:44 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Test for Suspected Outliers using Dixon Test	1
Dixon's Outlier Test for Group2X	
Number of data = 20	
10% critical value: 0.401	
5% critical value: 0.45	
1% critical value: 0.535	
1. 37.867 is a Potential Outlier (Upper Tail)	
Test Statistic: 0.383	
For 10% significance level, 37.867 is not an outlier.	
For 5% significance level, 37.867 is not an outlier.	
For 1% significance level, 37.867 is not an outlier.	
2. 1.5 is a Potential Outlier (Lower Tail)	
Test Statistic: 0.198	
For 10% significance level, 1.5 is not an outlier.	
For 5% significance level, 1.5 is not an outlier.	
For 1% significance level, 1.5 is not an outlier.	

7.1.2 Rosner's Test for Univariate Data

1. Click **Outliers/Estimates ► Univariate ►Rosner.**

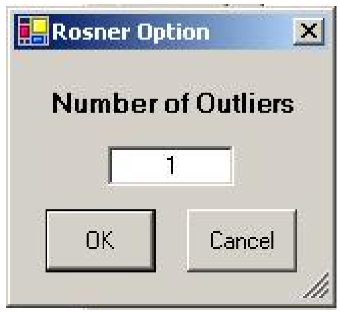


2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.

- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- Click on the "**Options**" button.



- The default number for suspected outliers is "**1**." In order to use this test, the user has to obtain an initial guess about the number of outliers that may be present in the data set. This can be done by using graphical displays such as a Q-Q plot. On this graphical Q-Q plot, higher observations that are well separated from the rest of the data may be considered to be potential or suspected outliers.

- Click "**OK**" to continue or "**Cancel**" to cancel the Outliers tests.

Output for Rosner Test.

	Rosner Outlier Test for Selected Variables
User Selected Options	
Date/Time of Computation	7/10/2007 3:56:51 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Test for 'N' Suspected Outliers using Rosner	1

Rosner's Outlier Test for X

Number of data: 53

Number of suspected outliers: 1

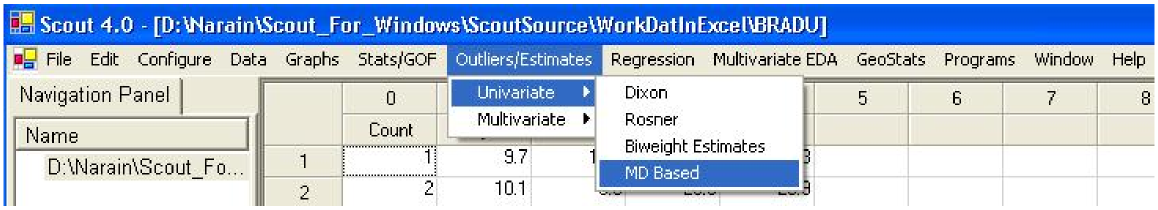
#	Mean	sd	Potential outlier	Test value	Critical value (5%)	Critical value (1%)
1	51.10	43.37	121.11	1.61	3.151	3.504

5% Significance Level, there is no Potential Outlier

For 1% Significance Level, there is no Potential Outlier

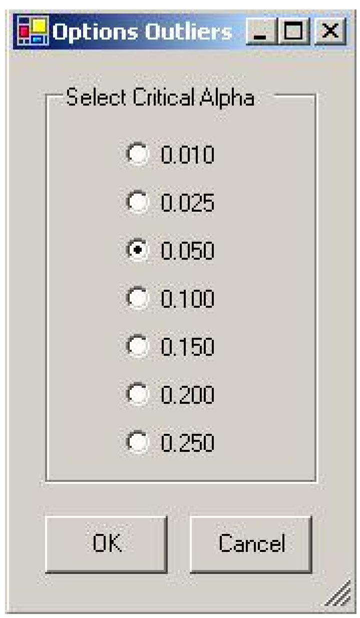
7.1.3 MD-Based (Grubbs Test) Test for Univariate Data

1. Click **Outliers/Estimates ► Univariate ►MD-Based.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**."



- Click "**OK**" to continue or "**Cancel**" to cancel the Outliers tests.

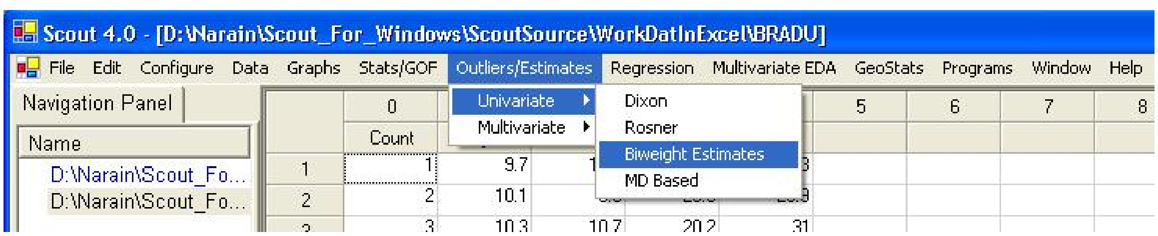
Output example: The data set "BRADU.xls" was used for the univariate MD-based (Grubbs) test. The theoretical maximum MD at the selected critical alpha is calculated and compared to the maximum MD obtained from the data set.

Output for MD-Based Grubbs Test.

	Univariate MD Based (Grubbs Test) Outlier Analysis
User Selected Options	
Date/Time of Computation	1/7/2008 4:10:24 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
x1	
No Outliers Present	
Initial Conditions	
Max MD	Max MD(0.05)
5.796	10.78

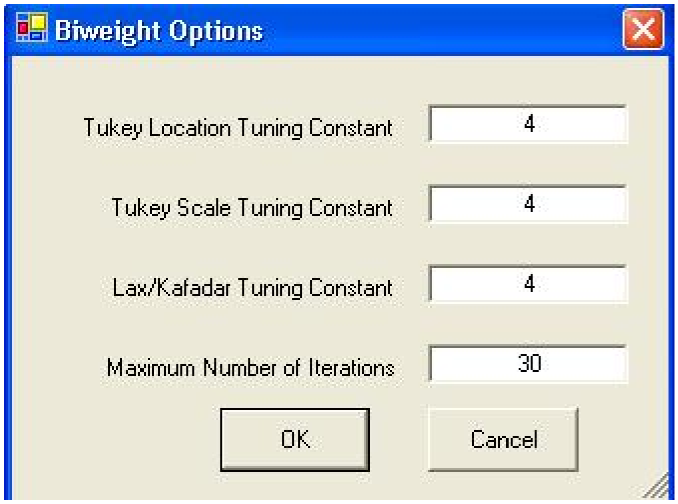
7.1.4 Biweight Estimate for Univariate Data

1. Click **Outliers/Estimates ► Univariate ►Biweight.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**."



- Click "**OK**" continue or "**Cancel**" to cancel the Outliers tests.

Output example: The estimates of location and scale were computed using the Tukey's bisquare function and the Kafadar biweight function.

Output for Biweight Estimates.

	Univariate Biweight Outlier Analysis
User Selected Options	
Date/Time of Computation	1/7/2008 4:14:02 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Number of Iterations	30
Tukey Location Tuning Constant	4
Tukey Scale Tuning Constant	4
Kafadar Scale Tuning Constant	4

Robust Biweight Estimates

				Tukey				Kafadar	
		Classical		One-Step		Final		Final	
Variable	Obs No.	Location	Scale	Location	Scale	Location	Scale	Location	Scale
x1	75	3.207	3.653	1.56	1.377	1.513	1.313	1.621	2.709

7.2 Robust Estimation and Identification of Multiple Multivariate Outliers

A myriad of classical and robust outlier identification procedures are available in the literature. Most of procedures covering about the last three decades of research in the area of robust estimation and outlier identification methods have been incorporated in Scout. For the sake of comparison and completeness, some classical methods are also available in Scout. A list of articles covering some of those research procedures is provided in the references.

Several formal graphical method comparison tools have been incorporated in Scout. Specifically, in both the outlier module and the Regression module, the user can pick several methods and Scout will produce graphical comparison displays of those methods. Some examples illustrating those methods are included in the User Guide. Several benchmark data sets from the literature have been used throughout this user guide.

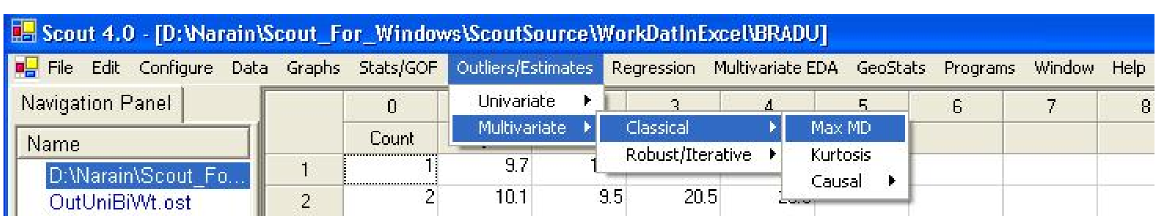
7.2.1 Classical Outlier Testing

Due to historical importance (Wilk (1963)) and for the sake of completeness, classical outlier methods have also been incorporated in Scout. The Classical outlier module offers two tests for discordances: multivariate kurtosis and Mahalanobis distances (sometimes called generalized distances). Multivariate kurtosis is also useful as a test for deviation from normality in one or more dimensions. Both of those tests assume that the data represent a random sample from a multivariate (p-dimensional, $p \geq 1$) normal population.

7.2.1.1 Mahalanobis' Distances

The classical Mardia's multivariate kurtosis (Mardia 1970, 1974, and Schwager and Margolin 1982) outlier (and multinormality) test and the MD test (Ferguson 1961a, 1961b and Barnett and Lewis 1994) have been incorporated into Scout. The generalized distance (MD-based) test is a multivariate extension of a univariate Grubb's test (Grubbs 1950). Scout also computes robustified multivariate kurtosis, skewness, and the largest MD. As can be seen below, outliers have a huge influence (impact) on those statistics.

1. Click **Outliers/Estimates ► Multivariate ► Classical ► Max MDs.**

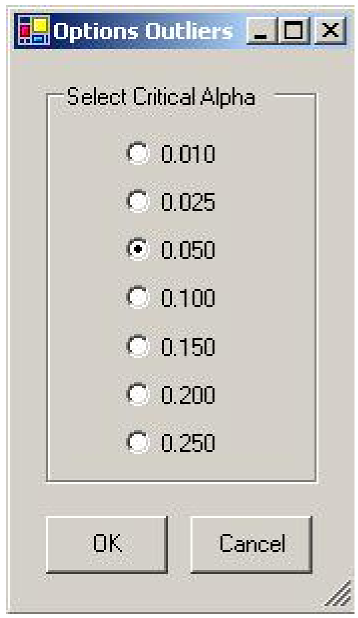


2. The "**Select Variables**" screen (Section 3.4) will appear.

- Select two or more variables from the "**Select Variables**" screen.

- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- Click on "**Options**."



- Select the required "**Critical Alpha**" and click "**OK**" to continue or "**Cancel**" to cancel the Outliers tests.

Output for Max MDs test for outliers.
Data Set used: Bradu (From Hawkins Kaas, and Bradu, 1984 article).

	Classical Sequential Outlier Test Based Upon Maximum Mahalanobis Distance (Max-MDs)		
User Selected Options			
Date/Time of Computation	3/4/2008 8:42:40 AM		
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU		
Full Precision	OFF		
Number of Observations	75		
Number of Variables	4		
Classical Mean Vector			
y	x1	x2	x3
1.279	3.207	5.597	7.231
Classical Standard Deviation Vector			
y	x1	x2	x3
3.493	3.653	8.239	11.74
Classical Covariance S Matrix			
y	x1	x2	x3
12.2	9.477	20.39	31.03
9.477	13.34	28.47	41.24
20.39	28.47	67.88	94.67
31.03	41.24	94.67	137.8
		Determinant	1906
		Log of Determinant	7.553
Eigenvalues of Classical Covariance S Matrix			
Eval 1	Eval 2	Eval 3	Eval 4
0.914	1.688	5.538	223.1
Critical Alpha	0.05		
Outlier Summary			
Obs No.	Max MD	Max MD (0.(	
14	43.7	17.43	
12	27.75	17.39	
11	34.26	17.34	
13	56.98	17.29	

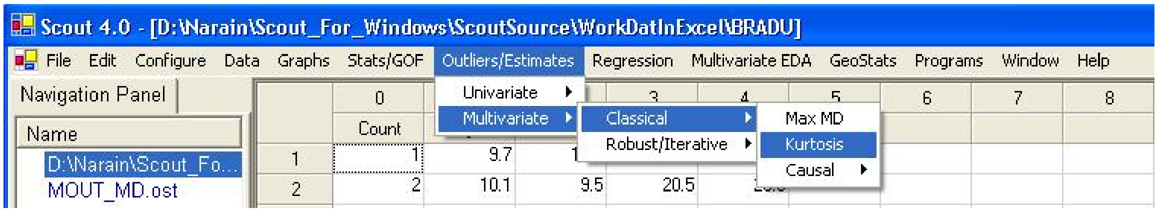
Result: Observations 11, 12, 13, and 14 were identified as outliers. The classical method with a classical start could not identify the first 10 observations as outliers.

7.2.1.2 Multivariate Kurtosis

Mardia's multivariate kurtosis is an extension of the univariate kurtosis and, thus, may also be used as a univariate outlier test. Multivariate kurtosis is also used to test multivariate normality.

Those tests, as incorporated in Scout, are sequential (repeated, without using any previously identified discordant values). The process stops when no further outliers are found. For the multivariate kurtosis test, the discordant observation identified is that point which has the largest generalized distance from the sample classical mean vector.

1. Click **Outliers/Estimates ► Multivariate ► Classical ► Kurtosis.**

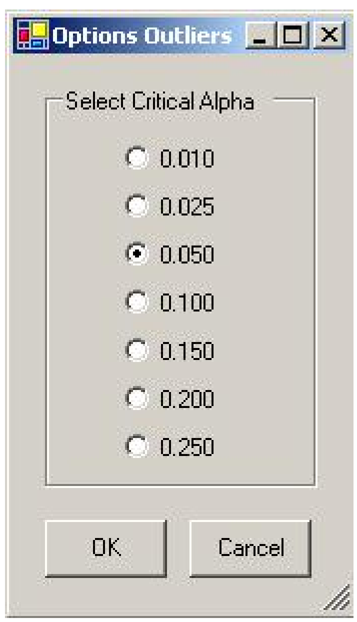


2. The "**Select Variables**" screen (Section 3.4) will appear.

- Select two or more variables from the "**Select Variables**" screen.

- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- Click on "**Options**."



- Select the required "**Critical Alpha**" and click on "**OK**" to continue or "**Cancel**" to cancel the Outliers tests.

Output for Kurtosis test for outliers.
Data Set used: Bradu.

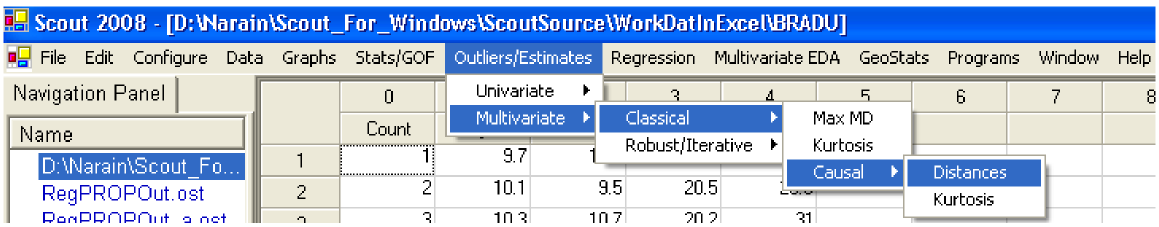
			Classical Sequential Outlier Test Based Upon Multivariate Kurtosis	
		User Selected Options		
		Date/Time of Computation	3/4/2008 8:44:50 AM	
		From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU	
		Full Precision	OFF	
		Number of Observations	75	
		Number of Variables	4	
		Classical Mean Vector		
y	x1	x2	x3	
1.279	3.207	5.597	7.231	
		Classical Standard Deviation Vector		
y	x1	x2	x3	
3.493	3.653	8.239	11.74	
		Classical Covariance S Matrix		
y	x1	x2	x3	
12.2	9.477	20.39	31.03	
9.477	13.34	28.47	41.24	
20.39	28.47	67.88	94.67	
31.03	41.24	94.67	137.8	
		Determinant	1906	
		Log of Determinant	7.553	
		Eigenvalues of Classical Covariance S Matrix		
Eval 1	Eval 2	Eval 3	Eval 4	
0.914	1.688	5.538	223.1	
		Critical Alpha	0.05	
Outlier Summary				
Obs No.	Kurtosis		Kurtosis (0.05)	
14	53.97		25.2	
12	38.26		25.16	
11	43.57		25.1	
13	59.16		25.19	

Result: Once again, only the observations 11, 12, 13, and 14 were identified as outliers. The other 10 outliers could not be identified due to masking effects.

7.2.1.3 Identifying Causal Variables

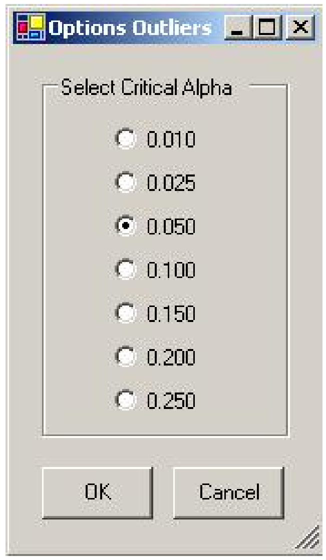
Once an outlier test has been performed, the user may wish to identify the variables (if any) which are responsible for each discordant observation. This can be done by selecting the "Causal Variables" option from the pull-down menu. However, there are several other methods (e.g., Q-Q plot of individual variables, bivariate scatter plot with tolerance ellipsoid) available in Scout that can also be used to identify variables that might cause an observation to be an outlier. The details of this classical method can be found in (Garner, et al. (1991a and 1991b). This method retests each discordant observation with one variable excluded at a time. Thus, each discordant observation is tested p times using all subsets of p-1 of the variables. A variable is listed as causal only if absence of that variable prevents rejection of the outlier. Although this procedure is based on iterations of rigorous tests of hypothesis, the user should consider its results only as general guidance and not as definitive proof of the cause. This method also requires some additional research.

1. Click **Outliers/Estimates ► Multivariate ► Classical ► Casual ► Distances or Kurtosis.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

- Select two or more variables from the "**Select Variables**" screen.
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**."



- Select the required "**Critical Alpha**" and click "**OK**" to continue "**Cancel**" to cancel the Causal test.

Output for Causal Variables using the MD test for outliers.
Output for MD test for outliers.
Data Set used: Bradu.

	Classical Causal Variable Routine Using MDs
User Selected Options	
Date/Time of Computation	1/8/2008 7:47:44 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Number of Observations	75
Number of Columns	4
Critical Alpha	0.05
Max MD (4,0.05)	17.43
Max MD (3,0.05)	15.55

New Data Matrix (Outlier Rows)

	y	x1	x2	x3
14	0.1	11	34	34
12	-0.4	12	23	37
11	-0.2	11	24	35
13	0.7	12	26	34

Obs No.	MD_Distance	P Value
14	43.7	1
12	27.75	1
11	34.26	1
13	56.98	1

Row 14

x2 is a Causal Variable

Obs No.	Observed	Predicted
14	34	24.15

Row 12

y is a Causal Variable

Obs No.	Observed	Predicted
12	-0.4	11.33

Row 11

y is a Causal Variable

Obs No.	Observed	Predicted
11	-0.2	10.38

x3 is a Causal Variable

Obs No.	Observed	Predicted
11	35	29.58

Row 13

y is a Causal Variable

Obs No.	Observed	Predicted
13	0.7	9.81

x1 is a Causal Variable

Obs No.	Observed	Predicted
13	12	11.32

x2 is a Causal Variable

Obs No.	Observed	Predicted
13	26	24.24

x3 is a Causal Variable

Obs No.	Observed	Predicted
13	34	32.54

Results: Observations 14, 12, 11 and 13 are identified as outliers with variable “x2” in 14, variable “y” in 12, variable “y” in 11 and all variables in observation 13 as potential causal variables. The predicted value is obtained by using regression with the causal variable as the dependent variable and other variables as the independent variables.

Output for Causal Variables using the Kurtosis test for outliers.

	Classical Causal Variable Rountine Using Kurtosis
User Selected Options	
Date/Time of Computation	1/16/2008 11:45:48 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Number of Observations	75
Number of Columns	4
Critical Alpha	0.05
Full Cutoff	25.2002
Reg. Cutoff	16.1643

New Data Matrix (Outlier Rows)

y	x1	x2	x3
0.1	11	34	34
-0.4	12	23	37
-0.2	11	24	35
0.7	12	26	34

Obs No.	Kurtosis	P Value
14	53.9678865248759	N/A
12	38.2586440611149	N/A
11	43.5700927137082	N/A
13	59.1622093139548	N/A

Row 14

x2 is a Causal Variable

Obs No.	Observed	Predicted
14	34	17.3214183128282

Row 12

y is a Causal Variable

Obs No.	Observed	Predicted
12	-0.4	11.9998996724309

Row 11

y is a Causal Variable

Obs No.	Observed	Predicted
11	-0.2	11.6785030135424

x3 is a Causal Variable

Obs No.	Observed	Predicted
11	35	23.8400404984235

Row 13

y is a Causal Variable

Obs No.	Observed	Predicted
13	0.7	12.0185193167481

x3 is a Causal Variable

Obs No.	Observed	Predicted
13	34	26.7004459115258

7.2.2 Robust Outlier Testing

Detection of multiple (one or more) anomalies in multivariate data sets is a complex problem. Considerable research has been performed on this topic. Many classical and robust outlier identification methods have been developed since early 1970s. Mahalanobis distances (MDs or Mds) play the key role in identifying the outlying observations. As is well known, multiple outliers tend to influence the estimates (even robust estimates) of the means, variances, covariances, and the MDs significantly. Therefore, the MDs get distorted by the same observations that they are supposed to find.

In an effort to identify the best possible method(s) to identify outliers, the developers of Scout considered several classical as well as robust outlier identification and robust estimation methods.

Scientists dealing with the multivariate data realize that there is no substitute for the graphical display of their data. Therefore, the developers of Scout have put extra emphasis on formal graphical displays of multivariate data sets. The Scout outlier module offers the following methods to identify outliers in multivariate data sets.

- Classical (using the Max MD as proposed by Wilks in 1963)
- Sequential Classical
- Huber
- Extended Minimum Covariance Determinant (MCD)
- Proposed (PROP)
- Multivariate Trimming (MVT)

The critical values used for the Max MD statistics are obtained using the Bonferroni inequality and the scaled beta distribution: (n-1)*(n-1) Beta (p/2,(n- p-1)/2)/n of MDs. Often, because of the computational ease, an approximate chi-square distribution with p degrees of freedom is used for the distribution of MDs. The difference between the two options is significant, especially when the dimensionality, p, is large ($p = 4$ is large enough for the sample of size $n = 100$). For details, one can refer to Singh (1993). For comparison sake, both distributional options have been incorporated into Scout.

All of the outlier methods listed above (except for the classical method) are iterative. The MCD method (Rousseeuw and Van Driessen (1999)) has been extended to accommodate other options. For example, instead of obtaining the MCD by minimizing the determinant of a covariance matrix based upon $h = [(n+p+1/2)]$ observations, one can choose other values for h. The objective is to find "real and only real outliers," and not to identify outliers using the maximum break down point option. The multivariate trimming (MVT) method (Devlin, S.J, et al. (1981)) is also available in Scout to identify outliers.

The Huber and PROP methods represent M-estimation methods based upon the Huber (Huber, 1981) influence function and the PROP (Singh, 1993) influence function, respectively. The iterative process reduces the influence of potential outliers iteratively. This is especially true for the PROP influence function. The convergence is normally achieved in less than 10 iterations. A default of 20 iterations is used in the Scout software. In those iterative procedures, initial estimates of the population mean vector and the variance covariance matrix are required. Several initial estimate methods are available in Scout. Specifically, classical maximum likelihood estimates (MLEs) and various robust estimates (e.g., median, MAD or IQR; median, OKG; median, KG; and MCD) are available as initial start estimates. The use of robust initial estimates is recommended for improved and more resistant estimators of the population parameters at the final iteration. The use of the PROP influence function with an initial robust start (e.g., OKG estimate of covariance matrix) seems to be very effective in identifying multivariate outliers and computing robust and resistant estimates of the mean vector and the covariance matrices.

1. The sequential classical method performs the classical procedure (with a classical initial start) by iteratively removing outliers at each iteration. The MDs are compared to the Max (MD) critical value (e.g., for $\alpha = 0.1$). That is, there is a hard rejection point (= Max (MD)) for outliers. This procedure suffers from masking effects when multiple outliers are present. This masking effect can be reduced if an initial robust start is chosen instead of the classical initial start. This is illustrated in the graphical comparison section in the following.

2. The Huber procedure uses the Huber influence function (Huber 1981) which assigns unit weight to the observations coming from the central part of the distribution and reduced weight coming from the tails of the underlying distributions. However, outliers always leave some influence on Huber estimates.

3. The MCD method uses the minimum covariance determinant (MCD) method of Rousseeuw and Leroy (1987). Its objective is to find a set of observations in the data set whose covariance matrix has the lowest determinant. The fast algorithm to compute the minimum covariance determinant estimator has been incorporated into Scout.

4. The PROP procedure uses a smooth redescending influence function and uses the cut-off points from the distribution of the MDs. Both options, the Beta distribution and the chi-square distribution, are available in Scout. The use of beta distribution is recommended. Using this influence function, the extreme outliers coming from the tails of the contaminating distribution get almost negligible weights. This procedure provides an automatic way of dealing with the outlying observations present in a multivariate data set. A tuning constant, "c," can be used to make the influence function more resistant to outliers. In most applications, c = 1 works very well. However, when proportion of discordant observations increases, smaller values of c (< 1.0) are recommended. This procedure also works very effectively in estimating the principal components (PCs), including the low variance PCs. Probability plots of the PCs based upon this influence function are good enough to reveal all kinds of outliers, including those which might be inflating variances and covariances inappropriately and those which might be violating the correlation structure imposed by the bulk of the data.

5. Robust procedures based upon multivariate trimming (MVT) often work well in estimating the robust PCs and revealing discordant observations. A fixed proportion, p (0.05, 0.1, 0.2 etc.), of the observations with the largest values of MDs is temporarily set aside and the estimates of the mean vector and variance covariance matrix are recomputed based on the remaining n (1 - p) observations. This iterative procedure, which requires computations of those MDs at each step, stops as soon as the desired degree of accuracy has been achieved.

As mentioned before, another important aspect of robust outlier detection is obtaining an initial estimate to start the iterative robust procedures. Scout provides several procedures to compute the initial estimates of the mean vector and the covariance matrix:

- Classical
- Sequential Classical
- Robust Median/MAD
- OKG (Orthogonalized Kettering Gnanadesikan, Maronna-Zamar, 2002)
- KG (Kettering Gnanadesikan, 1972)
- MCD

The classical method uses the mean vector and the variance-covariance matrix as the initial estimate.

In the outlier module, the sequential classical method computes mean vector and the covariance matrix iteratively by removing outliers (observations exceeding the Max (MD) at each iteration.

The Robust Median/MAD (Median Absolute Deviation) method uses the median vector as the initial estimate of the location vector. For the dispersion matrix, the classical variance covariance matrix is used, but the diagonal elements are replaced by the simple robust estimate of the variance, given by $\left(\frac{MAD}{0.6745}\right)^2$.

If the MAD is equal to (or approximately equal to) 0 (as in Fisher's Iris data set), then the diagonal elements are replaced by the respective, $\left(\frac{IQR}{1.355}\right)^2$, where IQR is computed separately for each variable. This is called the IQR fix in Scout.

The OKG (Orthogonalized Kettering Gnanadesikan) method uses the median vector as the initial estimate of the location vector. In practice, such a KG covariance matrix may not be positive definite (can yield even negative eigen values). The dispersion matrix of the Kettering Gnanadesikan can be orthogonalized using the procedure described by Maronna and Zamar (2002) to get a positive definite dispersion matrix. This procedure is also available in Scout.

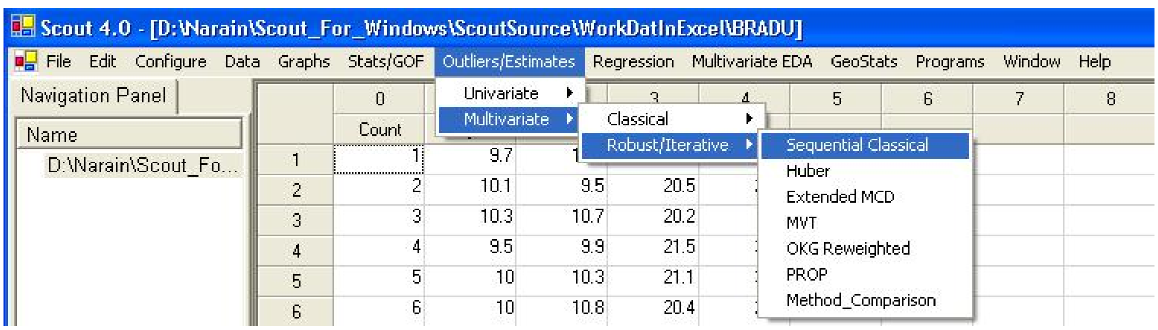
For the MCD method, the objective is to find a subset of some specified size, h ($n/2 \le h \le n$), which will minimize the determinant of the covariance matrix based upon that subset of size h. The subset of size h minimizing the determinant of the covariance matrix is termed as the best subset. The positive integer, h, is known as coverage or half sample. The most commonly used and default value of h is $[(n+p+1)/2]$ = largest integer contained in $(n+p+1)/2$.

The BD point of an MCD estimate is given by the fraction $(n-h+1)/n$. There is a direct relation between the coverage value, h, and the BD point of the MCD estimates. Higher values of h yield estimates with the lower BD point. The use of the default value of coverage, h, roughly identifies the optimal (~ about 50%) number of outliers.

The MCD method can also be used to obtain the initial robust estimates of the location and scatter matrix.

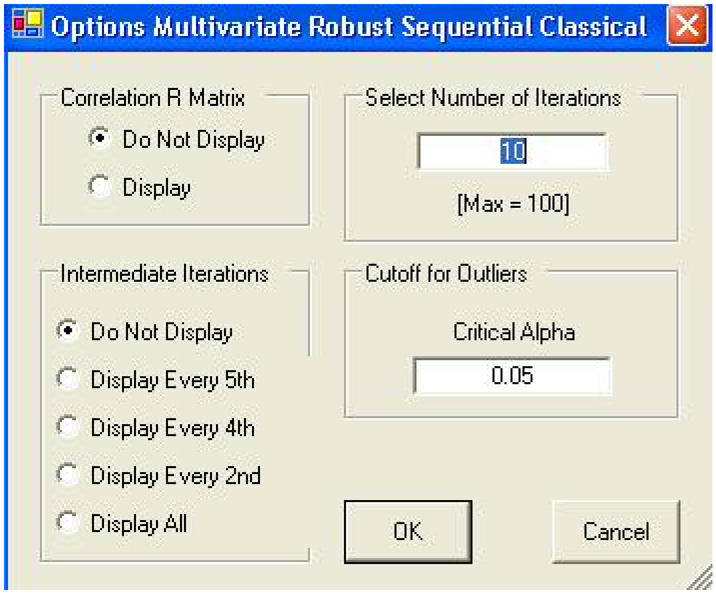
7.2.2.1 Sequential Classical

1. Click **Outlier/Estimates ► Multivariate ► Robust ►Sequential Classical.**



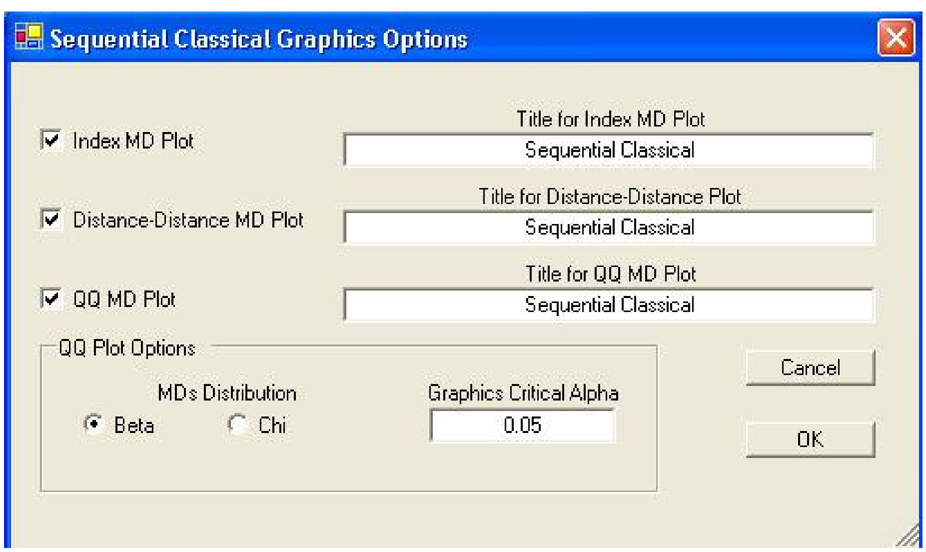
2. The "**Select Variables**" screen will appear (Section 3.4).

- Select the variables from the screen.
- If various groups are available and the analysis is to be done for those groups, then group variables can be selected from the drop-down list by clicking on the arrow below the "**Group by Variable**" button.
- Click the "**Options**" button for various options.



- Specify if the correlation matrix of the final robust estimate is to be displayed or not. The default option is "**Do Not Display**."
- Specify the number of iterations to be computed. The default is "**10**."
- Specify if the intermediate iterations be displayed or not. The default is "**Do not display**."
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click the "**Graphics**" button for the graphics options and check the three check boxes to get the following screen.



- Specify the “**Title for Index MD Plot**.” This is an index plot of the robust distances obtained using the sequential outlier estimates.

- Specify the “**Title for Distance-Distance Plot**.” This is a plot of the classical Mahalanobis against the robust distances obtained using the sequential outlier estimates.

- Specify the “**Title for Q-Q MD Plot**.” Select the distribution required for the “**Q-Q Plot**” and the “**Graphics Critical Alpha**” for identifying the outliers.

  Note*: The “****Graphics Critical Alpha****” should match the “****Critical Alpha****” from the* ***Options Multivariate Robust Sequential Classical*** *window to obtain the same outliers. The user should type suitable titles related to the data set.*

- Click “**OK**” to continue or “**Cancel**” to cancel the options.

* Click “**OK**” to continue or “**Cancel**” to cancel the sequential classical procedure.

Output example: The data set “**BRADU.xls**” was used for Sequential Classical. The outliers are removed (down weighted from 1 to 0) at the end of each iteration and the location and scale estimates are calculated at the end of each iteration.

Output for Sequential Outliers method.
Data Set used: Bradu.

	Multivariate Robust Sequential Classical Outlier Analysis
User Selected Options	
Date/Time of Computation	3/4/2008 9:04:35 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Display Correlation R Matrix	Do Not Display Correlation R matrix
Number of Iterations	10
Show Intermediate Results	Do Not Display Intermediate Results
Title for Index Plot	Sequential Classical
Title for Distance-Distance Plot	Sequential Classical
Title for QQ Plot	Sequential Classical
Graphics Critical Alpha	0.05
MDs Distribution	Beta

Number of Observations	75
Number of Selected Variables	4
Max Squared MD (0.05)	17.43

Classical Mean Vector

y	x1	x2	x3
1.279	3.207	5.597	7.231

Classical Covariance S Matrix

y	x1	x2	x3
12.2	9.477	20.39	31.03
9.477	13.34	28.47	41.24
20.39	28.47	67.88	94.67
31.03	41.24	94.67	137.8

Determinant	1906
Log of Determinant	7.553

Eigenvalues of Classical Covariance S Matrix

Eval 1	Eval 2	Eval 3	Eval 4
0.914	1.688	5.538	223.1

3 Outliers were found using Classical Method
4 Outliers were found using Sequential Classicall

Output for Sequential Outliers method (continued).

Squared MDs			
Obs	Classical	Sequential	Weights
1	6.026	5.682	1
2	6.729	6.588	1
3	7.209	7.197	1
4	6.321	9.265	1
5	6.339	6.751	1
6	7.036	6.755	1
7	8.235	8.706	1
8	6.714	6.788	1
9	6.445	7.52	1
10	7.346	7.215	1
11	18.61	333.7	0
12	27.74	366.4	0
13	14.79	310.2	0
14	43.7	512.8	0
15	3.386	5.109	1
16	4.783	5.689	1
17	2.004	2.496	1
18	0.751	0.764	1
19	1.408	1.934	1
20	2.556	2.954	1
21	1.281	2.178	1
22	2.578	3.282	1
23	1.216	2.757	1
24	1.321	3.455	1
25	0.658	1.394	1
26	1.413	4.544	1
27	2.492	5.615	1
28	0.765	1.829	1
29	0.364	1.577	1
30	2.793	4.472	1
31	3.395	3.219	1
32	1.772	2.704	1
33	0.967	3.078	1
34	1.475	1.797	1
35	1.58	1.622	1
36	1.008	3.666	1
37	3.368	4.751	1

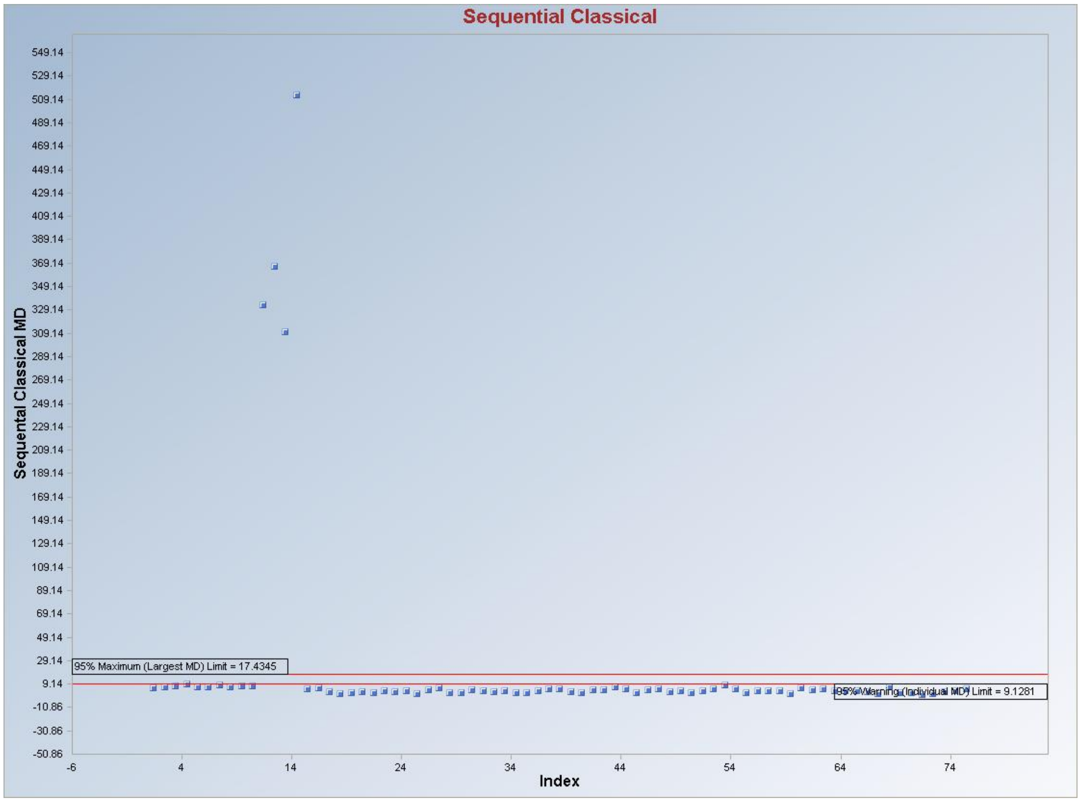
(The complete output table is not shown.)

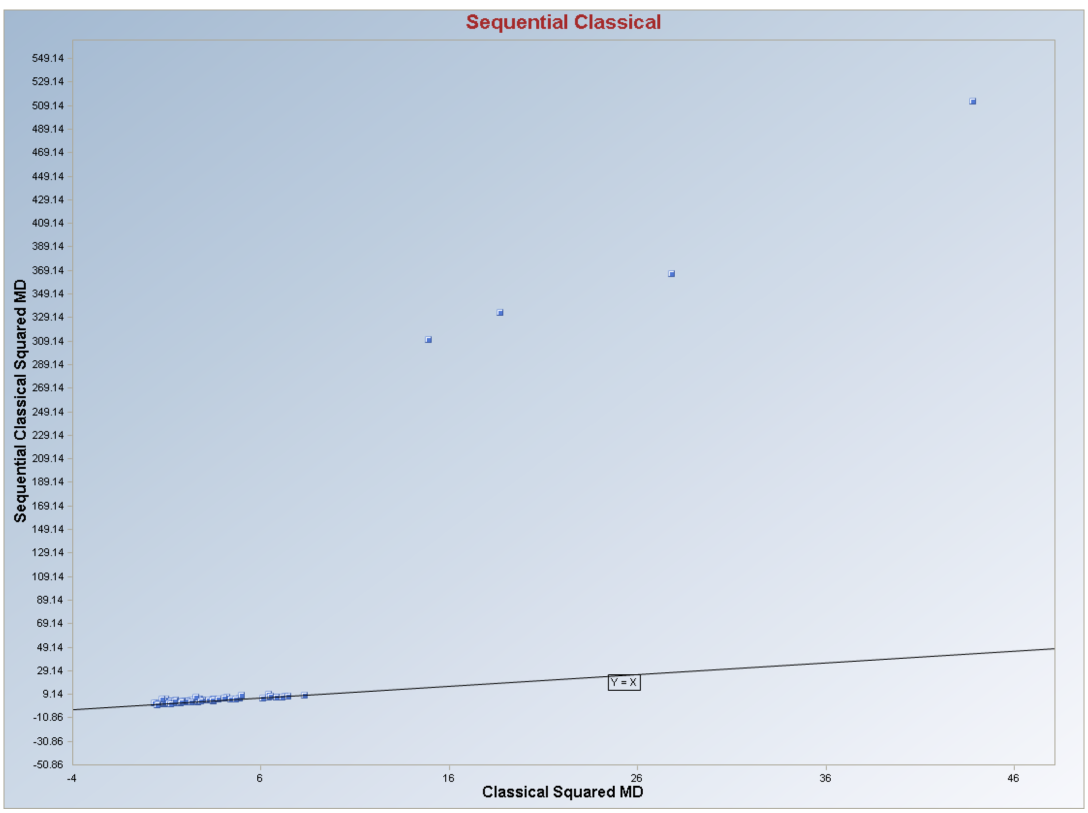
Output for Sequential Outliers method (continued).

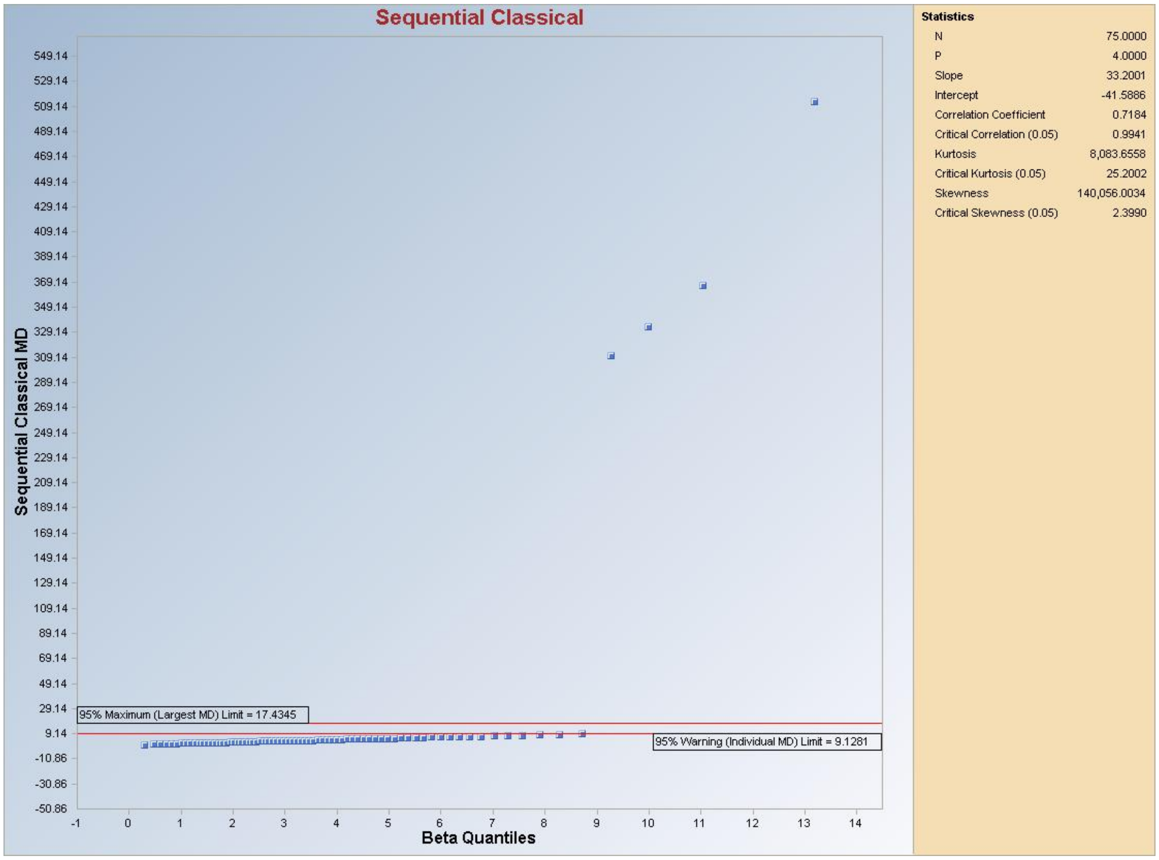
Sequential Classical Estimates					
		Sequential Classical Mean Vector			
y	x1	x2	x3		
1.348	2.739	4.406	5.666		
		Sequential Classical Covariance S Matrix			
y	x1	x2	x3		
12.8	10.63	23.09	34.89		
10.63	9.938	19.57	29.69		
23.09	19.57	43.69	64.82		
34.89	29.69	64.82	99.08		
		Determinant	56.61		
		Log of Determinant	4.036		
		Classical Kurtosis	53.97		
		Sequential Kurtosis	8084		

Results: Four (4) observations (11, 12, 13 and 14), with squared distances greater than the Max (squared MD), were given zero (0) weights (hard rejection) and were considered to be outliers.

Graphical Output for Sequential Outliers method (continued).



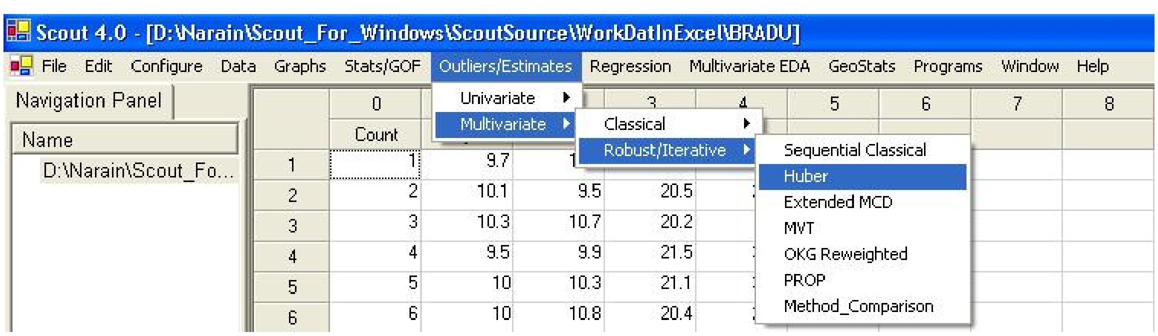




Graphical Interpretation: The observations between the "**Warning (Individual MD) Limit**" and "**Maximum (Largest MD) Limit**" lines represent borderline outliers and may require further investigation. The Warning Limit represents the critical value from the scaled beta distribution of the MDs at a specified level of significance (here, 0.95), and the Maximum (Largest MD) Limit represents the critical value of the Max (MD) obtained using the Bonferroni inequality (details in Singh, 1993).

7.2.2.2 Huber

1. Click **Outlier/Estimates ► Multivariate ► Robust ►Huber.**

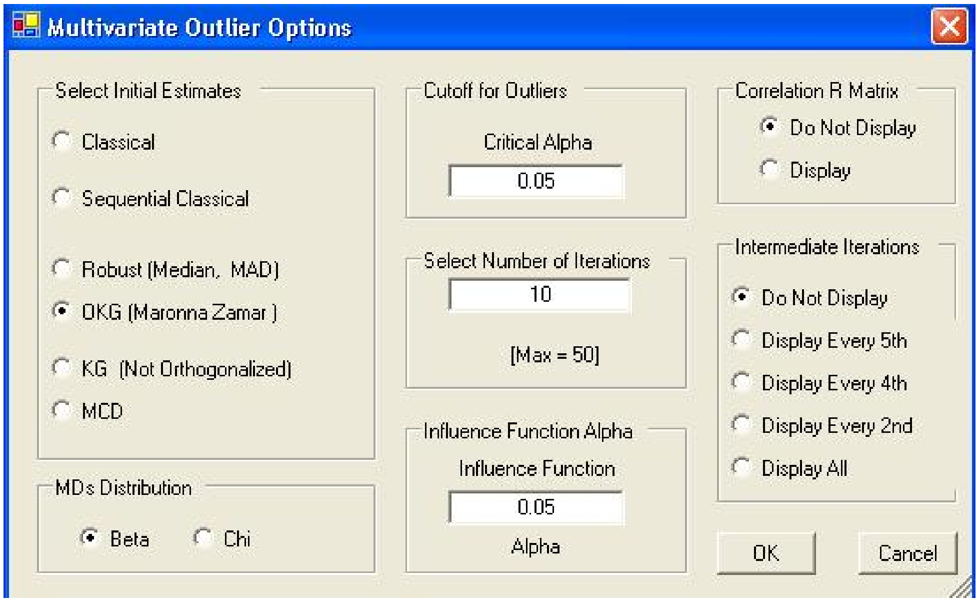


2. The "**Select Variables**" screen will appear (Section 3.4).

- Select the variables from the screen.

- If various groups are available and an analysis is to be done for those groups, then group variables can be selected from the drop-down list by clicking on the arrow below the "**Group by Variable**" button.

- Click the "**Options**" button for various options.



  - Specify the "**Initial Estimates**" to start the Huber iterative procedure. The default is "**OKG (Maronna Zamar)**."

  - Specify the distribution for the Mahalanobis distances in the "**MDs Distribution**." The default is "**Beta**."

  - Specify the "**Critical Alpha**," the cutoff for outliers. The default is "**0.05**."

  - Specify the "**Number of Iterations**." The default is "**10**."

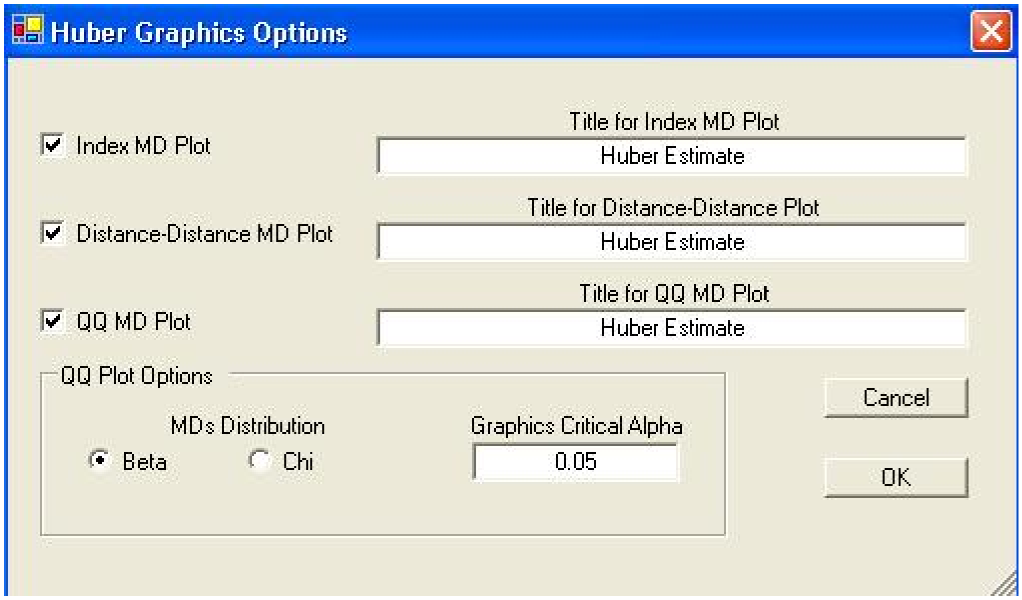
  - Specify the "**Influence Function Alpha**" for the Huber weighting process. The default is "**0.05**."

  - Specify "**Correlation R Matrix**." The default is "**Do Not Display**."

  - Specify "**Intermediate Iterations**." The default is "**Do Not Display**."

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click the "**Graphs**" button for various options.



- Specify the "**Title for Index MD Plot**." This is an index plot of the robust distances obtained using the Huber estimates.

- Specify the "**Title for Distance-Distance Plot**." This is a plot of the classical Mahalanobis against the robust distances obtained using the Huber estimates.

- Specify the "**Title for Q-Q MD Plot**." Select the distribution required for the "**Q-Q Plot**" and the "**Graphics Critical Alpha**" for identifying the outliers.

  Note: *The "**Graphics Critical Alpha**" should match the "**Critical Alpha**" from the **Multivariate Outlier Options** window to obtain the same outliers.*

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

• Click "**OK**" to continue or "**Cancel**" to cancel the Huber procedure.

Output example: The data set "**BRADU.xls**" was used for the Huber method. It has 75 observations and four variables. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the Huber influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were then calculated.

Output for Huber outliers method.
Data Set used: Bradu.

	Huber Multivariate Outlier Analysis
User Selected Options	
Date/Time of Computation	1/8/2008 1:21:20 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Critical Alpha	0.05
Influence Function Alpha	0.05
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Display Correlation R Matrix	Do Not Display Correlation R matrix
Distribution of Squared MDs	Beta Distribution
Number of Iterations	10
Show Intermediate Results	Do Not Display Intermediate Results
Title for Index Plot	Huber Estimate
Title for Distance-Distance Plot	Huber Estimate
Title for QQ Plot	Huber Estimate
Graphics Critical Alpha	0.05
MDs Distribution	Beta

Number of Observations	75
Number of Selected Variables	4

Critical Values

Max Squared MD (0.05)	17.43
Individual Squared MD (0.05)	9.128
Multivariate Kurtosis(0.05)	25.2
Influence Fn. Squared MD (0.05)	9.128

Classical Mean Vector

y	x1	x2	x3
1.279	3.207	5.597	7.231

Classical Covariance S Matrix

y	x1	x2	x3
12.2	9.477	20.39	31.03
9.477	13.34	28.47	41.24
20.39	28.47	67.88	94.67
31.03	41.24	94.67	137.8

Determinant	1906.26080388262

Output for Huber outliers method (continued).

Eigenvalues for Classical Covariance S Matrix

Eval 1	Eval 2	Eval 3	Eval 4
0.914	1.688	5.538	223.1

Median Vector

y	x1	x2	x3
0.1	1.8	2.2	2.1

MAD/0.6745 Vector Representing Standard Deviation

y	x1	x2	x3
0.89	1.927	1.631	1.779

OKG Mean Vector

y	x1	x2	x3
1.258	-7.552	-6.052	2.521

Robust OKG (Maronna Zamar) Covariance S Matrix

y	x1	x2	x3
0.598	0.243	0.115	-0.231
0.243	2.856	0.108	0.241
0.115	0.108	2.175	0.00402
-0.231	0.241	0.00402	2.34
		Determinant	7.84553740487618

Robust OKG Eigenvalues

Eval 1	Eval 2	Eval 3	Eval 4
0.53	2.149	2.316	2.975

Final Weighted Mean Vector

y	x1	x2	x3
1.332	2.849	4.68	6.033

Final Covariance S Matrix

y	x1	x2	x3
12.76	10.57	22.94	34.67
10.57	10.14	20.08	30.37
22.94	20.08	45	66.52
34.67	30.37	66.52	101.4
		Determinant	119.506396724439

Output for Huber outliers method (continued).

		Observations with Squared MDs greater than 17.43 may be considered as potential outliers!		
Observation Number	Classical Squared MD	Initial Squared MD	Final Squared MD	Final Weights
1	6.026	654.4	5.709	1
2	6.729	699.8	6.635	1
3	7.209	761.9	7.183	1
4	6.321	769.1	7.201	1
5	6.339	765.6	6.216	1
6	7.036	701.8	6.7	1
7	8.235	740.2	8.57	1
8	6.714	692.4	6.71	1
9	6.445	740.8	6.606	1
10	7.346	719.1	7.218	1
11	18.61	691.2	160.8	0.239
12	27.74	732.6	182.7	0.225
13	14.79	712.8	148.1	0.25
14	43.7	907	270.9	0.184
15	3.386	1.944	3.909	1
16	4.783	2.228	5.488	1
17	2.004	2.761	2.463	1
18	0.751	0.303	0.789	1
19	1.408	0.675	1.918	1
20	2.556	1.234	2.828	1
21	1.281	1.165	1.867	1
22	2.578	1.322	3.273	1
23	1.216	2.84	1.843	1
24	1.321	1.542	2.856	1
25	0.658	3.846	1.261	1
26	1.413	2.229	2.699	1
27	2.492	2.387	3.599	1
28	0.765	1.029	1.64	1
29	0.364	1.612	1.135	1
30	2.793	1.945	4.316	1
31	3.395	1.721	3.276	1
32	1.772	2.478	2.6	1
33	0.967	1.674	2.028	1
34	1.475	4.461	1.824	1

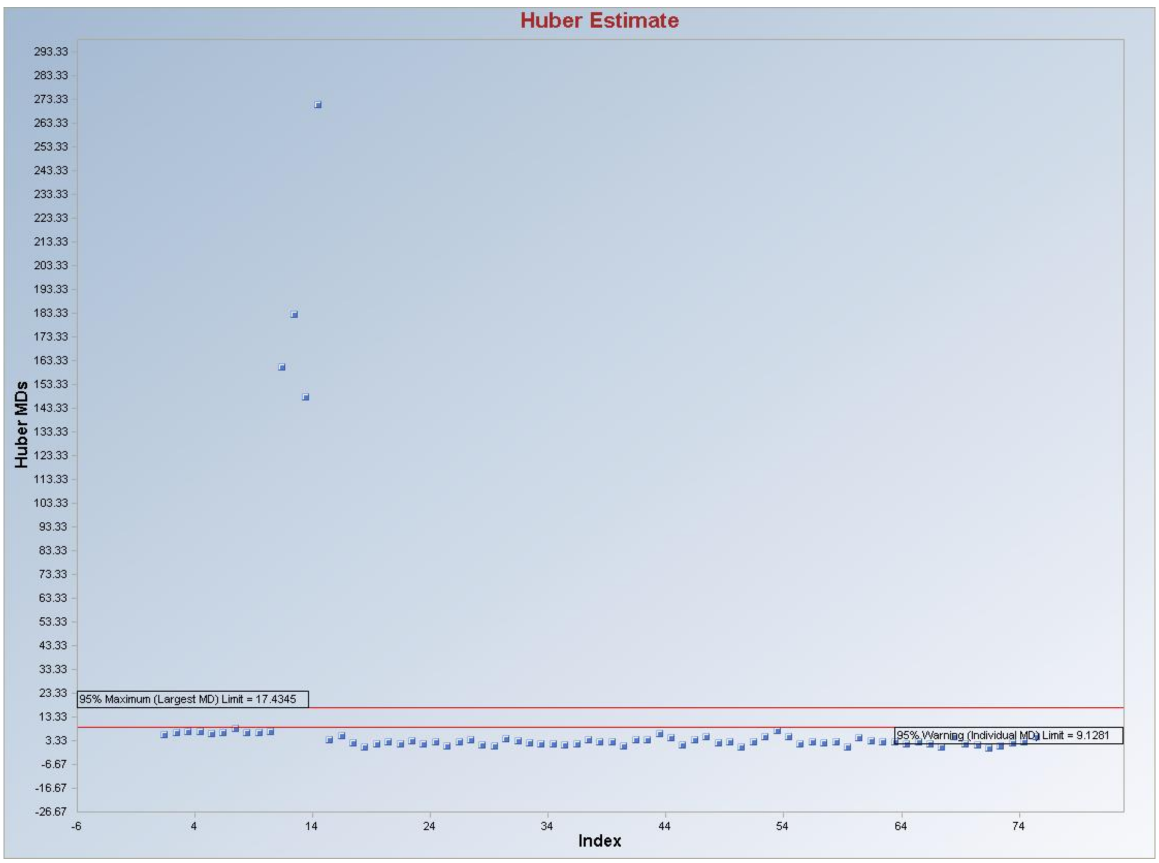
Output for Huber outliers method (continued).

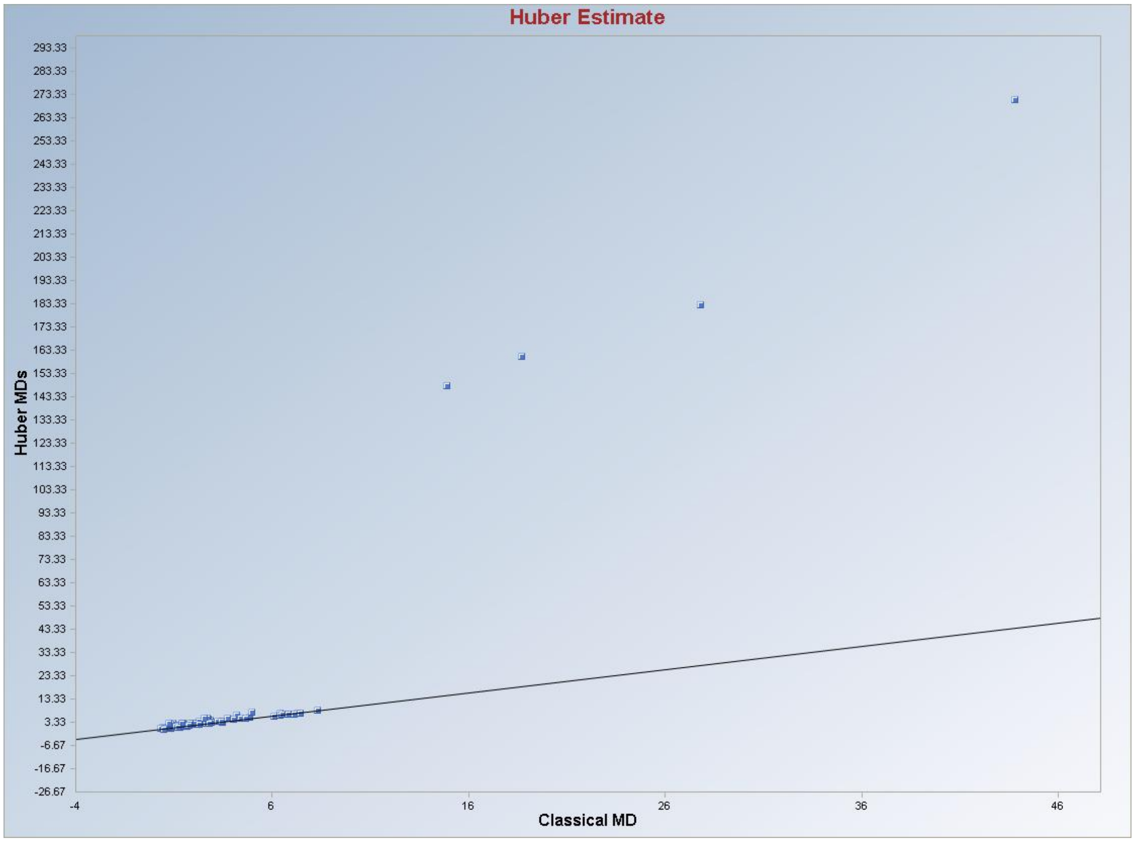
35	1.58	0.923	1.561	1			
36	1.008	2.086	1.945	1			
37	3.368	2.047	3.842	1			
38	0.889	2.826	3.046	1			
39	1.935	3.883	2.76	1			
40	1.237	1.463	1.314	1			
41	3.139	1.386	3.637	1			
42	3.358	2.147	3.89	1			
43	4.108	2.526	6.291	1			
44	2.686	2.505	4.734	1			
45	1.278	3.85	1.671	1			
46	2.225	1.865	3.851	1			
47	4.332	5.896	4.947	1			
48	2.184	1.054	2.641	1			
49	2.552	1.422	2.927	1			
50	0.278	0.856	0.74	1			
51	1.858	1.936	2.959	1			
52	4.587	3.562	4.898	1			
53	4.889	4.616	7.612	1			
54	2.604	2.131	4.971	1			
55	1.622	1.159	2.19	1			
56	1.898	1.137	3.037	1			
57	0.773	2.385	2.402	1			
58	1.988	1.208	2.9	1			
59	0.463	0.523	0.681	1			
60	3.945	2.628	4.791	1			
61	2.806	1.957	3.26	1			
62	0.675	3.98	3.022	1			
63	1.757	2.344	2.934	1			
64	1.112	1.602	1.943	1			
65	1.336	1.909	3.045	1			
66	1.696	3.273	1.913	1			
67	0.445	1.752	0.688	1			
68	2.462	4.222	5.049	1			
69	1.154	2.636	1.83	1			
70	1.042	1.609	1.423	1			
71	0.413	0.114	0.381	1			
72	1.114	0.732	1.117	1			
73	2.207	1.311	2.576	1			

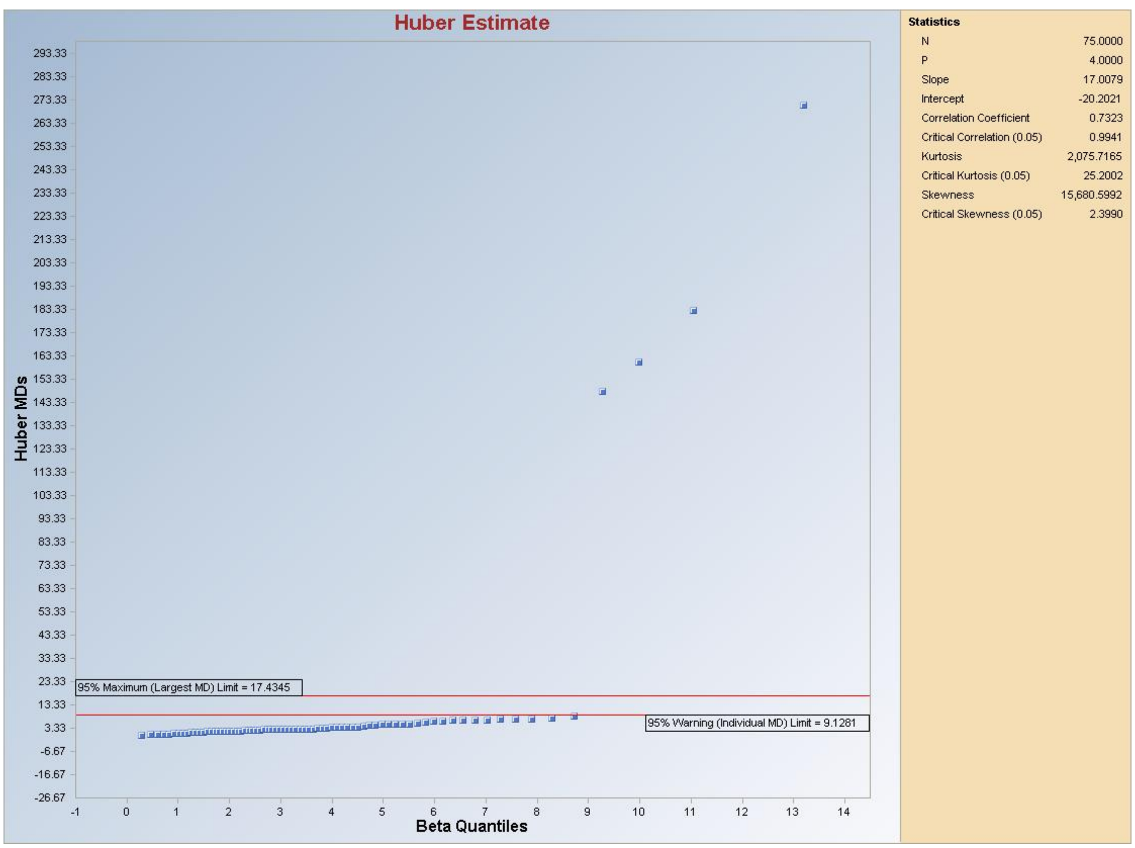
Output for Huber outliers method (continued).

74	2.742	2.635	3.069	1			
75	3.632	2.62	5.12	1			
	Classical	Initial	Final				
Multivariate Kurtosis	53.97	101431	2076				

Results: Four (4) observations (11, 12, 13 and 14), with squared distances greater than the squared MD 17.43, were given weights between 0 and 1 (soft rejection) and were considered to be outliers. Note that due to masking effects, the Huber method did not identify the remaining 10 outliers present in this data set.



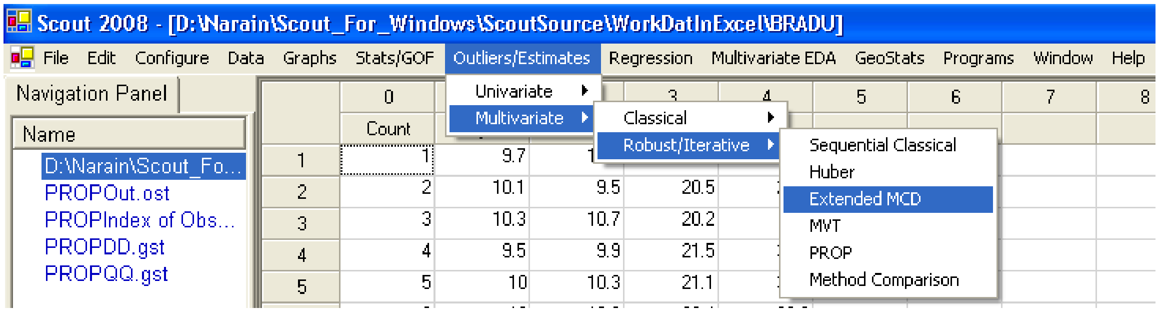

Huber Estimate
Huber MDs
Classical MD


Huber Estimate
Huber MDs
Beta Quantiles
95% Maximum (Largest MD) Limit = 17.4345
95% Warning (Individual MD) Limit = 9.1281
Statistics
N 75.0000
P 4.0000
Slope 17.0079
Intercept -20.2021
Correlation Coefficient 0.7323
Critical Correlation (0.05) 0.9941
Kurtosis 2,075.7165
Critical Kurtosis (0.05) 25.2002
Skewness 15,680.5992
Critical Skewness (0.05) 2.3990

Graphical Interpretation: Observations (if any) between the "**Warning (Individual MD) Limit**" and "**Maximum (Largest MD) Limit**" lines may require further investigation. Those observations have reduced weights between 0 and 1.

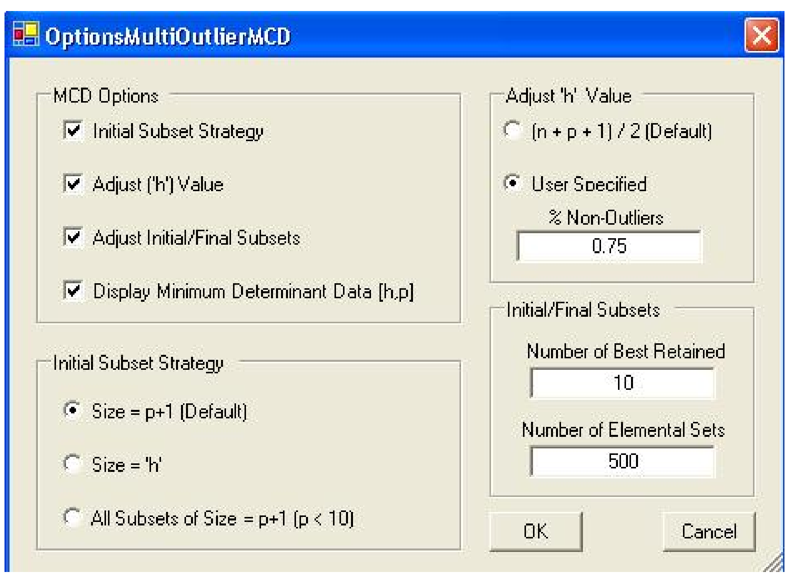
7.2.2.3 Extended MCD

1. Click **Outlier/Estimates ► Multivariate ► Robust ►Extended MCD.**



2. The "**Select Variables**" screen will appear (Section 3.4).

- Select the variables from the screen.
- If various groups are available and analysis is to be done for those groups, then group variables can be selected from the drop-down list by clicking on the arrow below the "**Group by Variable**" button.
- Click the "**Options"** button for various options.



  - When all of the checkboxes are checked in the "**MCD Options**," the options window looks like the one above.

- o "**Initial Subset Strategy**": this is used to specify the size of the initial subsets. It can be equal to the number of variable plus 1 (p + 1) or of size equal to h, the number of non-outliers.

- o Specify the "**Initial Subset Strategy.**" The default is "**Size = p+1**."

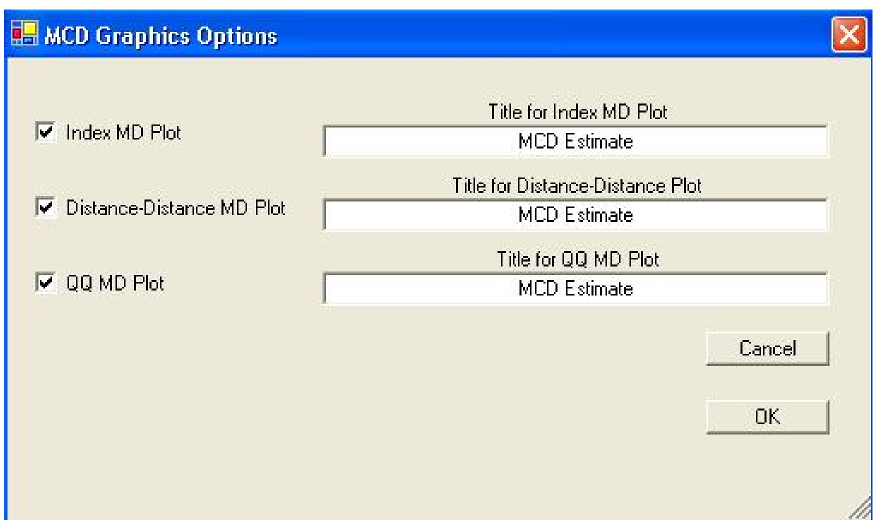
- o "**Adjust "h" Value**": this is used to specify the number of non-outliers. It can be equal to $\frac{n+p+1}{2}$ or a percentage.
  Specify the **Adjust "h" Value**. The default is "$\frac{n+p+1}{2}$."

- o "**Adjust Initial/Final Subsets**": this is used to specify the number of elemental subsets of size "p+1" or "h" to be used to start the C-Step operations of the MCD algorithm and the number of subsets with the lowest determinant of the scatter matrix to be retained to continue C-Steps until convergence.

- o Specify the "**Adjust Initial/Final Subsets**." The defaults are "**10**" and "**500**" respectively**.**

- o Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click the "**Graphs**" button for various options.



- o Specify the "**Title for Index MD Plot**." This is an index plot of the robust distances obtained using the MCD estimates.

- o Specify the "**Title for Distance-Distance Plot**." This is a plot of the classical Mahalanobis against the robust distances obtained using the MCD estimates.

- o Specify the "**Title for Q-Q MD Plot**."

- o Click “**OK**” to continue or “**Cancel**” to cancel the options.

- Click “**OK**” to continue or “**Cancel**” to cancel the MCD procedure.

Output example: The data set “**BRADU.xls**” was used for the MCD method. It has 75 observations and four variables. The MCD estimates of location and scale were obtained using the best “h” subset. Then this scale estimate was adjusted for multi-normality. Using this estimate, outliers were obtained and weighted (hard weighting) accordingly. The weighted mean vector and the weighted covariance matrix were then calculated.

Output for MCD outliers method.
Data Set used: Bradu.

	Multivariate Robust MCD Outlier Analysis
User Selected Options	
Date/Time of Computation	1/8/2008 4:52:34 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Initial Elemental Subset Size	500 Random Subsets of Initial Size of 'p + 1' will be used
User Selected 'h' Value	'h' Value of '(n * 0.75' will be used
r Selected Number of Initial Subsets	500
ed Number of Intermediate Subsets	500
d Number of Best Retained Subsets	10
Minimum Determinant Data (h,p)	Will Not be Display in Output
Title for Index Plot	MCD Estimate
Title for Distance-Distance Plot	MCD Estimate
Title for QQ Plot	MCD Estimate
Graphics Critical Alpha	0.05

Number of Observations	75
Number of Variables	4
Size of Elemental Subsets	5
Number of Initial Subsets	500
Number of Intermediate Subsets	500
'h' Value	56
Breakdown Value	0.2667
Chisquare with 4 DoF (0.975)	11.1510
Unsquared Chisquare with 4 DoF (0.975)	3.3393
Chisquare with 4 DoF (0.50)	3.3531

Classical Mean Vector

y	x1	x2	x3
1.279	3.207	5.597	7.231

Classical Covariance S Matrix

y	x1	x2	x3
12.2	9.477	20.39	31.03
9.477	13.34	28.47	41.24
20.39	28.47	67.88	94.67
31.03	41.24	94.67	137.8
			Determinant 1906.26080388263

Output for MCD outliers method (continued).

Best 'h' Subset of 56 Observations																			
15,	16,	17,	18,	19,	20,	21,	22,	23,	24,	25,	26,	27,	28,	29,	30,	31,	32,	33,	34
35,	36,	37,	38,	39,	40,	41,	42,	44,	45,	46,	48,	49,	50,	51,	54,	55,	56,	57,	58
59,	60,	61,	62,	63,	64,	65,	66,	67,	69,	70,	71,	72,	73,	74,	75				

MCD Mean Vector

y	x1	x2	x3
-0.0893	1.554	1.861	1.707

MCD Covariance S Matrix

y	x1	x2	x3
0.303	0.153	-0.00848	-0.138
0.153	1.036	0.0987	0.167
-0.00848	0.0987	1.04	0.158
-0.138	0.167	0.158	1.084

MCD Determinant 0.282788956390238

Adjustment Factor for Multinormality 1.41188216564302

Adjusted MCD Covariance S Matrix

y	x1	x2	x3
0.428	0.215	-0.012	-0.194
0.215	1.463	0.139	0.236
-0.012	0.139	1.468	0.222
-0.194	0.236	0.222	1.53

Adjusted MCD Determinant 1.12371519856147

Obs	Unsquared I	Weights
1	32.13	0
2	33.3	0
3	34.66	0
4	34.76	0
5	34.71	0
6	33.2	0
7	34.19	0
8	33.13	0
9	34.18	0
10	33.79	0
11	31.17	0

Output for MCD outliers method (continued).

12	32.15	0		
13	31.63	0		
14	35.3	0		
15	1.932	1		
16	1.927	1		
17	1.651	1		
18	0.699	1		
19	1.227	1		
20	1.794	1		
21	1.656	1		
22	1.854	1		
23	1.831	1		
24	1.546	1		
25	1.838	1		
26	1.672	1		
27	2.048	1		
28	1.176	1		
29	1.14	1		
30	1.863	1		
31	1.623	1		
32	1.603	1		
33	1.394	1		
34	2.036	1		
35	1.642	1		
36	1.491	1		
37	1.796	1		
38	1.811	1		
39	2.122	1		
40	1.074	1		
41	1.773	1		
42	1.663	1		
43	2.331	1		
44	1.925	1		
45	2.003	1		
46	1.688	1		
47	2.904	1		
48	1.617	1		
49	1.84	1		
50	1.241	1		

Output for MCD outliers method (continued).

52	2.336	1
53	3.05	1
54	1.953	1
55	1.208	1
56	1.398	1
57	1.584	1
58	1.5	1
59	1.055	1
60	1.934	1
61	1.987	1
62	2	1
63	1.641	1
64	1.677	1
65	1.885	1
66	1.736	1
67	1.246	1
68	2.339	1
69	1.568	1
70	2.123	1
71	0.912	1
72	0.815	1
73	1.761	1
74	1.639	1
75	2.137	1

Observations with Unsquared MDs greater than 11.15
may be considered as potential outliers!

Weighted Mean Vector

y	x1	x2	x3
-0.0738	1.538	1.78	1.687

Weighted Covariance S Matrix

y	x1	x2	x3
0.317	0.0587	0.00186	-0.105
0.0587	1.132	0.0508	0.117
0.00186	0.0508	1.152	0.141
-0.105	0.117	0.141	1.07

Weighted S Determinant 0.410144007183786

Output for MCD outliers method (continued).

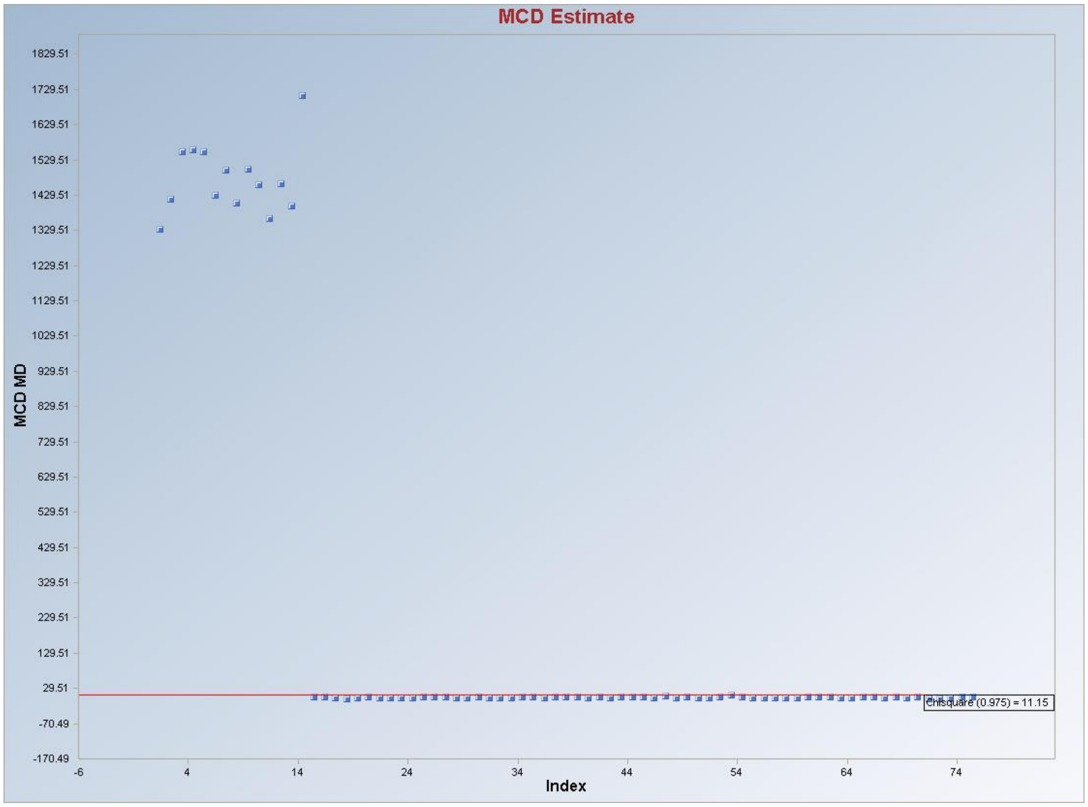
Squared MDs					
Obs	MCD MDs	Weights	Classical MD		
1	36.51	0	6.026		
2	37.66	0	6.729		
3	39.4	0	7.209		
4	39.48	0	6.321		
5	39.42	0	6.339		
6	37.81	0	7.036		
7	38.75	0	8.235		
8	37.53	0	6.714		
9	38.77	0	6.445		
10	38.22	0	7.346		
11	36.94	0	18.61		
12	38.24	0	27.74		
13	37.4	0	14.79		
14	41.37	0	43.7		
15	2.137	1	3.386		
16	2.298	1	4.783		
17	1.968	1	2.004		
18	0.794	1	0.751		
19	1.336	1	1.408		
20	2.198	1	2.556		
21	1.988	1	1.281		
22	1.929	1	2.578		
23	1.943	1	1.216		
24	1.776	1	1.321		
25	2.064	1	0.658		
26	2.042	1	1.413		
27	2.286	1	2.492		
28	1.268	1	0.765		
29	1.269	1	0.364		
30	2.111	1	2.793		
31	1.723	1	3.395		
32	1.919	1	1.772		
33	1.633	1	0.967		
34	2.35	1	1.475		
35	2.026	1	1.58		
36	1.845	1	1.008		
37	2.114	1	3.368		

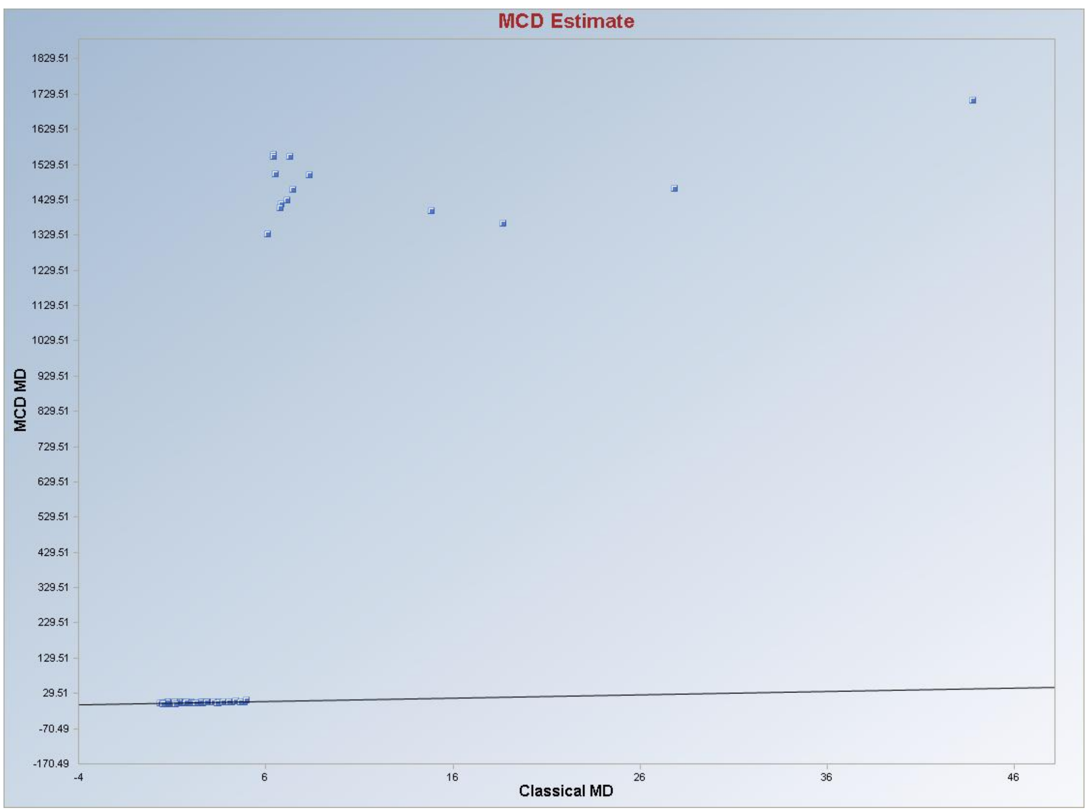
Output for MCD outliers method (continued).

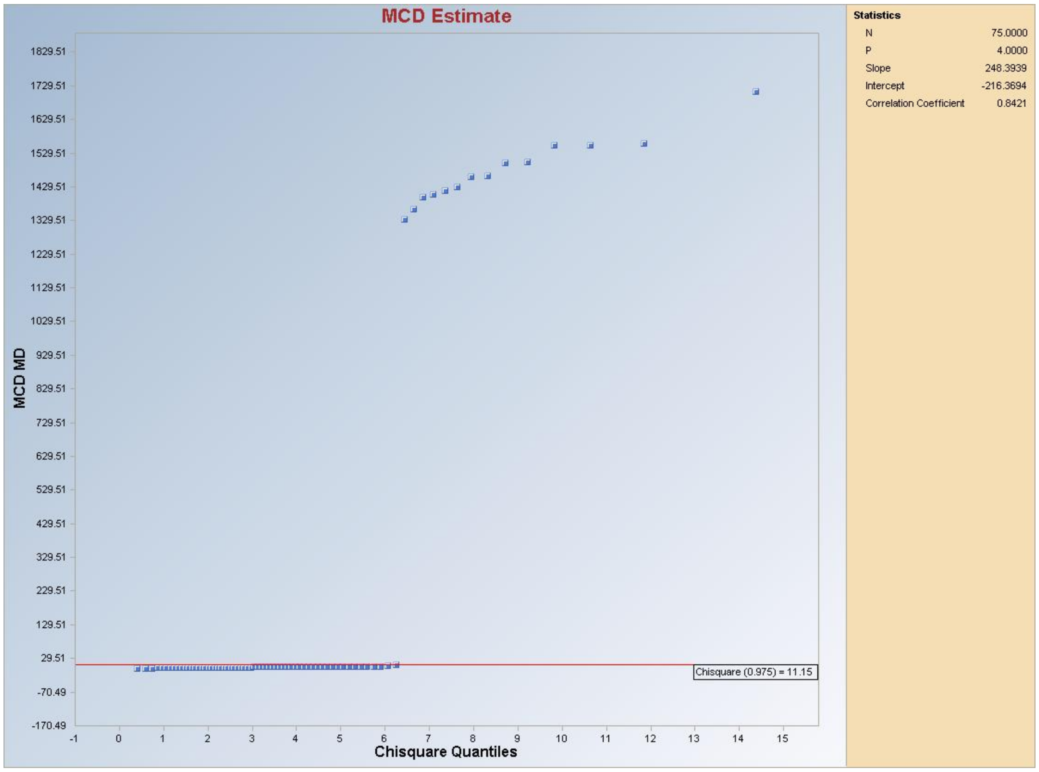
40	1.322	1	1.237		
41	2.046	1	3.139		
42	1.967	1	3.358		
43	2.471	1	4.108		
44	2.192	1	2.686		
45	2.103	1	1.278		
46	1.901	1	2.225		
47	2.909	1	4.332		
48	1.925	1	2.184		
49	2.269	1	2.552		
50	1.539	1	0.278		
51	1.858	1	1.858		
52	2.34	1	4.587		
53	3.148	1	4.889		
54	2.231	1	2.604		
55	1.377	1	1.622		
56	1.648	1	1.898		
57	1.82	1	0.773		
58	1.756	1	1.988		
59	1.291	1	0.463		
60	2.342	1	3.945		
61	2.244	1	2.806		
62	2.254	1	0.675		
63	1.923	1	1.757		
64	1.986	1	1.112		
65	2.082	1	1.336		
66	2.112	1	1.696		
67	1.362	1	0.445		
68	2.483	1	2.462		
69	1.758	1	1.154		
70	2.173	1	1.042		
71	1.133	1	0.413		
72	0.937	1	1.114		
73	1.723	1	2.207		
74	2.032	1	2.742		
75	2.23	1	3.632		
			Classical Kurtosis	53.97	
			MCD Kurtosis	9.072E+11	

Results: Fourteen (14) observations (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14), with squared distances greater than the Max squared MCD MDs, were given a weight of zero (0) (hard rejection) and were considered to be outliers. Also, note the large value of the MCD kurtosis.

Output for MCD outliers method (continued).



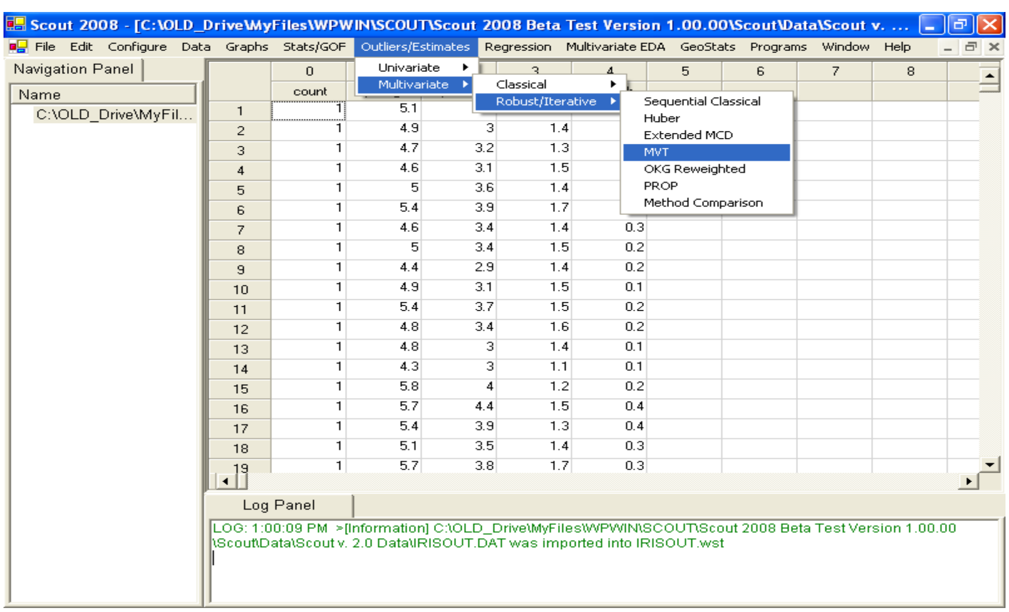




Graphical Interpretation: The observations greater than "the **chi-square (0.975)**" line may be considered to be potential outliers. Those observations have weights of zero (0). To be consistent with the literature, on the graphs generated for MCD method, chi-square quantiles (and not beta quantiles) have been used.

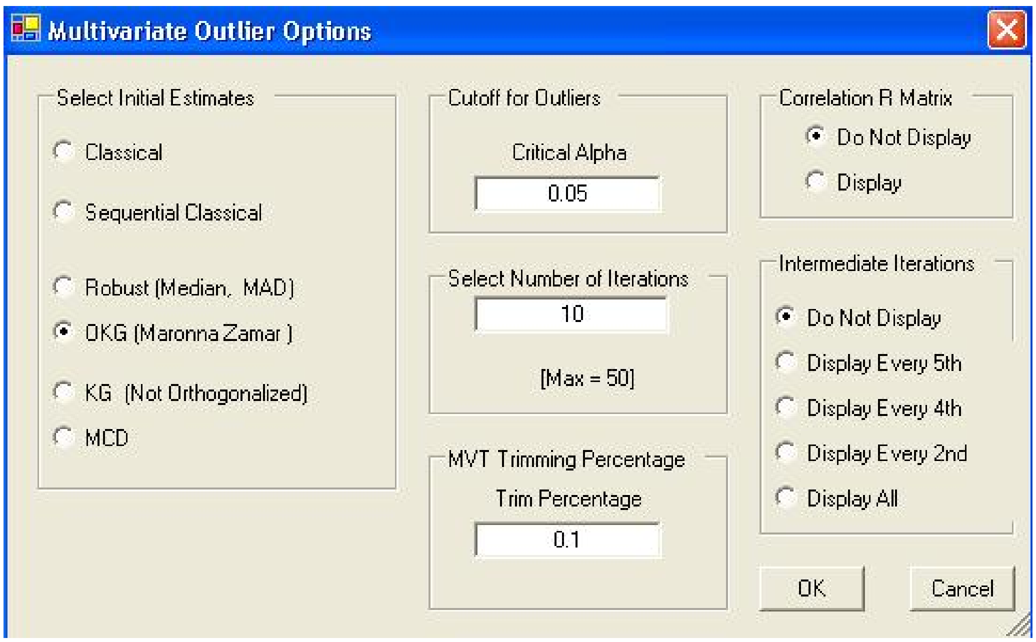
7.2.2.4 MVT

1. Click **Outlier/Estimates ► Multivariate ► Robust ►MVT.**



2. The "**Select Variables**" screen will appear (Section 3.4).

- Select the variables from the screen.
- If various groups are available and an analysis is to be done for those groups, then group variables can be selected from the drop-down list by clicking on the arrow below the "**Group by Variable**" button.
- Click the "**Options**" button for various options.

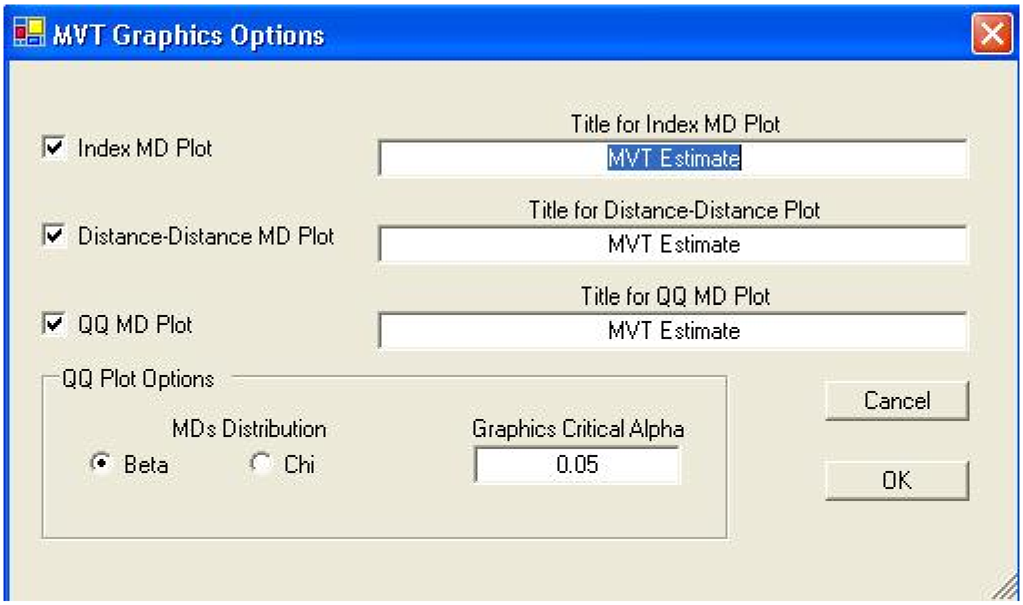


  - Specify the "**Select Initial Estimates**" to start the Huber iterative procedure. The default is "**OKG (Maronna Zamar)**."
  - Specify the distribution for Mahalanobis distances in the "**MDs Distribution**." The default is "**Beta**."
  - Specify the "**Critical Alpha**," the cutoff for outliers. The default is "**0.05**."
  - Specify the "**Select Number of Iterations**." The default is "**10**."
  - Specify the "**MVT Trimming Percentage**" for the trimming process. The default is "**0.05**."
  - Specify whether or not to display the "**Correlation R Matrix**." The default is "**Do Not Display**."

- Specify the "**Intermediate Iterations**." The default is "**Do Not Display**."

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click the "**Graphs**" button for various options.



- Specify the "**Title for Index MD Plot**." This is an index plot of the robust distances obtained using the MVT estimates.

- Specify the "**Title for Distance-Distance Plot**." This is a plot of the classical Mahalanobis against the robust distances obtained using the MVT estimates.

- Specify the "**Title for Q-Q MD Plot**."

- Select the distribution required for the "**Q-Q Plot**" and the "**Graphics Critical Alpha**" for identifying the outliers.

  Note: *The "**Graphics Critical Alpha**" should match the "**Critical Alpha**" from the **Multivariate Outlier Options** window to obtain the same outliers.*

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the MVT procedure.

Output example: The data set “**BRADU.xls**” was used for the Huber method. It has 75 observations and four variables. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the Huber influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were then calculated.

Output for MVT outliers method.
Data Set used: Bradu.

	MVT Multivariate Outlier Analysis
User Selected Options	
Date/Time of Computation	1/8/2008 6:25:49 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Critical Alpha	0.05
Trimming Percentage	10%
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Display Correlation R Matrix	Do Not Display Correlation R matrix
Number of Iterations	10
Show Intermediate Results	Do Not Display Intermediate Results
Title for Index Plot	MVT Estimate
Title for Distance-Distance Plot	MVT Estimate
Title for QQ Plot	MVT Estimate
Graphics Critical Alpha	0.05
MDs Distribution	Beta

Number of Observations	75
Number of Selected Variables	4

Critical Values

Max Squared MD (0.05)	17.43
Individual Squared MD (0.05)	9.128
Multivariate Kurtosis(0.05)	25.2

Classical Mean Vector

y	x1	x2	x3
1.279	3.207	5.597	7.231

Classical Covariance S Matrix

y	x1	x2	x3
12.2	9.477	20.39	31.03
9.477	13.34	28.47	41.24
20.39	28.47	67.88	94.67
31.03	41.24	94.67	137.8

Determinant 1906.26080388262

Output for MVT outliers method (continued).

Eigenvalues for Classical Covariance S Matrix

Eval 1	Eval 2	Eval 3	Eval 4
0.914	1.688	5.538	223.1

Median Vector

y	x1	x2	x3
0.1	1.8	2.2	2.1

MAD/0.6745 Vector Representing Standard Deviation

y	x1	x2	x3
0.89	1.927	1.631	1.779

OKG Mean Vector

y	x1	x2	x3
1.258	-7.552	-6.052	2.521

Robust OKG (Maronna Zamar) Covariance S Matrix

y	x1	x2	x3
0.598	0.243	0.115	-0.231
0.243	2.856	0.108	0.241
0.115	0.108	2.175	0.00402
-0.231	0.241	0.00402	2.34
		Determinant	7.84553740487618

Robust OKG Eigenvalues

Eval 1	Eval 2	Eval 3	Eval 4
0.53	2.149	2.316	2.975

Final Weighted Mean Vector

y	x1	x2	x3
0.968	2.418	3.672	4.566

Final Covariance S Matrix

y	x1	x2	x3
9.88	8.16	17.43	26.43
8.16	7.891	14.78	22.53
17.43	14.78	32.71	48.33
26.43	22.53	48.33	74.4
		Determinant	43.7076080629344

Output for MVT outliers method (continued).

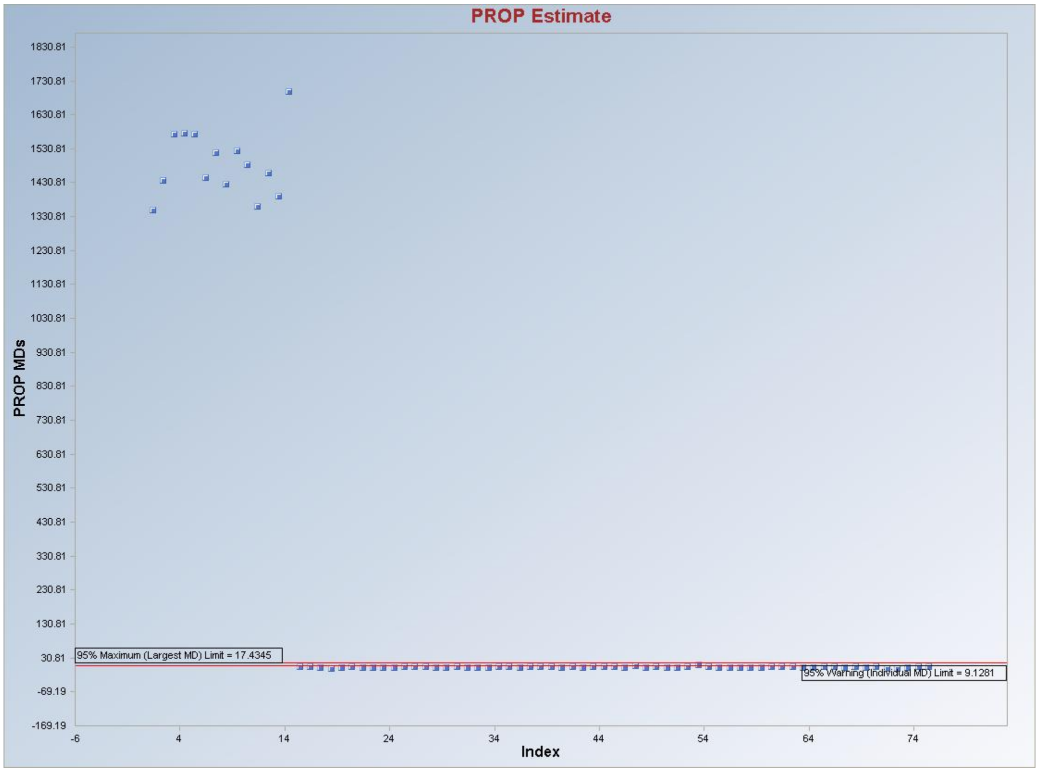
		Observations with Squared MDs greater than 17.43 may be considered as potential outliers!		
Observation Number	Classical Squared MD	Initial Squared MD	Final Squared MD	Final Weights
1	6.026	654.4	8.004	1
2	6.729	699.8	9.167	1
3	7.209	761.9	9.825	1
4	6.321	769.1	13.1	0
5	6.339	765.6	9.918	1
6	7.036	701.8	9.236	1
7	8.235	740.2	11.12	0
8	6.714	692.4	8.997	1
9	6.445	740.8	10.86	0
10	7.346	719.1	9.925	1
11	18.61	691.2	353.1	0
12	27.74	732.6	387.7	0
13	14.79	712.8	325.5	0
14	43.7	907	527.9	0
15	3.386	1.944	4.865	1
16	4.783	2.228	5.779	1
17	2.004	2.761	2.357	1
18	0.751	0.303	0.692	1
19	1.408	0.675	1.821	1
20	2.556	1.234	2.838	1
21	1.281	1.165	2.22	1
22	2.578	1.322	3.12	1
23	1.216	2.84	2.614	1
24	1.321	1.542	3.501	1
25	0.658	3.846	1.335	1
26	1.413	2.229	4.579	1
27	2.492	2.387	5.45	1
28	0.765	1.029	1.737	1
29	0.364	1.612	1.561	1
30	2.793	1.945	4.399	1
31	3.395	1.721	3.116	1
32	1.772	2.478	2.552	1
33	0.967	1.674	2.962	1
34	1.475	4.461	1.718	1

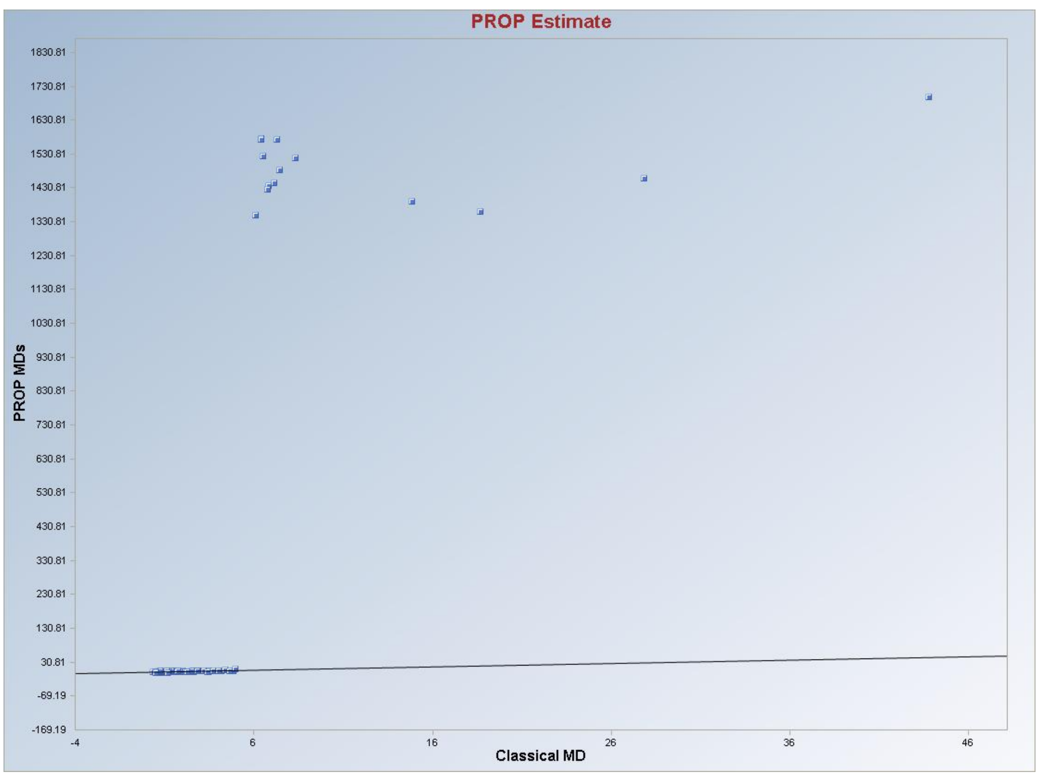
Output for MVT outliers method (continued).

41	3.139	1.386	3.732	1		
42	3.358	2.147	4.224	1		
43	4.108	2.526	6.397	1		
44	2.686	2.505	5.298	1		
45	1.278	3.85	1.564	1		
46	2.225	1.865	3.955	1		
47	4.332	5.896	5.06	1		
48	2.184	1.054	2.574	1		
49	2.552	1.422	3.11	1		
50	0.278	0.856	1.747	1		
51	1.858	1.936	3.817	1		
52	4.587	3.562	4.813	1		
53	4.889	4.616	8.281	1		
54	2.604	2.131	5.445	1		
55	1.622	1.159	2.062	1		
56	1.898	1.137	3.033	1		
57	0.773	2.385	3.679	1		
58	1.988	1.208	3.251	1		
59	0.463	0.523	1.106	1		
60	3.945	2.628	5.978	1		
61	2.806	1.957	3.985	1		
62	0.675	3.98	4.866	1		
63	1.757	2.344	2.945	1		
64	1.112	1.602	3.478	1		
65	1.336	1.909	3.353	1		
66	1.696	3.273	2.272	1		
67	0.445	1.752	1.194	1		
68	2.462	4.222	6.71	1		
69	1.154	2.636	2.238	1		
70	1.042	1.609	1.476	1		
71	0.413	0.114	0.331	1		
72	1.114	0.732	1.016	1		
73	2.207	1.311	2.601	1		
74	2.742	2.635	3.807	1		
75	3.632	2.62	5.231	1		
	Classical	Initial	Final			
Multivariate Kurtosis	53.97	101431	8820			

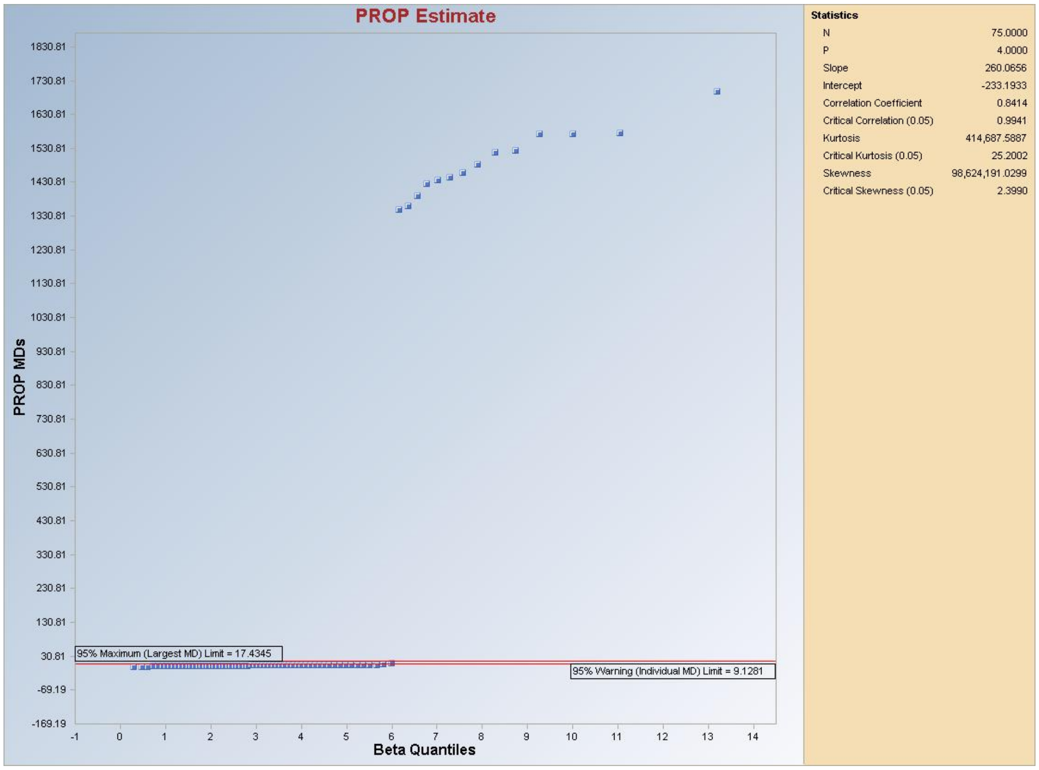
Results: Seven (7) (= 10% of 75) observations (4, 7, 9, 11, 12, 13 and 14), with squared distances greater than the Max (squared MD), were each given a zero (0) weight (hard rejection) and were considered to be outliers. In order to identify all of the 14 outliers, one has to use higher trimming percentages, such as 20%.

Output for MVT outliers method (continued).





Output for MVT outliers method (continued).

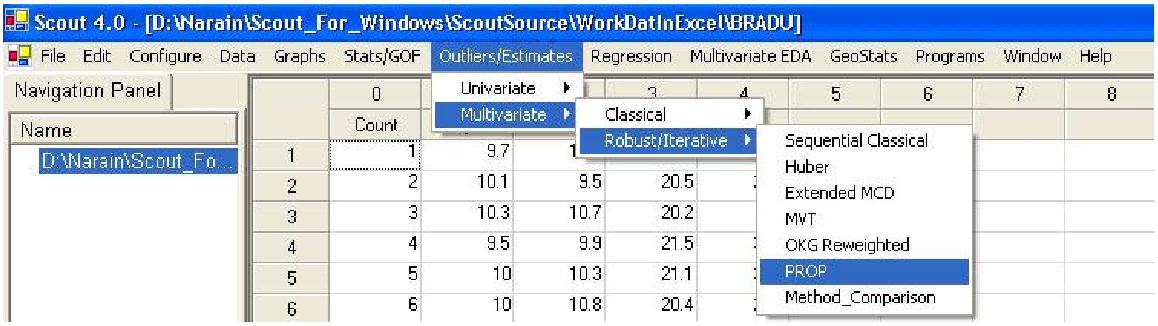


Graphical Interpretation: As before, the observations between the "**Warning (Individual MD) Limit**" and "**Maximum (Largest MD) Limit**" lines may also represent potential outliers.

Note*: Many times in practice, depending upon the trimming percentage value, this method may assign "0" weights (may find more outliers than actual outliers present in the data set) to some non-outlying observations with MDs smaller than the Max (MDs). In order to overcome this* problem*, at the final iteration, Scout compares the MVT MDs with the critical value Max (MDs), and the observations with MDs less than the critical value Max (MDs) are reassigned full unit weight. Estimates of the mean vector and the covariance matrix are then recomputed.*

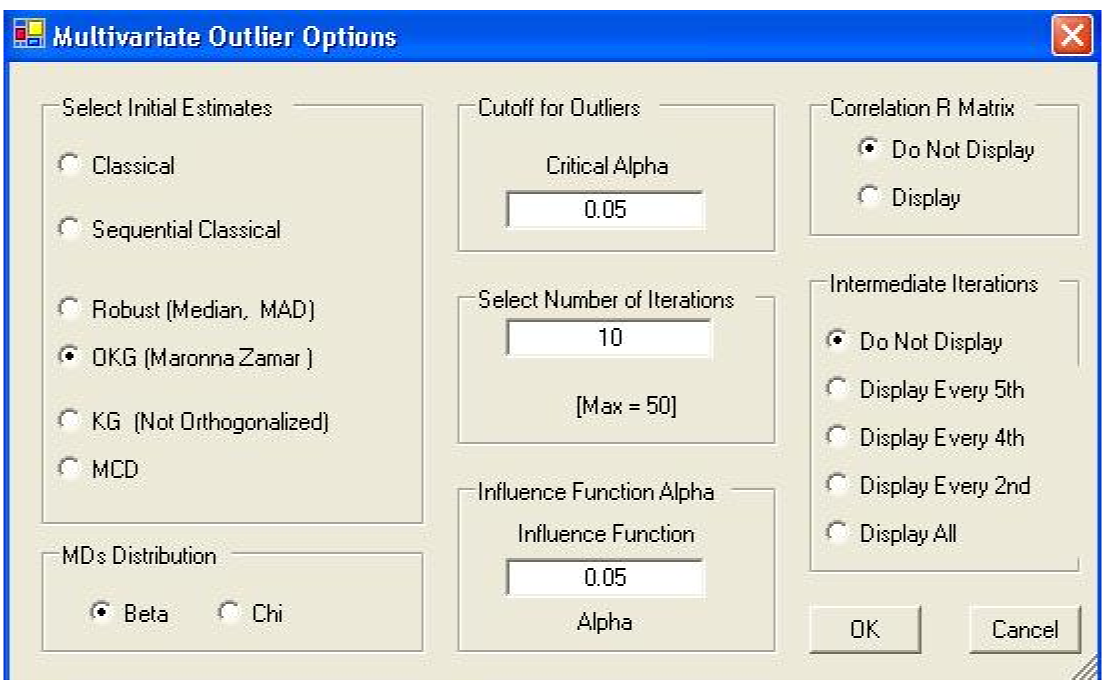
7.2.2.5 PROP

1. Click **Outlier/Estimates ► Multivariate ► Robust ►PROP.**



2. The "**Select Variables**" screen will appear (Section 3.4).

- Select the variables from the screen.
- If various groups are available and an analysis is to be done for those groups, then group variables can be selected from the drop-down list by clicking on the arrow below the "**Group by Variable**" button.
- Click the "**Options**" button for various options.



- Specify the initial estimates listed in the "**Select Initial Estimates**" to start the PROP iterative procedure. The default is "**OKG (Maronna Zamar)**."

- Specify the distribution for the Mahalanobis distances in the "**MDs Distribution**." The default is "**Beta**."

- Specify the "**Critical Alpha**," the cutoff for outliers. The default is "**0.05**."

- Specify the "**Select Number of Iterations**." The default is "**10**."

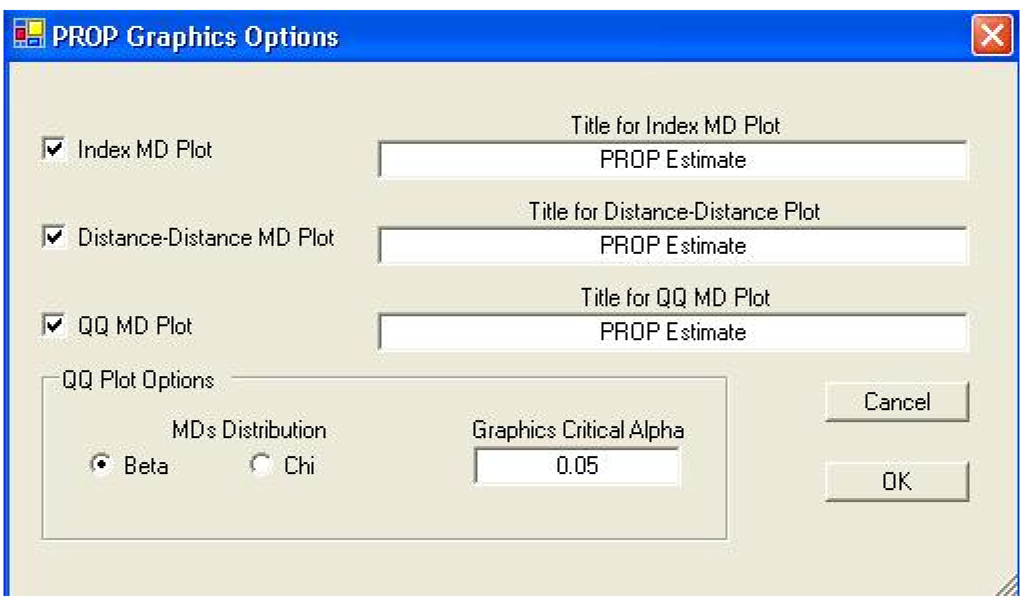
- Specify the "**Influence Function Alpha**" for the Huber weighting process. The default is "**0.05**."

- Specify whether or not to display the "**Correlation R Matrix**." The default is "**Do Not Display**."

- Specify "**Intermediate Iterations**." The default is "**Do Not Display**."

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click the "**Graphs**" button for various options.



- Specify the "**Title for Index MD Plot**." This is an index plot of the robust distances obtained using the PROP estimates.

- Specify the "**Title for Distance-Distance Plot**." This is a plot of the classical Mahalanobis against the robust distances obtained using the PROP estimates.

- Specify the "**Title for Q-Q MD Plot**."

- Select the distribution required for the "**Q-Q Plot Options**" and the "**Graphics Critical Alpha**" for identifying the outliers.

  Note*: The "****Graphics Critical Alpha****" should match the "****Critical Alpha****" from the outlier* ***Multivariate Outlier Options*** *window to obtain the same outliers.*

- Click "**OK**" to continue or "**Cancel**" to cancel the PROP procedure.

• Click "**OK**" to continue or "**Cancel**" to cancel the computing.

Output example: The data set "**BRADU.xls**" was used for the PROP method. It has 75 observations and four variables. The initial estimates of location and scale for each group were the median vector and the scale matrix obtained from the OKG method. The outliers were found using the PROP influence function and the observations were given weights accordingly. The weighted mean vector and the weighted covariance matrix were then calculated.

Output for PROP outliers method.
Data Set used: Bradu

	PROP Multivariate Outlier Analysis
User Selected Options	
Date/Time of Computation	1/8/2008 1:54:07 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF
Critical Alpha	0.05
Influence Function Alpha	0.05
Initial Estimates	Robust Median Vector and OKG (Maronna-Zamar) Matrix
Display Correlation R Matrix	Do Not Display Correlation R matrix
Distribution of Squared MDs	Beta Distribution
Number of Iterations	10
Show Intermediate Results	Do Not Display Intermediate Results
Title for Index Plot	PROP Estimate
Title for Distance-Distance Plot	PROP Estimate
Title for QQ Plot	PROP Estimate
Graphics Critical Alpha	0.05
MDs Distribution	Beta

Number of Observations	75
Number of Selected Variables	4

Critical Values	
Max Squared MD (0.05)	17.43
Individual Squared MD (0.05)	9.128
Multivariate Kurtosis(0.05)	25.2
Influence Fn. Squared MD (0.05)	9.128

Classical Mean Vector

y	x1	x2	x3
1.279	3.207	5.597	7.231

Classical Covariance S Matrix

y	x1	x2	x3
12.2	9.477	20.39	31.03
9.477	13.34	28.47	41.24
20.39	28.47	67.88	94.67
31.03	41.24	94.67	137.8

Determinant 1906.26080388262

Output for PROP outliers method (continued).

Eigenvalues for Classical Covariance S Matrix			
Eval 1	Eval 2	Eval 3	Eval 4
0.914	1.688	5.538	223.1

Median Vector			
y	x1	x2	x3
0.1	1.8	2.2	2.1

MAD/0.6745 Vector Representing Standard Deviation			
y	x1	x2	x3
0.89	1.927	1.631	1.779

OKG Mean Vector			
y	x1	x2	x3
1.258	-7.552	-6.052	2.521

Robust OKG (Maronna Zamar) Covariance S Matrix			
y	x1	x2	x3
0.598	0.243	0.115	-0.231
0.243	2.856	0.108	0.241
0.115	0.108	2.175	0.00402
-0.231	0.241	0.00402	2.34
		Determinant	7.84553740487618

Robust OKG Eigenvalues			
Eval 1	Eval 2	Eval 3	Eval 4
0.53	2.149	2.316	2.975

Final Weighted Mean Vector			
y	x1	x2	x3
-0.0776	1.544	1.787	1.679

Final Covariance S Matrix			
y	x1	x2	x3
0.315	0.0675	0.0111	-0.117
0.0675	1.129	0.0364	0.136
0.0111	0.0364	1.146	0.161
-0.117	0.136	0.161	1.057
		Determinant	0.388282776749715

Output for PROP outliers method (continued).

	Observations with Squared MDs greater than 17.43 may be considered as potential outliers!			
Observation Number	Classical Squared MD	Initial Squared MD	Final Squared MD	Final Weights
1	6.026	654.4	1350	1.861E-16
2	6.729	699.8	1437	5.659E-17
3	7.209	761.9	1574	9.216E-18
4	6.321	769.1	1578	8.734E-18
5	6.339	765.6	1574	9.190E-18
6	7.036	701.8	1446	4.962E-17
7	8.235	740.2	1521	1.843E-17
8	6.714	692.4	1428	6.359E-17
9	6.445	740.8	1524	1.774E-17
10	7.346	719.1	1483	3.026E-17
11	18.61	691.2	1360	1.616E-16
12	27.74	732.6	1459	4.184E-17
13	14.79	712.8	1393	1.033E-16
14	43.7	907	1699	1.893E-18
15	3.386	1.944	4.666	1
16	4.783	2.228	5.315	1
17	2.004	2.761	3.841	1
18	0.751	0.303	0.627	1
19	1.408	0.675	1.769	1
20	2.556	1.234	4.802	1
21	1.281	1.165	3.964	1
22	2.578	1.322	3.721	1
23	1.216	2.84	3.999	1
24	1.321	1.542	3.126	1
25	0.658	3.846	4.222	1
26	1.413	2.229	4.15	1
27	2.492	2.387	5.235	1
28	0.765	1.029	1.709	1
29	0.364	1.612	1.619	1
30	2.793	1.945	4.703	1
31	3.395	1.721	2.993	1
32	1.772	2.478	3.73	1
33	0.967	1.674	2.75	1
34	1.475	4.461	5.473	1

Output for PROP outliers method (continued).

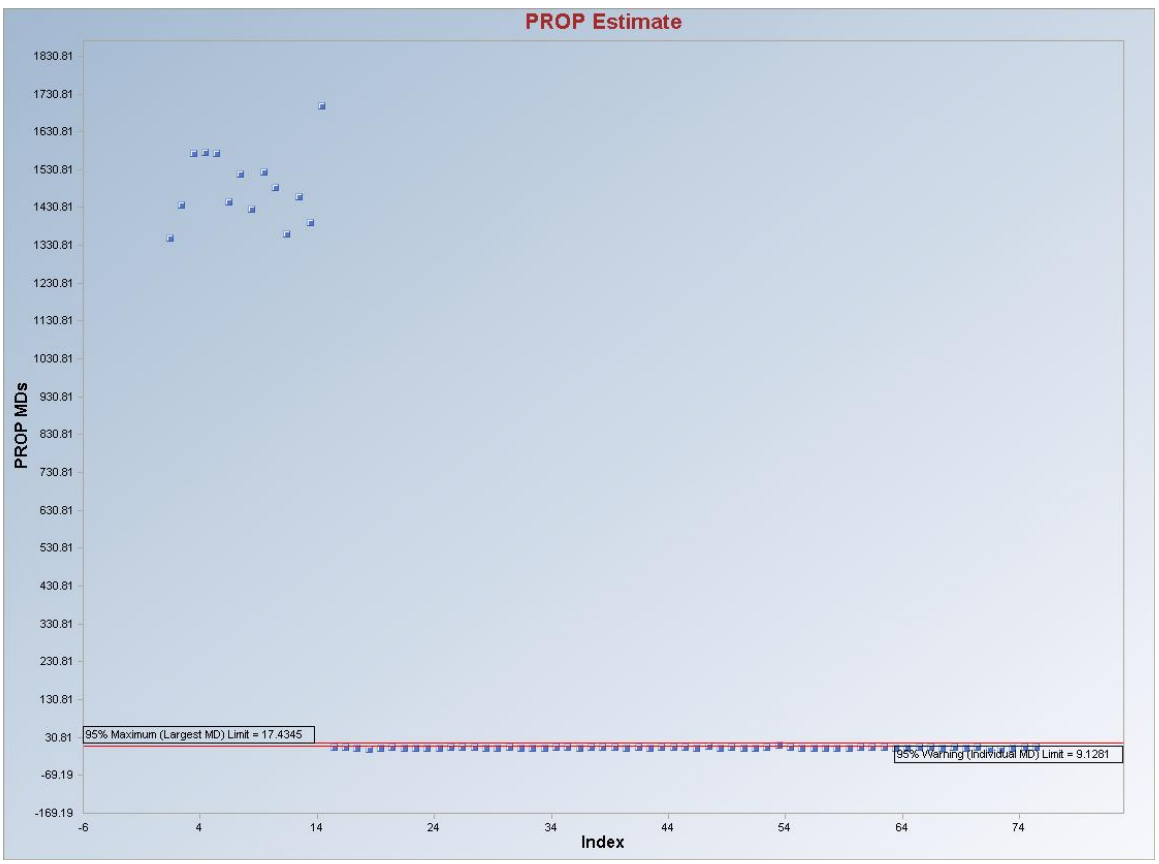
35	1.58	0.923	4.093	1			
36	1.008	2.086	3.372	1			
37	3.368	2.047	4.44	1			
38	0.889	2.826	4.44	1			
39	1.935	3.883	4.769	1			
40	1.237	1.463	1.735	1			
41	3.139	1.386	4.151	1			
42	3.358	2.147	3.949	1			
43	4.108	2.526	6.062	1			
44	2.686	2.505	4.854	1			
45	1.278	3.85	4.4	1			
46	2.225	1.865	3.711	1			
47	4.332	5.896	8.531	1			
48	2.184	1.054	3.712	1			
49	2.552	1.422	5.112	1			
50	0.278	0.856	2.346	1			
51	1.858	1.936	3.421	1			
52	4.587	3.562	5.498	1			
53	4.889	4.616	10.85	0.698			
54	2.604	2.131	4.937	1			
55	1.622	1.159	1.877	1			
56	1.898	1.137	2.77	1			
57	0.773	2.385	3.327	1			
58	1.988	1.208	3.164	1			
59	0.463	0.523	1.691	1			
60	3.945	2.628	5.543	1			
61	2.806	1.957	5.029	1			
62	0.675	3.98	5.127	1			
63	1.757	2.344	3.823	1			
64	1.112	1.602	3.936	1			
65	1.336	1.909	4.628	1			
66	1.696	3.273	4.468	1			
67	0.445	1.752	1.921	1			
68	2.462	4.222	6.464	1			
69	1.154	2.636	3.062	1			
70	1.042	1.609	5.021	1			
71	0.413	0.114	1.278	1			
72	1.114	0.732	0.872	1			
73	2.207	1.311	3.196	1			

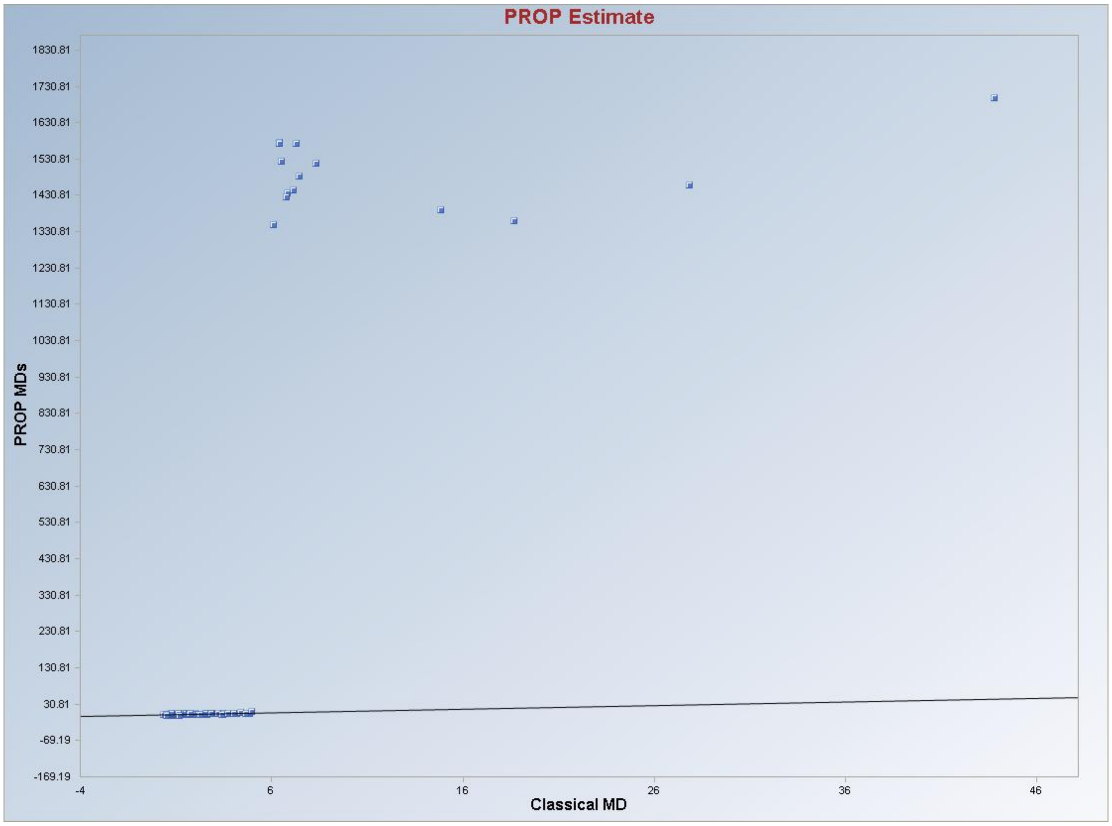
Output for PROP outliers method (continued).

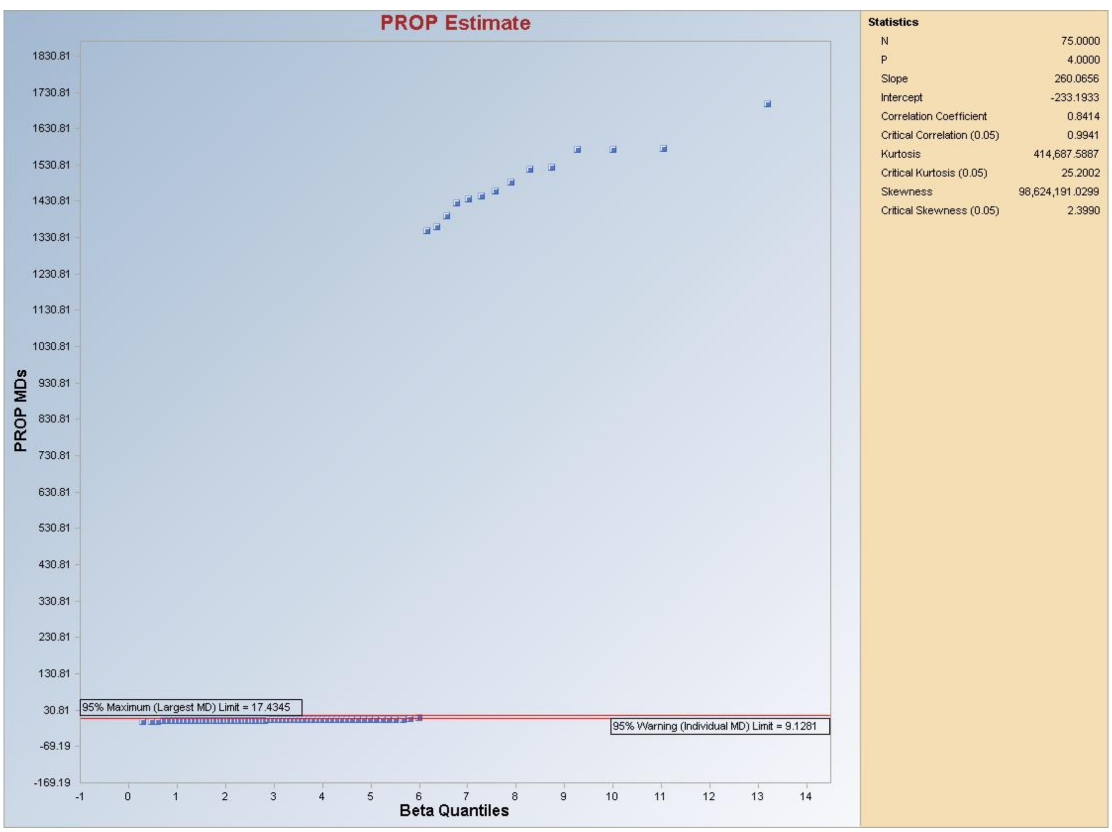
74	2.742	2.635	4.099	1			
75	3.632	2.62	5.414	1			
	Classical	Initial	Final				
Multivariate Kurtosis	53.97	101431	414688				

Results: Fourteen (14) observations (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14), with squared distances greater than the Max (squared MD), were each assigned an almost zero (0) weight (soft rejection) and may be considered to be outliers. Another observation (#53) also received a reduced (<1) weight.

Note: *PROP estimates with or without the 14 outliers are in close agreement with the classical estimates without the 14 outliers.*




PROP Estimate
PROP MDs
Classical MD


PROP Estimate
PROP MDs
Beta Quantiles
95% Maximum (Largest MD) Limit = 17.4345
95% Warning (Individual MD) Limit = 9.1281
Statistics
N 75.0000
P 4.0000
Slope 260.0656
Intercept -233.1933
Correlation Coefficient 0.8414
Critical Correlation (0.05) 0.9941
Kurtosis 414,687.5887
Critical Kurtosis (0.05) 25.2002
Skewness 98,624,191.0299
Critical Skewness (0.05) 2.3990

Graphical Interpretation: the observations (e.g., #53) between the "**Warning (Individual MD) Limit**" and the "**Maximum (Largest MD) Limit**" received a reduced (<1) weight.

Note 1*: If the initial estimate of the covariance matrix is not positive definite, then a warning message (in orange) is displayed, so that one of the other options, which yields a positive definite covariance matrix, can be selected. This is illustrated as follows.*

Data Set used: Stackloss.
Select Initial Estimates: Robust MAD/Median.

Classical Mean Vector

Stack-Loss	Air-Flow	Temp.	Acid-Conc
17.52	60.43	21.1	86.29

Classical Covariance S Matrix

Stack-Loss	Air-Flow	Temp.	Acid-Conc
103.5	85.76	28.15	21.79
85.76	84.06	22.66	24.57
28.15	22.66	9.99	6.621
21.79	24.57	6.621	28.71
		Determinant	62844.8794181548

Eigenvalues for Classical Covariance S Matrix

Eval 1	Eval 2	Eval 3	Eval 4
1.96	7.194	22.96	194.1

Median Vector

Stack-Loss	Air-Flow	Temp.	Acid-Conc
15	58	20	87

MAD/0.6745 Vector Representing Standard Deviation

Stack-Loss	Air-Flow	Temp.	Acid-Conc
5.93	5.93	2.965	4.448

Robust MAD Covariance S Matrix

Stack-Loss	Air-Flow	Temp.	Acid-Conc
35.17	85.76	28.15	21.79
85.76	35.17	22.66	24.57
28.15	22.66	8.792	6.621
21.79	24.57	6.621	19.78
		Determinant	171530.124177133

Robust MAD Eigenvalues

Eval 1	Eval 2	Eval 3	Eval 4
-50.94	-2.115	11.32	140.6

Initial Covariance Matrix is Not Positive Definite!
Please use other options to compute the Initial Covariance Matrix!

Note 2*: If any of the elements of the MAD/0.6745 vector is less than 10^{-5}, then a fix, called the IQR Fix, is used. In such cases, the variability measure, MAD/0.6745, is replaced by IQR/1.355. This is illustrated as follows.*

Classical Mean Vector

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
5.006	3.428	1.462	0.246

Classical Covariance S Matrix

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.124	0.0992	0.0164	0.0103
0.0992	0.144	0.0117	0.0093
0.0164	0.0117	0.0302	0.00607
0.0103	0.0093	0.00607	0.0111
		Determinant	2.11308767598396E-06

Eigenvalues for Classical Covariance S Matrix

Eval 1	Eval 2	Eval 3	Eval 4
0.00903	0.0268	0.0369	0.236

Median Vector

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
5	3.4	1.5	0.2

MAD/0.6745 Vector Representing Standard Deviation

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.297	0.371	0.148	0

Adjusted by IQR/1.35 MAD/0.6745 Vector

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.297	0.371	0.148	0.0741

IQR_Fix

Robust MAD Covariance S Matrix

sp-le~th-1	sp-width-1	pt-le~th-1	pt-width-1
0.0879	0.0992	0.0164	0.0103
0.0992	0.137	0.0117	0.0093
0.0164	0.0117	0.022	0.00607
0.0103	0.0093	0.00607	0.00549
		Determinant	1.27592503630283E-07

7.2.3 Method Comparisons

The Method Comparison module (available in the Outliers/Estimates drop-down menu) is a formal graphical method of comparing various classical and robust outlier identification methods incorporated in Scout. Specifically, selected classical and robust prediction and tolerance ellipsoids (contour ellipses) are drawn on two-dimensional scatter plots of selected variables. The main objective of this module is to compare the effectiveness of the various outlier methods included in Scout.

Those contour plots are displayed at the same two levels as the horizontal lines (warning limit and maximum limit) displayed on the Q-Q Plots of the MDs. The individual (Indv-MD) contour (prediction ellipsoid) corresponds to the inner ellipsoid given by the probability statement:

$$p\left(d_i^2 \le d_{ind}^2\right) \le (1-\alpha);\ \text{i} = 1, 2, \ldots, \text{n}.$$

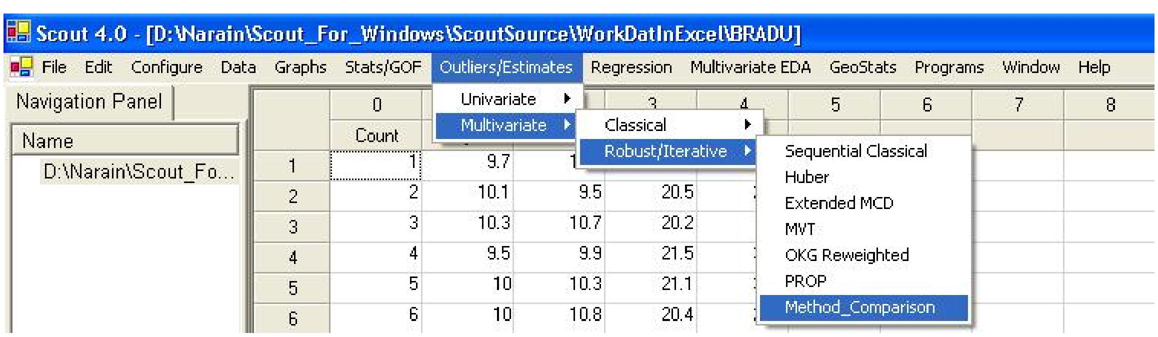
The Simultaneous (Max-MD) outer contour ellipsoid corresponds to a tolerance ellipsoid given by the probability statement:

$$p\left(d_i^2 \le d_\alpha^b\right) \le (1-\alpha);\ \text{i} = 1, 2, \ldots, \text{n}$$

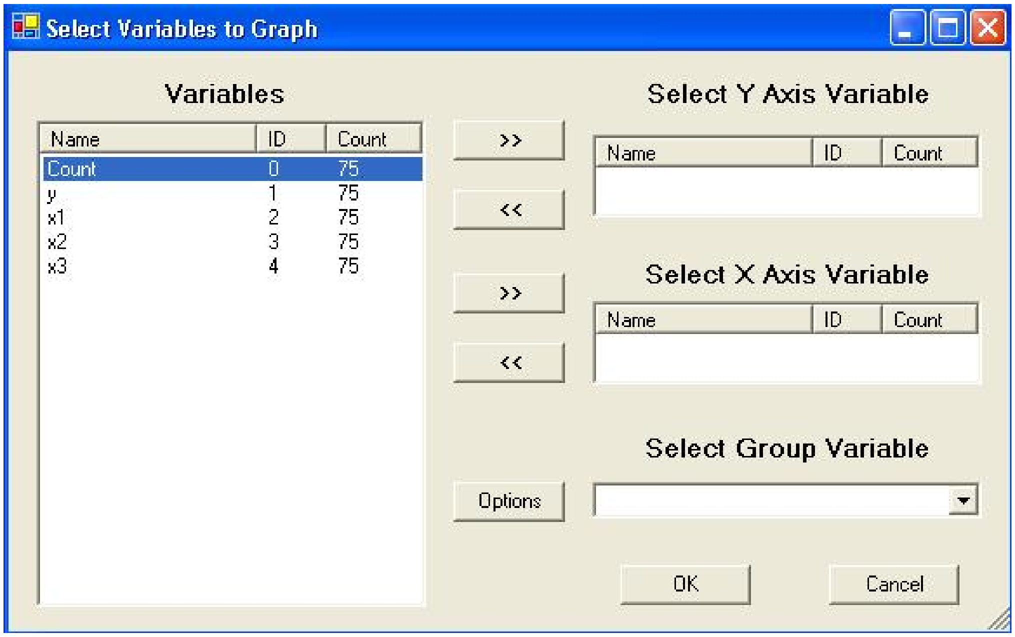
For details, refer to Singh (1993), and Singh and Nocerino (1995). The plots based upon the classical MDs accommodate outliers as a part of the same population and often fail to identify all of the outliers present in the data set. The outlying observations are more prominent on the contour plots obtained using robustified distanced and estimates. Observations falling outside the outer ellipse (tolerance ellipsoid) are outliers; whereas, the observations lying between the inner (prediction ellipsoid) and the outer ellipses may be also represent potential outliers.

If the data set is categorized by a group column, then the contour ellipses (prediction or tolerance ellipsoids) can be drawn separately for each of the groups included in the data set. The plots shown here are obtained using some well-known data sets.

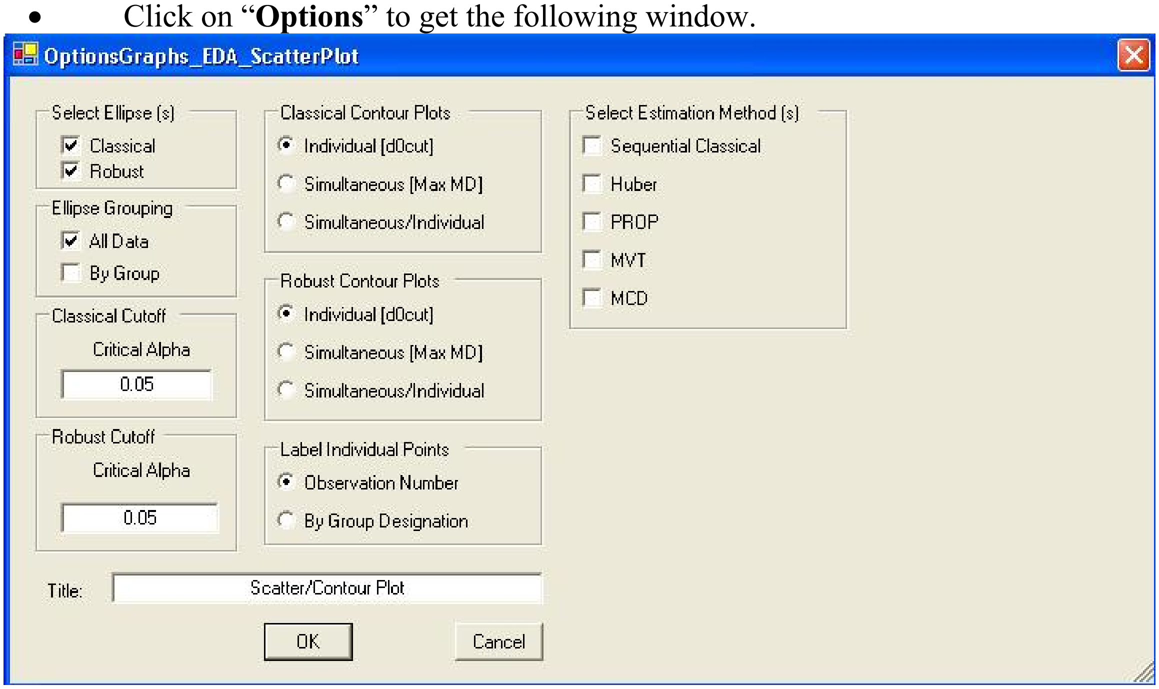
- Click **Outlier/Estimates ► Multivariate ► Robust ►Method Comparison.**



- The following variable selection screen appears.

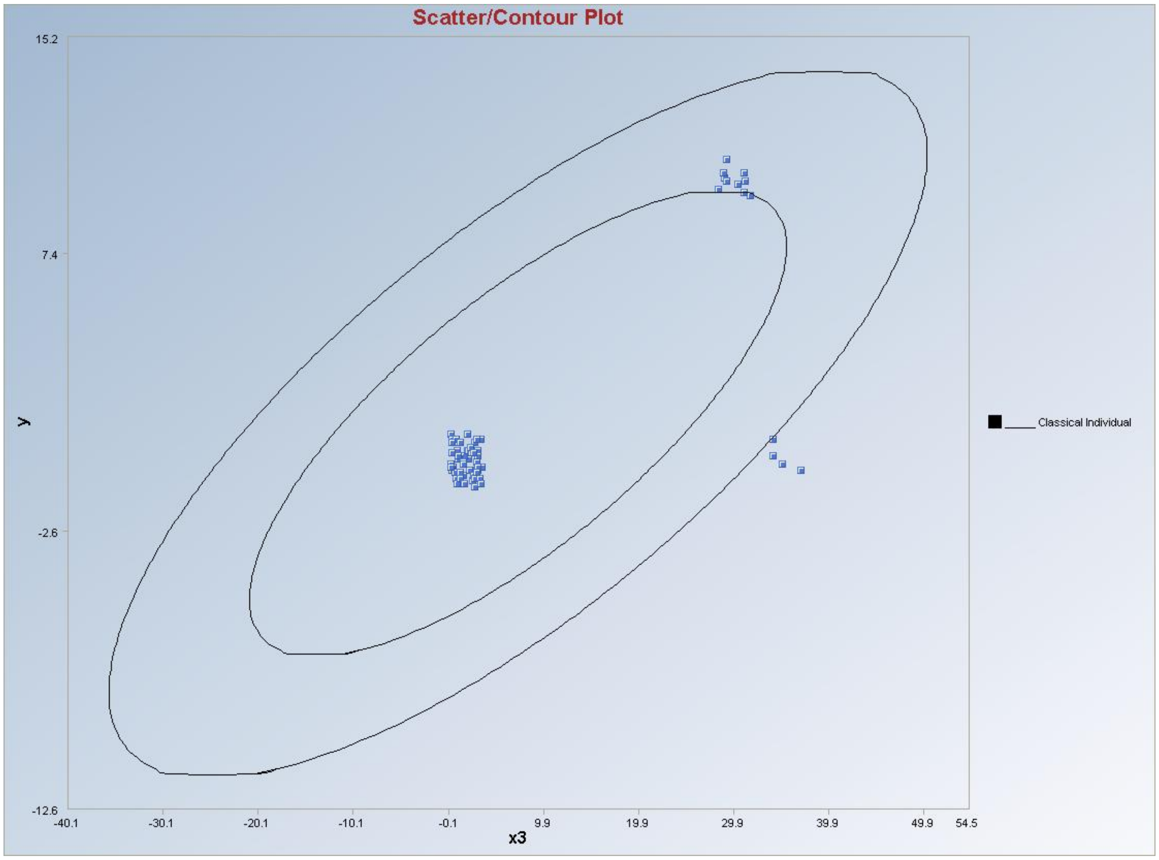


- Specify the variable for Y-axis under "**Select Y Axis Variable**."
- Specify the variable for X-axis under "**Select X Axis Variable**."
- Specify the group variable in the "**Select Group Variable**" drop-down if a group variable is present in the data set.
- Click on "**Options**" to get the following window.



- Click the "**Simultaneous/Individual**" radio button in the "**Classical Contour Plots**" box. Click "**OK**" to continue or "**Cancel**" to cancel the options window.

Data Set used: Bradu. Both classical prediction and simultaneous ellipsoids are drawn.

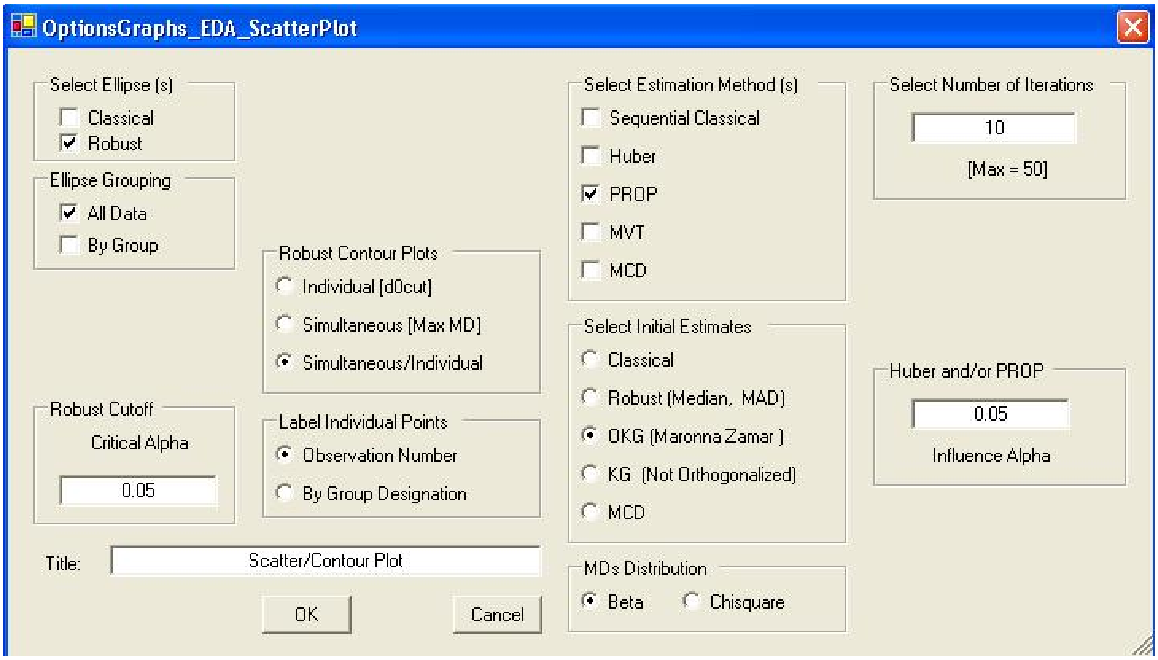


- Open the options window and uncheck the "**Classical**" option in the "**Select Ellipse(s)**" box.

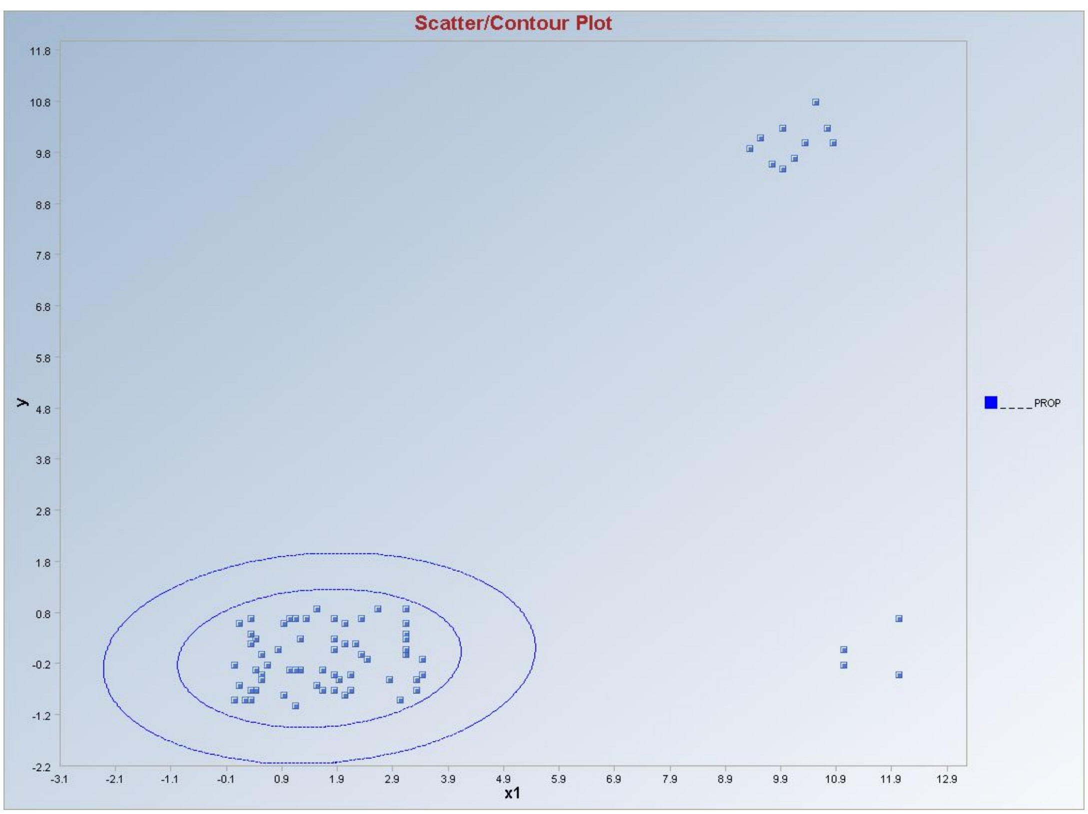
- Click the "**Simultaneous/Individual**" radio button in the "**Robust Contour Plots**" box.

- Specify any of the estimation methods from the "**Select Estimation Method(s)**" box; e.g., **PROP**.

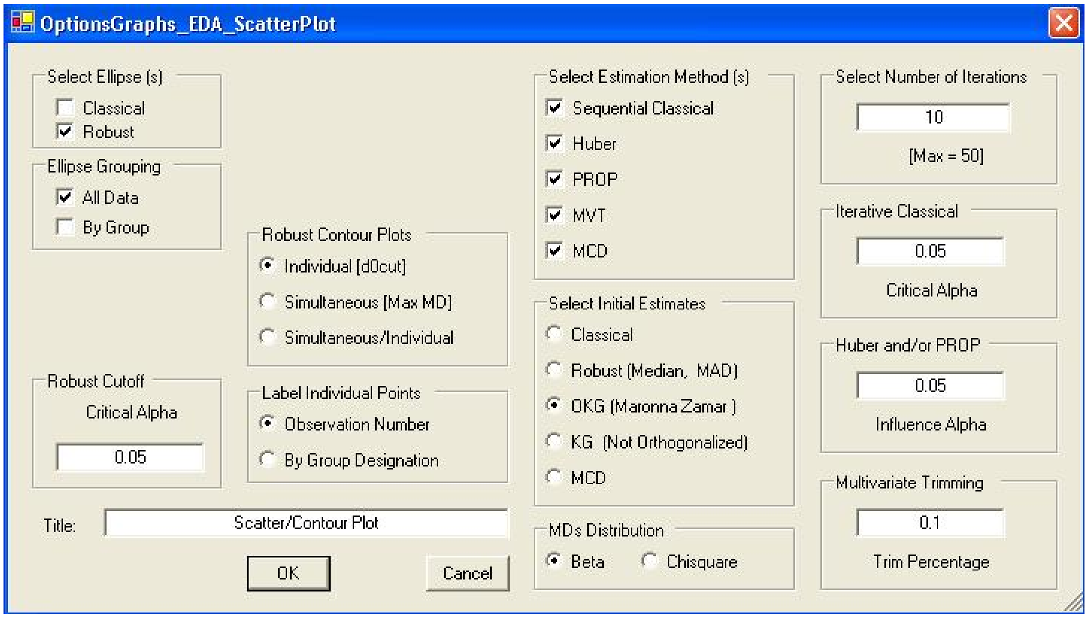
- Specify the preferred "**Robust Cutoff Critical Alpha**," "**Select Initial Estimates**," "**Select Number of Iterations**," "**MDs Distribution**" and "**Influence Alpha**." Click "**OK**" to continue for the graph.



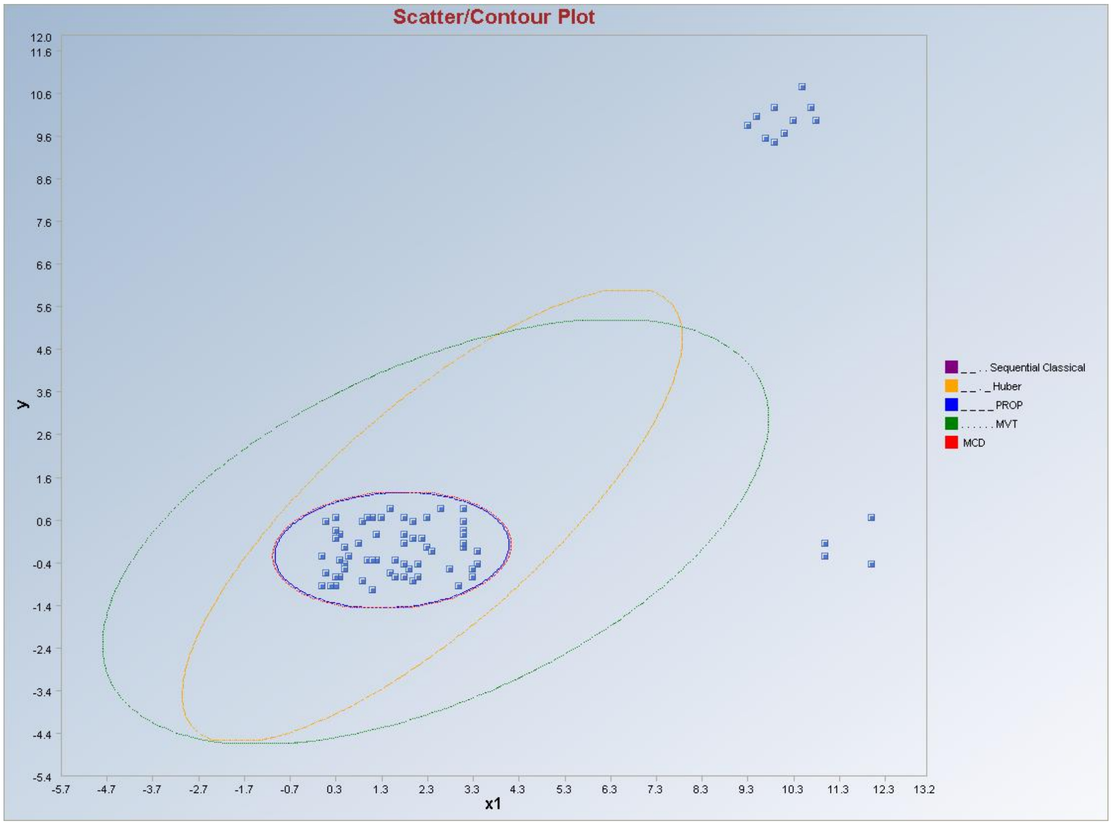
Data Set used: Bradu. Both inner and outer ellipsoids are drawn using the PROP method.



- Comparing various robust methods.

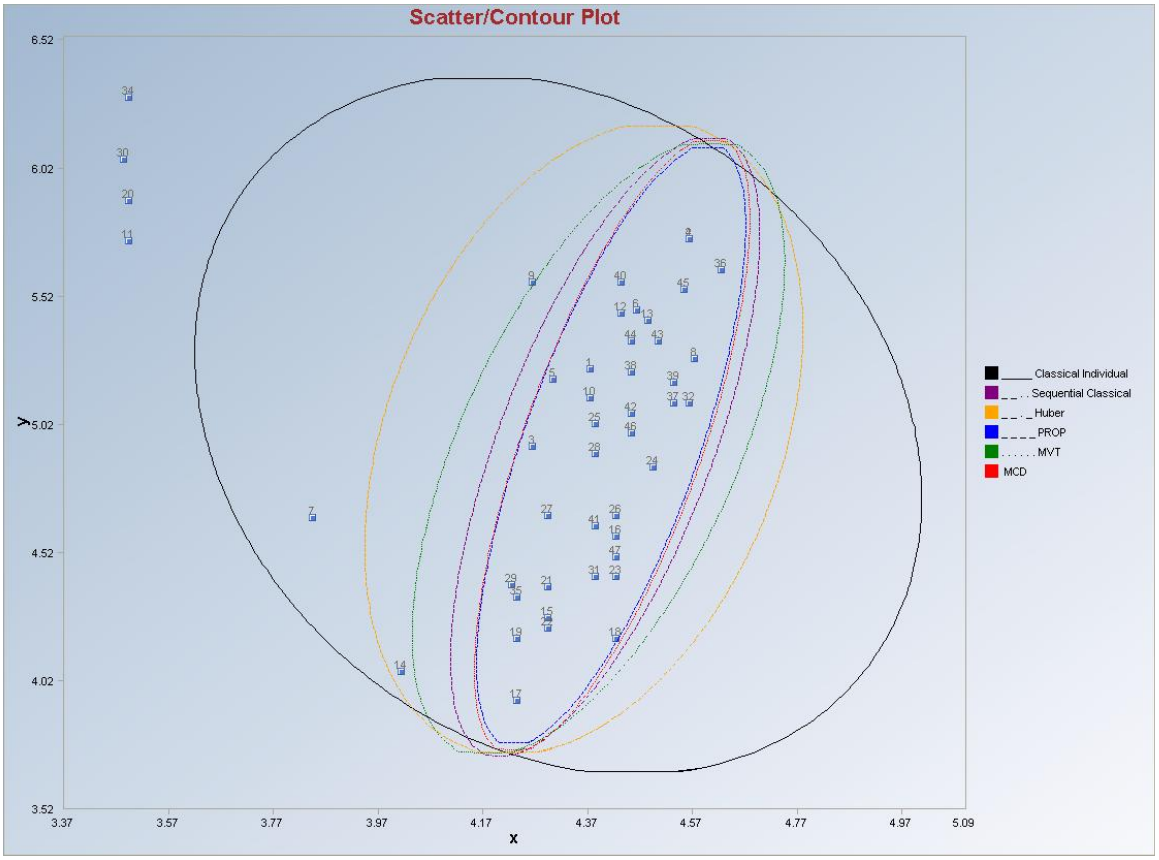


Data Set used: Bradu



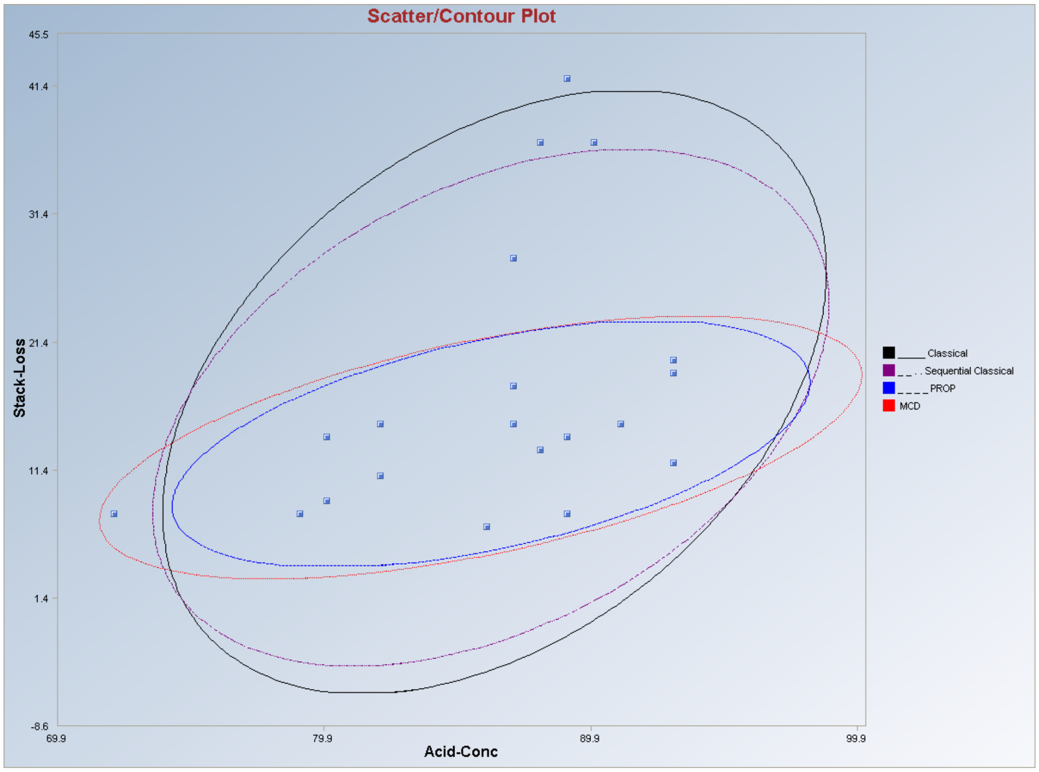
Sequential Classical (with robust OKG initial start); the PROP and MCD ellipses overlap each other.
***Note**: The "**Select Initial Estimates**," "**MDs Distribution**," "**Number of Iterations**" and "**Robust Cutoff**" remain the same for the sequential classical, Huber, PROP and MVT methods.*

Data Set used: Star Cluster.

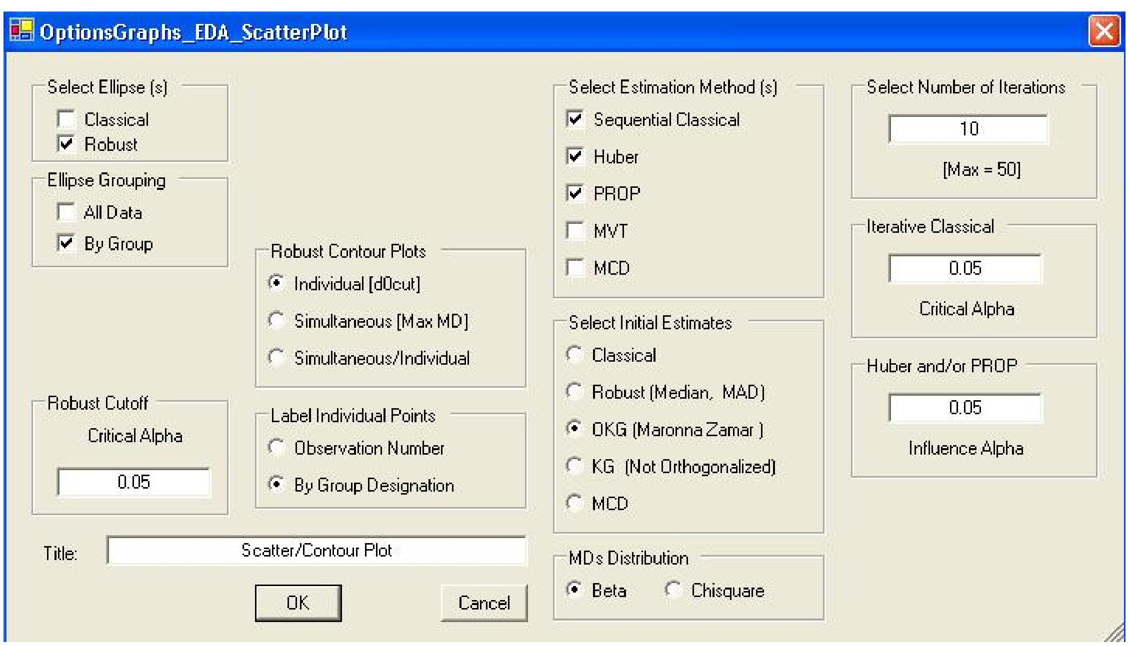


***Note**: In this example, all of the methods (including the Huber method), except for the classical method, identified the four main outliers present in the data set.*

Data Set used: Stackloss.



- Example with Group variable in the variable selection screen.



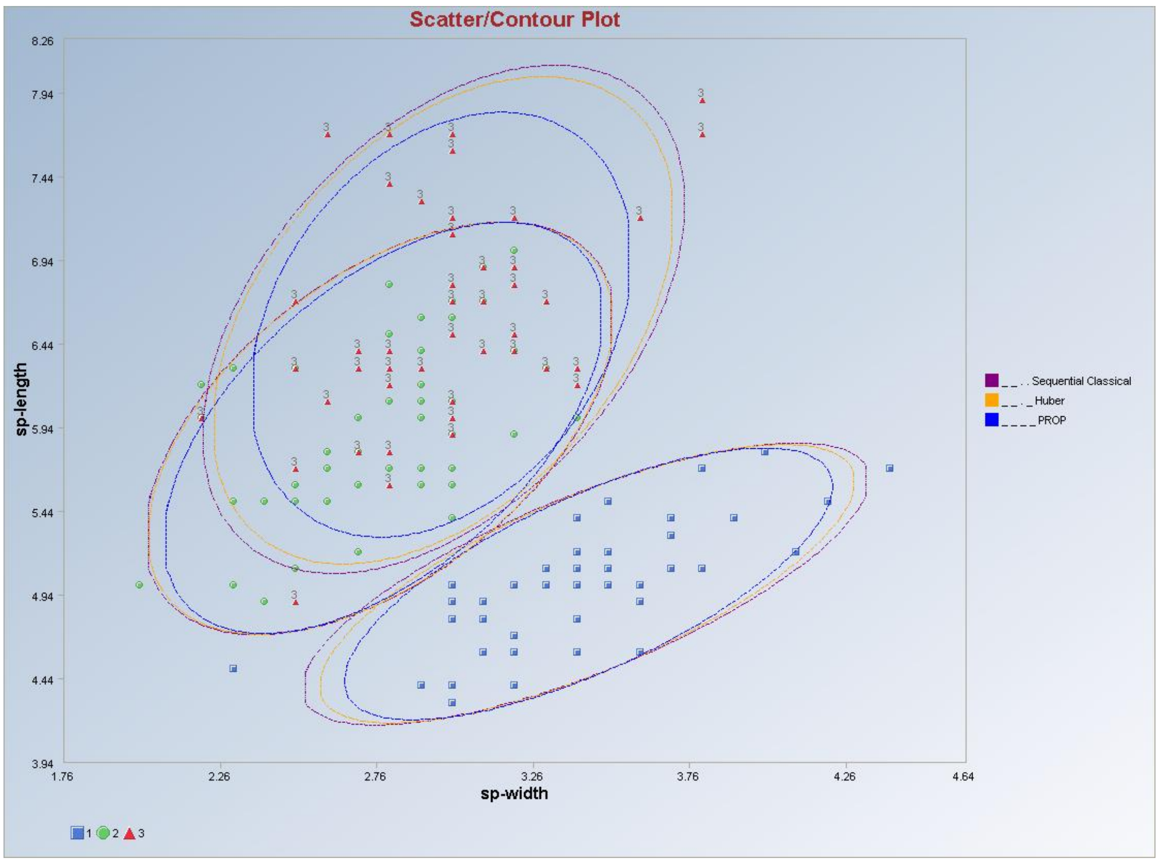
- Check “**By Group**” in the “**Ellipse Grouping**” box and uncheck “**All Data**.”

- Click on the “**By Group Designation**” radio button in the “**Label Individual Points**” box.

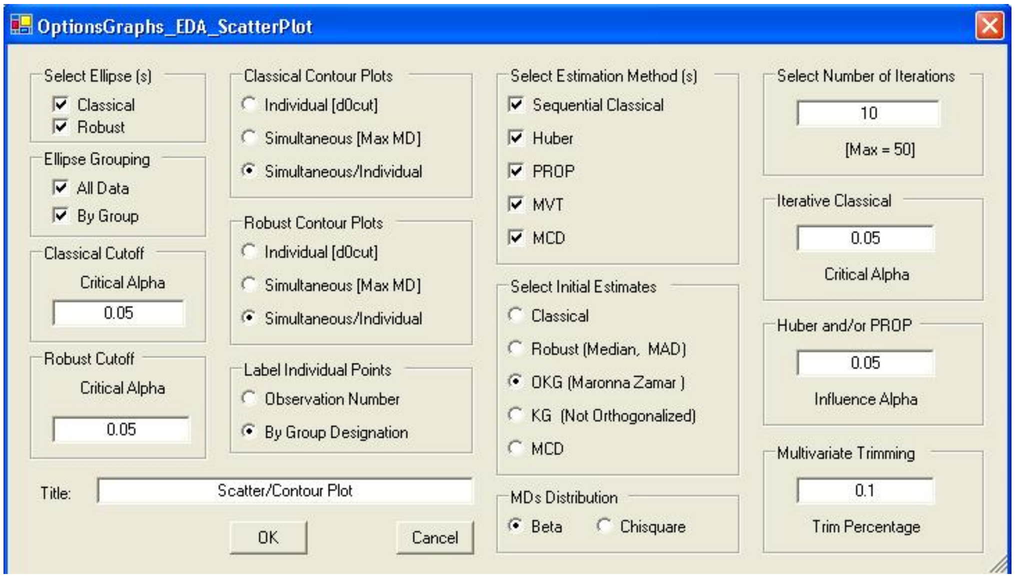
- Specify the preferred contour plots and the estimation methods.

- Click “**OK**” to continue or “**Cancel**” to cancel the options.

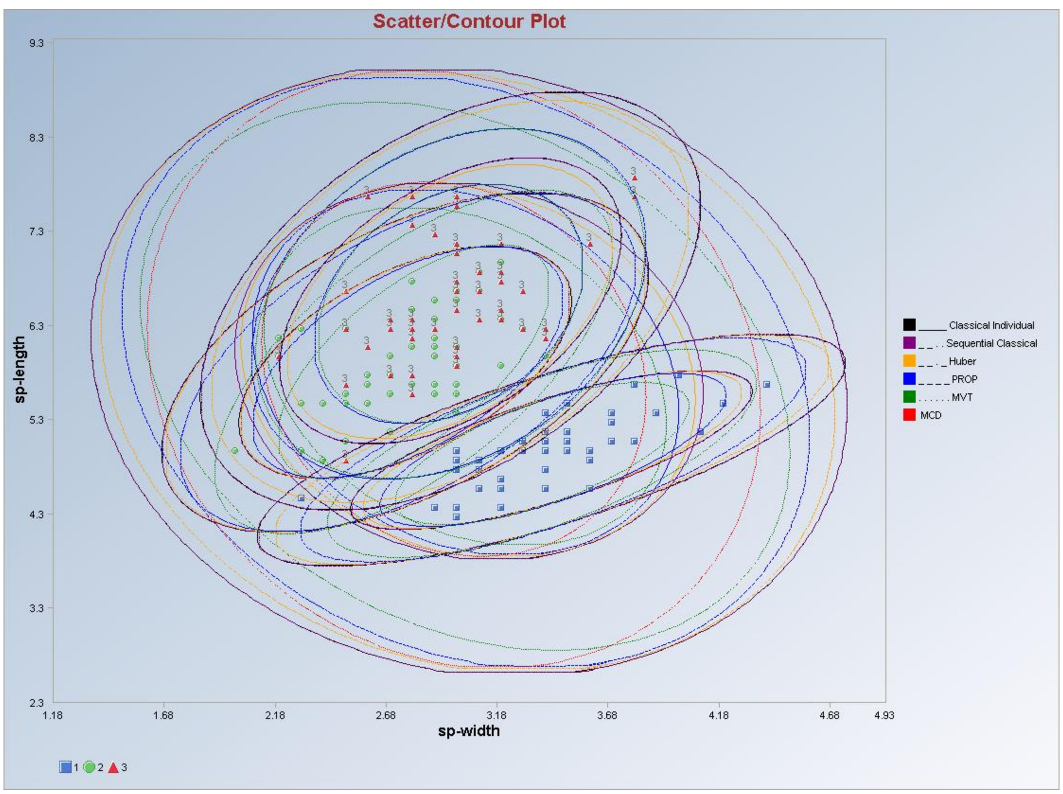
Data Set used: Fulliris (Fisher 1936 data set with 3 species).



Note: The user may select all of the options available in the options window. But, this selection will result in a busy (with overlapping ellipses) and cluttered graph which is difficult to understand. The user should select useful options from all available options.



Data Set used: Fulliris (Fisher 1936 data set with 3 species).



References

Alqallaf, F.A., Konis, K.P., Martin, R.D., and Zamar, R.H. (2002). "Scalable Robust Covariance and Correlation Estimates for Data Mining." In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Edmonton.

Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data, New York, J. Wiley and Sons.

Bernholt, T., and Fischer, P. (2004). "The Complexity of Computing the MCD-Estimator," Theoretical Computer Science, 326, 383-398.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975). "Robust Estimation and Outlier Detection with Correlation Coefficients," Biometrika, 62, 531-545.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981). "Robust Estimation of Dispersion Matrices and Principal Components," Journal of the American Statistical Association, 76, 354-362.

Dixon, W.J. (1953). "Processing Data for Outliers." Biometrics 9: 74-89.

Fung, W. (1993). "Unmasking Outliers and Leverage Points: a Confirmation," Journal of the American Statistical Association, 88, 515-519.

Ferguson, T.S. (1961a). "On the Rejection of Outliers," Proc., Fourth Berkeley Symposium. Math. Statist., 1, 253-287.

Ferguson, T.S. (1961b). "Rules of Rejection of Outliers," Rev., Inst. Int. de Statist, 3, 29-43.

Garner, F.C., Stapanian, M.A., and Fitzgerald, K.E. (1991a). "Finding Causes of Outliers in Multivariate Environmental Data," Journal of Chemometrics, 5, 241-248.

Garner, F.C., Stapanian, M.A., Fitzgerald, K.E., Flatman, G.T., and Englund, E.J. (1991b). "Properties of Two Multivariate Outlier Tests," Communications in Statistics, 20, 667-687.

Grubbs, F.E. (1950). "Sample Criterion for Testing Outlying Observations," Ann. Math. Statist., 21, 27-58.

Gnanadesikan, R., and Kettenring, J.R. (1972). "Robust Estimates, Residuals, and Outlier Detection with Multi-response Data," Biometrics, 28, 81-124.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, New York.

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984). "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," Technometrics, 26, 197-208.

Huber, P.J. (1981). Robust Statistics, John Wiley and Sons, NY.

Hubert, M., Rousseeuw, P.J., and Vanden Branden, K. (2005). "ROBPCA: A New Approach to Robust Principal Component Analysis," Technometrics, 47, 64-79.

Kafadar, K. (1982). "A Biweight Approach to the One-Sample Problem," Journal of the American Statistical Association, 77, 416-424.

Lax, D.A. (1985). "Robust Estimators of Scale: Finite Sample Performance in Long-Tailed Symmetric Distributions," Journal of the American Statistical Association, 80, 736-741.

Mardia, K.V. (1970). "Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies," Biometrika, 57, 519-530.

Mardia, K.V. (1974). "Applications of Some Measures of Multivariate Skewness Kurtosis in Testing Normality and Robustness Studies," Sankhya, 36, 115-128.

Maronna, R.A., and Zamar, R.H. (2002). "Robust Estimates of Location and Dispersion for High-Dimensional Data sets," Technometrics, 44, 307-317.

Mehrotra, D.V. (1995). "Robust Elementwise Estimation of a Dispersion Matrix," Biometrics, 51, 1344-1351.

Mosteller, F., and Tukey, J.W. (1977). Data Analysis and Regression, Addison-Wesley Reading, MA.

Pena, D., and Prieto, F.J. (2001). "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," Technometrics, 286-299.

ProUCL 4.00.04. (2009). "ProUCL Version 4.00.04 User Guide." The software ProUCL 4.00.04 can be downloaded from the web site at: http://www.epa.gov/esd/tsc/software.htm.

ProUCL 4.00.04. (2009). "ProUCL Version 4.00.04 Technical Guide." The software ProUCL 4.00.04 can be downloaded from the web site at: http://www.epa.gov/esd/tsc/software.htm.

Rocke, D.M., and Woodruff, D.L. (1996). “Identification of Outliers in Multivariate Data,” Journal of the American Statistical Association, 91, 1047-1061.

Rocke, D.M., and Woodruff, D.L. (1997). “Robust Estimation of Multivariate Location and Shape,” Journal of Statistical Planning and Inference, 57, pp. 245-255.

Rousseeuw, P.J., and van Zomeren, B.C. (1990). “Unmasking Multivariate Outliers and Leverage Points,” Journal of the American Statistical Association, 85, 633-651.

Rosner, B. (1975). “On Detection of Many Outliers,” Technometrics, 17, 221-227.

Rousseeuw, P.J., and Leroy, A.M. (1987). Robust Regression and Outlier Detection, John Wiley and Sons, NY.

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

Schwager, S.J., Margolin, B.H. (1982). Detection of Multivariate Normal Outliers, Ann. Statist., 10, 943-954.

Scout. (2002). A Data Analysis Program, Technology Support Project, USEPA, NERL-LV, Las Vegas, Nevada.

Scout. (2008). Technical Guide under preparation.

Singh, A. (1993). Omnibus Robust Procedures for Assessment of Multivariate Normality and Detection of Multivariate Outliers, In Multivariate Environmental Statistics, Patil G.P. and Rao, C.R., Editors, pp. 445-488, Elsevier Science Publishers.

Singh, A., and Nocerino, J.M. (1995). Robust Procedures for the Identification of Multiple Outliers, Handbook of Environmental Chemistry, Statistical Methods, Vol. 2. G, pp. 229-277, Springer Verlag, Germany.

Sinha, B.K. (1984). “Detection of Multivariate Outliers in Elliptically Symmetric Distributions,” Anal. Statist., 12, 1558-1565.

Stefanski, L.A., and Boos, D.D. (2002). “The Calculus of M–estimators,” The American Statistician, 56, 29-38.

Tukey, J.W. (1977). Exploratory Data Analysis, Addison-Wesley Publishing Company, Reading, MA.

Wilks, S.S. (1963). “Multivariate Statistical Outliers,” Sankhya, 25, 407-426.

Chapter 8

QA/QC

Issues related to the reliability of data are often grouped under the general heading of "quality assurance and quality control" (QA/QC), a description that captures the idea that data quality can not only be documented but can also be controlled through appropriate practices and procedures. Even with the most stringent and costly controls, data will never be perfect: errors are inevitable as samples are collected, prepared and analyzed.

One goal of QA/QC is to quantify these errors so that subsequent statistical analysis and interpretation can take them into account. A second goal is to monitor the errors so that spurious or biased data can be recognized and, if possible, corrected. A third goal is to provide information that can be used to improve sampling practices and analytical procedures so that the impact of errors can be minimized. Scout offers QA/QC methods for data with and without non-detects. Kaplan-Meier (KM) estimates of mean and standard deviation are used for data with non-detects.

Scout also allows the user to test the behavior of "Site/Test" data against "Background/Training" data. In this module the statistics and estimates are computed using the "Background/Training" data and then graphs and the charts are produced for the whole data set which is inclusive of "Background/Training" data and "Site/Test" data. The important requirement for this module is that there should be a column which indicates the various groups which can be considered as the "Site/Test" data.

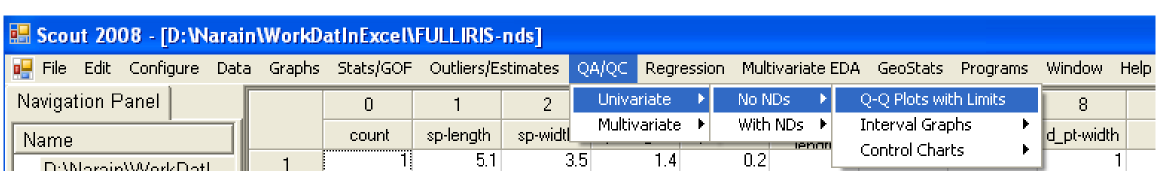
8.1 Univariate QA/QC

Scout offers several univariate procedures to achieve the goals specified above. They include Q-Q Plots with Limits, Interval Graphs and Control Charts. Classical and robust methods have been incorporated in this module.

8.1.1 No Non-detects

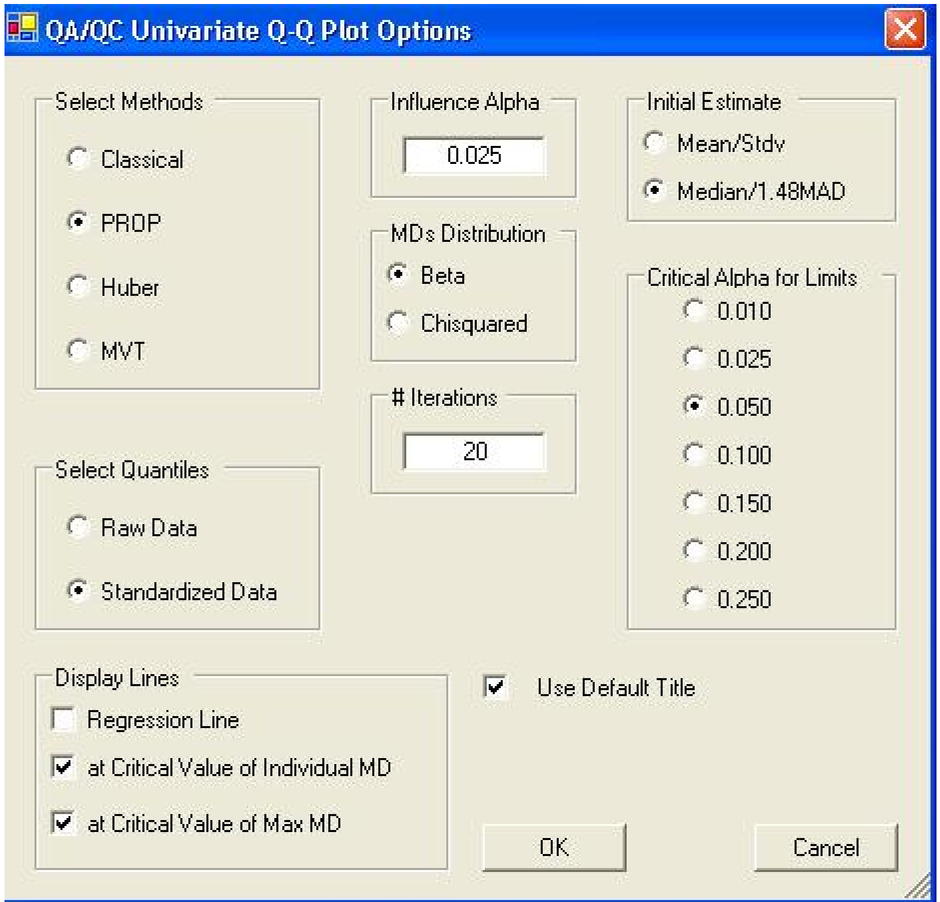
8.1.1.1 Q-Q Plots with Limits

1. Click on **QA/QC ► Univariate ► No NDs ► Q-Q Plots with Limits.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

- Click on the "**Options**" button for the options window.



  - Specify the method for computing the quantiles in "**Select Methods**." The default method is "**PROP**."

  - The robust methods need various input parameters like "**Influence Alpha**" or "**Trimming Percentage**," "**Initial Estimates**," "**MDs Distribution**," and "**# Iterations.**"

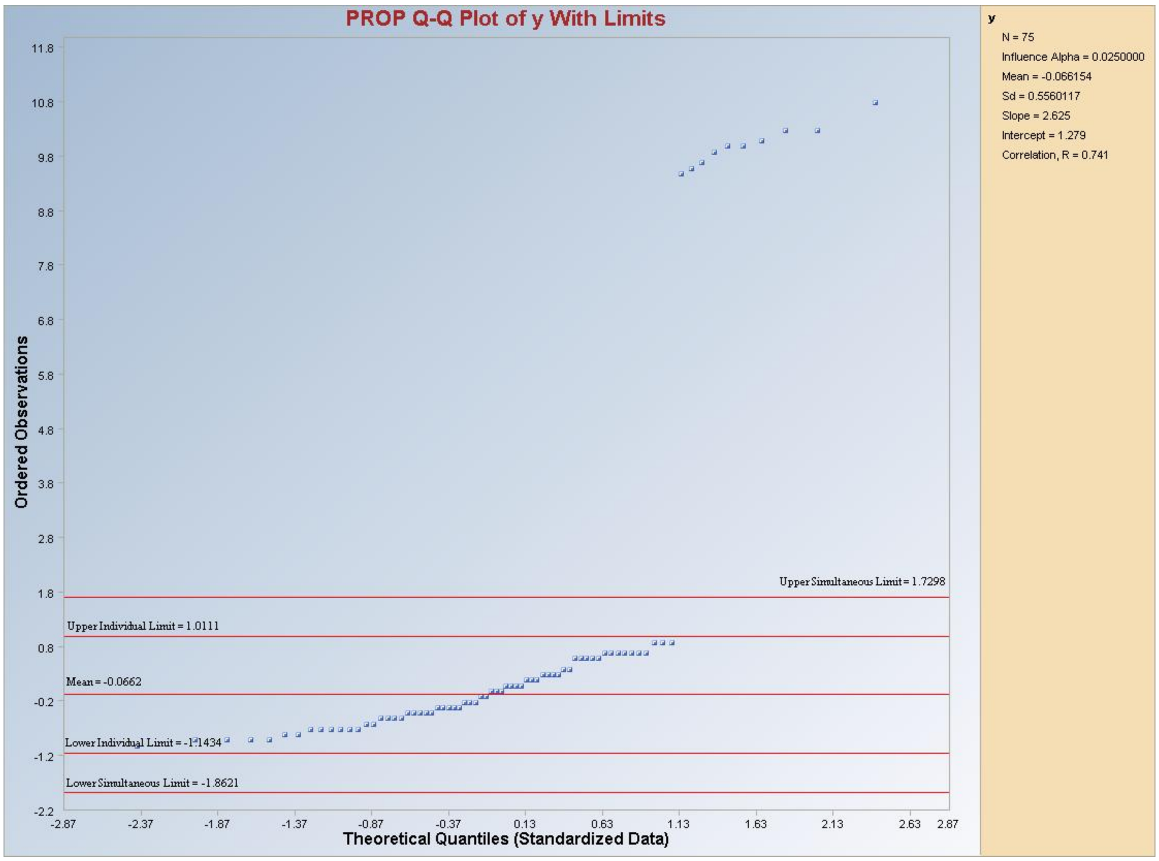
  - Specify the "**Critical Alpha for Limits**" for identifying the outliers. Default is "**0.05**."

  - Specify the quantiles for the X-axis using the "**Select Quantiles**" option and options for displaying the regression lines.

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the Q-Q Plots with limits.

Output example: The data set "**Bradu.xls**" was used for the Q-Q Plot. The options used were the default options.

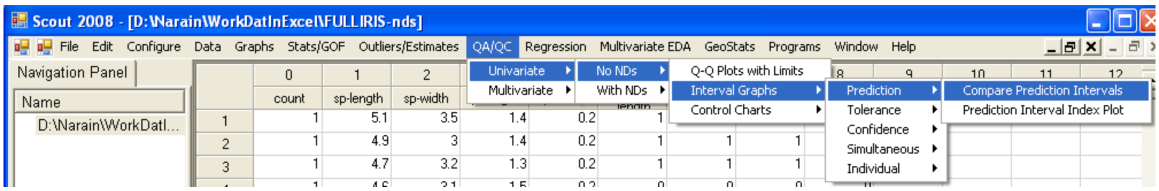


***Note**: The observations outside the simultaneous limits are considered as outliers.*

8.1.1.2 Interval Graphs

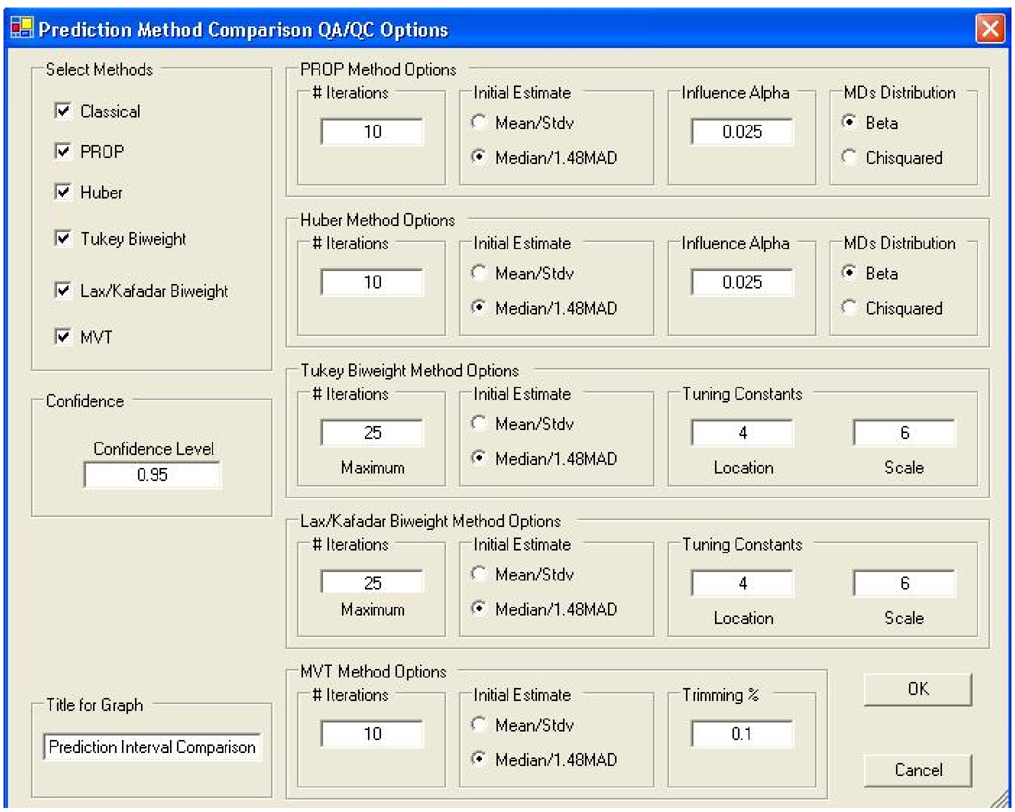
8.1.1.2.1 Compare Intervals

1. Click on **QA/QC ► Univariate ► No NDs ► Interval Graphs ►Prediction, Tolerance, Confidence, Simultaneous or Individual ► Compare Intervals.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

   - Click on the "**Options**" button for the options window.

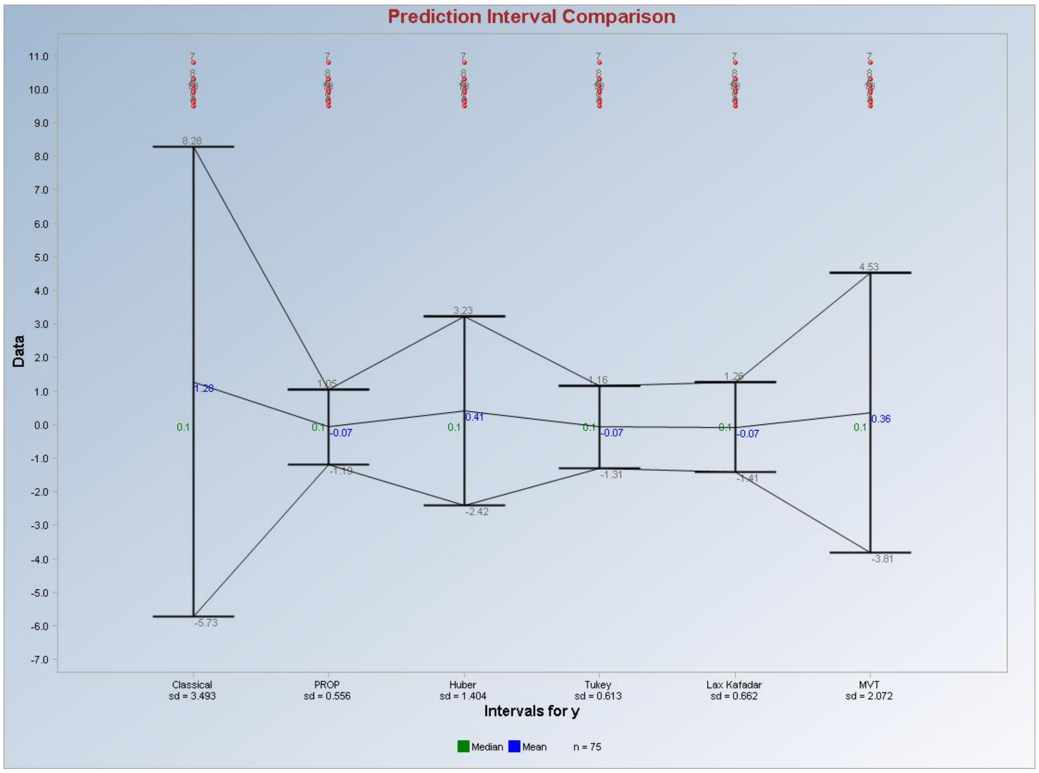


      - Select the methods to compare in "Select Methods" box. By default, all methods are selected.
      - Specify the various input parameters for the selected methods.
      - Click "**OK**" to continue or "**Cancel**" to cancel the options.

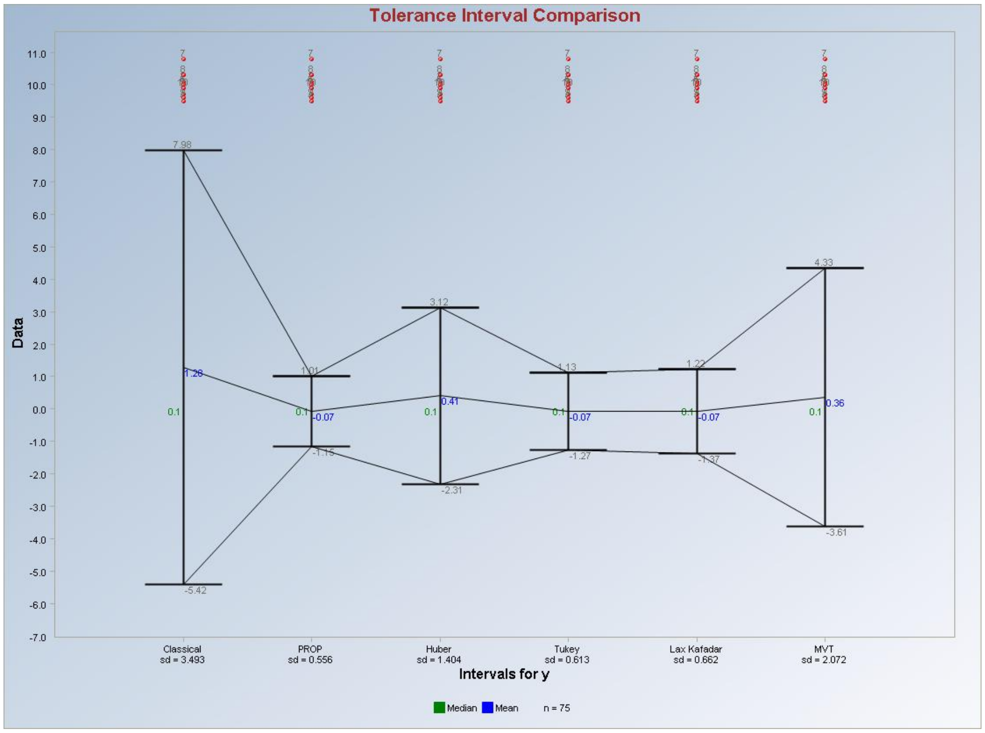
- Click on "**OK**" to continue or "**Cancel**" to cancel the intervals comparison.

Output example: The data set "**Bradu.xls**" was used for the Interval Comparison. The options used were the default options.

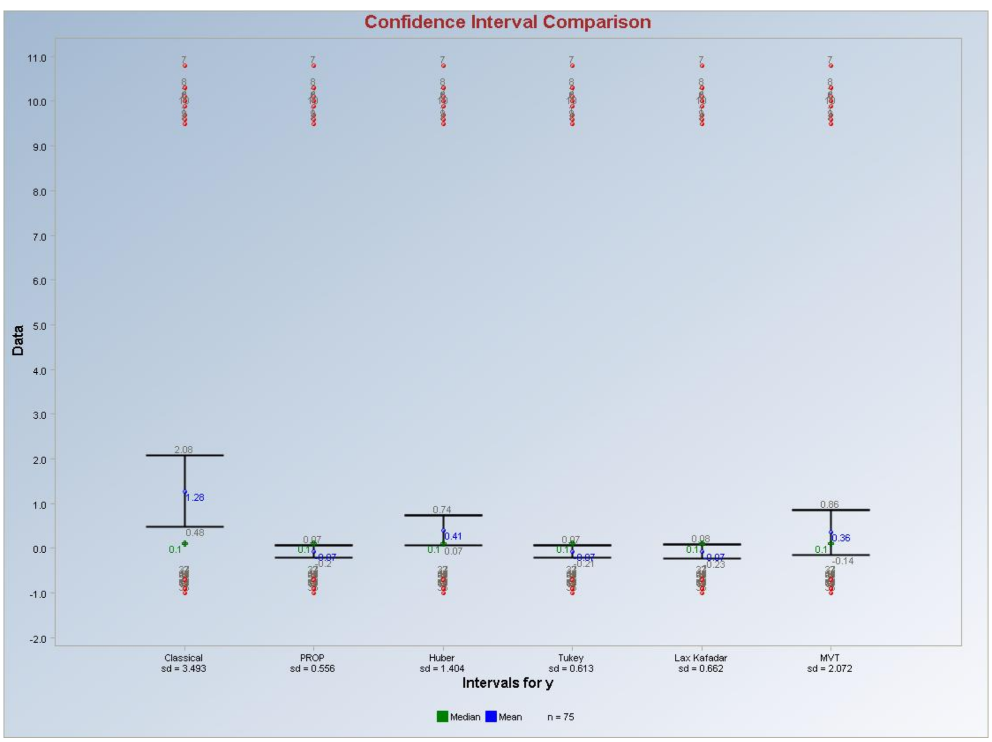
Output for the Prediction Interval Comparison.



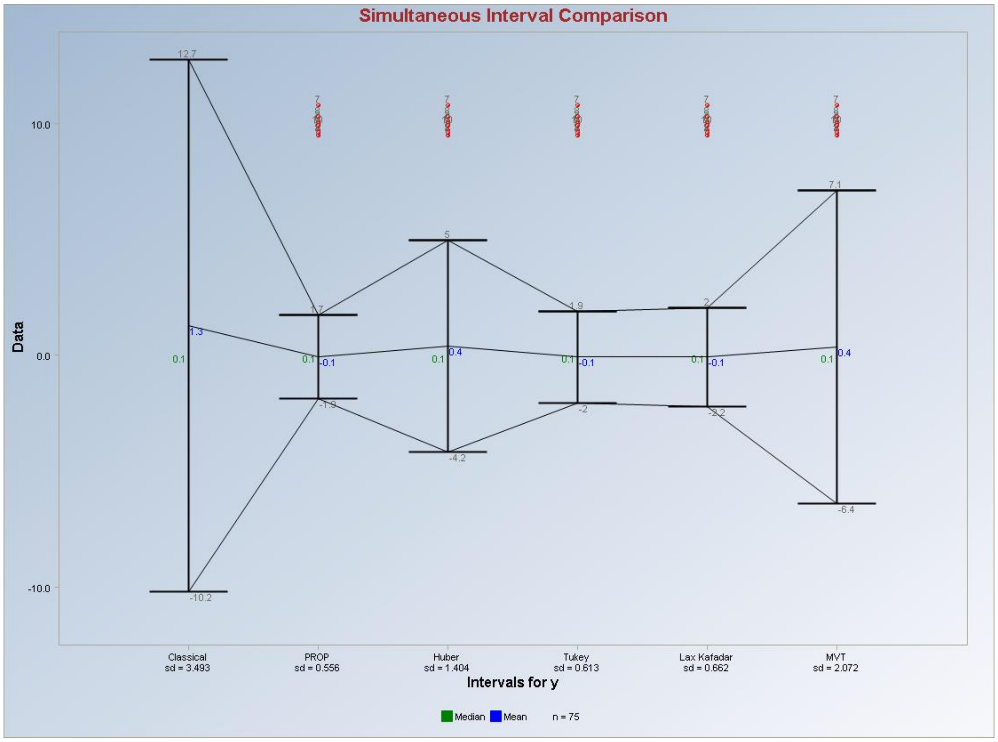
Output for the Tolerance Interval Comparison.



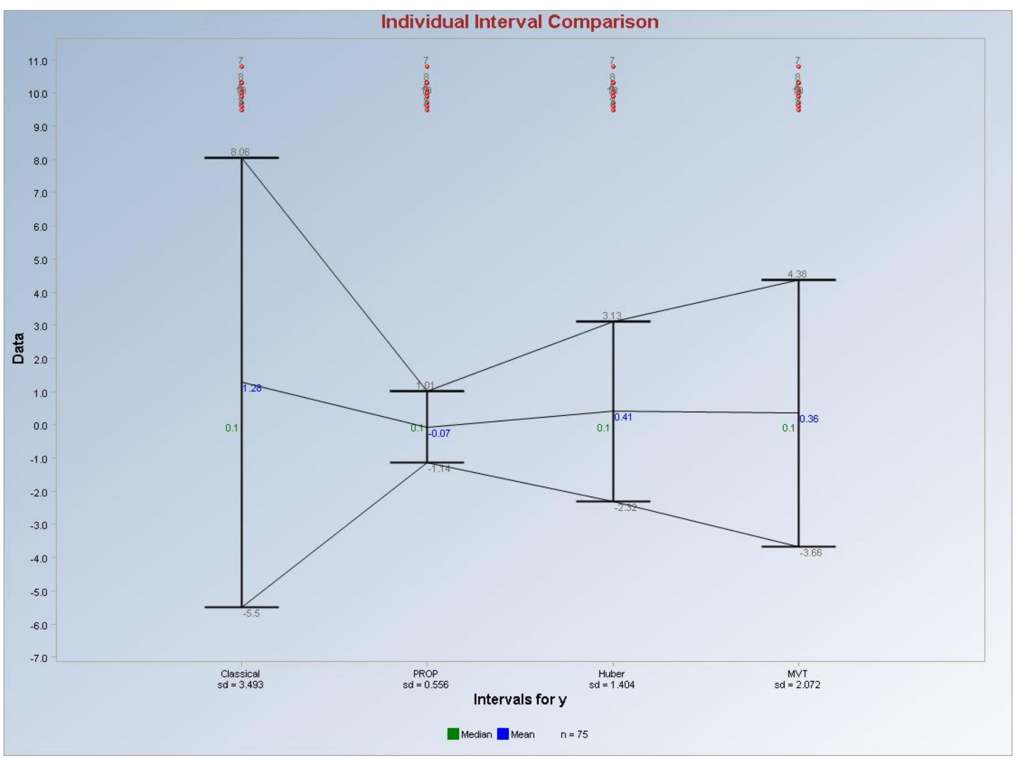
Output for the Confidence Interval Comparison.



Output for the Simultaneous Interval Comparison.

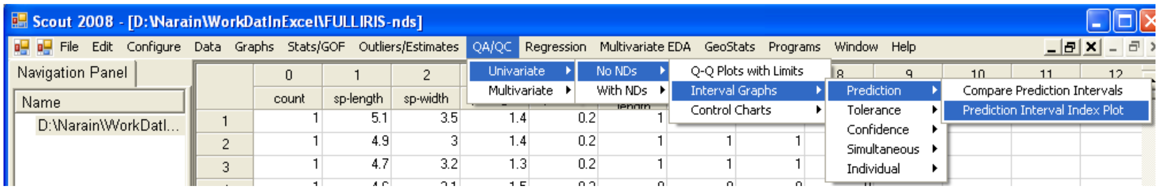


Output for the Individual Interval Comparison.



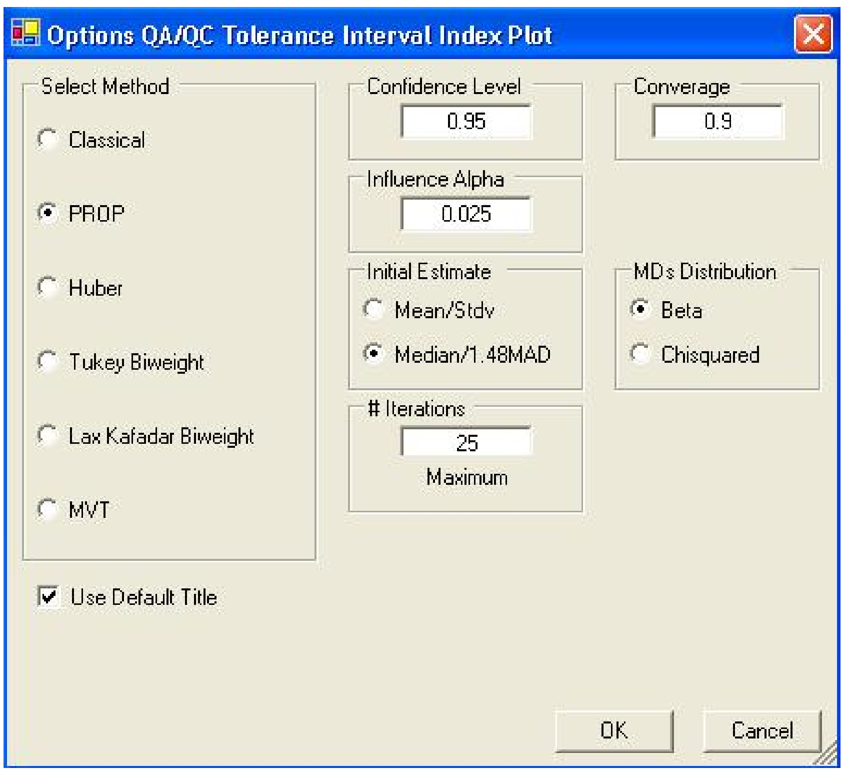
8.1.1.2.2 Intervals Index Plots

1. Click on **QA/QC ► Univariate ► No NDs ► Interval Graphs ►Prediction, Tolerance, Confidence, Simultaneous or Individual ► Intervals Index Plots.**



2. The “**Select Variables**” screen (Section 3.4) will appear.

- Click on the “**Options**” button for the options window.

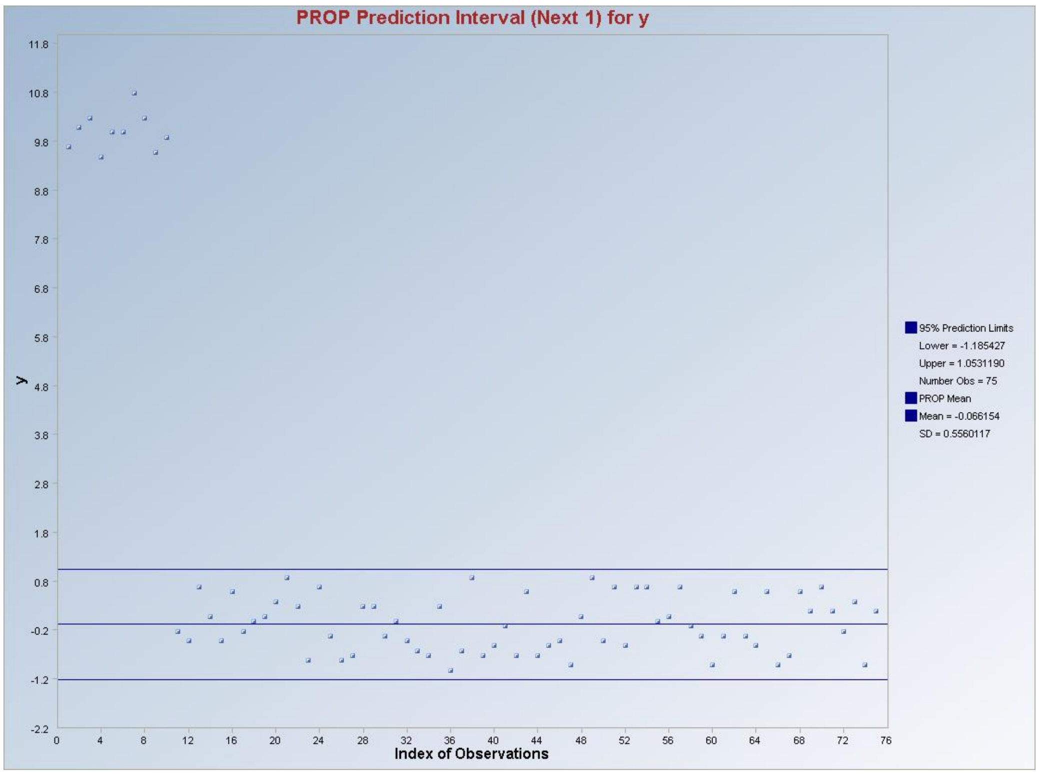


  - Select one of the methods for the interval in “**Select Methods**” box. By default, “**PROP**” is selected.
  - Specify the various input parameters for the selected method.
  - Click “**OK**” to continue or “**Cancel**” to cancel the options.

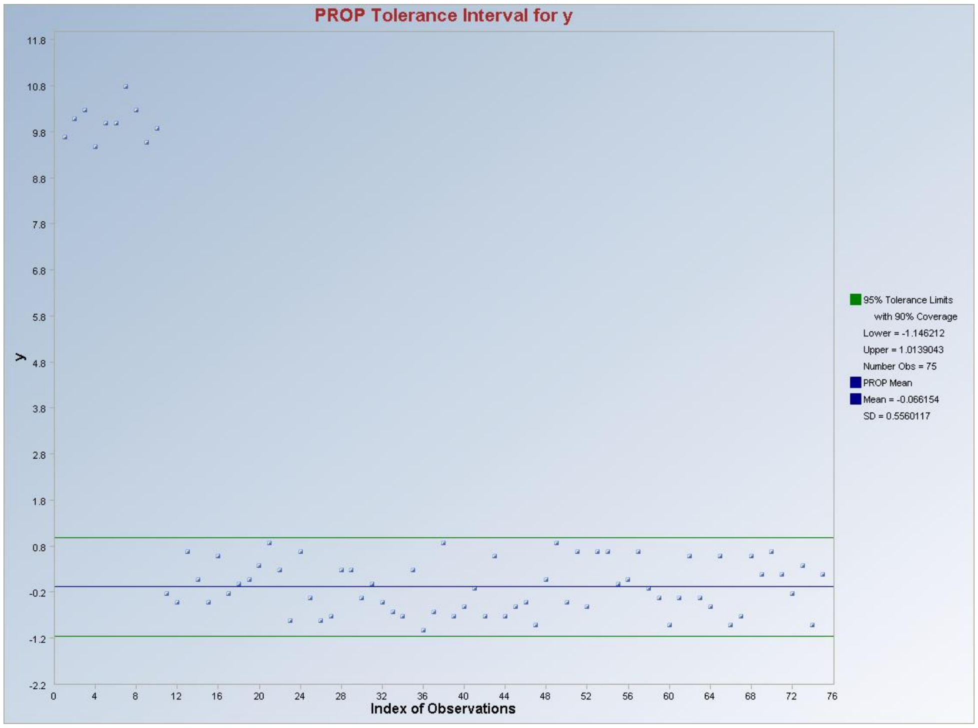
- Click on “**OK**” to continue or “**Cancel**” to cancel the intervals comparison.

Output example: The data set “**Bradu.xls**” was used for the Interval Index Plots. The options used were the default options.

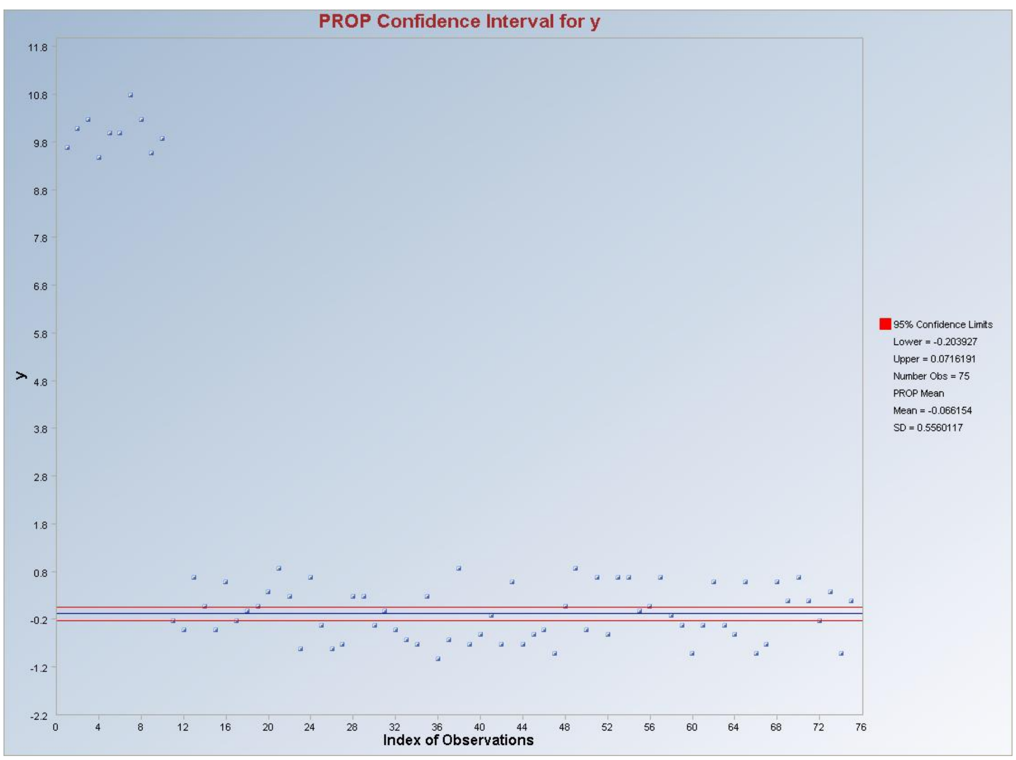
Output for the Prediction Interval Index Plot.



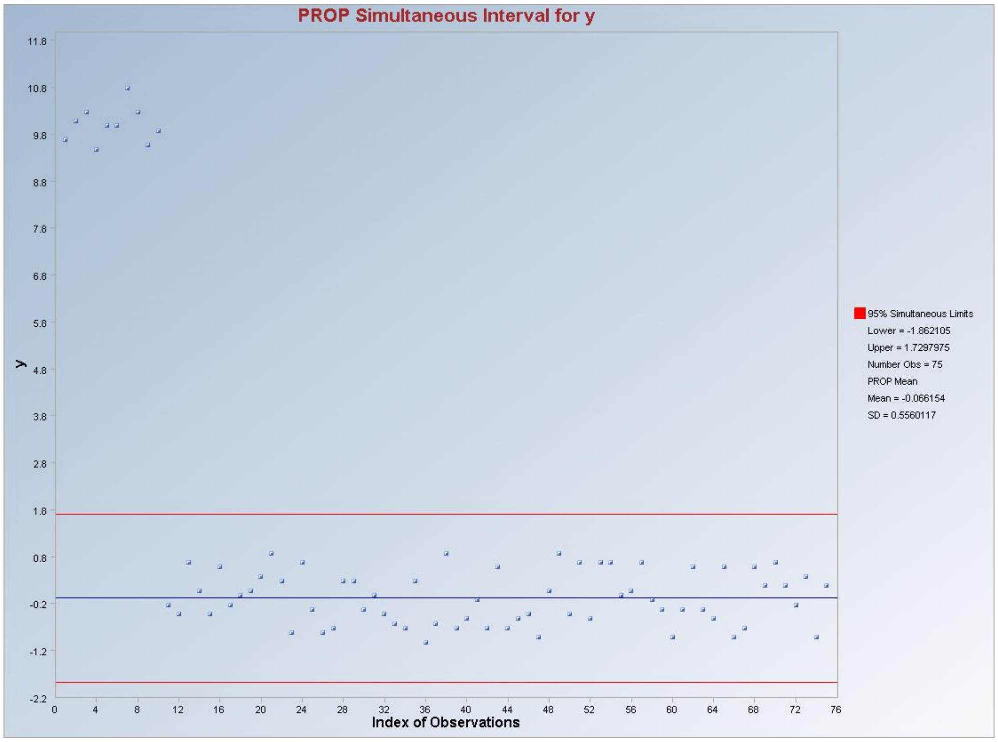
Output for the Tolerance Interval Index Plot.



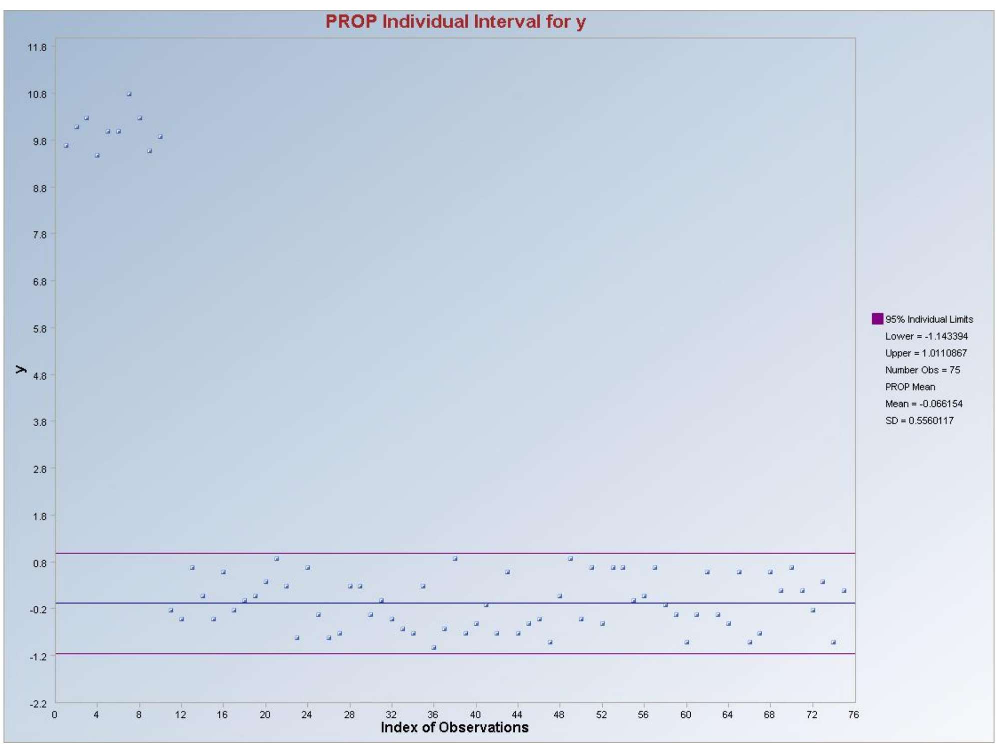
Output for the Confidence Interval Index Plot.



Output for the Simultaneous Interval Index Plot.



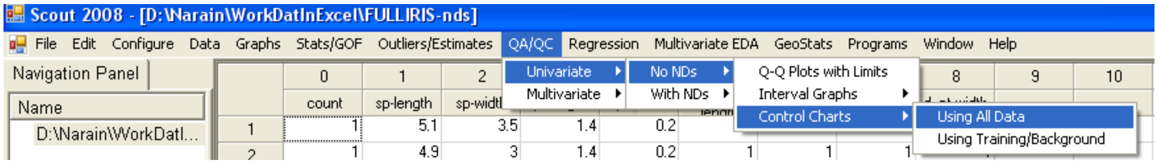
Output for the Individual Interval Index Plot.



8.1.1.3 Control Charts

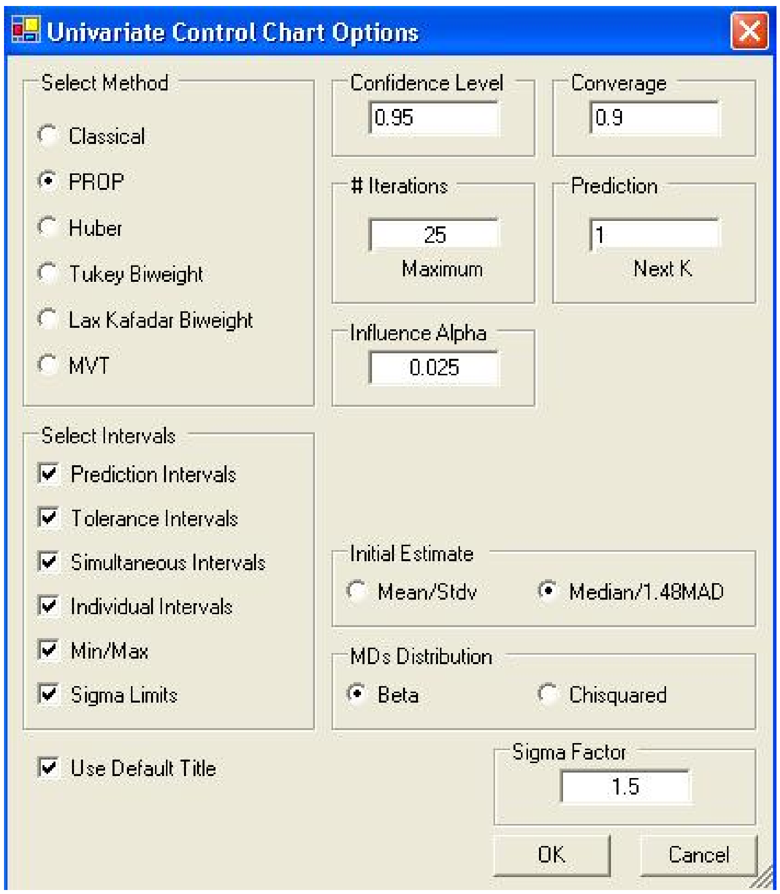
8.1.1.3.1 Using All Data

1. Click on **QA/QC ► Univariate ► No NDs ► Control Charts ► Using All Data.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

- Click on the "**Options**" button for the options window.

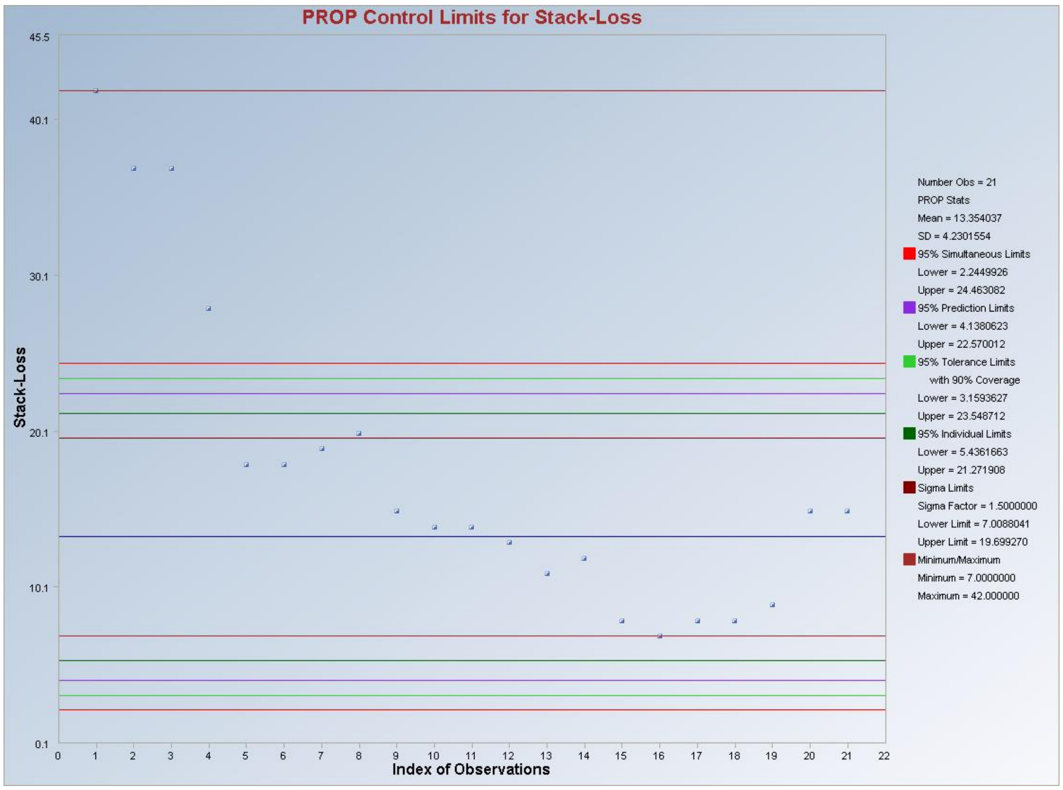


- Select one of the methods for the interval in "**Select Methods**" box. By default, "**PROP**" is selected.
- Specify the various intervals in the "**Select Intervals**" box.

- Specify the various options for the selected method and intervals.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

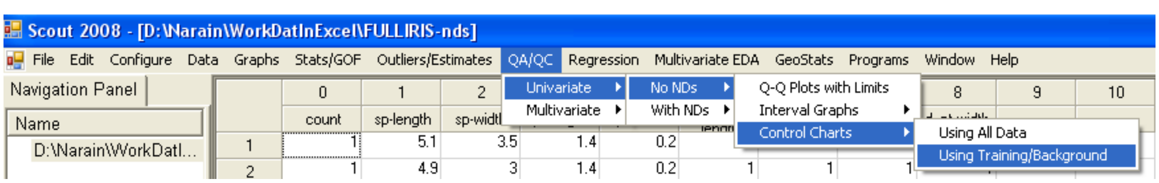
- Click on "**OK**" to continue or "**Cancel**" to cancel the control charts.

Output example: The data set "**Stackloss.xls**" was used for the Control Charts. The options used were the default options.

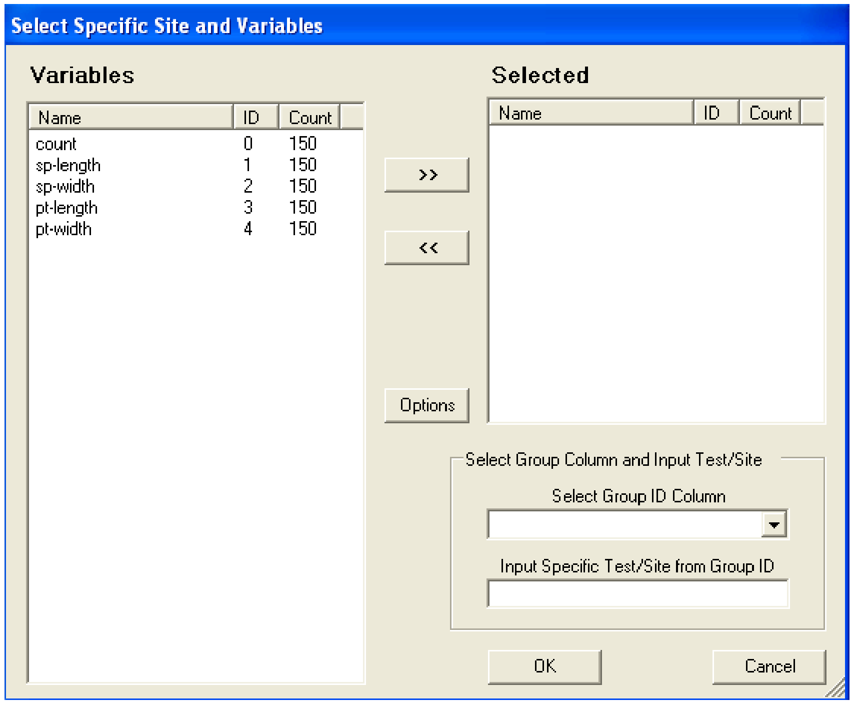


8.1.1.3.2 Using Training/Background

1. Click on **QA/QC ► Univariate ► No NDs ► Control Charts ► Training/Background.**



2. The "**Select Variables**" screen will appear.

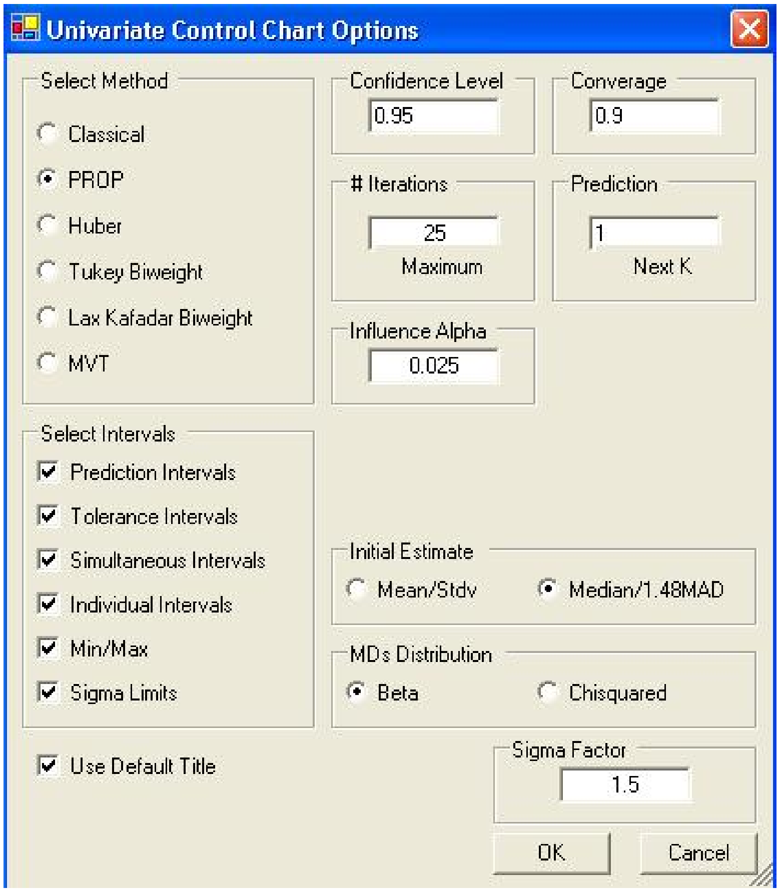


- Select the variable of interest.

- Select the group variable using the "**Select Group ID Column**" drop-down bar.

- Input the group name/number of the variable which is considered as the test set in the "**Input Specific Test/Site from Group ID**" box.

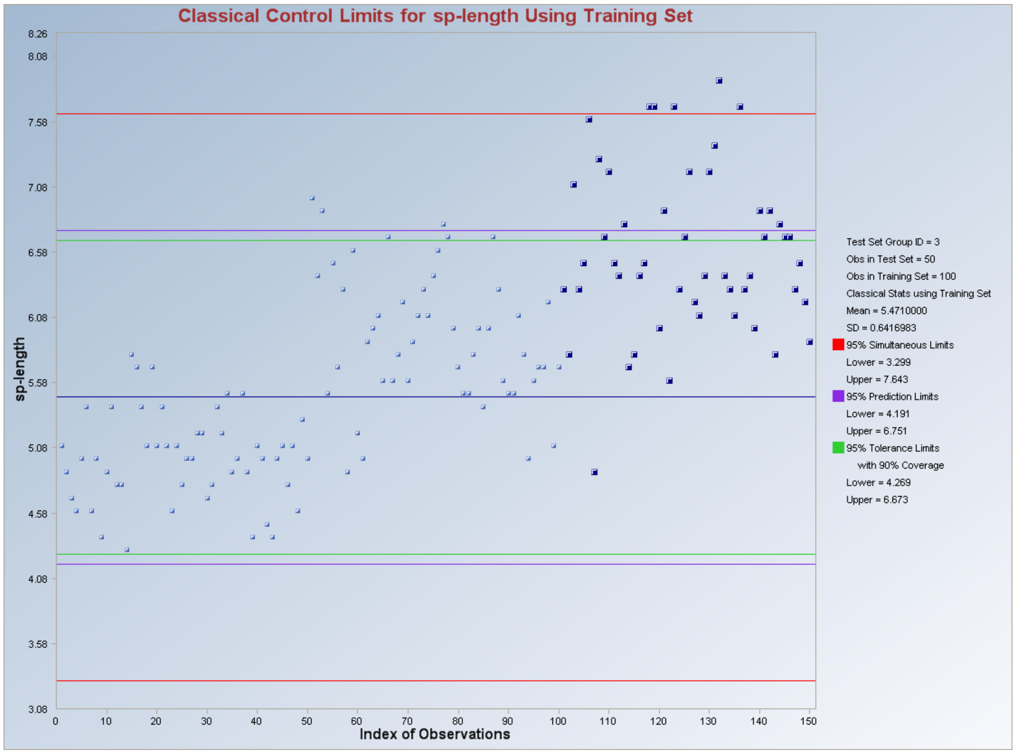
- Click on the "**Options**" button for the options window.



- Select one of the methods for the interval in "**Select Methods**" box. By default, "**PROP**" is selected.
- Specify the various intervals in the "**Select Intervals**" box.
- Specify the various options for the selected method and intervals.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the control charts.

Output example: The data set "**FullIris.xls**" was used for the Control Charts. The options used were the default options.

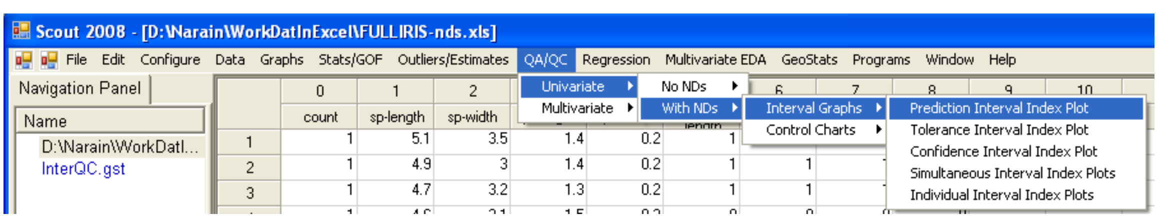


***Note**: The observations in "**dark blue**" and "**bigger point marks**" are from group 3. The intervals are calculated using the observations from group 1 and 2 only which were used as the "**Training/Background**" set for this data set.*

8.1.2 With Non-detects

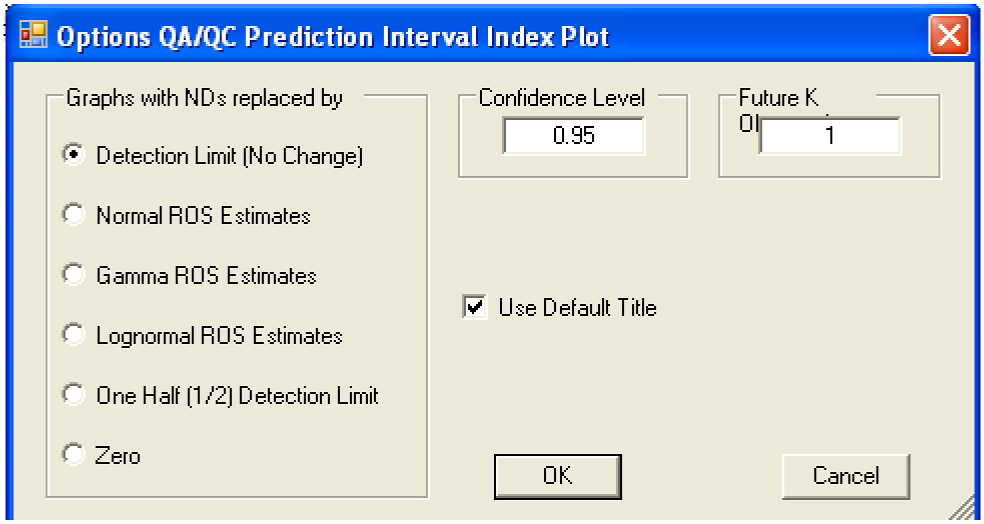
8.1.2.1 Interval Graphs

1. Click on **QA/QC ► Univariate ► With NDs ► Interval Graphs ►Prediction, Tolerance, Confidence, Simultaneous or Individual.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

   - Click on the "**Options**" button for the options window.

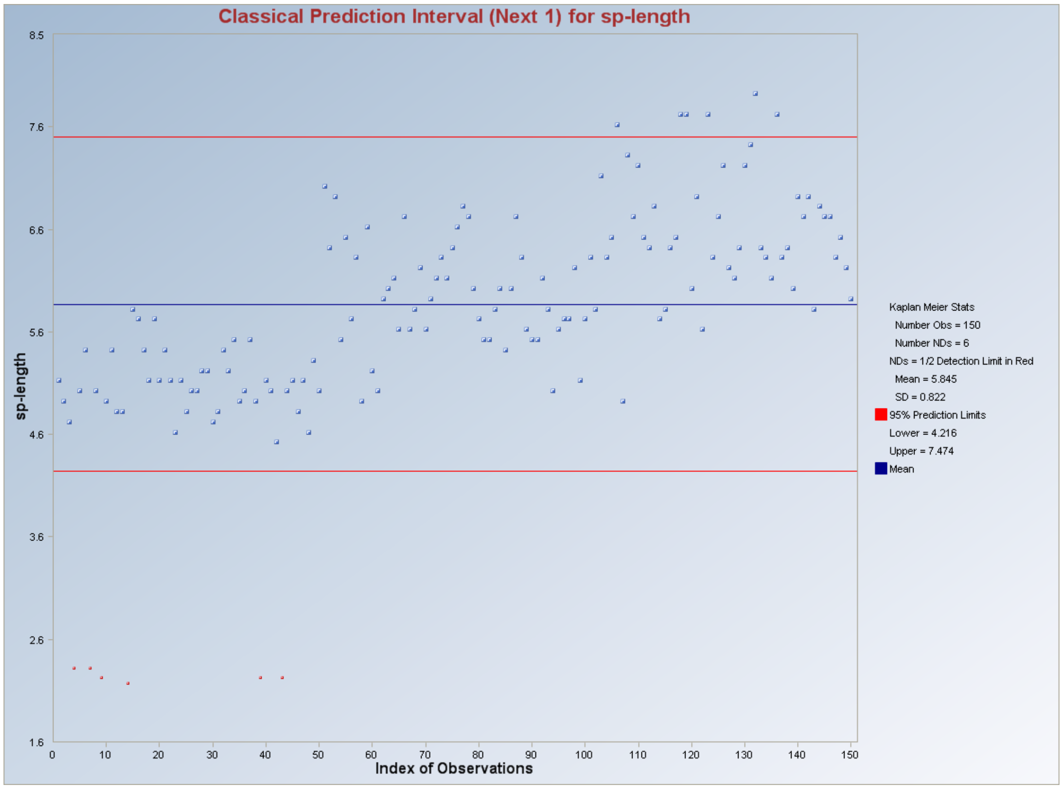


- Select the method to replace the non-detects with in "**Graphs with NDs replaced by**" box. Default method is "**Detection Limit.**"
- Specify the various input parameters for the selected method.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the intervals comparison.

Output example: The data set "**FullIris.xls**" was used for the Interval Index Plots. The options used were the default options.

Output for the Prediction Interval Index Plot with NDs.

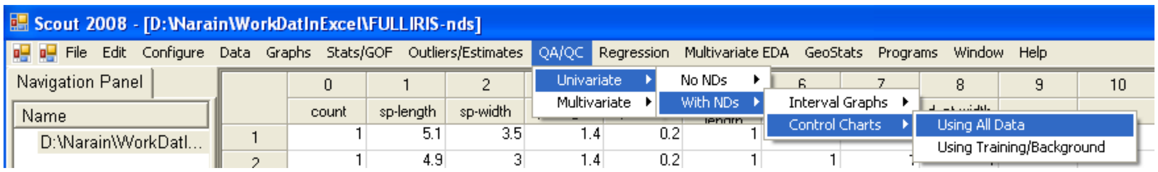


***Note**: The non-detect observations are replaced by "**One Half (1/2) Detection Limit**" indicated by the red points.*

8.1.2.2 Control Charts

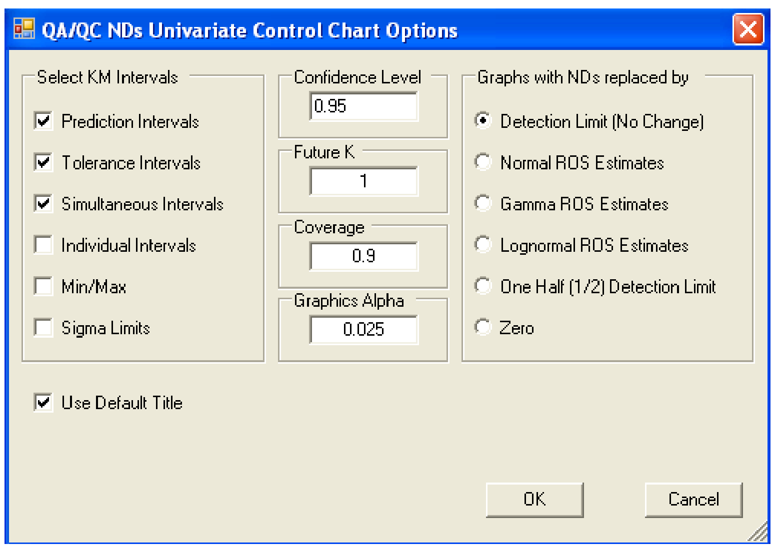
8.1.2.2.1 Using All Data

1. Click on **QA/QC ► Univariate ► With NDs ► Control Charts ► Using All Data.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

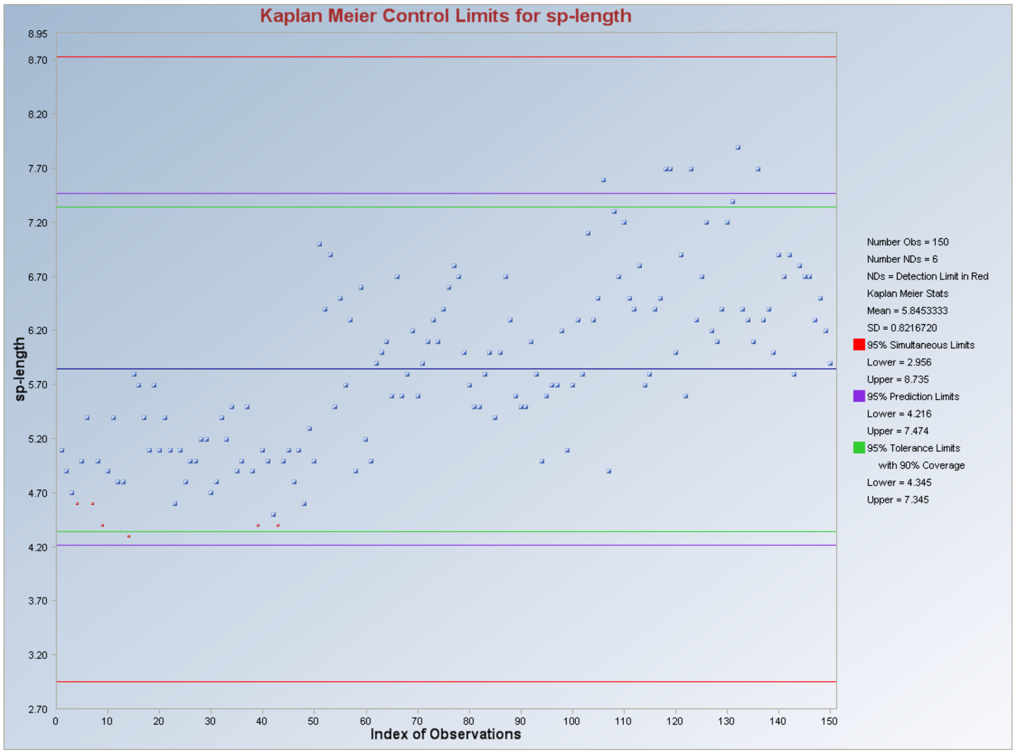
- Click on the "**Options**" button for the options window.



  - Select the intervals to be displayed on the control chart from the "**Select KM Intervals**" box.
  - Specify the various parameters for the selected intervals.
  - Select the method to replace the non-detects with in "**Graphs with NDs replaced by**" box. Default method is "**Detection Limit.**"
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the control charts.

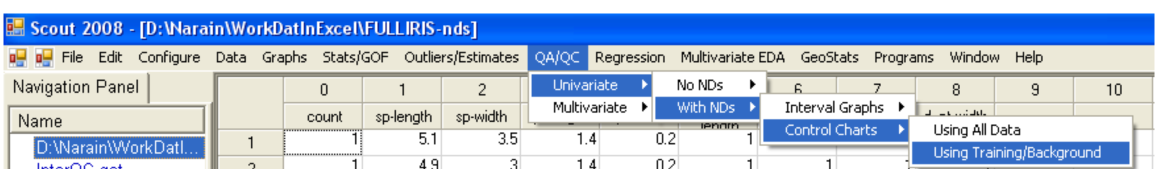
Output example: The data set “**FullIRIS.xls**” was used for the Control Charts. The options used were the default options.



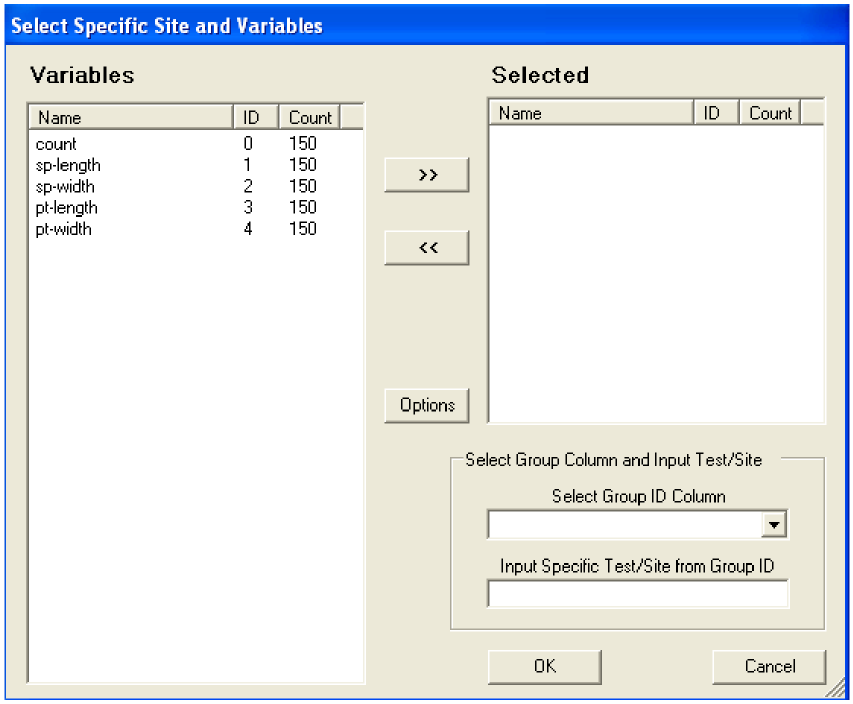
*Note: The non-detect observations are replaced by “**Detection Limit**” indicated by the” **red points**.”*

8.1.2.2.2 Using Training/Background

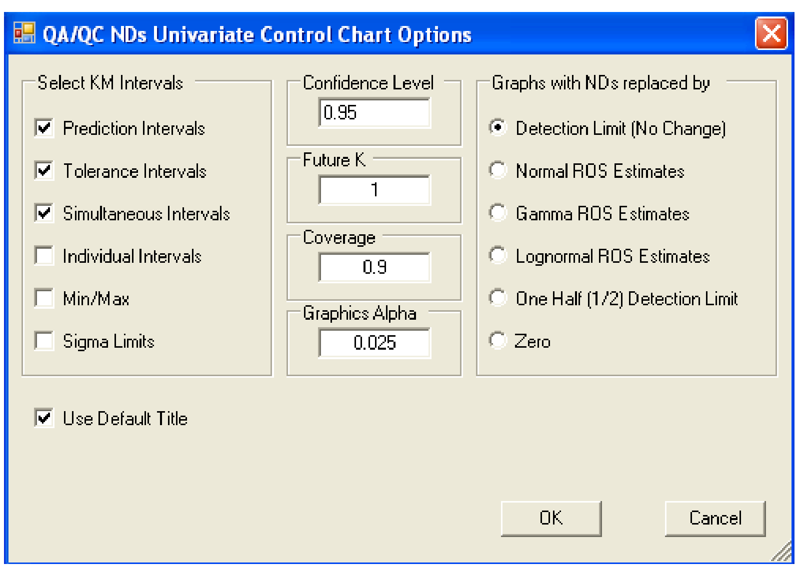
1. Click on **QA/QC ► Univariate ► With NDs ► Control Charts ► Training/Background.**



2. The “**Select Variables**” screen will appear.



- Select the variable of interest.
- Select the group variable using the "**Select Group ID Column**" drop-down bar.
- Input the group name/number of the variable which is considered as the test set in the "**Input Specific Test/Site from Group ID**" box.
- Click on the "**Options**" button for the options window.



- Select the intervals to be displayed on the control chart from the "**Select KM Intervals**" box.

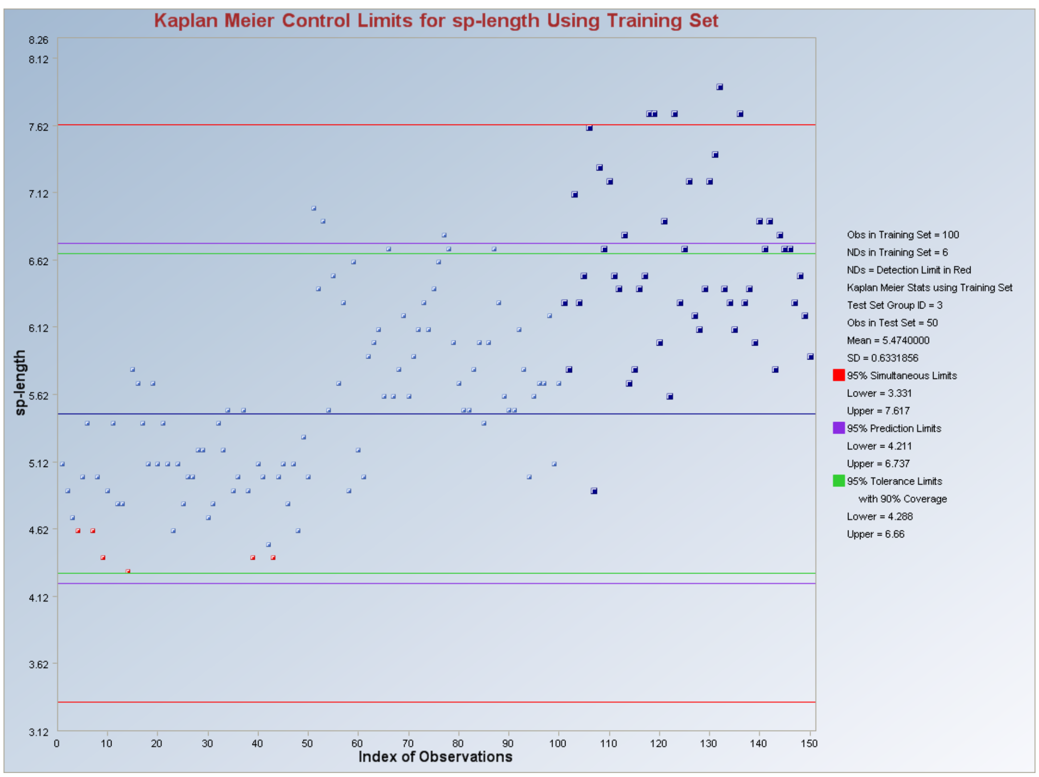
- Specify the various parameters for the selected intervals.

- Select the method to replace the non-detects with in "**Graphs with NDs replaced by**" box. Default method is "**Detection Limit**."

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the control charts.

Output example: The data set "**FullIris.xls**" was used for the Control Charts. The options used were the default options.



Note: *The observations in "**dark blue**" and "**bigger points**" are from group 3. The intervals are calculated using the observations from group 1 and 2 only which were used as the "**Training/Background**" set for this data set. The "**red points**" indicate non-detects at the detection limit.*

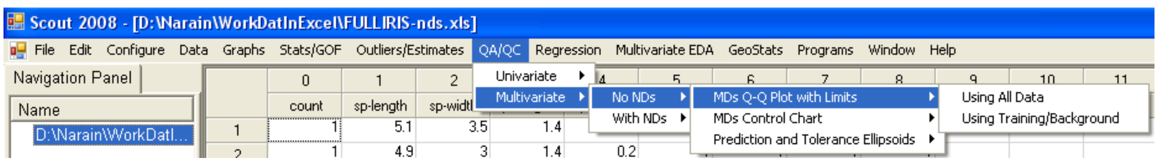
8.2 Multivariate QA/QC

Several classical and robust multivariate procedures are available in the QA/QC module of Scout. The multivariate, with non-detects module uses the Kaplan Meier estimates. This QA/QC module includes MDs Q-Q Plots with limits, MDs Control Charts and Prediction and Tolerance ellipsoids. The robust methods include in this module are explained in Chapter 7.

8.2.1 No Non-detects

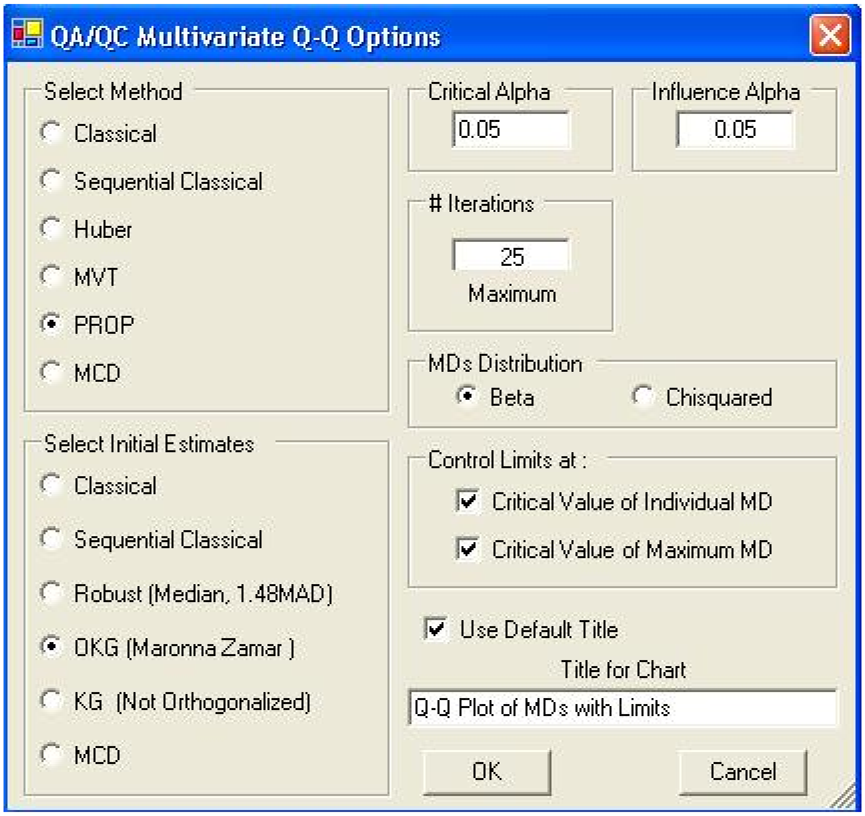
8.2.1.1 MDs Q-Q Plots with Limits

1. Click on **QA/QC ► Multivariate ► No NDs ► MDs Q-Q Plots with Limits.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

- Click on the "**Options**" button for the options window.



  - Specify the method for computing the quantiles in "**Select Methods**." The default method is "**PROP**."

  - The robust methods need various input parameters like "**Influence Alpha**" or "**Trimming Percentage**," "**Initial Estimates**," "**MDs Distribution**," "**MDs Distribution**" and "**# Iterations.**"

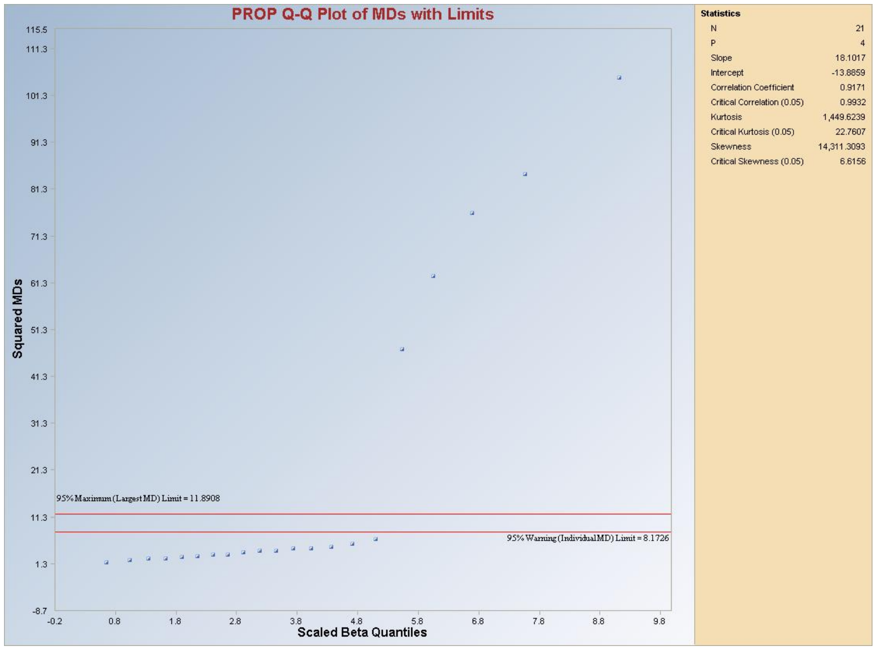
  - Specify the "**Critical Alpha for Limits**" for identifying the outliers. Default is "**0.05**."

  - Specify the lines for control limits bys using the "**Control Limits at**:" option. Both options are unchecked as default.

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the MDs Q-Q Plots with limits.

Output example (Using All Data): The data set "**Stackloss.xls**" was used for the Q-Q Plot. The options used were the default options.



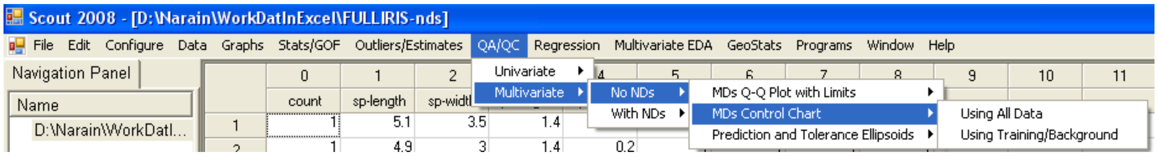
***Note**: The observations above the maximum limit line are considered as outliers.*

Output example (Using Training/Background): The data set "**FullIris.xls**" was used for the Q-Q Plot. The options used were the default options.

***Note**: The observations in "**dark blue**" and "**bigger points**" are from group 3. The estimates of mean vector and the covariance matrix are calculated using the observations from group 1 and 2 only which were used as the "**Training/Background**" set for this data set.*

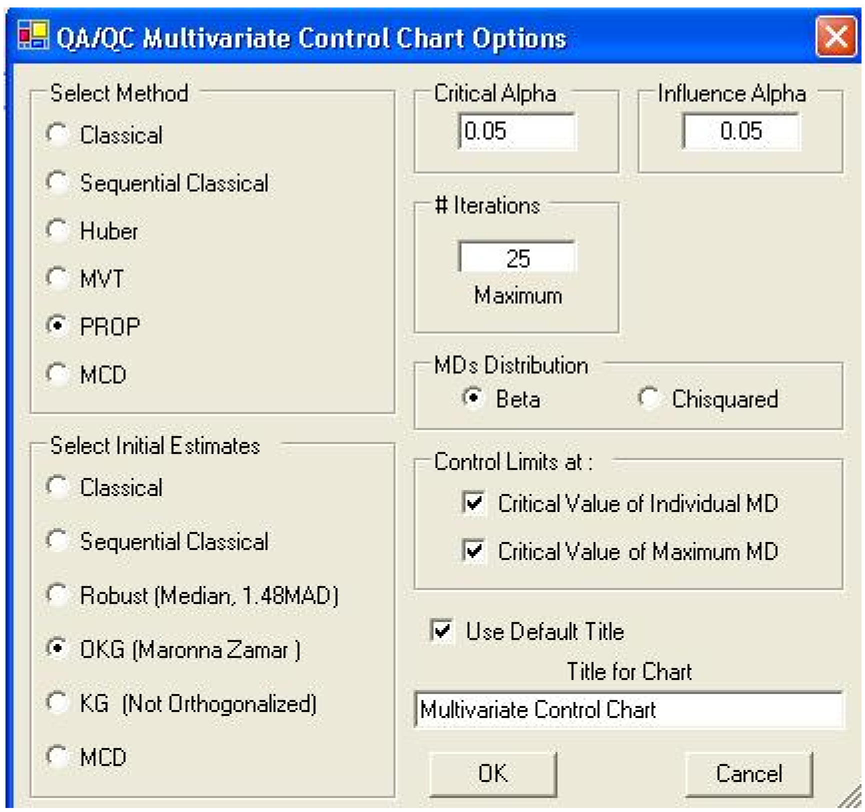
8.2.1.2 MDs Control Chart

1. Click on **QA/QC ► Multivariate ► No NDs ► MDs Control Chart.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

- Click on the "**Options**" button for the options window.

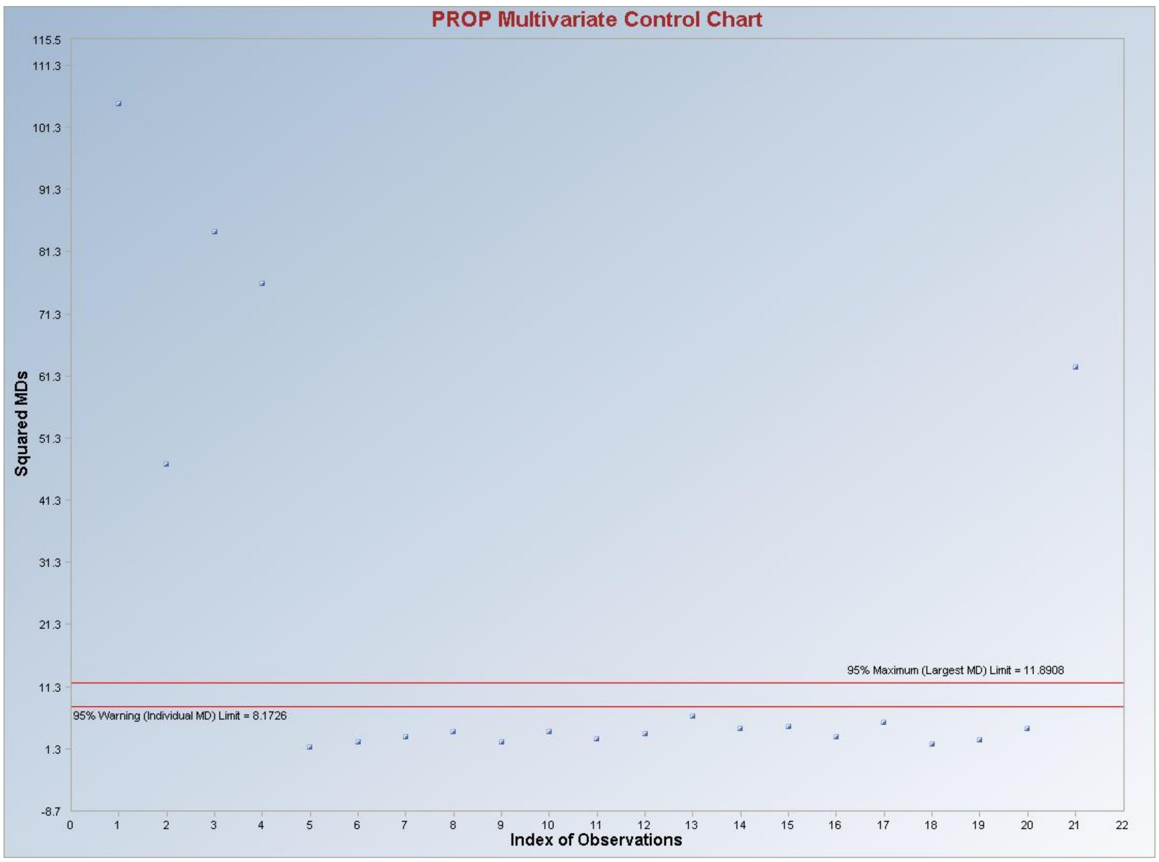


  - Specify the method for computing the quantiles in "**Select Methods**." The default method is "**PROP**."
  - The robust methods need various input parameters like "**Influence Alpha**" or "**Trimming Percentage**," "**Initial Estimates**," "**MDs Distribution**," "**MDs Distribution**" and "**# Iterations.**"
  - Specify the "**Critical Alpha for Limits**" for identifying the outliers. Default is "**0.05**."

- Specify the lines for control limits bys using the “**Control Limits at**:” option. Both options are unchecked as default.
- Click “**OK**” to continue or “**Cancel**” to cancel the options.

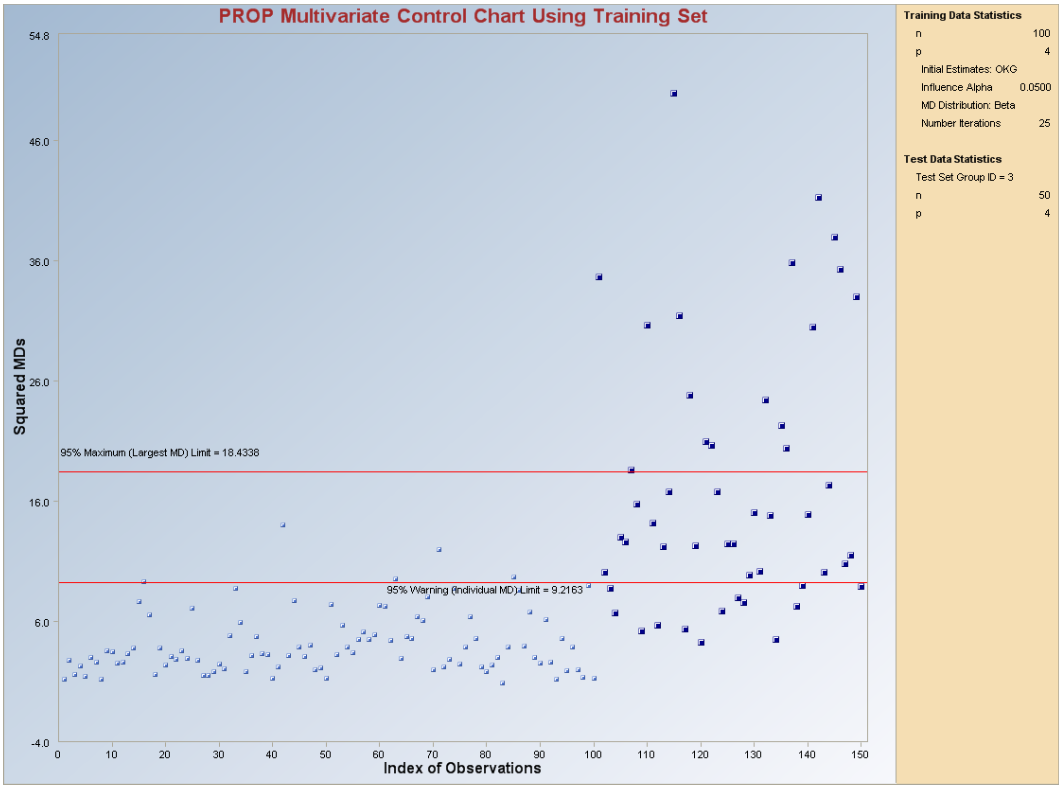
- Click on “**OK**” to continue or “**Cancel**” to cancel the MDs Control Charts.

Output example (Using All Data): The data set “**Stackloss.xls**” was used for the MDs Control Charts. The options used were the default options.



Note*: The observations above the maximum limit line are considered as outliers.*

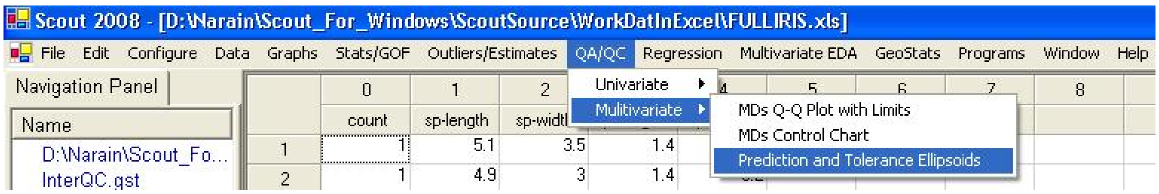
Output example (Using Training/Background): The data set "**FullIris.xls**" was used for the MDs Control Charts. The options used were the default options.



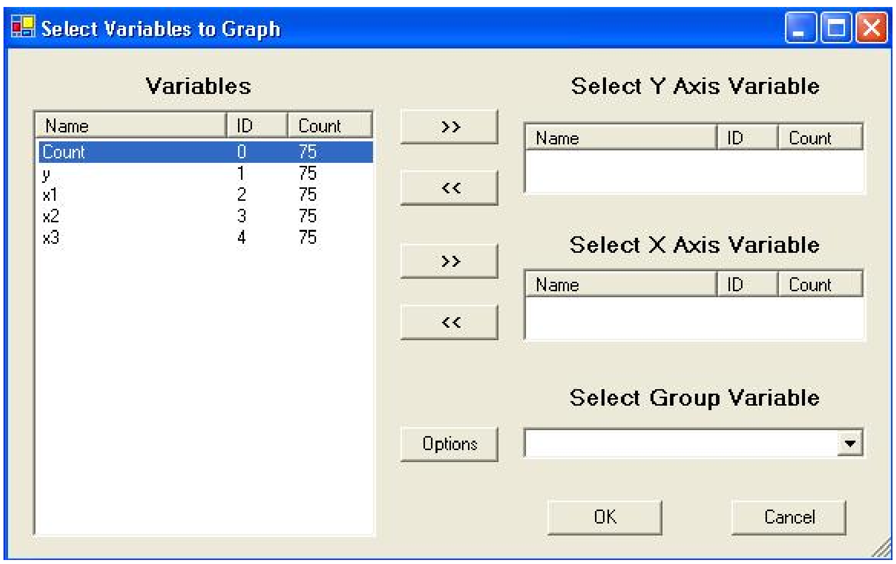
Note: *The observations in "**dark blue**" and "**bigger points**" are from group 3. The estimates of mean vector and the covariance matrix are calculated using the observations from group 1 and 2 only which were used as the "**Training/Background**" set for this data set.*

8.2.1.3 Prediction and Tolerance Ellipsoids

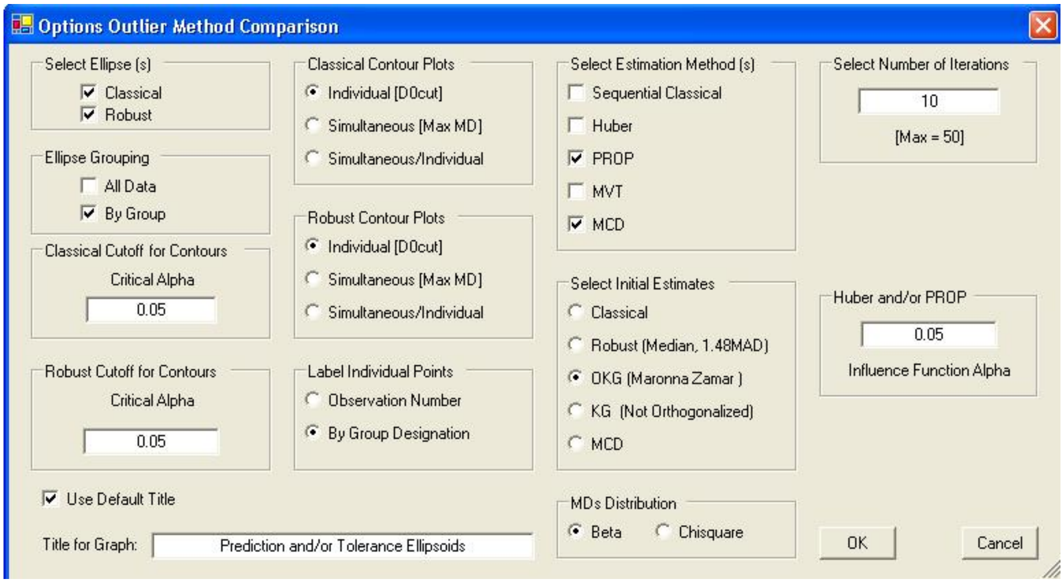
1. Click on **QA/QC ► Multivariate ► No NDs ► Prediction and Tolerance Ellipsoids.**



2. The following "**Select Variables**" screen will appear.



- Click on the "**Options**" button for the options window.

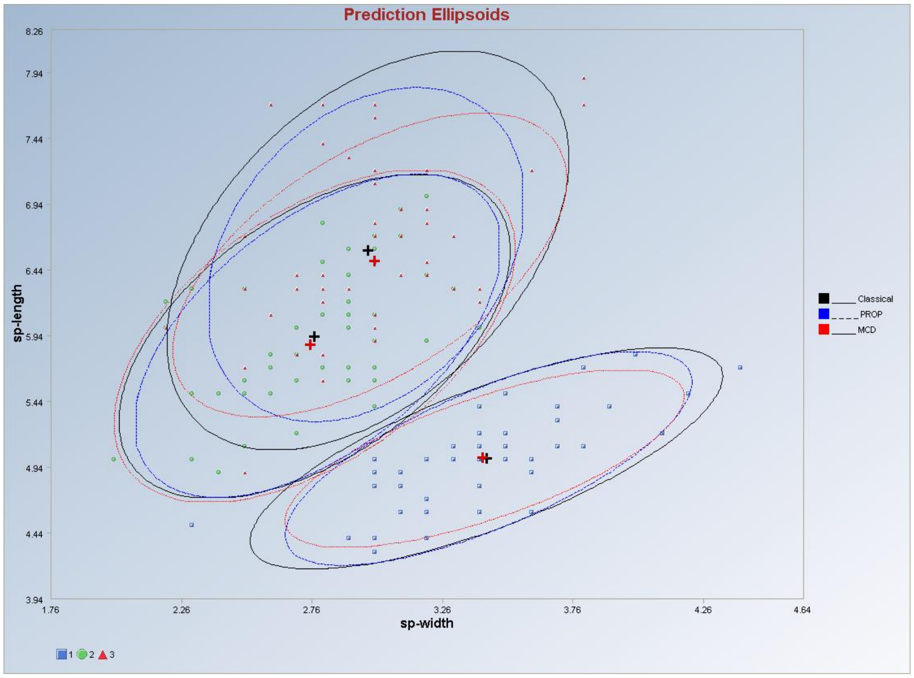


- Specify the required options. These options are discussed in **Section 7.2.3**. The user has an option for drawing the ellipsoids by groups if the observations are from different groups.

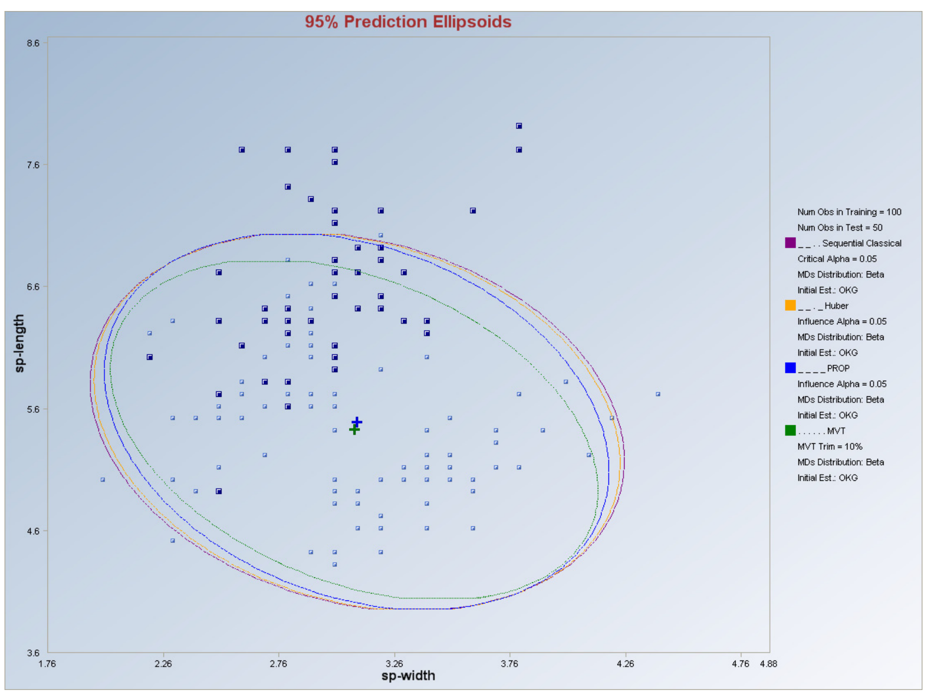
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

* Click on "**OK**" to continue or "**Cancel**" to cancel the Ellipsoids.

Output example (Using All Data): The data set "**Fulliris.xls**" was used for the Ellipsoids. The options used are shown in the options window screenshot above. The ellipses are being drawn by groups.



Output example (Using Background/Training): The data set "**Fulliris.xls**" was used for the Ellipsoids.

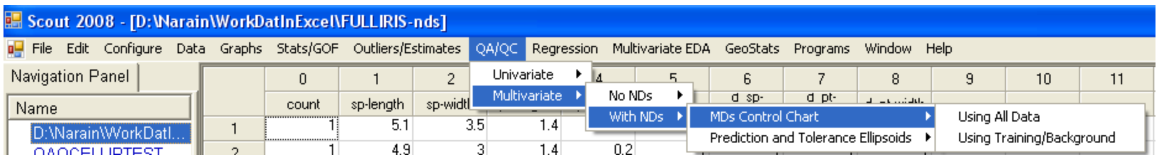


Note: *The observations in "**dark blue**" and "**bigger points**" are from group 3. The estimates of mean vector and the covariance matrix are calculated using the observations from group 1 and 2 only which were used as the "**Training/Background**" set for this data set.*

8.2.2 With Non-detects

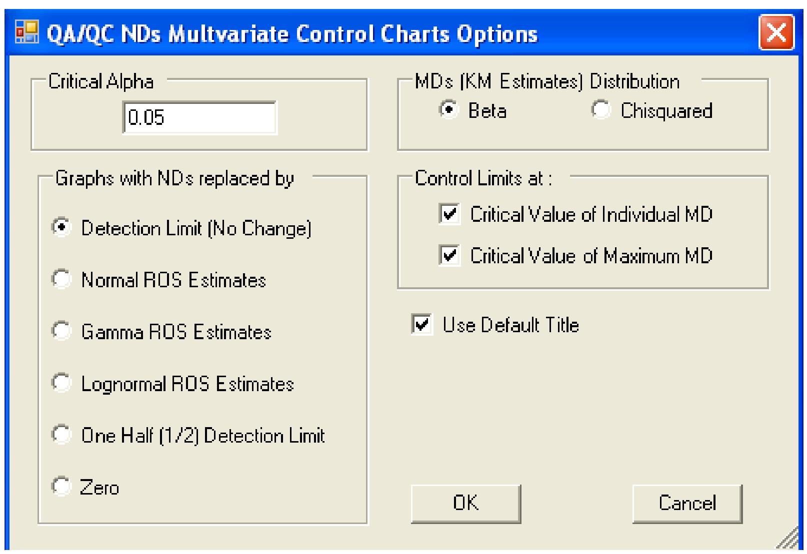
8.2.2.1 MDs Control Charts

1. Click on **QA/QC ► Multivariate ► With NDs ► MDs Control Charts.**



2. The "**Select Variables**" screen (Section 3.4) will appear.

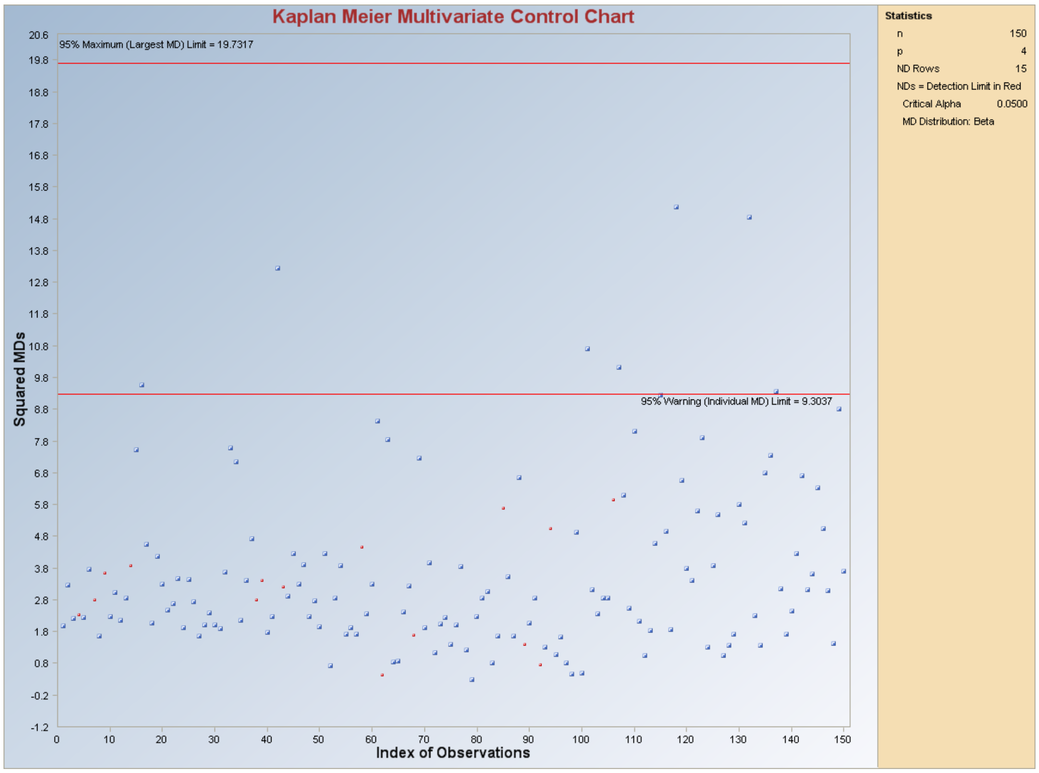
- Click on the "**Options**" button for the options window.



- Specify the "**Critical Alpha**" for identifying the outliers. Default is "**0.05**."
- Specify the distribution for the distances using "**MDs (KM Estimates) Distribution**" box.
- Specify the lines for control limits bys using the "**Control Limits at:**" option. Both options are unchecked as default.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

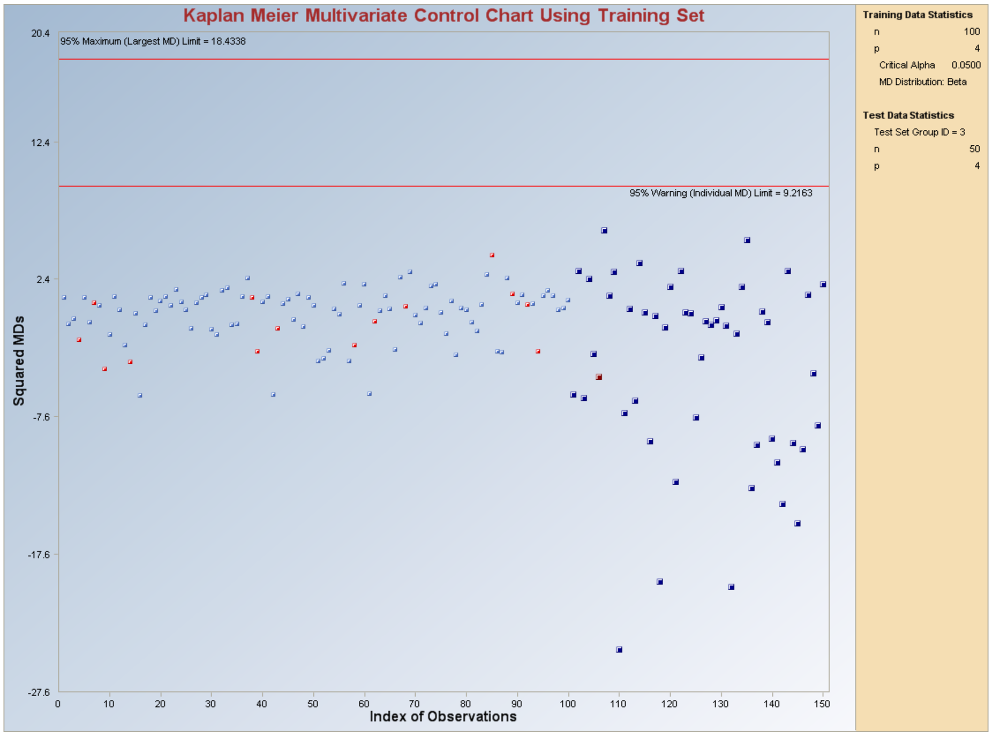
- Click on "**OK**" to continue or "**Cancel**" to cancel the MDs Control Charts.

Output example (Using All Data): The data set "**FullIris.xls**" was used for the control chart. The options used were the default options.



Note*: The observations above the maximum limit line are considered as outliers. The non-detect observations are in "**red point marks**."*

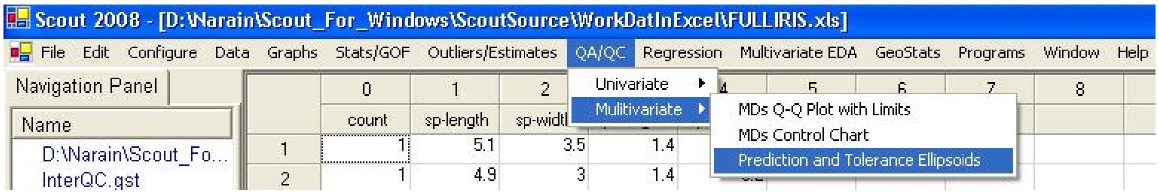
Output example (Using Training/Background): The data set "**FullIris.xls**" was used control charts. The options used were the default options.



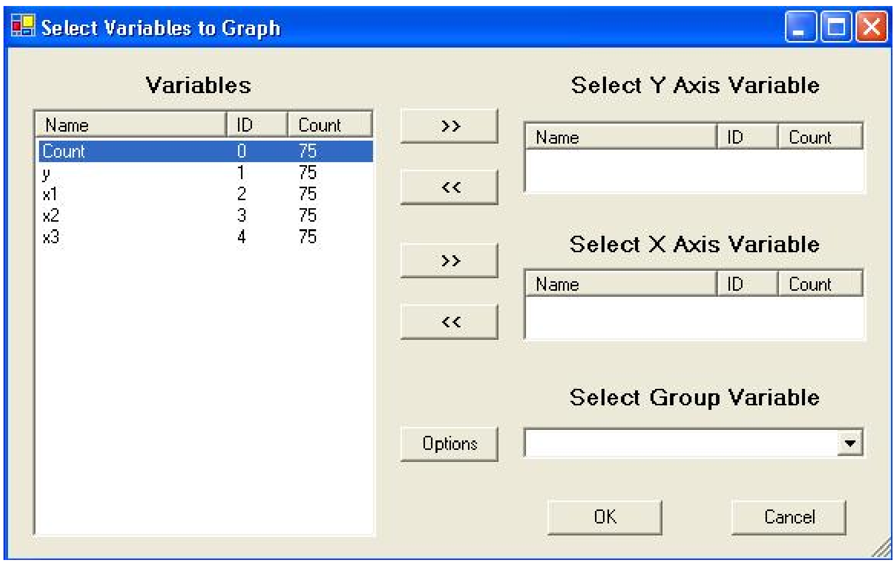
***Note**: The observations in "**dark blue**" and "**bigger points**" are from group 3. The estimates of mean vector and the covariance matrix are calculated using the observations from group 1 and 2 only which were used as the "**Training/Background**" set for this data set. The non-detect observations are in "**red point marks**."*

8.2.2.2 Prediction and Tolerance Ellipsoids

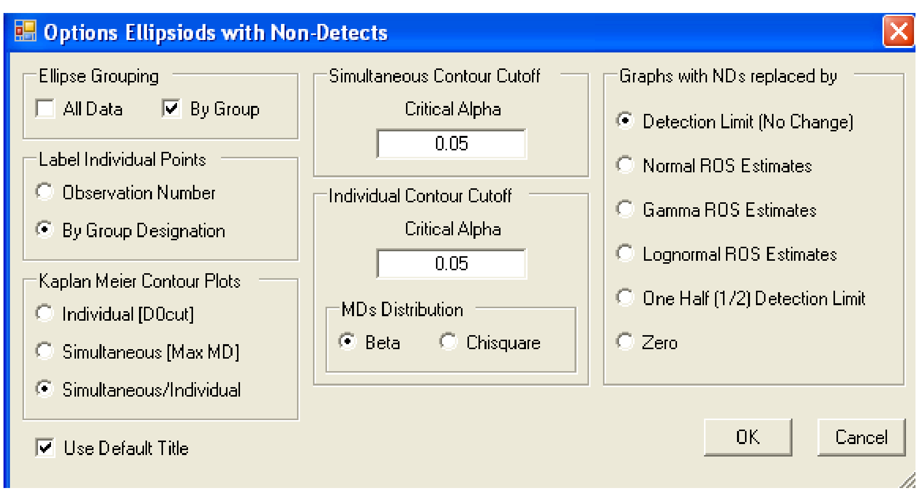
1. Click on **QA/QC ► Multivariate ► With NDs ► Prediction and Tolerance Ellipsoids.**



2. The following "**Select Variables**" screen will appear.



- Click on the "**Options**" button for the options window.

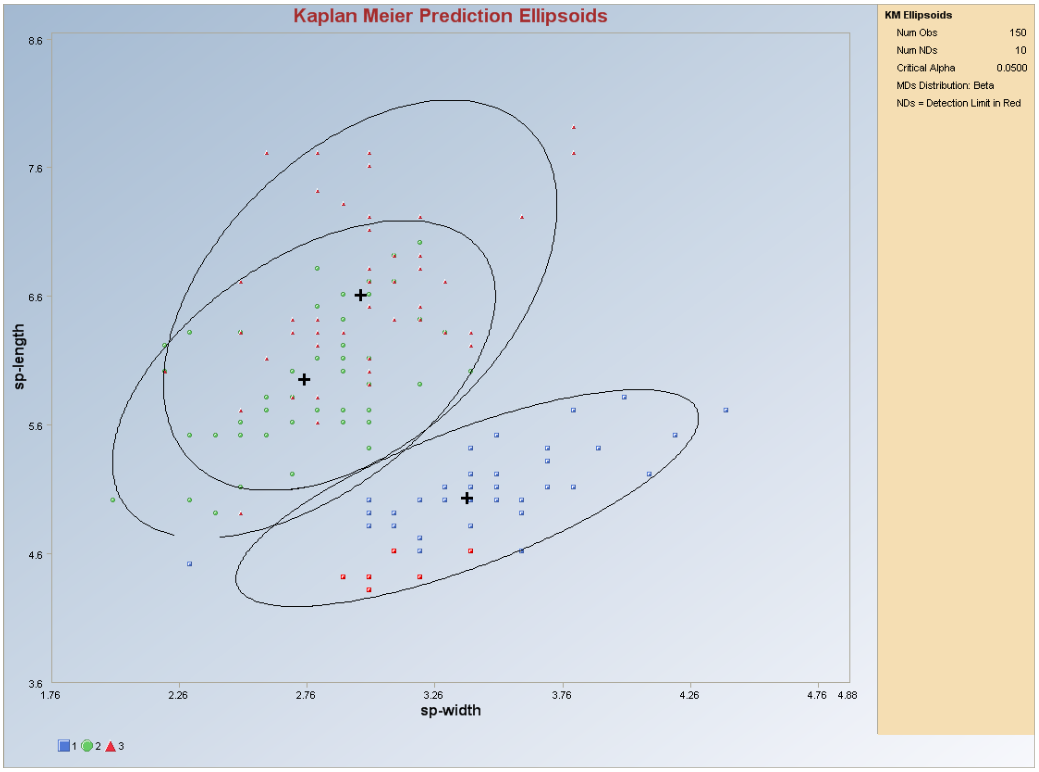


- Specify the required options. These options include "**Kaplan Meier Contour Plots**", "Critical Alpha(s)," "**MDs Distributions**" for the contours and "**Graphs with NDs replaced by**" option. The user has an option for drawing the ellipsoids by groups if the observations are from different groups.

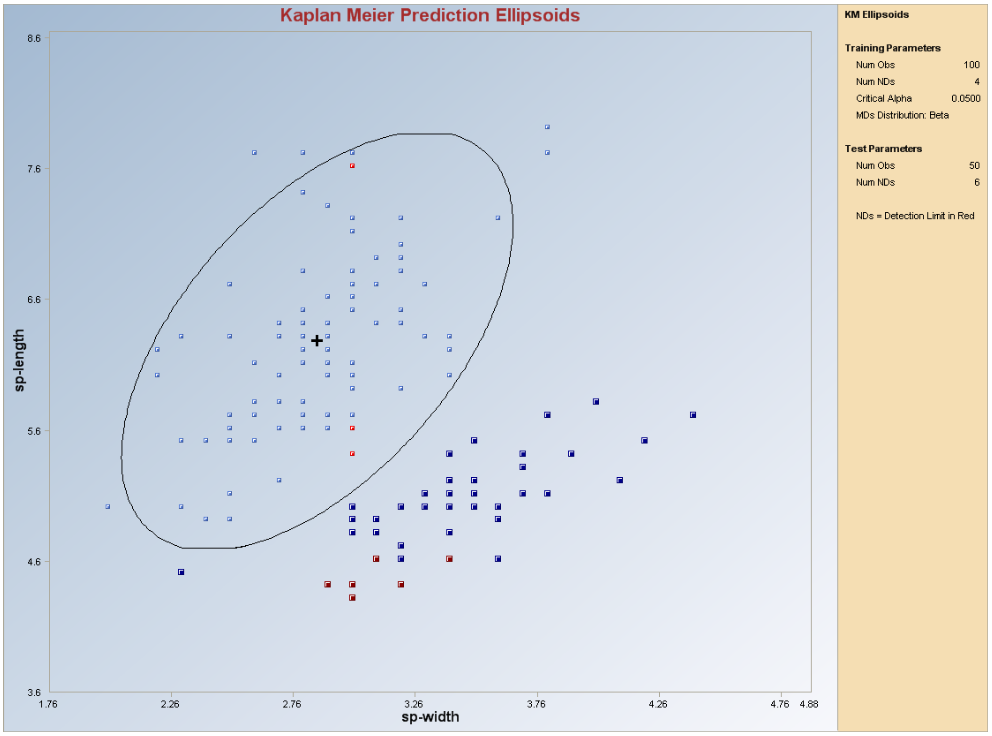
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the Ellipsoids.

Output example (Using All Data): The data set "**Fulliris.xls**" was used for the Ellipsoids. The options used are shown in the options window screenshot above. The ellipses are being drawn by groups.



Output example (Using Background/Training): The data set "**Fulliris.xls**" was used for the Ellipsoids.



***Note**: The observations in "**dark blue**" and "**bigger points**" are from group 3. The estimates of mean vector and the covariance matrix are calculated using the observations from group 1 and 2 only which were used as the "**Training/Background**" set for this data set.*