THE POWER TO KNOW.

SAS/STAT® 9.2 User's Guide

The ROBUSTREG Procedure

(Book Excerpt)

SAS® Documentation

This document is an individual chapter from *SAS/STAT® 9.2 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

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

1st electronic book, March 2008
2nd electronic book, February 2009

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

Chapter 74

The ROBUSTREG Procedure

Contents

Overview: ROBUSTREG Procedure

The main purpose of robust regression is to detect outliers and provide resistant (stable) results in the presence of outliers. In order to achieve this stability, robust regression limits the influence of outliers. Historically, three classes of problems have been addressed with robust regression techniques:

- problems with outliers in the y-direction (response direction)
- problems with multivariate outliers in the x-space (i.e., outliers in the covariate space, which are also referred to as leverage points)
- problems with outliers in both the y-direction and the x-space

Many methods have been developed in response to these problems. However, in statistical applications of outlier detection and robust regression, the methods most commonly used today are Huber M estimation, high breakdown value estimation, and combinations of these two methods. The ROBUSTREG procedure in SAS 9.2 provides four such methods: M estimation, LTS estimation, S estimation, and MM estimation.

1. M estimation was introduced by Huber (1973), and it is the simplest approach both computationally and theoretically. Although it is not robust with respect to leverage points, it is still used extensively in analyzing data for which it can be assumed that the contamination is mainly in the response direction.
2. Least trimmed squares (LTS) estimation is a high breakdown value method introduced by Rousseeuw (1984). The breakdown value is a measure of the proportion of contamination that an estimation method can withstand and still maintain its robustness. The performance of this method was improved by the FAST-LTS algorithm of Rousseeuw and Van Driessen (2000).
3. S estimation is a high breakdown value method introduced by Rousseeuw and Yohai (1984). With the same breakdown value, it has a higher statistical efficiency than LTS estimation.
4. MM estimation, introduced by Yohai (1987), combines high breakdown value estimation and M estimation. It has both the high breakdown property and a higher statistical efficiency than S estimation.

Features

The main features of the ROBUSTREG procedure are as follows:

- offers four estimation methods: M, LTS, S, and MM

- provides 10 weight functions for M estimation
- provides robust R^2 and deviance for all estimates
- provides asymptotic covariance and confidence intervals for regression parameter with the M, S, and MM methods
- provides robust Wald and F tests for regression parameters with the M and MM methods
- provides outlier and leverage-point diagnostics
- supports parallel computing for S and LTS estimates
- produces fit plots and diagnostic plots by using ODS Graphics

Getting Started: ROBUSTREG Procedure

The following examples demonstrate how you can use the ROBUSTREG procedure to fit a linear regression model and obtain outlier and leverage-point diagnostics.

M Estimation

This example shows how you can use the ROBUSTREG procedure to do M estimation, which is a commonly used method for outlier detection and robust regression when contamination is mainly in the response direction.

```
data stack;
   input  x1 x2 x3 y exp$ @@;
datalines;
80  27   89  42   e1   80  27   88  37   e2
75  25   90  37   e3   62  24   87  28   e4
62  22   87  18   e5   62  23   87  18   e6
62  24   93  19   e7   62  24   93  20   e8
58  23   87  15   e9   58  18   80  14  e10
58  18   89  14  e11   58  17   88  13  e12
58  18   82  11  e13   58  19   93  12  e14
50  18   89   8  e15   50  18   86   7  e16
50  19   72   8  e17   50  19   79   8  e18
50  20   80   9  e19   56  20   82  15  e20
70  20   91  15  e21
;
```

The data set stack is the well-known stackloss data set presented by Brownlee (1965). The data describe the operation of a plant for the oxidation of ammonia to nitric acid and consist of 21 four-dimensional observations. The explanatory variables for the response stackloss (y) are the rate of operation (x1), the cooling water inlet temperature (x2), and the acid concentration (x3).

The following ROBUSTREG statements analyze the data:

```
proc robustreg data=stack;
   model y = x1 x2 x3 / diagnostics leverage;
   id    exp;
   test  x3;
run;
```

By default, the procedure does M estimation with the bisquare weight function, and it uses the median method for estimating the scale parameter. The MODEL statement specifies the covariate effects. The DIAGNOSTICS option requests a table for outlier diagnostics, and the LEVERAGE option adds leverage-point diagnostic results to this table for continuous covariate effects. The ID statement specifies that the variable exp is used to identify each observation (experiment) in this table. If the ID statement is omitted, the observation number is used to identify the observations. The TEST statement requests a test of significance for the covariate effects specified. The results of this analysis are displayed in the following figures.

Figure 74.1 Model Fitting Information and Summary Statistics

The ROBUSTREG Procedure

Model Information

Data Set	WORK.STACK
Dependent Variable	y
Number of Independent Variables	3
Number of Observations	21
Method	M Estimation

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
x1	53.0000	58.0000	62.0000	60.4286	9.1683	5.9304
x2	18.0000	20.0000	24.0000	21.0952	3.1608	2.9652
x3	82.0000	87.0000	89.5000	86.2857	5.3586	4.4478
y	10.0000	15.0000	19.5000	17.5238	10.1716	5.9304

Figure 74.1 displays the model fitting information and summary statistics for the response variable and the continuous covariates. The columns labeled Q1, Median, and Q3 provide the lower quantile, median, and upper quantile, respectively. The column labeled MAD provides a robust estimate of the univariate scale, which is computed as the standardized median absolute deviation (MAD). See Huber (1981, p. 108) for more details about the standardized MAD. The columns labeled Mean and Standard Deviation provide the usual mean and standard deviation. A large difference between the standard deviation and the MAD for a variable indicates some extreme values for this variable. In the stackloss data, the stackloss (response y) has the biggest difference between the standard deviation and the MAD.

Figure 74.2 Model Parameter Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-42.2854	9.5045	-60.9138	-23.6569	19.79	<.0001
x1	1	0.9276	0.1077	0.7164	1.1387	74.11	<.0001
x2	1	0.6507	0.2940	0.0744	1.2270	4.90	0.0269
x3	1	-0.1123	0.1249	-0.3571	0.1324	0.81	0.3683
Scale	1	2.2819					

Figure 74.2 displays the table of robust parameter estimates, standard errors, and confidence limits. The row labeled Scale provides a point estimate of the scale parameter in the linear regression model, which is obtained by the median method. See the section "M Estimation" on page 5666 for more information about scale estimation methods. For the stackloss data, M estimation yields the fitted linear model:

$$\hat{y} = -42.2845 + 0.9276\text{x1} + 0.6507\text{x2} - 0.1123\text{x3}$$

Figure 74.3 Diagnostics

Diagnostics

Obs	exp	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
1	e1	2.2536	5.5284	*	1.0995	
2	e2	2.3247	5.6374	*	-1.1409	
3	e3	1.5937	4.1972	*	1.5604	
4	e4	1.2719	1.5887		3.0381	*
21	e21	2.1768	3.6573	*	-4.5733	*

Figure 74.3 displays outlier and leverage-point diagnostics. Standardized robust residuals are computed based on the estimated parameters. Both the Mahalanobis distance and the robust MCD distance are displayed. Outliers and leverage points, identified with asterisks, are defined by the standardized robust residuals and robust MCD distances that exceed the corresponding cutoff values displayed in the diagnostics summary. Observations 4 and 21 are outliers because their standardized robust residuals exceed the cutoff value in absolute value. The procedure detects 4 observations with high leverage. Leverage points (points with high leverage) with smaller standardized robust residuals than the cutoff value in absolute value are called good leverage points; others are called bad leverage points. Observation 21 is a bad leverage point.

Two particularly useful plots for revealing outliers and leverage points are a scatter plot of the standardized robust residuals against the robust distances (RDPLOT) and a scatter plot of the robust distances against the classical Mahalanobis distances (DDPLOT).

For the stackloss data, the following statements produce the RDPLOT in Figure 74.4 and the DDPLOT in Figure 74.5. The histogram and the normal quantile-quantile plots for the standardized robust residuals are also created with the HISTOGRAM and QQPLOT options in the PROC ROBUSTREG statement. See Figure 74.6 and Figure 74.7.

```
ods graphics on;

proc robustreg data=stack
        plots=(rdplot ddplot reshistogram resqqplot);
   model y = x1 x2 x3;
run;

ods graphics off;
```

Figure 74.4 RDPLOT for Stackloss Data

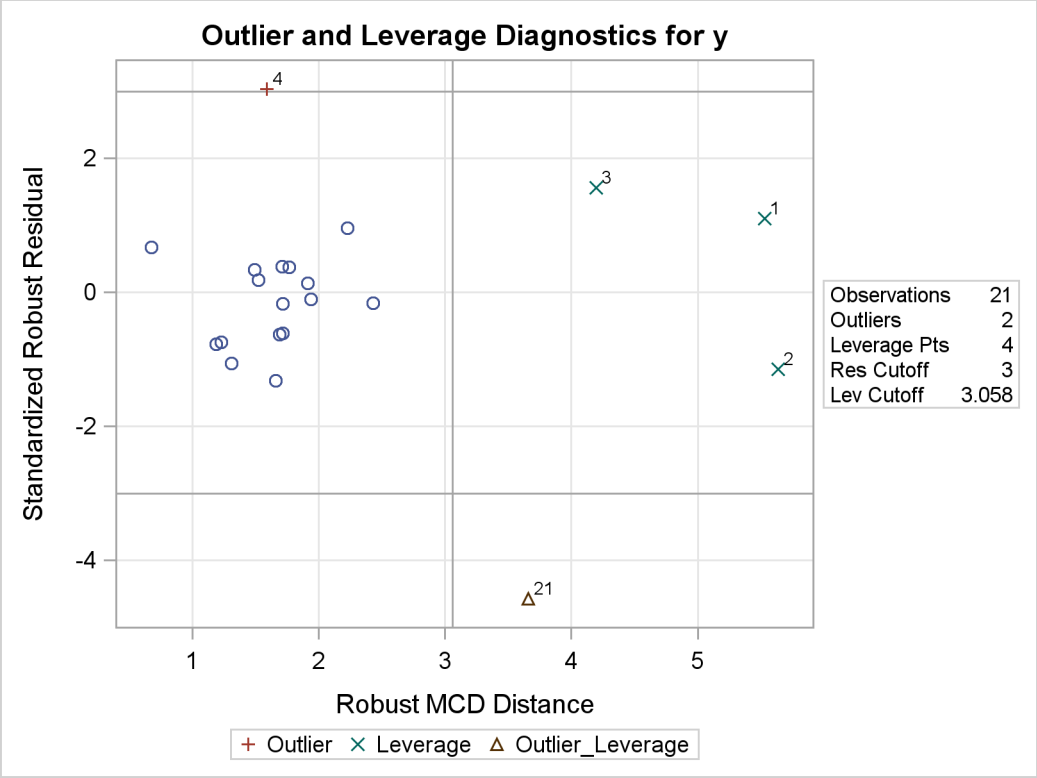


Figure 74.5 DDPLOT for Stackloss Data

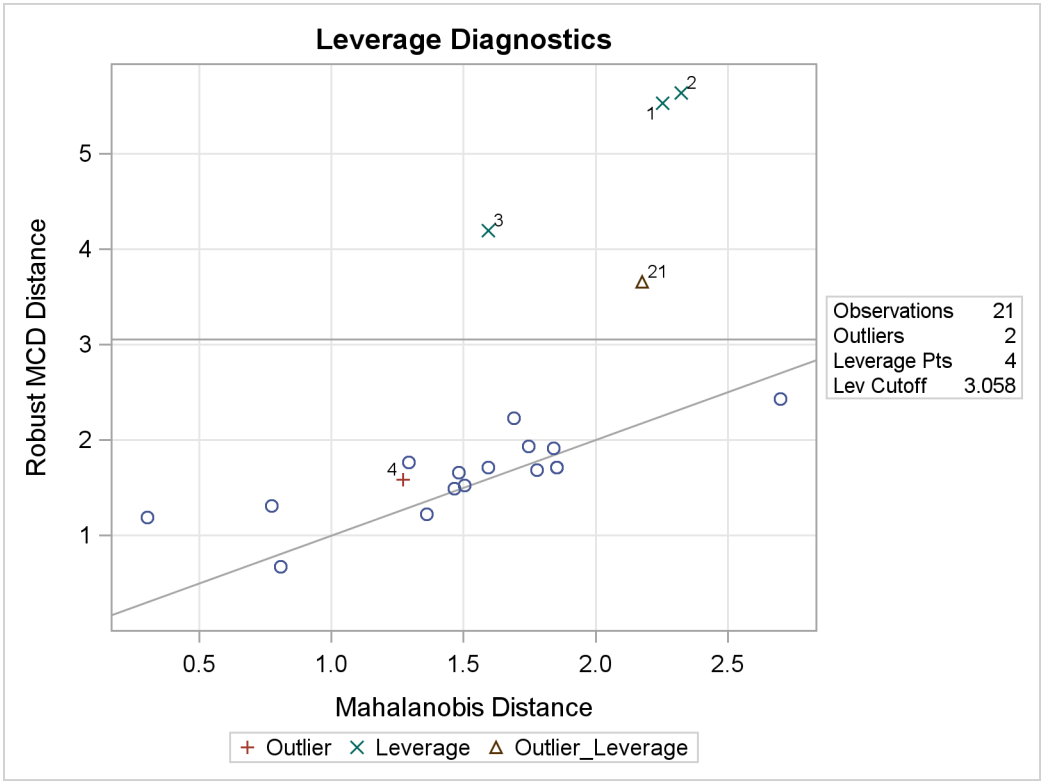


Figure 74.6 Histogram

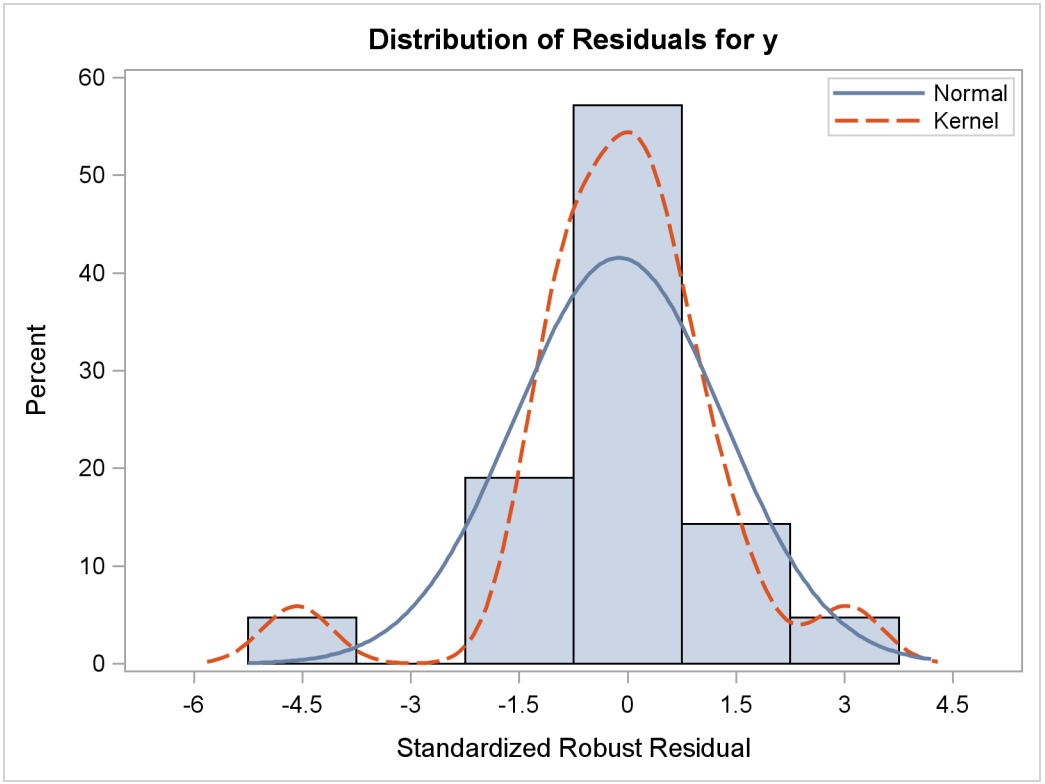
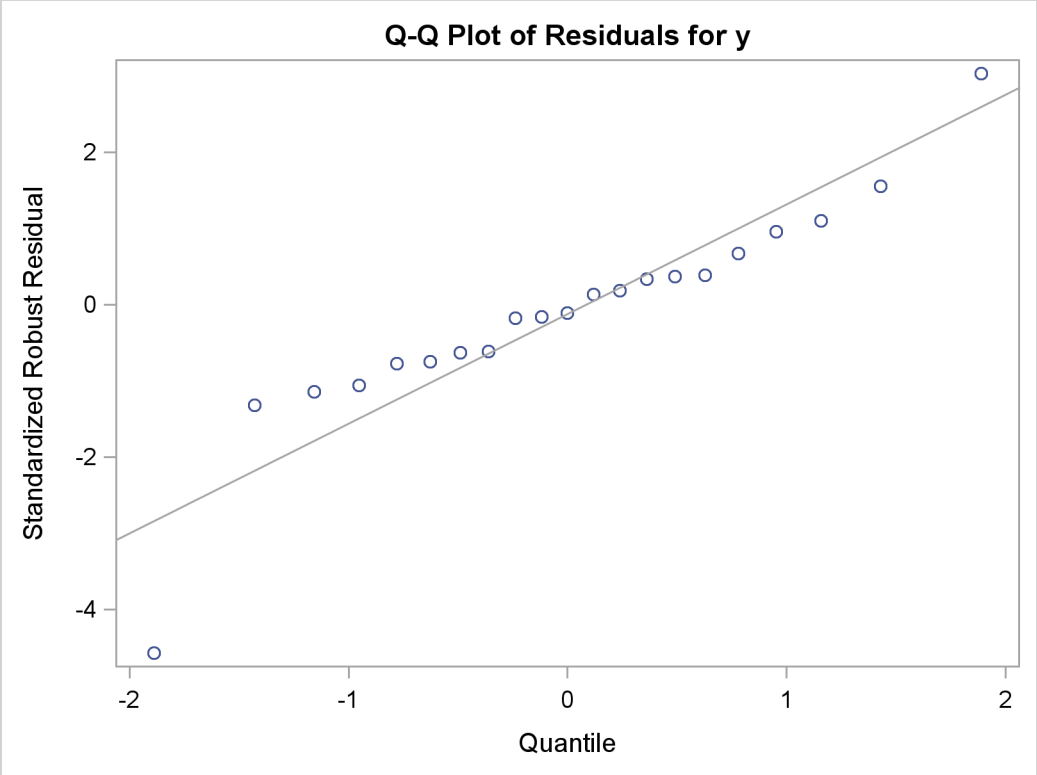


Figure 74.7 Q-Q Plot



These plots are helpful in identifying outliers as well as good and bad high leverage points.

These graphical displays are requested by specifying the ODS GRAPHICS statement and the PLOTS= option in the PROC statement. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For specific information about the graphics available in the ROBUSTREG procedure, see the section "ODS Graphics" on page 5686.

Figure 74.8 Goodness-of-Fit Statistics

Goodness-of-Fit

Statistic	Value
R-Square	0.6659
AICR	29.5231
BICR	36.3361
Deviance	125.7905

Figure 74.8 displays robust versions of goodness-of-fit statistics for the model. You can use the robust information criteria, AICR and BICR, for model selection and comparison. For both AICR and BICR, the lower the value, the more desirable the model.

Figure 74.9 Test of Significance

Robust Linear Tests

Test

Test	Test Statistic	Lambda	DF	Chi-Square	Pr > ChiSq
Rho	0.9378	0.7977	1	1.18	0.2782
Rn2	0.8092		1	0.81	0.3683

Figure 74.9 displays the test results requested by the TEST statement. The ROBUSTREG procedure conducts two robust linear tests, the ρ test and the R_n^2 test. See the section "Linear Tests" on page 5672 for information about how the procedure computes test statistics and the correction factor lambda. Due to the large p-values for both tests, you can conclude that the effect x3 is not significant at the 5% level.

For the bisquare weight function, the default tuning constant, $c = 4.685$, is chosen to yield a 95% asymptotic efficiency of the M estimates with the Gaussian distribution. See the section "M Estimation" on page 5666 for details. The smaller the constant c, the lower the asymptotic efficiency but the sharper the M estimate as an outlier detector. For the stackloss data set, you could consider using a sharper outlier detector.

In the following invocation of the ROBUSTREG procedure, a smaller constant, $c = 3.5$, is used. This tuning constant corresponds to an efficiency close to 85%. See Chen and Yin (2002) for relations between the tuning constant and asymptotic efficiency of M estimates.

```
proc robustreg method=m(wf=bisquare(c=3.5)) data=stack;
   model y = x1 x2 x3 / diagnostics leverage;
   id    exp;
   test  x3;
run;
```

Figure 74.10 Model Parameter Estimates

The ROBUSTREG Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-37.1076	5.4731	-47.8346	-26.3805	45.97	<.0001
x1	1	0.8191	0.0620	0.6975	0.9407	174.28	<.0001
x2	1	0.5173	0.1693	0.1855	0.8492	9.33	0.0022
x3	1	-0.0728	0.0719	-0.2138	0.0681	1.03	0.3111
Scale	1	1.4265					

Figure 74.10 displays the table of robust parameter estimates, standard errors, and confidence limits with the constant $c = 3.5$.

The refitted linear model is

$$\hat{y} = -37.1076 + 0.8191x1 + 0.5173x2 - 0.0728x3$$

Figure 74.11 Diagnostics

Diagnostics

Obs	exp	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
1	e1	2.2536	5.5284	*	4.2719	*
2	e2	2.3247	5.6374	*	0.7158	
3	e3	1.5937	4.1972	*	4.4142	*
4	e4	1.2719	1.5887		5.7792	*
21	e21	2.1768	3.6573	*	-6.2727	*

Figure 74.11 displays outlier and leverage-point diagnostics with the constant $c = 3.5$. Besides observations 4 and 21, observations 1 and 3 are also detected as outliers.

LTS Estimation

If the data are contaminated in the x-space, M estimation does not do well. The following example shows how you can use LTS estimation to deal with this situation.

```
data hbk;
   input index$ x1 x2 x3 y @@;
datalines;
1   10.1  19.6  28.3   9.7     39  2.1  0.0  1.2  -0.7
2    9.5  20.5  28.9  10.1     40  0.5  2.0  1.2  -0.5
3   10.7  20.2  31.0  10.3     41  3.4  1.6  2.9  -0.1
4    9.9  21.5  31.7   9.5     42  0.3  1.0  2.7  -0.7
5   10.3  21.1  31.1  10.0     43  0.1  3.3  0.9   0.6
6   10.8  20.4  29.2  10.0     44  1.8  0.5  3.2  -0.7
7   10.5  20.9  29.1  10.8     45  1.9  0.1  0.6  -0.5
8    9.9  19.6  28.8  10.3     46  1.8  0.5  3.0  -0.4
9    9.7  20.7  31.0   9.6     47  3.0  0.1  0.8  -0.9
10   9.3  19.7  30.3   9.9     48  3.1  1.6  3.0   0.1
11  11.0  24.0  35.0  -0.2     49  3.1  2.5  1.9   0.9
12  12.0  23.0  37.0  -0.4     50  2.1  2.8  2.9  -0.4
13  12.0  26.0  34.0   0.7     51  2.3  1.5  0.4   0.7
14  11.0  34.0  34.0   0.1     52  3.3  0.6  1.2  -0.5
15   3.4   2.9   2.1  -0.4     53  0.3  0.4  3.3   0.7
16   3.1   2.2   0.3   0.6     54  1.1  3.0  0.3   0.7
17   0.0   1.6   0.2  -0.2     55  0.5  2.4  0.9   0.0
18   2.3   1.6   2.0   0.0     56  1.8  3.2  0.9   0.1
19   0.8   2.9   1.6   0.1     57  1.8  0.7  0.7   0.7
```

```
20   3.1   3.4   2.2   0.4     58  2.4  3.4  1.5  -0.1
21   2.6   2.2   1.9   0.9     59  1.6  2.1  3.0  -0.3
22   0.4   3.2   1.9   0.3     60  0.3  1.5  3.3  -0.9
23   2.0   2.3   0.8  -0.8     61  0.4  3.4  3.0  -0.3
24   1.3   2.3   0.5   0.7     62  0.9  0.1  0.3   0.6
25   1.0   0.0   0.4  -0.3     63  1.1  2.7  0.2  -0.3
26   0.9   3.3   2.5  -0.8     64  2.8  3.0  2.9  -0.5
27   3.3   2.5   2.9  -0.7     65  2.0  0.7  2.7   0.6
28   1.8   0.8   2.0   0.3     66  0.2  1.8  0.8  -0.9
29   1.2   0.9   0.8   0.3     67  1.6  2.0  1.2  -0.7
30   1.2   0.7   3.4  -0.3     68  0.1  0.0  1.1   0.6
31   3.1   1.4   1.0   0.0     69  2.0  0.6  0.3   0.2
32   0.5   2.4   0.3  -0.4     70  1.0  2.2  2.9   0.7
33   1.5   3.1   1.5  -0.6     71  2.2  2.5  2.3   0.2
34   0.4   0.0   0.7  -0.7     72  0.6  2.0  1.5  -0.2
35   3.1   2.4   3.0   0.3     73  0.3  1.7  2.2   0.4
36   1.1   2.2   2.7  -1.0     74  0.0  2.2  1.6  -0.9
37   0.1   3.0   2.6  -0.6     75  0.3  0.4  2.6   0.2
38   1.5   1.2   0.2   0.9
;
```

The data set hbk is an artificial data set generated by Hawkins, Bradu, and Kass (1984). Both ordinary least squares (OLS) estimation and M estimation (not shown here) suggest that observations 11 to 14 are outliers. However, these four observations were generated from the underlying model, whereas observations 1 to 10 were contaminated. The reason that OLS estimation and M estimation do not pick up the contaminated observations is that they cannot distinguish good leverage points (observations 11 to 14) from bad leverage points (observations 1 to 10). In such cases, the LTS method identifies the true outliers.

The following statements invoke the ROBUSTREG procedure with the LTS estimation method:

```
proc robustreg data=hbk fwls method=lts;
   model y = x1 x2 x3 / diagnostics leverage;
   id index;
run;
```

Figure 74.12 Model Fitting Information and Summary Statistics

The ROBUSTREG Procedure

Model Information

Data Set	WORK.HBK
Dependent Variable	y
Number of Independent Variables	3
Number of Observations	75
Method	LTS Estimation

Figure 74.12 *continued*

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
x1	0.8000	1.8000	3.1000	3.2067	3.6526	1.9274
x2	1.0000	2.2000	3.3000	5.5973	8.2391	1.6309
x3	0.9000	2.1000	3.0000	7.2307	11.7403	1.7791
y	−0.5000	0.1000	0.7000	1.2787	3.4928	0.8896

Figure 74.12 displays the model fitting information and summary statistics for the response variable and independent covariates.

Figure 74.13 LTS Profile

LTS Profile

Total Number of Observations	75
Number of Squares Minimized	57
Number of Coefficients	4
Highest Possible Breakdown Value	0.2533

Figure 74.13 displays information about the LTS fit, which includes the breakdown value of the LTS estimate. The breakdown value is a measure of the proportion of contamination that an estimation method can withstand and still maintain its robustness. In this example the LTS estimate minimizes the sum of 57 smallest squares of residuals. It can still estimate the true underlying model if the remaining 18 observations are contaminated. This corresponds to the breakdown value around 0.25, which is set as the default.

Figure 74.14 LTS Parameter Estimates

LTS Parameter Estimates

Parameter	DF	Estimate
Intercept	1	−0.3431
x1	1	0.0901
x2	1	0.0703
x3	1	−0.0731
Scale (sLTS)	0	0.7451
Scale (Wscale)	0	0.5749

Figure 74.14 displays parameter estimates for covariates and scale. Two robust estimates of the scale parameter are displayed. See the section "Final Weighted Scale Estimator" on page 5675 for how these estimates are computed. The weighted scale estimator (Wscale) is a more efficient estimator of the scale parameter.

Figure 74.15 Diagnostics

Diagnostics

Obs	index	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
1	1	1.9168	29.4424	*	17.0868	*
3	2	1.8558	30.2054	*	17.8428	*
5	3	2.3137	31.8909	*	18.3063	*
7	4	2.2297	32.8621	*	16.9702	*
9	5	2.1001	32.2778	*	17.7498	*
11	6	2.1462	30.5892	*	17.5155	*
13	7	2.0105	30.6807	*	18.8801	*
15	8	1.9193	29.7994	*	18.2253	*
17	9	2.2212	31.9537	*	17.1843	*
19	10	2.3335	30.9429	*	17.8021	*
21	11	2.4465	36.6384	*	0.0406	
23	12	3.1083	37.9552	*	−0.0874	
25	13	2.6624	36.9175	*	1.0776	
27	14	6.3816	41.0914	*	−0.7875	

Figure 74.15 displays outlier and leverage-point diagnostics. The ID variable index is used to identify the observations. If you do not specify this ID variable, the observation number is used to identify the observations. However, the observation number depends on how the data are read. The first 10 observations are identified as outliers, and observations 11 to 14 are identified as good leverage points.

Figure 74.16 Final Weighted LS Estimates

Parameter Estimates for Final Weighted Least Squares Fit

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	−0.1805	0.1044	−0.3852	0.0242	2.99	0.0840
x1	1	0.0814	0.0667	−0.0493	0.2120	1.49	0.2222
x2	1	0.0399	0.0405	−0.0394	0.1192	0.97	0.3242
x3	1	−0.0517	0.0354	−0.1210	0.0177	2.13	0.1441
Scale	0	0.5572					

Figure 74.16 displays the final weighted least squares estimates. These estimates are least squares estimates computed after deleting the detected outliers.

Syntax: ROBUSTREG Procedure

The following statements are available in PROC ROBUSTREG:

PROC ROBUSTREG < *options* > **;**
BY *variables* **;**
CLASS *variables* **;**
ID *variables* **;**
MODEL *response*=< *effects* > < / *options* > **;**
OUTPUT < *OUT=SAS-data-set* > < *options* > **;**
PERFORMANCE < *options* > **;**
TEST *effects* **;**
WEIGHT *variable* **;**

The PROC ROBUSTREG statement invokes the procedure. The METHOD= option in the PROC ROBUSTREG statement selects one of the four estimation methods, M, LTS, S, and MM. By default, Huber M estimation is used. The MODEL statement is required and specifies the variables used in the regression. Main effects and interaction terms can be specified in the MODEL statement, as in the GLM procedure (Chapter 39, "The GLM Procedure.") The CLASS statement specifies which explanatory variables are treated as categorical. These variables are allowed in the MODEL statement only for M estimation and not for other estimation methods. The ID statement names variables to identify observations in the outlier diagnostics tables. The WEIGHT statement identifies a variable in the input data set whose values are used to weight the observations. The OUTPUT statement creates an output data set containing final weights, predicted values, and residuals. The TEST statement requests robust linear tests for the model parameters. The PERFORMANCE statement tunes the performance of the procedure by using single or multiple processors available on the hardware. In one invocation of PROC ROBUSTREG, multiple OUTPUT and TEST statements are allowed.

PROC ROBUSTREG Statement

PROC ROBUSTREG < *options* > **;**

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

COVOUT
saves the estimated covariance matrix in the OUTEST= data set. This option is not supported for LTS estimation.

DATA=*SAS-data-set*
specifies the input SAS data set used by PROC ROBUSTREG. By default, the most recently created SAS data set is used.

FWLS

requests that final weighted least squares estimates be computed. These estimates are equivalent to the least squares estimates after the detected outliers are deleted.

INEST=*SAS-data-set*

specifies an input SAS data set that contains initial estimates for all the parameters in the model. See the section "INEST= Data Set" on page 5683 for a detailed description of the contents of the INEST= data set.

ITPRINT

displays the iteration history for the iteratively reweighted least squares algorithm used by M and MM estimation. You can also use this option in the MODEL statement.

NAMELEN=*n*

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

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sorting order for the levels of the classification variables (specified in the CLASS statement). This ordering determines which parameters in the model correspond to each level in the data. The following table explains how PROC ROBUSTREG interprets values of the ORDER= option.

Table 74.1 Options for Order

Value of ORDER=	Levels Sorted By
DATA	order of appearance in the input data set
FORMATTED	formatted value
FREQ	descending frequency count; levels with the most observations come first in the order
INTERNAL	unformatted value

By default, ORDER=FORMATTED. For FORMATTED and INTERNAL, the sort order is machine dependent. For more information about sorting order, refer to the chapter titled "The SORT Procedure" in the *Base SAS Procedures Guide*.

OUTEST=*SAS-data-set*

specifies an output SAS data set containing the parameter estimates, and, if the COVOUT option is specified, the estimated covariance matrix. See the section "OUTEST= Data Set" on page 5684 for a detailed description of the contents of the OUTEST= data set.

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

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

specifies options that control details of the plots. If you have enabled ODS GRAPHICS but do not specify the PLOTS= option, then PROC ROBUSTREG produces the robust fit plot by default when the model includes a single continuous independent variable.

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

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

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

ALL
creates all appropriate plots.

DDPLOT<(LABEL=ALL | LEVERAGE | NONE | OUTLIER)>
creates a plot of robust distance against Mahalanobis distance. See the section "Leverage Point and Outlier Detection" on page 5683 for details about robust distance. The LABEL= option specifies how the points on this plot are to be labeled, as summarized by the following table.

Table 74.2 Options for Label

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

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

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

FITPLOT<(NOLIMITS)>
creates a plot of robust fit against the single independent continuous variable specified in the model. You can request this plot when only a single independent continuous variable is specified in the model. Confidence limits are added on the plot by default. The NOLIMITS option suppresses these limits.

HISTOGRAM
creates a histogram for the standardized robust residuals. The histogram is superimposed with a normal density curve and a kernel density curve.

NONE
suppresses all plots.

QQPLOT
creates the normal quantile-quantile plot for the standardized robust residuals.

RDPLOT<(LABEL=ALL | LEVERAGE | NONE | OUTLIER)>
creates the plot of standardized robust residual against robust distance. See the section "Leverage Point and Outlier Detection" on page 5683 for details about robust distance. The LABEL= option specifies a label method for points on this plot. These label methods are described in Table 74.2.

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

SEED=*number*

specifies the seed for the random number generator used to randomly select the subgroups and subsets for LTS and S estimation. By default or if you specify zero, the ROBUSTREG procedure generates a random seed.

METHOD=*method type* **<(** *options* **)>**

specifies the estimation method and *options* specify some additional options for the estimation method. PROC ROBUSTREG provides four estimation methods: M estimation, LTS estimation, S estimation, and MM estimation. The default method is M estimation.

NOTE: Since the LTS and S methods use subsampling algorithms, these methods are not suitable in an analysis with categorical independent variables specified in the CLASS statement. These methods are not suitable in an analysis with continuous independent variables that have only a few unequal values or a few unequal values within one BY group. This also applies to the initial LTS and S estimates in the MM method. In summary, if the model includes categorical independent variables or continuous independent variables with a few unequal values, the M method is recommended.

Options with METHOD=M

With METHOD=M, you can specify the following additional *options*:

ASYMPCOV=*H1* | *H2* | *H3*

specifies the type of asymptotic covariance computed for the M estimate. The three types are described in the section "Asymptotic Covariance and Confidence Intervals" on page 5671. By default, ASYMPCOV= H1.

CONVERGENCE=*criterion***<(EPS=***value***)>**

specifies a convergence criterion for the M estimate. The three criteria listed in the following table are available.

Table 74.3 Options to Specify Convergence Criteria

Type	Option
coefficient	CONVERGENCE=COEF
residual	CONVERGENCE=RESID
weight	CONVERGENCE=WEIGHT

By default, CONVERGENCE = COEF. You can specify the precision of the convergence criterion with the EPS= option. By default, EPS=1.E−8.

MAXITER=*n*

sets the maximum number of iterations during the parameter estimation. By default, MAXITER=1000.

SCALE=*scale type | value*

specifies the scale parameter or a method for estimating the scale parameter. These methods and options are summarized in the following table.

Table 74.4 Options to Specify Scale

Scale	Option	Default d
Fixed constant	SCALE=*value*	
Huber estimate	SCALE=HUBER<(D=d)>	2.5
Median estimate	SCALE=MED	
Tukey estimate	SCALE=TUKEY<(D=d)>	2.5

By default, SCALE = MED.

WF | WEIGHTFUNCTION=*function type*

specifies the weight function used for the M estimate. The ROBUSTREG procedure provides 10 weight functions, which are listed in the following table. You can specify the parameters in these functions with the A=, B=, and C= options. These functions are described in the section "M Estimation" on page 5666. The default weight function is bisquare.

Table 74.5 Options to Specify Weight Functions

Weight Function	Option	Default a, b, c
andrews	WF=ANDREWS<(C=c)>	1.339
bisquare	WF=BISQUARE<(C=c)>	4.685
cauchy	WF=CAUCHY<(C=c)>	2.385
fair	WF=FAIR<(C=c)>	1.4
hampel	WF=HAMPEL<(<A=a> <B=b> <C=c>)>	2, 4, 8
huber	WF=HUBER<(C=c)>	1.345
logistic	WF=LOGISTIC<(C=c)>	1.205
median	WF=MEDIAN<(C=c)>	0.01
talworth	WF=TALWORTH<(C=c)>	2.795
welsch	WF=WELSCH<(C=c)>	2.985

Options with METHOD=LTS

With METHOD=LTS, you can specify the following additional *options*:

CSTEP=*n*

specifies the number of C-steps for the LTS estimate. See the section "LTS Estimate" on page 5673 for how the default value is determined.

IADJUST=ALL | NONE

requests (IADJUST=ALL) or suppresses (IADJUST=NONE) the intercept adjustment for all estimates in the LTS algorithm. By default, the intercept adjustment is used for data sets with less than 10000 observations. See the section "Algorithm" on page 5674 for details.

H=*n*

specifies the quantile for the LTS estimate. See the section "LTS Estimate" on page 5673 for how the default value is determined.

NBEST=*n*

specifies the number of best solutions kept for each subgroup during the computation of the LTS estimate. The default number is 10, which is the maximum number allowed.

NREP=*n*

specifies the number of repeats of least squares fit in subgroups during the computation of the LTS estimate. See the section "LTS Estimate" on page 5673 for how the default number is determined.

SUBANALYSIS

requests a display of the subgrouping information and parameter estimates within subgroups. This option generates the following ODS tables.

Table 74.6 ODS Tables Available with SUBANALYSIS Option

ODS Table Name	Description
BestEstimates	Best final estimates for LTS
BestSubEstimates	Best estimates for each subgroup
CStep	C-step information for LTS
Groups	Grouping information for LTS

SUBGROUPSIZE=*n*

specifies the data set size of the subgroups in the computation of the LTS estimate. The default number is 300.

Options with METHOD=S

With METHOD=S, you can specify the following additional *options*:

ASYMPCOV=*H1* | *H2* | *H3* | *H4*

specifies the type of asymptotic covariance computed for the S estimate. The four types are described in the section "Asymptotic Covariance and Confidence Intervals" on page 5678. By default, ASYMPCOV= H4.

CHIF= TUKEY | YOHAI

specifies the χ function for the S estimate. PROC ROBUSTREG provides two χ functions, Tukey's bisquare function and Yohai's optimal function, which you can request with CHIF=TUKEY and CHIF=YOHAI, respectively. The default is Tukey's bisquare function.

EFF=*value*

specifies the efficiency (as a fraction) for the S estimate. The parameter k_0 in the χ function is determined by this efficiency. The default efficiency is determined such that

the consistent S estimate has the breakdown value of 25%. This option is overwritten by the K0= option if both of them are used.

K0=*value*

specifies the $k0$ parameter in the χ function of the S estimate. For CHIF=TUKEY, the default is 1.548. For CHIF=YOHAI, the default is 0.66. These default values correspond to a 50% breakdown value of the consistent S estimate.

MAXITER=*n*

sets the maximum number of iterations for computing the scale parameter of the S estimate. By default, MAXITER=1000.

NREP=*n*

specifies the number of repeats of subsampling in the computation of the S estimate. See the section "Algorithm" on page 5676 for how the default number of repeats is determined.

NOREFINE

suppresses the refinement for the S estimate. See the section "Algorithm" on page 5676 for details.

SUBSETSIZE=*n*

specifies the size of the subset for the S estimate. See the section "Algorithm" on page 5676 for how its default value is determined.

TOLERANCE=*value*

specifies the tolerance for the S estimate of the scale. The default value is 0.001.

Options with METHOD=MM

With METHOD=MM, you can specify the following additional *options*:

ASYMPCOV=*H1* | *H2* | *H3* | *H4*

specifies the type of asymptotic covariance computed for the MM estimate. The four types are described in the section "Details: ROBUSTREG Procedure" on page 5666. By default, ASYMPCOV= H4.

BIASTEST< (ALPHA=*number***) >**

requests the bias test for the final MM estimate. See the section "Bias Test" on page 5680 for details about this test.

CHIF= TUKEY | YOHAI

selects the χ function for the MM estimate. PROC ROBUSTREG provides two χ functions: Tukey's bisquare function and Yohai's optimal function, which you can request with CHIF=TUKEY and CHIF=YOHAI, respectively. The default is Tukey's bisquare function. This χ function is also used by the initial S estimate if you specify the INITEST=S option.

CONVERGENCE=*criterion***<(EPS=***number***)>**
specifies a convergence criterion for the MM estimate. The three criteria listed in the following table are available.

Table 74.7 Options to Specify Convergence Criteria

Type	Option
coefficient	CONVERGENCE=COEF
residual	CONVERGENCE=RESID
weight	CONVERGENCE=WEIGHT

By default, CONVERGENCE = COEF. You can specify the precision of the convergence criterion with the EPS= option. By default, EPS=1.E−8.

EFF=*value*
specifies the efficiency (as a fraction) for the MM estimate. The parameter k_1 in the χ function is determined by this efficiency. The default efficiency is set to 0.85, which corresponds to $k_1 = 3.440$ for CHIF=TUKEY or $k_1 = 0.868$ for CHIF=YOHAI.

INITH=*n*
specifies the integer h for the initial LTS estimate used by the MM estimator. See the section "Algorithm" on page 5680 for how to specify h and how the default is determined.

INITEST=LTS | S
specifies the initial estimator for the MM estimator. By default, the LTS estimator with its default settings is used as the initial estimator for the MM estimator.

K0=*number*
specifies the parameter k_0 in the χ function for the MM estimate. For CHIF=TUKEY, the default is $k_0 = 2.9366$. For CHIF=YOHAI, the default is $k_0 = 0.7405$. These default values correspond to the 25% breakdown value of the MM estimator.

MAXITER=*n*
sets the maximum number of iterations during the parameter estimation. By default, MAXITER=1000.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC ROBUSTREG to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

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

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

For more information about the BY statement, see *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the *Base SAS Procedures Guide*.

CLASS Statement

CLASS *variables* **;**

Explanatory variables that are classification variables rather than quantitative numeric variables must be listed in the CLASS statement. For each explanatory variable listed in the CLASS statement, indicator variables are generated for the levels assumed by the CLASS variable. If the CLASS statement is used, it must appear before the MODEL statement.

ID Statement

ID *variables* **;**

When the diagnostics table is requested with the DIAGNOSTICS option in the MODEL statement, the variables listed in the ID statement are displayed in addition to the observation number. These variables can be used to identify each observation. If the ID statement is omitted, the observation number is used to identify the observations.

MODEL Statement

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

Main effects and interaction terms can be specified in the MODEL statement, as in the GLM procedure (Chapter 39, "The GLM Procedure.") Classification variables are not allowed in the MODEL statement when you specify MM estimation or LTS estimation by using the METHOD= option in the PROC statement.

The optional *label*, which must be a valid SAS name, is used to label the model in the OUTEST data set.

You can specify the following options for the model fit.

ALPHA=*value*
specifies the significance level for the confidence intervals for regression parameters. The value must be between 0 and 1. By default, ALPHA=0.05.

CORRB
produces the estimated correlation matrix of the parameter estimates.

COVB
produces the estimated covariance matrix of the parameter estimates.

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

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

ITPRINT
displays the iteration history for the iteratively reweighted least squares algorithm used by M and MM estimation. You can also use this option in the PROC statement.

LEVERAGE<(<CUTOFF=*value* | **CUTOFFALPHA=***value***> <QUANTILE=***n***>)>**

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

Since the MCD algorithm uses subsampling, it is not suitable to apply the leverage-point analysis to continuous variables that have only a few nonzero values or a few nonzero values within one BY group.

NOGOODFIT
suppresses the computation of goodness-of-fit statistics.

NOINT
specifies no-intercept regression.

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

OUTPUT Statement

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

The OUTPUT statement creates an output SAS data set containing statistics calculated after fitting the model. At least one specification of the form *keyword=name* is required.

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

The following specifications can appear in the OUTPUT statement:

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

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

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

LEVERAGE specifies a variable to indicate leverage points. To include this variable in the OUTPUT data set, you must specify the LEVERAGE option in the PROC statement. See the section "Leverage Point and Outlier Detection" on page 5683 for how to define LEVERAGE.

OUTLIER specifies a variable to indicate outliers. See the section "Leverage Point and Outlier Detection" on page 5683 for how to define OUTLIER.

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

RESIDUAL | R specifies a variable to contain the unstandardized residuals

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

SRESIDUAL | SR specifies a variable to contain the standardized residuals

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

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

WEIGHT specifies a variable to contain the computed final weights.

PERFORMANCE Statement

The PERFORMANCE statement is used to change default options that affect the performance of PROC ROBUSTREG and to request tables that show the performance options in effect and timing details. See Chen (2002) for some empirical results.

PERFORMANCE < *options* > **;**

The following options are available:

CPUCOUNT=*1-1024*

CPUCOUNT=*ACTUAL*

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

DETAILS

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

THREADS

enables multithreaded computation. This option overrides the SAS system option THREADS | NOTHREADS.

NOTHREADS

disables multithreaded computation. This option overrides the SAS system option THREADS | NOTHREADS.

TEST Statement

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

With M estimation and MM estimation, the TEST statement provides a means of obtaining a test for the canonical linear hypothesis concerning the parameters of the tested effects:

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

where q is the total number of parameters of the tested effects.

PROC ROBUSTREG provides two kinds of robust tests: the ρ test and the R_n^2 test. They are described in the section "Details: ROBUSTREG Procedure" on page 5666. No test is available for LTS and S estimation.

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

WEIGHT Statement

WEIGHT *variable* **;**

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

If you want to use fixed weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. The values of the WEIGHT variable can be nonintegral and are not truncated. Observations with nonpositive or missing values for the weight variable do not contribute to the fit of the model.

Details: ROBUSTREG Procedure

This section describes the statistical and computational aspects of the ROBUSTREG procedure. The following notation is used throughout this section.

Let $X = (x_{ij})$ denote an $n \times p$ matrix, $y = (y_1, \ldots, y_n)^T$ denote a given n-vector of responses, and $\theta = (\theta_1, \ldots, \theta_p)^T$ denote an unknown p-vector of parameters or coefficients whose components are to be estimated. The matrix X is called the design matrix. Consider the usual linear model

$$y = X\theta + e$$

where $e = (e_1, \ldots, e_n)^T$ is an n-vector of unknown errors. It is assumed that (for a given X) the components e_i of e are independent and identically distributed according to a distribution $L(\cdot/\sigma)$, where σ is a scale parameter (usually unknown). Often $L(\cdot) \approx \Phi(\cdot)$, the standard normal distribution function. The vector of residuals for a given value of $\hat{\theta}$ is denoted by $r = (r_1, \ldots, r_n)^T$ and the ith row of the matrix X is denoted by x_i^T.

M Estimation

M estimation in the context of regression was first introduced by Huber (1973) as a result of making the least squares approach robust. Although M estimators are not robust with respect to leverage points, they are popular in applications where leverage points are not an issue.

Instead of minimizing a sum of squares of the residuals, a Huber-type M estimator $\hat{\theta}_M$ of θ minimizes a sum of less rapidly increasing functions of the residuals:

$$Q(\theta) = \sum_{i=1}^{n} \rho(\frac{r_i}{\sigma})$$

where $r = y - X\theta$. For the ordinary least squares estimation, ρ is the square function, $\rho(z) = z^2$.

If σ is known, then by taking derivatives with respect to θ, $\hat{\theta}_{\boldsymbol{M}}$ is also a solution of the system of p equations:

$$\sum_{i=1}^{n} \psi(\frac{r_i}{\sigma}) x_{ij} = 0, \ j = 1, \ldots, p$$

where $\psi = \frac{\partial \rho}{\partial z}$. If ρ is convex, $\hat{\theta}_{\boldsymbol{M}}$ is the unique solution.

The ROBUSTREG procedure solves this system by using iteratively reweighted least squares (IRLS). The weight function $w(x)$ is defined as

$$w(z) = \frac{\psi(z)}{z}$$

The ROBUSTREG procedure provides 10 kinds of weight functions through the WEIGHTFUNCTION= option in the MODEL statement. Each weight function corresponds to a ρ function. See the section "Weight Functions" on page 5668 for a complete discussion. You can specify the scale parameter σ with the SCALE= option in the PROC statement.

If σ is unknown, both θ and σ are estimated by minimizing the function

$$Q(\theta, \sigma) = \sum_{i=1}^{n} [\rho(\frac{r_i}{\sigma}) + a]\sigma, \ a > 0$$

The algorithm proceeds by alternately improving $\hat{\theta}$ in a location step and $\hat{\sigma}$ in a scale step.

For the scale step, three methods are available to estimate σ, which you can select with the SCALE= option.

1. (SCALE=HUBER<(D=d)>) Compute $\hat{\sigma}$ by the iteration

$$(\hat{\sigma}^{(m+1)})^2 = \frac{1}{nh} \sum_{i=1}^{n} \chi_d(\frac{r_i}{\hat{\sigma}^{(m)}})(\hat{\sigma}^{(m)})^2$$

where

$$\chi_d(x) = \begin{cases} x^2/2 & \text{if } |x| < d \\ d^2/2 & \text{otherwise} \end{cases}$$

is the Huber function and $h = \frac{n-p}{n}(d^2 + (1-d^2)\Phi(d) - 0.5 - d\sqrt{2\pi}e^{-\frac{1}{2}d^2})$ is the Huber constant (refer to Huber 1981, p. 179). You can specify d with the D= option. By default, $d = 2.5$.

2. (SCALE=TUKEY<(D=d)>) Compute $\hat{\sigma}$ by solving the supplementary equation

$$\frac{1}{n-p}\sum_{i=1}^{n}\chi_d(\frac{r_i}{\sigma}) = \beta$$

where

$$\chi_d(x) = \begin{cases} \frac{3x^2}{d^2} - \frac{3x^4}{d^4} + \frac{x^6}{d^6} & \text{if } |x| < d \\ 1 & \text{otherwise} \end{cases}$$

Here $\psi = \frac{1}{6}\chi_1'$ is Tukey's bisquare function, and $\beta = \int \chi_d(s)d\Phi(s)$ is the constant such that the solution $\hat{\sigma}$ is asymptotically consistent when $L(\cdot/\sigma) = \Phi(\cdot)$ (refer to Hampel et al. 1986, p. 149). You can specify d with the D= option. By default, $d = 2.5$.

3. (SCALE=MED) Compute $\hat{\sigma}$ by the iteration

$$\hat{\sigma}^{(m+1)} = \text{median}\{|y_i - x_i^T\hat{\theta}^{(m)}|/\beta_0, i = 1, \ldots, n\}$$

where $\beta_0 = \Phi^{-1}(.75)$ is the constant such that the solution $\hat{\sigma}$ is asymptotically consistent when $L(\cdot/\sigma) = \Phi(\cdot)$ (refer to Hampel et al. 1986, p. 312).

Note that SCALE = MED is the default.

Algorithm

The basic algorithm for computing M estimates for regression is iteratively reweighted least squares (IRLS). As the name suggests, a weighted least squares fit is carried out inside an iteration loop. For each iteration, a set of weights for the observations is used in the least squares fit. The weights are constructed by applying a weight function to the current residuals. Initial weights are based on residuals from an initial fit. The ROBUSTREG procedure uses the unweighted least squares fit as a default initial fit. The iteration terminates when a convergence criterion is satisfied. The maximum number of iterations is set to 1000. You can specify the weight function and the convergence criteria.

Weight Functions

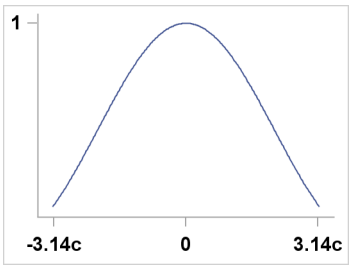
You can specify the weight function for M estimation with the WEIGHTFUNCTION= option. The ROBUSTREG procedure provides 10 weight functions. By default, the procedure uses the bisquare

weight function. In most cases, M estimates are more sensitive to the parameters of these weight functions than to the type of the weight function. The median weight function is not stable and is seldom recommended in data analysis; it is included in the procedure for completeness. You can specify the parameters for these weight functions. Except for the hampel and median weight functions, default values for these parameters are defined such that the corresponding M estimates have 95% asymptotic efficiency in the location model with the Gaussian distribution (see Holland and Welsch 1977).

The following list shows the weight functions available. See Table 74.5 for the default values of the constants in these weight functions.

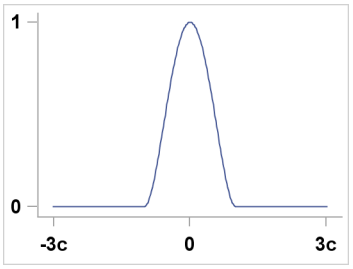
andrews

$$W(x,c) = \begin{cases} \frac{\sin(\frac{x}{c})}{\frac{x}{c}} & \text{if } |x| \leq \pi c \\ 0 & \text{otherwise} \end{cases}$$



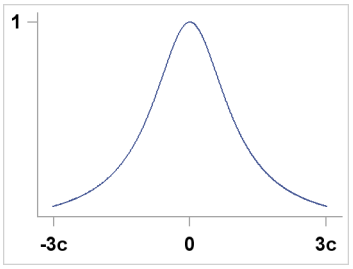
bisquare

$$W(x,c) = \begin{cases} (1-(\frac{x}{c})^2)^2 & \text{if } |x| < c \\ 0 & \text{otherwise} \end{cases}$$



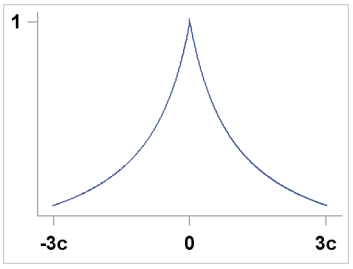
cauchy

$$W(x,c) = \frac{1}{1+(\frac{|x|}{c})^2}$$

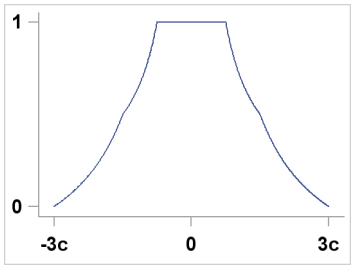


fair

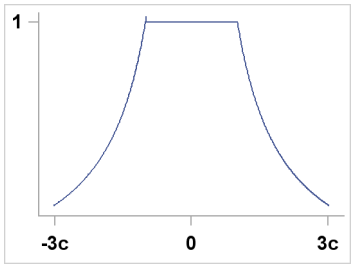
$$W(x,c) = \frac{1}{(1+\frac{|x|}{c})}$$



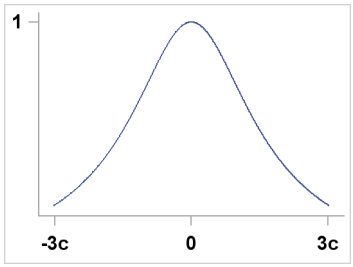
hampel $$W(x,a,b,c) = \begin{cases} 1 & |x| < a \\ \frac{a}{|x|} & a < |x| \le b \\ \frac{a}{|x|}\frac{c-|x|}{c-b} & b < |x| \le c \\ 0 & \text{otherwise} \end{cases}$$



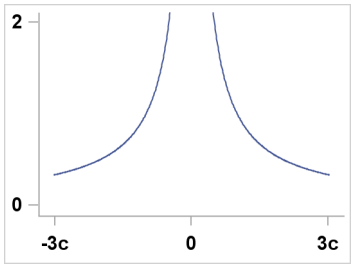
huber $$W(x,c) = \begin{cases} 1 & \text{if } |x| < c \\ \frac{c}{|x|} & \text{otherwise} \end{cases}$$



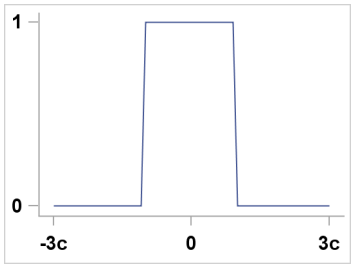
logistic $$W(x,c) = \frac{\tanh(\frac{x}{c})}{\frac{x}{c}}$$



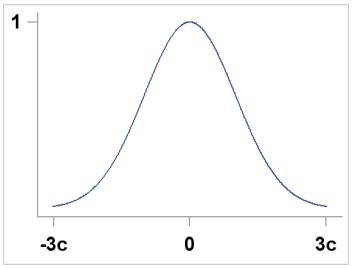
median $$W(x,c) = \begin{cases} \frac{1}{c} & \text{if } x = 0 \\ \frac{1}{|x|} & \text{otherwise} \end{cases}$$



talworth $$W(x,c) = \begin{cases} 1 & \text{if } |x| < c \\ 0 & \text{otherwise} \end{cases}$$



welsch $W(x,c) = \exp(-\frac{1}{2}(\frac{x}{c})^2)$



Convergence Criteria

The following convergence criteria are available in PROC ROBUSTREG:

- relative change in the coefficients (CONVERGENCE= COEF)
- relative change in the scaled residuals (CONVERGENCE= RESID)
- relative change in weights (CONVERGENCE= WEIGHT)

You can specify the criteria with the CONVERGENCE= option in the PROC statement. The default is CONVERGENCE= COEF.

You can specify the precision of the convergence criterion with the EPS= suboption. The default is EPS=1.E−8.

In addition to these convergence criteria, a convergence criterion based on scale-independent measure of the gradient is always checked. See Coleman et al. (1980) for more details. A warning is issued if this criterion is not satisfied.

Asymptotic Covariance and Confidence Intervals

The following three estimators of the asymptotic covariance of the robust estimator are available in PROC ROBUSTREG:

H1: $K^2 \dfrac{[1/(n-p)]\sum(\psi(r_i))^2}{[(1/n)\sum(\psi'(r_i))]^2}(X^TX)^{-1}$

H2: $K \dfrac{[1/(n-p)]\sum(\psi(r_i))^2}{[(1/n)\sum(\psi'(r_i))]}W^{-1}$

H3: $K^{-1}\dfrac{1}{(n-p)}\sum(\psi(r_i))^2 W^{-1}(X^TX)W^{-1}$

where $K = 1 + \frac{p}{n}\frac{\mathrm{Var}(\psi')}{(E\psi')^2}$ is a correction factor and $W_{jk} = \sum \psi'(r_i)x_{ij}x_{ik}$. Refer to Huber (1981, p. 173) for more details.

You can specify the asymptotic covariance estimate with the option ASYMPCOV= option. The ROBUSTREG procedure uses H1 as the default because of its simplicity and stability. Confidence intervals are computed from the diagonal elements of the estimated asymptotic covariance matrix.

R Square and Deviance

The robust version of R^2 is defined as

$$R^2 = \frac{\sum \rho(\frac{y_i-\hat{\mu}}{\hat{s}}) - \sum \rho(\frac{y_i-x_i^T\hat{\theta}}{\hat{s}})}{\sum \rho(\frac{y_i-\hat{\mu}}{\hat{s}})}$$

and the robust deviance is defined as the optimal value of the objective function on the σ^2 scale:

$$D = 2(\hat{s})^2 \sum \rho(\frac{y_i - x_i^T\hat{\theta}}{\hat{s}})$$

where $\rho' = \psi$, $\hat{\theta}$ is the M estimator of θ, $\hat{\mu}$ is the M estimator of location, and $\hat{s}$ is the M estimator of the scale parameter in the full model.

Linear Tests

Two tests are available in PROC ROBUSTREG for the canonical linear hypothesis

$$H_0: \quad \theta_j = 0, \quad j = i_1, \ldots, i_q$$

where q is the total number of parameters of the tested effects. The first test is a robust version of the F test, which is referred to as the ρ test. Denote the M estimators in the full and reduced model as $\hat{\theta}(0) \in \Omega_0$ and $\hat{\theta}(1) \in \Omega_1$, respectively. Let

$$\begin{aligned} Q_0 &= Q(\hat{\theta}(0)) = \min\{Q(\theta)|\theta \in \Omega_0\} \\ Q_1 &= Q(\hat{\theta}(1)) = \min\{Q(\theta)|\theta \in \Omega_1\} \end{aligned}$$

with

$$Q = \sum_{i=1}^{n} \rho(\frac{r_i}{\sigma})$$

The robust F test is based on the test statistic

$$S_n^2 = \frac{2}{q}[Q_1 - Q_0]$$

Asymptotically $S_n^2 \sim \lambda\chi_q^2$ under H_0, where the standardization factor is $\lambda = \int \psi^2(s)d\Phi(s) / \int \psi'(s)\ d\Phi(s)$ and Φ is the cumulative distribution function of the standard normal distribution. Large values of S_n^2 are significant. This test is a special case of the general τ test of Hampel et al. (1986, Section 7.2).

The second test is a robust version of the Wald test, which is referred to as R_n^2 test. The test uses a test statistic

$$R_n^2 = n(\hat{\theta}_{i_1}, \ldots, \hat{\theta}_{i_q}) H_{22}^{-1} (\hat{\theta}_{i_1}, \ldots, \hat{\theta}_{i_q})^T$$

where $\frac{1}{n}H_{22}$ is the $q \times q$ block (corresponding to $\theta_{i_1}, \ldots, \theta_{i_q}$) of the asymptotic covariance matrix of the M estimate $\hat{\theta}_M$ of θ in a p-parameter linear model.

Under H_0, the statistic R_n^2 has an asymptotic χ^2 distribution with q degrees of freedom. Large values of R_n^2 are significant. Refer to Hampel et al. (1986, Chapter 7) for more details.

Model Selection

When M estimation is used, two criteria are available in PROC ROBUSTREG for model selection. The first criterion is a counterpart of the Akaike (1974) AIC criterion for robust regression, and it is defined as

$$AICR = 2\sum_{i=1}^{n}\rho(r_{i:p}) + \alpha p$$

where $r_{i:p} = (y_i - x_i^T\hat{\theta})/\hat{\sigma}$, $\hat{\sigma}$ is a robust estimate of σ and $\hat{\theta}$ is the M estimator with p-dimensional design matrix.

As with AIC, α is the weight of the penalty for dimensions. The ROBUSTREG procedure uses $\alpha = 2E\psi^2/E\psi'$ (Ronchetti 1985) and estimates it by using the final robust residuals.

The second criterion is a robust version of the Schwarz information criteria (BIC), and it is defined as

$$BICR = 2\sum_{i=1}^{n}\rho(r_{i:p}) + p\log(n)$$

High Breakdown Value Estimation

The *breakdown value* of an estimator is defined as the smallest fraction of contamination that can cause the estimator to take on values arbitrarily far from its value on the uncontamined data. The breakdown value of an estimator can be used as a measure of the robustness of the estimator. Rousseeuw and Leroy (1987) and others introduced the following high breakdown value estimators for linear regression.

LTS Estimate

The least trimmed squares (LTS) estimate proposed by Rousseeuw (1984) is defined as the p-vector

$$\hat{\theta}_{LTS} = \arg\min_{\theta} Q_{LTS}(\theta)$$

where

$$Q_{LTS}(\theta) = \sum_{i=1}^{h} r_{(i)}^2$$

$r_{(1)}^2 \le r_{(2)}^2 \le \ldots \le r_{(n)}^2$ are the ordered squared residuals $r_i^2 = (y_i - x_i^T\theta)^2$, $i = 1, \ldots, n$, and h is defined in the range $\frac{n}{2} + 1 \le h \le \frac{3n+p+1}{4}$.

You can specify the parameter h with the H= option in the PROC statement. By default, $h = [\frac{3n+p+1}{4}]$. The breakdown value is $\frac{n-h}{n}$ for the LTS estimate.

The ROBUSTREG procedure computes LTS estimates by using the FAST-LTS algorithm of Rousseeuw and Van Driessen (2000). The estimates are often used to detect outliers in the data, which are then downweighted in the resulting weighted LS regression.

Algorithm

Least trimmed squares (LTS) regression is based on the subset of h observations (out of a total of n observations) whose least squares fit possesses the smallest sum of squared residuals. The coverage h can be set between $\frac{n}{2}$ and n. The LTS method was proposed by Rousseeuw (1984, p. 876) as a highly robust regression estimator with breakdown value $\frac{n-h}{n}$. The ROBUSTREG procedure uses the FAST-LTS algorithm given by Rousseeuw and Van Driessen (2000). The intercept adjustment technique is also used in this implementation. However, because this adjustment is expensive to compute, it is optional. You can use the IADJUST option in the PROC statement to request or suppress the intercept adjustment. By default, PROC ROBUSTREG does intercept adjustment for data sets with fewer than 10000 observations. The steps of the algorithm are described briefly as follows. Refer to Rousseeuw and Van Driessen (2000) for details.

1. The default h is $[\frac{3n+p+1}{4}]$, where p is the number of independent variables. You can specify any integer h with $[\frac{n}{2}] + 1 \leq h \leq [\frac{3n+p+1}{4}]$ with the H= option in the MODEL statement. The breakdown value for LTS, $\frac{n-h}{n}$, is reported. The default h is a good compromise between breakdown value and statistical efficiency.

2. If $p = 1$ (single regressor), the procedure uses the exact algorithm of Rousseeuw and Leroy (1987, p. 172).

3. If $p \geq 2$, the procedure uses the following algorithm. If $n < 2ssubs$, where *ssubs* is the size of the subgroups (you can specify *ssubs* by using the SUBGROUPSIZE= option in the PROC statement; by default, $ssubs = 300$), draw a random p-subset and compute the regression coefficients by using these p points (if the regression is degenerate, draw another p-subset). Compute the absolute residuals for all observations in the data set, and select the first h points with smallest absolute residuals. From this selected h-subset, carry out *nsteps* C-steps (Concentration step; see Rousseeuw and Van Driessen (2000) for details. You can specify *nsteps* with the CSTEP= option in the PROC statement; by default, $nsteps = 2$). Redraw p-subsets and repeat the preceding computing procedure *nrep* times, and then find the *nbsol* (at most) solutions with the lowest sums of h squared residuals. *nrep* can be specified with the NREP= option in the PROC statement. By default, NREP=min$\{500, \binom{n}{p}\}$. For small n and p, all $\binom{n}{p}$ subsets are used and the NREP= option is ignored (Rousseeuw and Hubert 1996). *nbsol* can be specified with the NBEST= option in the PROC statement. By default, NBEST=10. For each of these *nbsol* best solutions, take C-steps until convergence and find the best final solution.

4. If $n \geq 5ssubs$, construct 5 disjoint random subgroups with size *ssubs*. If $2ssubs < n < 5ssubs$, the data are split into at most four subgroups with *ssubs* or more observations in each subgroup, so that each observation belongs to a subgroup and the subgroups have roughly the same size. Let *nsubs* denote the number of subgroups. Inside each subgroup, repeat the procedure in step 3 $[\frac{nrep}{nsubs}]$ times and keep the *nbsol* best solutions. Pool the subgroups, yielding the merged set of size n_{merged}. In the merged set, for each of the $nsubs \times nbsol$ best solutions, carry out *nsteps* C-steps by using n_{merged} and $h_{merged} = [n_{merged}\frac{h}{n}]$ and keep the *nbsol* best solutions. In the full data set, for each of these *nbsol* best solutions, take C-steps by using n and h until convergence and find the best final solution.

R^2

The robust version of R^2 for the LTS estimate is defined as

$$R^2_{LTS} = 1 - \frac{s^2_{LTS}(X, y)}{s^2_{LTS}(\mathbf{1}, y)}$$

for models with the intercept term and as

$$R^2_{LTS} = 1 - \frac{s^2_{LTS}(X, y)}{s^2_{LTS}(\mathbf{0}, y)}$$

for models without the intercept term, where

$$s_{LTS}(X, y) = d_{h,n} \sqrt{\frac{1}{h} \sum_{i=1}^{h} r^2_{(i)}}$$

s_{LTS} is a preliminary estimate of the parameter σ in the distribution function $L(\cdot/\sigma)$.

Here $d_{h,n}$ is chosen to make s_{LTS} consistent, assuming a Gaussian model. Specifically,

$$\begin{aligned} d_{h,n} &= 1/\sqrt{1 - \frac{2n}{h c_{h,n}} \phi(1/c_{h,n})} \\ c_{h,n} &= 1/\Phi^{-1}(\frac{h+n}{2n}) \end{aligned}$$

with Φ and ϕ being the distribution function and the density function of the standard normal distribution, respectively.

Final Weighted Scale Estimator

The ROBUSTREG procedure displays two scale estimators, s_{LTS} and Wscale. The estimator Wscale is a more efficient scale estimator based on the preliminary estimate s_{LTS}, and it is defined as

$$\text{Wscale} = \sqrt{\frac{\sum_i w_i r_i^2}{\sum_i w_i - p}}$$

where

$$w_i = \begin{cases} 0 & \text{if } |r_i|/s_{LTS} > k \\ 1 & \text{otherwise} \end{cases}$$

You can specify k with the CUTOFF= option in the MODEL statement. By default, $k = 3$.

S Estimate

The S estimate proposed by Rousseeuw and Yohai (1984) is defined as the p-vector

$$\hat{\theta}_S = \arg\min_{\theta} S(\theta)$$

where the dispersion $S(\theta)$ is the solution of

$$\frac{1}{n-p}\sum_{i=1}^{n}\chi\left(\frac{y_i - x_i^T\theta}{S}\right) = \beta$$

Here β is set to $\int \chi(s)d\Phi(s)$ such that $\hat{\theta}_S$ and $S(\hat{\theta}_S)$ are asymptotically consistent estimates of θ and σ for the Gaussian regression model. The breakdown value of the S estimate is

$$\frac{\beta}{\max_s \chi(s)}$$

The ROBUSTREG procedure provides two choices for χ: Tukey's bisquare function and Yohai's optimal function.

Tukey's bisquare function, which you can specify with the option CHIF=TUKEY, is

$$\chi_{k_0}(s) = \begin{cases} 3(\frac{s}{k_0})^2 - 3(\frac{s}{k_0})^4 + (\frac{s}{k_0})^6, & \text{if } |s| \le k_0 \\ 1 & \text{otherwise} \end{cases}$$

The constant k_0 controls the breakdown value and efficiency of the S estimate. If you specify the efficiency by using the EFF= option, you can determine the corresponding k_0. The default k_0 is 2.9366 such that the breakdown value of the S estimate is 0.25 with a corresponding asymptotic efficiency for the Gaussian model of 75.9%.

The Yohai function, which you can specify with the option CHIF=YOHAI, is

$$\chi_{k_0}(s) = \begin{cases} \frac{s^2}{2} & \text{if } |s| \le 2k_0 \\ k_0^2[b_0 + b_1(\frac{s}{k_0})^2 + b_2(\frac{s}{k_0})^4 & \\ \quad +b_3(\frac{s}{k_0})^6 + b_4(\frac{s}{k_0})^8] & \text{if } 2k_0 < |s| \le 3k_0 \\ 3.25k_0^2 & \text{if } |s| > 3k_0 \end{cases}$$

where $b_0 = 1.792$, $b_1 = -0.972$, $b_2 = 0.432$, $b_3 = -0.052$, and $b_4 = 0.002$. If you specify the efficiency by using the EFF= option, you can determine the corresponding k_0. By default, k_0 is set to 0.7405 such that the breakdown value of the S estimate is 0.25 with a corresponding asymptotic efficiency for the Gaussian model of 72.7%.

Algorithm

The ROBUSTREG procedure implements the algorithm by Marazzi (1993) for the S estimate, which is a refined version of the algorithm proposed by Ruppert (1992). The refined algorithm is briefly described as follows.

Initialize iter $= 1$.

1. Draw a random q-subset of the total n observations and compute the regression coefficients by using these q observations (if the regression is degenerate, draw another q-subset), where $q \geq p$ can be specified with the SUBSIZE= option. By default, $q = p$.

2. Compute the residuals: $r_i = y_i - \sum_{j=1}^{p} x_{ij}\theta_j$ for $i = 1, \ldots, n$. If iter $= 1$, set $s^* = 2\text{median}\{|r_i|, i = 1, \ldots, n\}$; if $s^* = 0$, set $s^* = \min\{|r_i|, i = 1, \ldots, n\}$;
while $\sum_{i=1}^{n} \chi(r_i/s^*) > (n-p)\beta$, set $s^* = 1.5s^*$; go to step 3.
If iter > 1 and $\sum_{i=1}^{n} \chi(r_i/s^*) <= (n-p)\beta$, go to step 3; otherwise, go to step 5.

3. Solve for s the equation

$$\frac{1}{n-p}\sum_{i=1}^{n}\chi(r_i/s) = \beta$$

using an iterative algorithm.

4. If iter > 1 and $s > s^*$, go to step 5. Otherwise, set $s^* = s$ and $\theta^* = \theta$. If $s^* < TOLS$, return s^* and θ^*; otherwise, go to step 5.

5. If iter $< NREP$, set iter $=$ iter $+ 1$ and return to step 1; otherwise, return s^* and θ^*.

The ROBUSTREG procedure does the following refinement step by default. You can request that this refinement not be done by using the NOREFINE option in the PROC statement.

6. Let $\psi = \chi'$. Using the values s^* and θ^* from the previous steps, compute M estimates θ_M and σ_M of θ and σ with the setup for M estimation in the section "M Estimation" on page 5666. If $\sigma_M > s^*$, give a warning and return s^* and θ^*; otherwise, return σ_M and θ_M.

You can specify $TOLS$ with the TOLERANCE= option; by default, TOLERANCE=0.001. Alternately, you can specify $NREP$ with the NREP= option. You can also use the options NREP=NREP0 or NREP=NREP1 to determine $NREP$ according to the following table. NREP=NREP0 is set as the default.

Table 74.9 Default NREP

P	NREP0	NREP1
1	150	500
2	300	1000
3	400	1500
4	500	2000
5	600	2500
6	700	3000
7	850	3000
8	1250	3000
9	1500	3000
>9	1500	3000

R^2 and Deviance

The robust version of R^2 for the S estimate is defined as

$$R_S^2 = 1 - \frac{(n-p)S_p^2}{(n-1)S_\mu^2}$$

for the model with the intercept term and

$$R_S^2 = 1 - \frac{(n-p)S_p^2}{nS_0^2}$$

for the model without the intercept term, where S_p is the S estimate of the scale in the full model, S_μ is the S estimate of the scale in the regression model with only the intercept term, and S_0 is the S estimate of the scale without any regressor. The deviance D is defined as the optimal value of the objective function on the σ^2 scale:

$$D = S_p^2$$

Asymptotic Covariance and Confidence Intervals

Since the S estimate satisfies the first-order necessary conditions as the M estimate, it has the same asymptotic covariance as that of the M estimate. All three estimators of the asymptotic covariance for the M estimate in the section "Asymptotic Covariance and Confidence Intervals" on page 5671 can be used for the S estimate. Besides, the weighted covariance estimator H4 described in the section "Asymptotic Covariance and Confidence Intervals" on page 5681 is also available and is set as the default. Confidence intervals for estimated parameters are computed from the diagonal elements of the estimated asymptotic covariance matrix.

MM Estimation

MM estimation is a combination of high breakdown value estimation and efficient estimation, which was introduced by Yohai (1987). It has the following three steps.

1. Compute an initial (consistent) high breakdown value estimate $\hat{\theta}'$. The ROBUSTREG procedure provides two kinds of estimates as the initial estimate: the LTS estimate and the S estimate. By default, the LTS estimate is used because of its speed and high breakdown value. The breakdown value of the final MM estimate is decided by the breakdown value of the initial LTS estimate and the constant k_0 in the χ function. To use the S estimate as the initial estimate, you specify the INITEST=S option in the PROC statement. In this case, the breakdown value of the final MM estimate is decided only by the constant k_0. Instead of computing the LTS estimate or the S estimate as the initial estimate, you can also specify the initial estimate explicitly by using the INEST= option in the PROC statement. See the section "INEST= Data Set" on page 5683 for details.

2. Find $\hat{\sigma}'$ such that

$$\frac{1}{n-p}\sum_{i=1}^{n}\chi(\frac{y_i - x_i^T\hat{\theta}'}{\hat{\sigma}'}) = \beta$$

where $\beta = \int \chi(s)d\Phi(s)$.

The ROBUSTREG procedure provides two choices for χ: Tukey's bisquare function and Yohai's optimal function.

Tukey's bisquare function, which you can specify with the option CHIF=TUKEY, is

$$\chi_{k_0}(s) = \begin{cases} 3(\frac{s}{k_0})^2 - 3(\frac{s}{k_0})^4 + (\frac{s}{k_0})^6, & \text{if } |s| \le k_0 \\ 1 & \text{otherwise} \end{cases}$$

where k_0 can be specified with the K0= option. The default k_0 is 2.9366 such that the asymptotically consistent scale estimate $\hat{\sigma}'$ has the breakdown value of 25%.

Yohai's optimal function, which you can specify with the option CHIF=YOHAI, is

$$\chi_{k_0}(s) = \begin{cases} \frac{s^2}{2} & \text{if } |s| \le 2k_0 \\ k_0^2[b_0 + b_1(\frac{s}{k_0})^2 + b_2(\frac{s}{k_0})^4 & \\ \quad +b_3(\frac{s}{k_0})^6 + b_4(\frac{s}{k_0})^8] & \text{if } 2k_0 < |s| \le 3k_0 \\ 3.25k_0^2 & \text{if } |s| > 3k_0 \end{cases}$$

where $b_0 = 1.792$, $b_1 = -0.972$, $b_2 = 0.432$, $b_3 = -0.052$, and $b_4 = 0.002$. You can specify k_0 with the K0= option. The default k_0 is 0.7405 such that the asymptotically consistent scale estimate $\hat{\sigma}'$ has the breakdown value of 25%.

3. Find a local minimum $\hat{\theta}_{MM}$ of

$$Q_{MM} = \sum_{i=1}^{n}\rho(\frac{y_i - x_i^T\theta}{\hat{\sigma}'})$$

such that $Q_{MM}(\hat{\theta}_{MM}) \le Q_{MM}(\hat{\theta}')$. The algorithm for M estimation is used here.

The ROBUSTREG procedure provides two choices for ρ: Tukey's bisquare function and Yohai's optimal function.

Tukey's bisquare function, which you can specify with the option CHIF=TUKEY, is

$$\rho(s) = \chi_{k_1}(s) = \begin{cases} 3(\frac{s}{k_1})^2 - 3(\frac{s}{k_1})^4 + (\frac{s}{k_1})^6, & \text{if } |s| \le k_1 \\ 1 & \text{otherwise} \end{cases}$$

where k_1 can be specified with the K1= option. The default k_1 is 3.440 such that the MM estimate has 85% asymptotic efficiency with the Gaussian distribution.

Yohai's optimal function, which you can specify with the option CHIF=YOHAI, is

$$\rho(s) = \chi_{k_1}(s) = \begin{cases} \frac{s^2}{2} & \text{if } |s| \le 2k_1 \\ k_1^2[b_0 + b_1(\frac{s}{k_1})^2 + b_2(\frac{s}{k_1})^4 & \\ \quad +b_3(\frac{s}{k_1})^6 + b_4(\frac{s}{k_1})^8] & \text{if } 2k_1 < |s| \le 3k_1 \\ 3.25k_1^2 & \text{if } |s| > 3k_1 \end{cases}$$

where k_1 can be specified with the K1= option. The default k_1 is 0.868 such that the MM estimate has 85% asymptotic efficiency with the Gaussian distribution.

Algorithm

The initial LTS estimate is computed using the algorithm described in the section "LTS Estimate" on page 5673. You can control the quantile of the LTS estimate with the option INITH=h, where h is an integer between $[\frac{n}{2}] + 1$ and $[\frac{3n+p+1}{4}]$. By default, $h = [\frac{3n+p+1}{4}]$, which corresponds to a breakdown value of around 25%.

The initial S estimate is computed using the algorithm described in the section "S Estimate" on page 5676. You can control the breakdown value and efficiency of this initial S estimate by the constant k_0, which can be specified with the K0 option.

The scale parameter σ is solved by an iterative algorithm

$$(\sigma^{(m+1)})^2 = \frac{1}{(n-p)\beta} \sum_{i=1}^{n} \chi_{k_0}(\frac{r_i}{\sigma^{(m)}})(\sigma^{(m)})^2$$

where $\beta = \int \chi_{k_0}(s) d\Phi(s)$.

Once the scale parameter is computed, the iteratively reweighted least squares (IRLS) algorithm with fixed scale parameter is used to compute the final MM estimate.

Convergence Criteria

In the iterative algorithm for the scale parameter, the relative change of the scale parameter controls the convergence.

In the iteratively reweighted least squares algorithm, the same convergence criteria for the M estimate used before are used here.

Bias Test

Although the final MM estimate inherits the high breakdown value property, its bias due to the distortion of the outliers can be high. Yohai, Stahel, and Zamar (1991) introduced a bias test. The ROBUSTREG procedure implements this test when you specify the BIASTEST option in the PROC statement. This test is based on the initial scale estimate $\hat{\sigma}'$ and the final scale estimate $\hat{\sigma}'_1$, which is the solution of

$$\frac{1}{n-p} \sum_{i=1}^{n} \chi(\frac{y_i - x_i^T \hat{\theta}_{MM}}{\hat{\sigma}'_1}) = \beta$$

Let $\psi_{k_0}(z) = \frac{\partial \chi_{k_0}(z)}{\partial z}$ and $\psi_{k_1}(z) = \frac{\partial \chi_{k_1}(z)}{\partial z}$. Compute

$$\begin{aligned} \tilde{r}_i &= (y_i - x_i^T \hat{\theta}')/\hat{\sigma}' \quad \text{for } i = 1, \ldots, n \\ v_0 &= \frac{(1/n) \sum \psi'_{k_0}(\tilde{r}_i)}{(\hat{\sigma}'_1/n) \sum \psi_{k_0}(\tilde{r}_i)\tilde{r}_i} \end{aligned}$$

$$
\begin{aligned}
p_i^{(0)} &= \frac{\psi_{k_0}(\tilde{r}_i)}{(1/n)\sum \psi'_{k_0}(\tilde{r}_i)} \quad \text{for } i = 1, \ldots, n \\
p_i^{(1)} &= \frac{\psi_{k_1}(\tilde{r}_i)}{(1/n)\sum \psi'_{k_1}(\tilde{r}_i)} \quad \text{for } i = 1, \ldots, n \\
d^2 &= \frac{1}{n}\sum (p_i^{(1)} - p_i^{(0)})^2
\end{aligned}
$$

Let

$$T = \frac{2n(\hat{\sigma}'_1 - \hat{\sigma}')}{v_0 d^2 (\hat{\sigma}')^2}$$

Standard asymptotic theory shows that T approximately follows a χ^2 distribution with p degrees of freedom. If T exceeds the α quantile χ^2_α of the χ^2 distribution with p degrees of freedom, then the ROBUSTREG procedure gives a warning and recommends that you use other methods. Otherwise the final MM estimate and the initial scale estimate are reported. You can specify α with the ALPHA= option following the BIASTEST option. By default, ALPHA=0.99.

Asymptotic Covariance and Confidence Intervals

Since the MM estimate is computed as a M estimate with a fixed scale in the last step, the asymptotic covariance for the M estimate can be used here for the asymptotic covariance of the MM estimate. Besides the three estimators H1, H2, and H3 as described in the section "Asymptotic Covariance and Confidence Intervals" on page 5671, a weighted covariance estimator H4 is available:

$$\text{H4: } K^2 \frac{[1/(n-p)]\sum(\psi(r_i))^2}{[(1/n)\sum(\psi'(r_i))]^2} W^{-1}$$

where $K = 1 + \frac{p}{n}\frac{\text{Var}(\psi')}{(E\psi')^2}$ is the correction factor and $W_{jk} = \frac{1}{\bar{w}}\sum w_i x_{ij} x_{ik}$, $\bar{w} = \frac{1}{n}\sum w_i$.

You can specify these estimators with the option ASYMPCOV= [H1 | H2 | H3 | H4]. The ROBUSTREG procedure uses H4 as the default. Confidence intervals for estimated parameters are computed from the diagonal elements of the estimated asymptotic covariance matrix.

R Square and Deviance

The robust version of R^2 for the MM estimate is defined as

$$R^2 = \frac{\sum \rho(\frac{y_i - \hat{\mu}}{\hat{s}}) - \sum \rho(\frac{y_i - x_i^T\hat{\theta}}{\hat{s}})}{\sum \rho(\frac{y_i - \hat{\mu}}{\hat{s}})}$$

and the robust deviance is defined as the optimal value of the objective function on the σ^2 scale:

$$D = 2(\hat{s})^2 \sum \rho(\frac{y_i - x_i^T\hat{\theta}}{\hat{s}})$$

where $\rho' = \psi$, $\hat{\theta}$ is the MM estimator of θ, $\hat{\mu}$ is the MM estimator of location, and $\hat{s}$ is the MM estimator of the scale parameter in the full model.

Linear Tests

For MM estimation, the same ρ test and R_n^2 test used for M estimation can be used. See the section "Linear Tests" on page 5672 for details.

Model Selection

For MM estimation, the same two model selection methods used for M estimation can be used. See the section "Model Selection" on page 5673 for details.

Robust Distance

The ROBUSTREG procedure uses the robust multivariate location and scale estimates for leverage-point detection. The procedure computes a robust version of the Mahalanobis distance by using the minimum covariance determinant (MCD) method of Rousseeuw (1984).

Algorithm

PROC ROBUSTREG implements the algorithm given by Rousseeuw and Van Driessen (1999) for MCD, which is similar to the algorithm for LTS.

Robust Distance

The Mahalanobis distance is defined as

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

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{C}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x})$ are the empirical multivariate location and scale. Here $x_i = (x_{i1}, \ldots, x_{i(p-1)})^T$ does not include the intercept. The relation between the Mahalanobis distance $MD(x_i)$ and the hat matrix $H = (h_{ij}) = X(X^TX)^{-1}X^T$ is

$$h_{ii} = \frac{1}{n-1}MD_i^2 + \frac{1}{n}$$

The robust distance is defined as

$$RD(x_i) = [(x_i - T(X))^T C(X)^{-1}(x_i - T(X))]^{1/2}$$

where $T(X)$ and $C(X)$ are the robust multivariate location and scale obtained by MCD.

These distances are used to detect leverage points.

Leverage Point and Outlier Detection

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

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

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

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

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

You can specify the multiplier k of the cutoff value by using the CUTOFF= option in the MODEL statement. By default, $k = 3$. $\hat{\sigma}$ is estimated scale in the model.

An ODS table called DIAGNOSTICS contains these two variables.

INEST= Data Set

When you use the M or MM estimation, you can use the INEST= data set to specify initial estimates for all the parameters in the model. The INEST= option is ignored if you specify LTS or S estimation by using the METHOD=LTS or METHOD=S option or if you specify the INITEST= option after the METHOD=MM option in the PROC statement. The INEST= data set must contain the intercept variable (named Intercept) and all independent variables in the MODEL statement.

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

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

You can specify starting values for the iteratively reweighted least squares algorithm in the INEST= data set. The INEST= data set has the same structure as the OUTEST= data set but is not required to have all the variables or observations that appear in the OUTEST= data set. One simple use of the INEST= option is passing the previous OUTEST= data set directly to the next model as an INEST= data set, assuming that the two models have the same parameterization.

OUTEST= Data Set

The OUTEST= data set contains parameter estimates for the model. You can specify a label in the MODEL statement to distinguish between the estimates for different models used by the ROBUSTREG procedure. If the COVOUT option is specified, the OUTEST= data set also contains the estimated covariance matrix of the parameter estimates. Note that, if the ROBUSTREG procedure does not converge, the parameter estimates are set to missing in the OUTEST data set.

The OUTEST= data set contains all variables specified in the MODEL statement and the BY statement. One observation consists of parameter values for the model with the dependent variable having the value -1. If the COVOUT option is specified, there are additional observations containing the rows of the estimated covariance matrix. For these observations, the dependent variable contains the parameter estimate for the corresponding row variable. The following variables are also added to the data set:

MODEL	is a character variable containing the label of the MODEL statement, if present. Otherwise, the variable's value is blank.
NAME	is a character variable containing the name of the dependent variable for the parameter estimates or the name of the row for the covariance matrix estimates.
TYPE	is a character variable containing the type of the observation, either PARMS for parameter estimates or COV for covariance estimates.
METHOD	is a character variable containing the type of estimation method, either M estimation, LTS estimation, S estimation, or MM estimation.
STATUS	is a character variable containing the status of model fitting, either Converged, Warning, or Failed.
INTERCEPT	is a numeric variable containing the intercept parameter estimates and covariances.
SCALE	is a numeric variable containing the scale parameter estimates.

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

Computational Resources

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

For M estimation, the minimum working space (in bytes) needed is

$$3n + 2p^2 + 30p$$

If sufficient space is available, the input data set is also kept in memory; otherwise, the input data set is reread for computing the iteratively reweighted least squares estimates and the execution

time of the procedure increases substantially. For each reweighted least squares, $O(np^2 + p^3)$ multiplications and additions are required for computing the crossproduct matrix and its inverse. The $O(v)$ notation means that, for large values of the argument, v, $O(v)$ is approximately a constant times v.

Since the iteratively reweighted least squares algorithm converges very quickly (normally within fewer than 20 iterations), the computation of M estimates is fast.

LTS estimation is more expensive in computation. The minimum working space (in bytes) needed is

$$np + 12n + 4p^2 + 60p$$

The memory is mainly used to store the current data used by LTS for modeling. The LTS algorithm uses subsampling and spends much of its computing time on resampling and computing estimates for subsamples. Since it resamples if singularity is detected, it might take more time if the data set has serious singularities.

The MCD algorithm for leverage-point diagnostics is similar to the LTS algorithm.

ODS Table Names

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

Table 74.10 ODS Tables Produced by PROC ROBUSTREG

ODS Table Name	Description	Statement	Option
BestEstimates	Best final estimates for LTS	PROC	SUBANALYSIS
BestSubEstimates	Best estimates for each subgroup	PROC	SUBANALYSIS
BiasTest	Bias test for MM estimation	PROC	BIASTEST
ClassLevels	Classification variable levels	CLASS	default
CorrB	Parameter estimate correlation matrix	MODEL	CORRB
CovB	Parameter estimate covariance matrix	MODEL	COVB
CStep	C-step for LTS fitting	PROC	SUBANALYSIS
Diagnostics	Outlier diagnostics	MODEL	DIAGNOSTICS
DiagSummary	Summary of the outlier diagnostics	MODEL	default
GoodFit	R2, deviance, AIC, and BIC	MODEL	METHOD
InitLTSProfile	Profile for initial LTS estimate	PROC	METHOD
InitSProfile	Profile for initial S estimate	PROC	METHOD
IterHistory	Iteration history	PROC	ITPRINT
LTSEstimates	LTS parameter estimates	PROC	METHOD
LTSLocationScale	Location and scale for LTS	PROC	METHOD
LTSProfile	Profile for LTS estimator	PROC	METHOD
LTSRsquare	R2 for LTS estimate	PROC	METHOD
MMProfile	Profile for MM estimator	PROC	METHOD
ModelInfo	Model information	MODEL	default

Table 74.10 (continued)

ODS Table Name	Description	Statement	Option
NObs	Observations summary	PROC	default
ParameterEstimates	Parameter estimates	MODEL	default
ParameterEstimatesF	Final weighted LS estimates	PROC	FWLS
ParameterEstimatesR	Reduced parameter estimates	TEST	default
ParmInfo	Parameter indices	MODEL	default
SProfile	Profile for S estimator	PROC	METHOD
Groups	Groups for LTS fitting	PROC	SUBANALYSIS
SummaryStatistics	Summary statistics for model variables	MODEL	default
Tests	Results for tests	TEST	default

ODS Graphics

Graphical displays are important in robust regression and outlier detection. This section provides information about the basic ODS statistical graphics produced by the ROBUSTREG procedure.

If the model includes a single continuous independent variable, a plot of robust fit against this variable (FITPLOT) is provided by default. For diagnostics, two plots are particularly useful in revealing outliers and leverage points. The first is a scatter plot of the standardized robust residuals against the robust distances (RDPLOT). The second is a scatter plot of the robust distances against the classical Mahalanobis distances (DDPLOT). In addition to these two plots, a histogram and a quantile-quantile plot of the standardized robust residuals are also helpful.

These plots are controlled by the PLOTS= option in the PROC statement. You can specify more than one plot request with the PLOTS= option. Table 74.11 summarizes these requests.

In addition to the PLOTS= option, you must specify the ODS GRAPHICS statement. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

The names of the graphs that PROC ROBUSTREG generates are listed in Table 74.12, along with the required statements and options. The following subsections provide information about these graphs.

Table 74.11 Options for Plots

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

Fit Plot

When the model has a single independent continuous variable (with or without the intercept), the ROBUSTREG procedure automatically creates a plot of robust fit against this independent variable.

The following simple example is used to show the fit plot. The data, from Rousseeuw and Leroy (1987, Table 3), include the logarithm of surface temperature and the logarithm of light intensity for 47 stars in the direction of the constellation Cygnus.

```
data star;
   input index x y @@;
   label x = 'Log Temperature'
         y = 'Log Light Intensity';
datalines;
1   4.37  5.23     25  4.38  5.02
2   4.56  5.74     26  4.42  4.66
3   4.26  4.93     27  4.29  4.66
4   4.56  5.74     28  4.38  4.90
5   4.30  5.19     29  4.22  4.39
6   4.46  5.46     30  3.48  6.05
7   3.84  4.65     31  4.38  4.42
8   4.57  5.27     32  4.56  5.10
9   4.26  5.57     33  4.45  5.22
10  4.37  5.12     34  3.49  6.29
11  3.49  5.73     35  4.23  4.34
12  4.43  5.45     36  4.62  5.62
13  4.48  5.42     37  4.53  5.10
14  4.01  4.05     38  4.45  5.22
15  4.29  4.26     39  4.53  5.18
16  4.42  4.58     40  4.43  5.57
17  4.23  3.94     41  4.38  4.62
18  4.42  4.18     42  4.45  5.06
19  4.23  4.18     43  4.50  5.34
20  3.49  5.89     44  4.45  5.34
21  4.29  4.38     45  4.55  5.54
22  4.29  4.22     46  4.45  4.98
23  4.42  4.42     47  4.42  4.50
24  4.49  4.85
;
```

The fit plot in Figure 74.17, created with the following statements, plots the robust fit of the logarithm of light intensity with the MM method against the logarithm of the surface temperature.

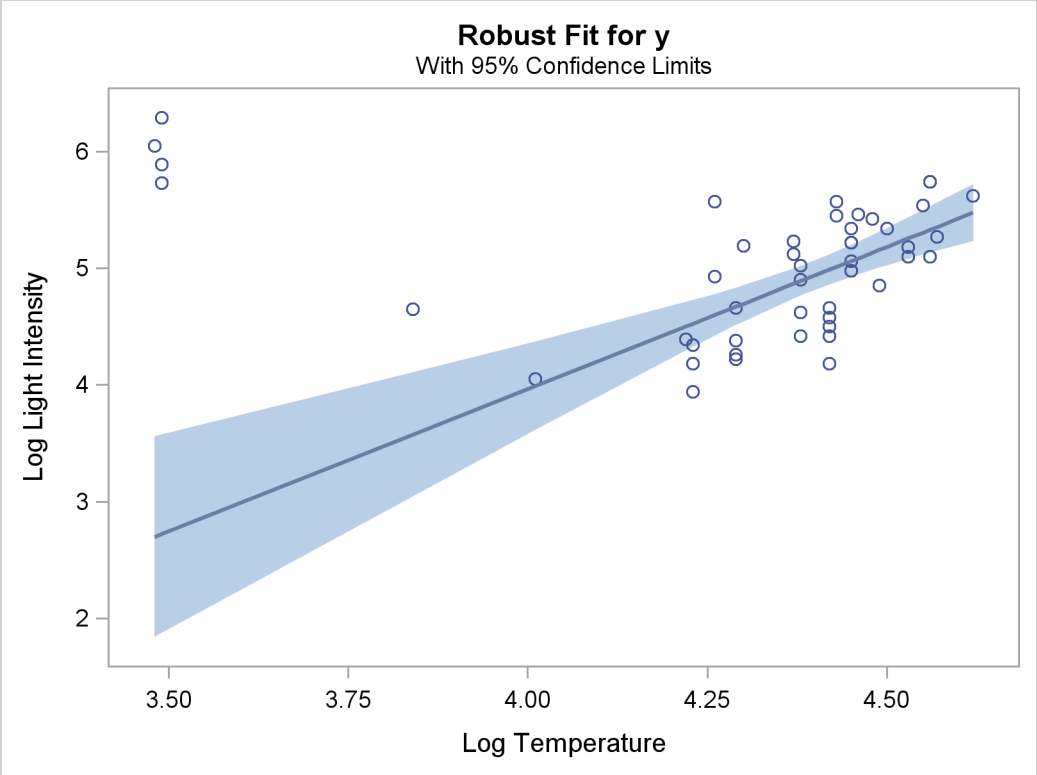
```
ods graphics on;

proc robustreg data=star method=mm ;
    model y = x;
run;

ods graphics off;
```

Confidence limits are added on the plot by default. The NOLIMITS option suppresses these limits.

Figure 74.17 Robust Fit




```
ods graphics on;

proc robustreg data=star method=mm plot=fitplot(nolimits);
    model y = x;
run;

ods graphics off;
```

Distance-Distance Plot

The distance-distance plot (DDPLOT) is mainly used for leverage-point diagnostics. It is a scatter plot of the robust distances against the classical Mahalanobis distances for the continuous independent variables. See the section "Leverage Point and Outlier Detection" on page 5683 for details about the robust distance. If there is a classification variable specified in the model, this plot is not created.

You can use the PLOT=DDPLOT option to request this plot. The following statements use the stack data set in the section "M Estimation" on page 5643 to create the single plot shown in Figure 74.5.

```
ods graphics on;

proc robustreg data=stack plot=ddplot;
   model y = x1 x2 x3;
run;

ods graphics off;
```

The reference lines represent the cutoff values. The diagonal line is also drawn to show the distribution of the distances. By default, all outliers and leverage points are labeled with observation numbers. To change the default, you can use the LABEL= option as described in Table 74.2.

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

Residual-Distance Plot

The residual-distance plot (RDPLOT) is used for both outlier and leverage-point diagnostics. It is a scatter plot of the standardized robust residuals against the robust distances. See the section "Leverage Point and Outlier Detection" on page 5683 for details about the robust distance. If there is a classification variable specified in the model, this plot is not created.

You can use the PLOT=RDPLOT option to request this plot. The following statements use the stack data set in the section "M Estimation" on page 5643 to create a single plot shown in Figure 74.4.

```
ods graphics on;

proc robustreg data=stack plot=rdplot;
   model y = x1 x2 x3;
run;

ods graphics off;
```

The reference lines represent the cutoff values. By default, all outliers and leverage points are labeled with observation numbers. To change the default, you can use the LABEL= option as described in Table 74.2.

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

Histogram and Q-Q Plot

PROC ROBUSTREG produces a histogram and a Q-Q plot for the standardized robust residuals. The histogram is superimposed with a normal density curve and a kernel density curve. Using the stack data set in the section "M Estimation" on page 5643, the following statements create the plots in Figure 74.6 and Figure 74.7.

```
ods graphics on;

proc robustreg data=stack plots=(histogram qqplot);
   model y = x1 x2 x3;
run;

ods graphics off;
```

ODS Graph Names

PROC ROBUSTREG assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. The names are listed in Table 74.12.

To request these graphs you must specify the ODS GRAPHICS statement in addition to the PLOTS= option described in Table 74.11. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

Table 74.12 ODS Graphics Produced by PROC ROBUSTREG

ODS Graph Name	Plot Description	Statement	PLOTS= Option
DDPlot	Robust distance vs. Mahalanobis distance	PROC	DDPLOT
FitPlot	Robust fit vs. independent variable	PROC	FITPLOT
Histogram	Histogram of standardized robust residuals	PROC	HISTOGRAM
QQPlot	Q-Q plot of standardized robust residuals	PROC	QQPLOT
RDPlot	Standardized robust residual vs. robust distance	PROC	RDPLOT

Examples: ROBUSTREG Procedure

Example 74.1: Comparison of Robust Estimates

This example contrasts several of the robust methods available in the ROBUSTREG procedure.

The following statements generate 1000 random observations. The first 900 observations are from a linear model, and the last 100 observations are significantly biased in the y-direction. In other words, 10% of the observations are contaminated with outliers.

```
data a (drop=i);
   do i=1 to 1000;
      x1=rannor(1234);
      x2=rannor(1234);
      e=rannor(1234);
      if i > 900 then y=100 + e;
      else y=10 + 5*x1 + 3*x2 + .5 * e;
      output;
   end;
run;
```

The following statements invoke PROC REG and PROC ROBUSTREG with the data set a.

```
proc reg data=a;
   model y = x1 x2;
run;
proc robustreg data=a method=m ;
   model y = x1 x2;
run;
proc robustreg data=a method=mm;
   model y = x1 x2;
run;
proc robustreg data=a method=s ;
   model y = x1 x2;
run;
proc robustreg data=a method=lts ;
   model y = x1 x2;
run;
```

The tables of parameter estimates generated with the ROBUSTREG procedure by using M estimation, MM estimation, S estimation, and LTS estimation are shown in Output 74.1.2, Output 74.1.3, Output 74.1.4, and Output 74.1.5, respectively. For comparison, the ordinary least squares (OLS) estimates produced by the REG procedure (Chapter 73, "The REG Procedure") are shown in Output 74.1.1. The four robust methods, M, MM, S, and LTS, correctly estimate the regression coefficients for the underlying model (10, 5, and 3), but the OLS estimate does not.

Output 74.1.1 OLS Estimates for Data with 10% Contamination

The REG Procedure
Model: MODEL1
Dependent Variable: y

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19.06712	0.86322	22.09	<.0001
x1	1	3.55485	0.86892	4.09	<.0001
x2	1	2.12341	0.83039	2.56	0.0107

Output 74.1.2 M Estimates for Data with 10% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.A
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	M Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	10.0024	0.0174	9.9683	10.0364	331908	<.0001
x1	1	5.0077	0.0175	4.9735	5.0420	82106.9	<.0001
x2	1	3.0161	0.0167	2.9834	3.0488	32612.5	<.0001
Scale	1	0.5780					

Output 74.1.3 MM Estimates for Data with 10% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.A
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	MM Estimation

Output 74.1.3 *continued*

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	10.0035	0.0176	9.9690	10.0379	323947	<.0001
x1	1	5.0085	0.0178	4.9737	5.0433	79600.6	<.0001
x2	1	3.0181	0.0168	2.9851	3.0511	32165.0	<.0001
Scale	0	0.6733					

Output 74.1.4 S Estimates for Data with 10% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.A
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	S Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	10.0055	0.0180	9.9703	10.0408	309917	<.0001
x1	1	5.0096	0.0182	4.9740	5.0452	76045.2	<.0001
x2	1	3.0210	0.0172	2.9873	3.0547	30841.3	<.0001
Scale	0	0.6721					

Output 74.1.5 LTS Estimates for Data with 10% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.A
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	LTS Estimation

Output 74.1.5 *continued*

LTS Parameter Estimates

Parameter	DF	Estimate
Intercept	1	10.0083
x1	1	5.0316
x2	1	3.0396
Scale (sLTS)	0	0.5880
Scale (Wscale)	0	0.5113

The next statements demonstrate that if the percentage of contamination is increased to 40%, the M method and the MM method with default options fail to estimate the underlying model. Output 74.1.6 and Output 74.1.7 display these estimates. However, by tuning the constant c for the M method and the constants INITH and K0 for the MM method, you can increase the breakdown values of the estimates and capture the right model. Output 74.1.8 and Output 74.1.9 display these estimates. Similarly, you can tune the constant EFF for the S method and the constant H for the LTS method and correctly estimate the underlying model with these methods. Results are not presented.

```
data b (drop=i);
   do i=1 to 1000;
      x1=rannor(1234);
      x2=rannor(1234);
      e=rannor(1234);
      if i > 600 then y=100 + e;
      else y=10 + 5*x1 +  3*x2 + .5 * e;
      output;
   end;
run;

proc robustreg data=b method=m ;
   model y = x1 x2;
run;

proc robustreg data=b method=mm;
   model y = x1 x2;
run;

proc robustreg data=b method=m(wf=bisquare(c=2));
   model y = x1 x2;
run;

proc robustreg data=b method=mm(inith=502 k0=1.8);
   model y = x1 x2;
run;
```

Output 74.1.6 M Estimates (Default Setting) for Data with 40% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.B
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	M Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	44.8991	1.5609	41.8399	47.9584	827.46	<.0001
x1	1	2.4309	1.5712	-0.6485	5.5104	2.39	0.1218
x2	1	1.3742	1.5015	-1.5687	4.3171	0.84	0.3601
Scale	1	56.6342					

Output 74.1.7 MM Estimates (Default Setting) for Data with 40% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.B
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	MM Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	43.0607	1.7978	39.5370	46.5844	573.67	<.0001
x1	1	2.7369	1.8140	-0.8185	6.2924	2.28	0.1314
x2	1	1.5211	1.7265	-1.8628	4.9049	0.78	0.3783
Scale	0	52.8496					

Output 74.1.8 M Estimates (Tuned) for Data with 40% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.B
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	M Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	10.0137	0.0219	9.9708	10.0565	209688	<.0001
x1	1	4.9905	0.0220	4.9473	5.0336	51399.1	<.0001
x2	1	3.0399	0.0210	2.9987	3.0811	20882.4	<.0001
Scale	1	1.0531					

Output 74.1.9 MM Estimates (Tuned) for Data with 40% Contamination

The ROBUSTREG Procedure

Model Information

Data Set	WORK.B
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	MM Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	10.0103	0.0213	9.9686	10.0520	221639	<.0001
x1	1	4.9890	0.0218	4.9463	5.0316	52535.9	<.0001
x2	1	3.0363	0.0201	2.9970	3.0756	22895.5	<.0001
Scale	0	1.8992					

When there are bad leverage points, the M method fails to estimate the underlying model no matter what constant c you use. In this case, other methods (LTS, S, and MM) in PROC ROBUSTREG, which are robust to bad leverage points, will correctly estimate the underlying model.

The following statements generate 1000 observations with 1% bad high leverage points.

```
data c (drop=i);
   do i=1 to 1000;
      x1=rannor(1234);
      x2=rannor(1234);
      e=rannor(1234);
      if i > 600 then y=100 + e;
      else y=10 + 5*x1 + 3*x2 + .5 * e;
      if i < 11 then x1=200 * rannor(1234);
      if i < 11 then x2=200 * rannor(1234);
      if i < 11 then y= 100*e;
      output;
   end;
run;
proc robustreg data=c method=mm(inith=502 k0=1.8);
   model y = x1 x2;
run;
proc robustreg data=c method=s(k0=1.8);
   model y = x1 x2;
run;
proc robustreg data=c method=lts(h=502);
   model y = x1 x2;
run;
```

Output 74.1.10 displays the MM estimates with initial LTS estimates, Output 74.1.11 displays the S estimates, and Output 74.1.12 displays the LTS estimates.

Output 74.1.10 MM Estimates for Data with 1% Leverage Points

The ROBUSTREG Procedure

Model Information

Data Set	WORK.C
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	MM Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	9.9820	0.0215	9.9398	10.0241	215369	<.0001
x1	1	5.0303	0.0206	4.9898	5.0707	59469.1	<.0001
x2	1	3.0222	0.0221	2.9789	3.0655	18744.9	<.0001
Scale	0	2.2134					

Output 74.1.11 S Estimates for Data with 1% Leverage Points

The ROBUSTREG Procedure

Model Information

Data Set	WORK.C
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	S Estimation

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	9.9808	0.0216	9.9383	10.0232	212532	<.0001
x1	1	5.0303	0.0208	4.9896	5.0710	58656.3	<.0001
x2	1	3.0217	0.0222	2.9782	3.0652	18555.7	<.0001
Scale	0	2.2094					

Output 74.1.12 LTS Estimates for Data with 1% Leverage Points

The ROBUSTREG Procedure

Model Information

Data Set	WORK.C
Dependent Variable	y
Number of Independent Variables	2
Number of Observations	1000
Method	LTS Estimation

LTS Parameter Estimates

Parameter	DF	Estimate
Intercept	1	9.9742
x1	1	5.0010
x2	1	3.0219
Scale (sLTS)	0	0.9952
Scale (Wscale)	0	0.5216

Example 74.2: Robust ANOVA

The classical analysis of variance (ANOVA) technique based on least squares assumes that the underlying experimental errors are normally distributed. However, data often contain outliers due to recording or other errors. In other cases, extreme responses occur when control variables in the experiments are set to extremes. It is important to distinguish these extreme points and determine

whether they are outliers or important extreme cases. You can use the ROBUSTREG procedure for robust analysis of variance based on M estimation. Typically, there are no high leverage points in a well-designed experiment, so M estimation is appropriate.

The following example shows how to use the ROBUSTREG procedure for robust ANOVA.

An experiment was carried out to study the effects of two successive treatments (T1, T2) on the recovery time of mice with certain diseases. Sixteen mice were randomly assigned into four groups for the four different combinations of the treatments. The recovery times (time) were recorded (in hours) as shown in the following data set recover.

```
data recover;
   input  T1 $ T2 $ time @@;
datalines;
0 0 20.2  0 0 23.9  0 0 21.9  0 0 42.4
1 0 27.2  1 0 34.0  1 0 27.4  1 0 28.5
0 1 25.9  0 1 34.5  0 1 25.1  0 1 34.2
1 1 35.0  1 1 33.9  1 1 38.3  1 1 39.9
;
```

The following statements invoke the GLM procedure (Chapter 39, "The GLM Procedure") for a standard ANOVA.

```
proc glm data=recover;
    class T1 T2;
    model time = T1 T2 T1*T2;
run;
```

Output 74.2.1 Overall ANOVA

The GLM Procedure

Dependent Variable: time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	209.9118750	69.9706250	1.86	0.1905
Error	12	451.9225000	37.6602083		
Corrected Total	15	661.8343750			

R-Square	Coeff Var	Root MSE	time Mean
0.317167	19.94488	6.136791	30.76875

Output 74.2.2 Model ANOVA

Source	DF	Type I SS	Mean Square	F Value	Pr > F
T1	1	81.4506250	81.4506250	2.16	0.1671
T2	1	106.6056250	106.6056250	2.83	0.1183
T1*T2	1	21.8556250	21.8556250	0.58	0.4609

Output 74.2.1 indicates that the overall model effect is not significant at the 10% level, and Output 74.2.2 indicates that neither treatment is significant at the 10% level.

The following statements invoke the ROBUSTREG procedure with the same model.

```
proc robustreg data=recover;
   class T1 T2;
   model time = T1 T2 T1*T2 / diagnostics;
   T1_T2: test T1*T2;
   output out=robout r=resid sr=stdres;
run;
```

Output 74.2.3 shows some basic information about the model and the response variable time.

Output 74.2.3 Model Fitting Information and Summary Statistics

The ROBUSTREG Procedure

Model Information

Data Set	WORK.RECOVER
Dependent Variable	time
Number of Independent Variables	2
Number of Continuous Independent Variables	0
Number of Class Independent Variables	2
Number of Observations	16
Method	M Estimation

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
time	25.5000	31.2000	34.7500	30.7688	6.6425	6.8941

The "Parameter Estimates" table in Output 74.2.4 indicates that the main effects of both treatments are significant at the 5% level.

Output 74.2.4 Model Parameter Estimates

Parameter Estimates

Parameter			DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept			1	36.7655	2.0489	32.7497	40.7814	321.98	<.0001
T1	0		1	-6.8307	2.8976	-12.5100	-1.1514	5.56	0.0184
T1	1		0	0.0000	.	.	.	.	.
T2	0		1	-7.6755	2.8976	-13.3548	-1.9962	7.02	0.0081
T2	1		0	0.0000	.	.	.	.	.
T1*T2	0	0	1	-0.2619	4.0979	-8.2936	7.7698	0.00	0.9490
T1*T2	0	1	0	0.0000	.	.	.	.	.
T1*T2	1	0	0	0.0000	.	.	.	.	.
T1*T2	1	1	0	0.0000	.	.	.	.	.
Scale			1	3.5346					

The reason for the difference between the traditional ANOVA and the robust ANOVA is explained by Output 74.2.5, which shows that the fourth observation is an outlier. Further investigation shows that the original value of 24.4 for the fourth observation was recorded incorrectly.

Output 74.2.6 displays the robust test results. The interaction between the two treatments is not significant. Output 74.2.7 displays the robust residuals and standardized robust residuals.

Output 74.2.5 Diagnostics

Diagnostics

Obs	Standardized Robust Residual	Outlier
4	5.7722	*

Output 74.2.6 Test of Significance

Robust Linear Tests

T1_T2

Test	Test Statistic	Lambda	DF	Chi-Square	Pr > ChiSq
Rho	0.0041	0.7977	1	0.01	0.9431
Rn2	0.0041		1	0.00	0.9490

Output 74.2.7 ROBUSTREG Output

Obs	T1	T2	time	resid	stdres
1	0	0	20.2	-1.7974	-0.50851
2	0	0	23.9	1.9026	0.53827
3	0	0	21.9	-0.0974	-0.02756
4	0	0	42.4	20.4026	5.77222
5	1	0	27.2	-1.8900	-0.53472
6	1	0	34.0	4.9100	1.38911
7	1	0	27.4	-1.6900	-0.47813
8	1	0	28.5	-0.5900	-0.16693
9	0	1	25.9	-4.0348	-1.14152
10	0	1	34.5	4.5652	1.29156
11	0	1	25.1	-4.8348	-1.36785
12	0	1	34.2	4.2652	1.20668
13	1	1	35.0	-1.7655	-0.49950
14	1	1	33.9	-2.8655	-0.81070
15	1	1	38.3	1.5345	0.43413
16	1	1	39.9	3.1345	0.88679

Example 74.3: Growth Study of De Long and Summers

Robust regression and outlier detection techniques have considerable applications to econometrics. The following example from Zaman, Rousseeuw, and Orhan (2001) shows how these techniques substantially improve the ordinary least squares (OLS) results for the growth study of De Long and Summers.

De Long and Summers (1991) studied the national growth of 61 countries from 1960 to 1985 by using OLS with the following data set growth.

```
data growth;
   input country$ GDP LFG EQP NEQ GAP @@;
datalines;
Argentin  0.0089 0.0118 0.0214 0.2286 0.6079
Austria   0.0332 0.0014 0.0991 0.1349 0.5809
Belgium   0.0256 0.0061 0.0684 0.1653 0.4109
Bolivia   0.0124 0.0209 0.0167 0.1133 0.8634
Botswana  0.0676 0.0239 0.1310 0.1490 0.9474
Brazil    0.0437 0.0306 0.0646 0.1588 0.8498
Cameroon  0.0458 0.0169 0.0415 0.0885 0.9333
Canada    0.0169 0.0261 0.0771 0.1529 0.1783
Chile     0.0021 0.0216 0.0154 0.2846 0.5402
Colombia  0.0239 0.0266 0.0229 0.1553 0.7695
CostaRic  0.0121 0.0354 0.0433 0.1067 0.7043
Denmark   0.0187 0.0115 0.0688 0.1834 0.4079
Dominica  0.0199 0.0280 0.0321 0.1379 0.8293
Ecuador   0.0283 0.0274 0.0303 0.2097 0.8205
ElSalvad  0.0046 0.0316 0.0223 0.0577 0.8414
Ethiopia  0.0094 0.0206 0.0212 0.0288 0.9805
Finland   0.0301 0.0083 0.1206 0.2494 0.5589
France    0.0292 0.0089 0.0879 0.1767 0.4708
Germany   0.0259 0.0047 0.0890 0.1885 0.4585
Greece    0.0446 0.0044 0.0655 0.2245 0.7924
Guatemal  0.0149 0.0242 0.0384 0.0516 0.7885
Honduras  0.0148 0.0303 0.0446 0.0954 0.8850
HongKong  0.0484 0.0359 0.0767 0.1233 0.7471
India     0.0115 0.0170 0.0278 0.1448 0.9356
Indonesi  0.0345 0.0213 0.0221 0.1179 0.9243
Ireland   0.0288 0.0081 0.0814 0.1879 0.6457
Israel    0.0452 0.0305 0.1112 0.1788 0.6816
Italy     0.0362 0.0038 0.0683 0.1790 0.5441
IvoryCoa  0.0278 0.0274 0.0243 0.0957 0.9207
Jamaica   0.0055 0.0201 0.0609 0.1455 0.8229
Japan     0.0535 0.0117 0.1223 0.2464 0.7484
Kenya     0.0146 0.0346 0.0462 0.1268 0.9415
Korea     0.0479 0.0282 0.0557 0.1842 0.8807
Luxembou  0.0236 0.0064 0.0711 0.1944 0.2863
Madagasc -0.0102 0.0203 0.0219 0.0481 0.9217
Malawi    0.0153 0.0226 0.0361 0.0935 0.9628
Malaysia  0.0332 0.0316 0.0446 0.1878 0.7853
Mali      0.0044 0.0184 0.0433 0.0267 0.9478
Mexico    0.0198 0.0349 0.0273 0.1687 0.5921
```

```
Morocco   0.0243 0.0281 0.0260 0.0540 0.8405
Netherla  0.0231 0.0146 0.0778 0.1781 0.3605
Nigeria  -0.0047 0.0283 0.0358 0.0842 0.8579
Norway    0.0260 0.0150 0.0701 0.2199 0.3755
Pakistan  0.0295 0.0258 0.0263 0.0880 0.9180
Panama    0.0295 0.0279 0.0388 0.2212 0.8015
Paraguay  0.0261 0.0299 0.0189 0.1011 0.8458
Peru      0.0107 0.0271 0.0267 0.0933 0.7406
Philippi  0.0179 0.0253 0.0445 0.0974 0.8747
Portugal  0.0318 0.0118 0.0729 0.1571 0.8033
Senegal  -0.0011 0.0274 0.0193 0.0807 0.8884
Spain     0.0373 0.0069 0.0397 0.1305 0.6613
SriLanka  0.0137 0.0207 0.0138 0.1352 0.8555
Tanzania  0.0184 0.0276 0.0860 0.0940 0.9762
Thailand  0.0341 0.0278 0.0395 0.1412 0.9174
Tunisia   0.0279 0.0256 0.0428 0.0972 0.7838
U.K.      0.0189 0.0048 0.0694 0.1132 0.4307
U.S.      0.0133 0.0189 0.0762 0.1356 0.0000
Uruguay   0.0041 0.0052 0.0155 0.1154 0.5782
Venezuel  0.0120 0.0378 0.0340 0.0760 0.4974
Zambia   -0.0110 0.0275 0.0702 0.2012 0.8695
Zimbabwe  0.0110 0.0309 0.0843 0.1257 0.8875
;
```

The regression equation they used is

$$GDP = \beta_0 + \beta_1 LFG + \beta_2 GAP + \beta_3 EQP + \beta_4 NEQ + \epsilon$$

where the response variable is the growth in gross domestic product per worker (GDP) and the regressors are labor force growth (LFG), relative GDP gap (GAP), equipment investment (EQP), and nonequipment investment (NEQ).

The following statements invoke the REG procedure (Chapter 73, "The REG Procedure") for the OLS analysis:

```
proc reg data=growth;
   model GDP  = LFG GAP EQP NEQ ;
run;
```

Output 74.3.1 OLS Estimates

The REG Procedure
Model: MODEL1
Dependent Variable: GDP

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	−0.01430	0.01028	−1.39	0.1697
LFG	1	−0.02981	0.19838	−0.15	0.8811
GAP	1	0.02026	0.00917	2.21	0.0313
EQP	1	0.26538	0.06529	4.06	0.0002
NEQ	1	0.06236	0.03482	1.79	0.0787

The OLS analysis shown in Output 74.3.1 indicates that GAP and EQP have a significant influence on GDP at the 5% level.

The following statements invoke the ROBUSTREG procedure with the default M estimation.

```
ods graphics on;

proc robustreg data=growth plots=all;
   model GDP  = LFG GAP EQP NEQ / diagnostics leverage;
   id country;
run;

ods graphics off;
```

Output 74.3.2 displays model information and summary statistics for variables in the model.

Output 74.3.2 Model Fitting Information and Summary Statistics

The ROBUSTREG Procedure

Model Information

Data Set	WORK.GROWTH
Dependent Variable	GDP
Number of Independent Variables	4
Number of Observations	61
Method	M Estimation

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation	MAD
LFG	0.0118	0.0239	0.0281	0.0211	0.00979	0.00949
GAP	0.5796	0.8015	0.8863	0.7258	0.2181	0.1778
EQP	0.0265	0.0433	0.0720	0.0523	0.0296	0.0325
NEQ	0.0956	0.1356	0.1812	0.1399	0.0570	0.0624
GDP	0.0121	0.0231	0.0310	0.0224	0.0155	0.0150

Output 74.3.3 displays the M estimates. Besides GAP and EQP, the robust analysis also indicates that NEQ is significant. This new finding is explained by Output 74.3.4, which shows that Zambia, the 60th country in the data, is an outlier. Output 74.3.4 also identifies leverage points based on the robust MCD distances; however, there are no serious high-leverage points in this data set.

Output 74.3.3 M Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.0247	0.0097	-0.0437	-0.0058	6.53	0.0106
LFG	1	0.1040	0.1867	-0.2619	0.4699	0.31	0.5775
GAP	1	0.0250	0.0086	0.0080	0.0419	8.36	0.0038
EQP	1	0.2968	0.0614	0.1764	0.4172	23.33	<.0001
NEQ	1	0.0885	0.0328	0.0242	0.1527	7.29	0.0069
Scale	1	0.0099					

Output 74.3.4 Diagnostics

Diagnostics

Obs	country	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
1	Argentin	2.6083	4.0639	*	-0.9424	
5	Botswana	3.4351	6.7391	*	1.4200	
8	Canada	3.1876	4.6843	*	-0.1972	
9	Chile	3.6752	5.0599	*	-1.8784	
17	Finland	2.6024	3.8186	*	-1.7971	
23	HongKong	2.1225	3.8238	*	1.7161	
27	Israel	2.6461	5.0336	*	0.0909	
31	Japan	2.9179	4.7140	*	0.0216	
53	Tanzania	2.2600	4.3193	*	-1.8082	
57	U.S.	3.8701	5.4874	*	0.1448	
58	Uruguay	2.5953	3.9671	*	-0.0978	
59	Venezuel	2.9239	4.1663	*	0.3573	
60	Zambia	1.8562	2.7135		-4.9798	*
61	Zimbabwe	1.9634	3.9128	*	-2.5959	

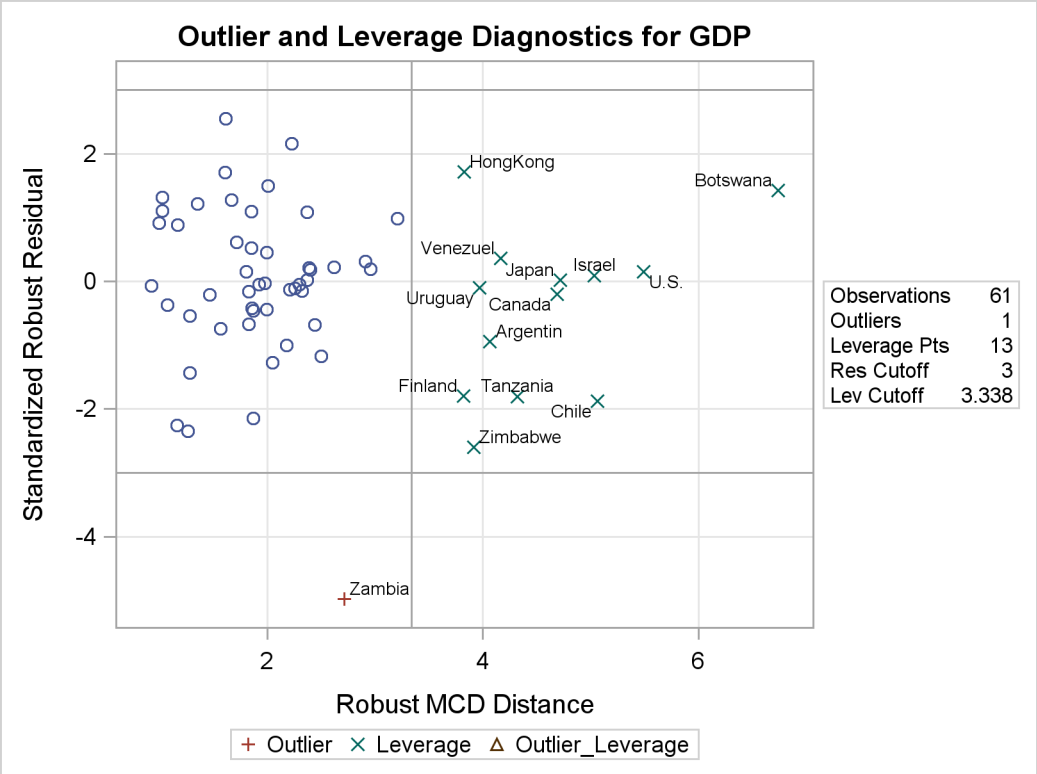
Figure 74.3.5 displays robust versions of goodness-of-fit statistics for the model.

Output 74.3.5 Goodness-of-Fit Statistics

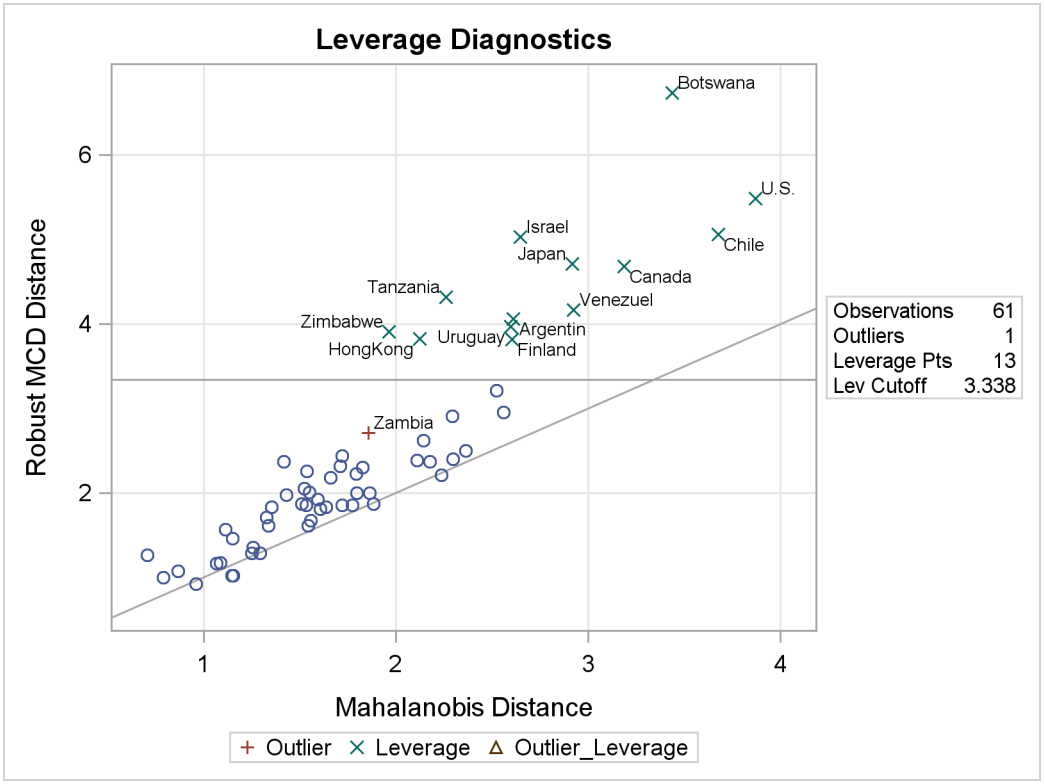
Goodness-of-Fit	
Statistic	**Value**
R-Square	0.3178
AICR	80.2134
BICR	91.5095
Deviance	0.0070

The PLOTS=ALL option generates four diagnostic plots. Figure 74.3.6 and Figure 74.3.7 are for outlier and leverage-point diagnostics. Figure 74.3.8 and Figure 74.3.9 are a histogram and a Q-Q plot of the standardized robust residuals, respectively.

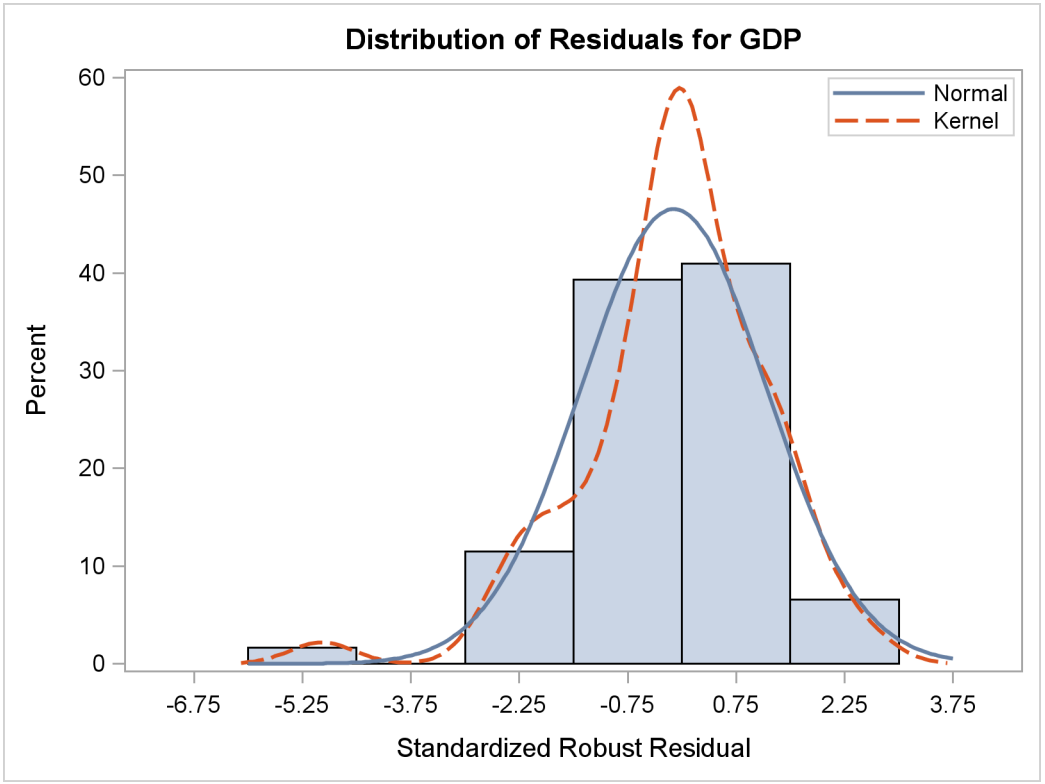
Output 74.3.6 RDPLOT for Stackloss Data



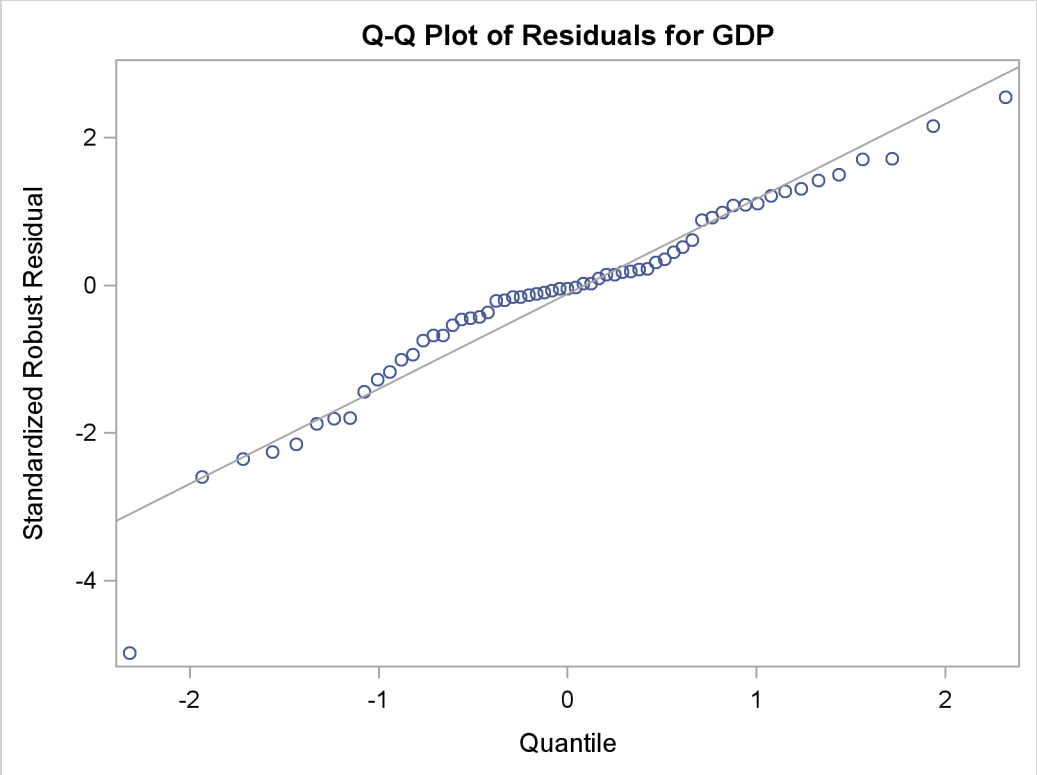
Output 74.3.7 DDPLOT for Stackloss Data



Output 74.3.8 Histogram



Output 74.3.9 Q-Q Plot



The following statements invoke the ROBUSTREG procedure with LTS estimation, which was used by Zaman, Rousseeuw, and Orhan (2001). The results are consistent with those of M estimation.

```
proc robustreg method=lts(h=33) fwls data=growth;
   model GDP  = LFG GAP EQP NEQ / diagnostics leverage ;
   id country;
run;
```

Output 74.3.10 LTS Estimates

The ROBUSTREG Procedure

LTS Parameter Estimates

Parameter	DF	Estimate
Intercept	1	-0.0249
LFG	1	0.1123
GAP	1	0.0214
EQP	1	0.2669
NEQ	1	0.1110
Scale (sLTS)	0	0.0076
Scale (Wscale)	0	0.0109

Output 74.3.10 displays the LTS estimates.

Output 74.3.11 Diagnostics and LTS R Square

Diagnostics

Obs	country	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
1	Argentin	2.6083	4.0639	*	-1.0715	
5	Botswana	3.4351	6.7391	*	1.6574	
8	Canada	3.1876	4.6843	*	-0.2324	
9	Chile	3.6752	5.0599	*	-2.0896	
17	Finland	2.6024	3.8186	*	-1.6367	
23	HongKong	2.1225	3.8238	*	1.7570	
27	Israel	2.6461	5.0336	*	0.2334	
31	Japan	2.9179	4.7140	*	0.0971	
53	Tanzania	2.2600	4.3193	*	-1.2978	
57	U.S.	3.8701	5.4874	*	0.0605	
58	Uruguay	2.5953	3.9671	*	-0.0857	
59	Venezuel	2.9239	4.1663	*	0.4113	
60	Zambia	1.8562	2.7135		-4.4984	*
61	Zimbabwe	1.9634	3.9128	*	-2.1201	

R-Square for LTS Estimation

R-Square	0.7418

Output 74.3.11 displays outlier and leverage-point diagnostics based on the LTS estimates.

Output 74.3.12 Final Weighted LS Estimates

Parameter Estimates for Final Weighted Least Squares Fit

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.0222	0.0093	-0.0405	-0.0039	5.65	0.0175
LFG	1	0.0446	0.1771	-0.3026	0.3917	0.06	0.8013
GAP	1	0.0245	0.0082	0.0084	0.0406	8.89	0.0029
EQP	1	0.2824	0.0581	0.1685	0.3964	23.60	<.0001
NEQ	1	0.0849	0.0314	0.0233	0.1465	7.30	0.0069
Scale	0	0.0116					

Output 74.3.12 displays the final weighted least squares estimates, which are identical to those reported in Zaman, Rousseeuw, and Orhan (2001).

References

Akaike, H. (1974), "A New Look at the Statistical Identification Model," *IEEE Trans. Automat Control*, 19, 716–723.

Brownlee, K.A. (1965), *Statistical Theory and Methodology in Science and Engineering*, 2nd ed., New York: John Wiley & Sons, Inc.

Chen, C. (2002), "Robust Regression and Outlier Detection with the ROBUSTREG Procedure," *Proceedings of the Twenty-seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Chen, C. and Yin, G. (2002), "Computing the Efficiency and Tuning Constants for M-Estimation," *Proceedings of the Joint Statistical Meetings*, 478–482.

Coleman, D. Holland, P., Kaden, N., Klema, V., and Peters, S.C. (1980), "A System of Subroutines for Iteratively Reweighted Least-Squares Computations," *ACM Transactions on Mathematical Software*, 6, 327–336.

De Long, J.B. and Summers, L.H. (1991), "Equipment Investment and Economic Growth," *Quarterly Journal of Economics*, 106, 445–501.

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

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.

Holland, P. and Welsch, R. (1977), "Robust Regression Using Interactively Reweighted Least-Squares," *Commun. Statist. Theor. Meth.*, 6, 813–827.

Huber, P.J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *Ann. Stat.*, 1, 799–821.

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

Marazzi, A. (1993), *Algorithm, Routines, and S Functions for Robust Statistics*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Ronchetti, E. (1985), "Robust Model Selection in Regression," *Statistics and Probability Letters*, 3, 21–23.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P.J. and Hubert, M. (1996), "Recent Development in PROGRESS," *Computational Statistics and Data Analysis*, 21, 67–85.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley & Sons, Inc.

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

Rousseeuw, P.J. and Van Driessen, K. (2000), "An Algorithm for Positive-Breakdown Regression Based on Concentration Steps," *Data Analysis: Scientific Modeling and Practical Application*, ed. W. Gaul, O. Opitz, and M. Schader, New York: Springer-Verlag, 335–346.

Rousseeuw, P.J. and Yohai, V. (1984), "Robust Regression by Means of S Estimators," in *Robust and Nonlinear Time Series Analysis*, ed. J. Franke, W. Härdle, and R.D. Martin, Lecture Notes in Statistics, 26, New York: Springer-Verlag, 256–274.

Ruppert, D. (1992), "Computing S Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, 1, 253–270.

Yohai V.J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," *Annals of Statistics*, 15, 642–656.

Yohai V.J., Stahel, W.A. and Zamar, R.H. (1991), "A Procedure for Robust Estimation and Inference in Linear Regression," in Stahel, W.A. and Weisberg, S.W., eds., *Directions in Robust Statistics and Diagnostics, Part II*, New York: Springer-Verlag.

Yohai, V.J. and Zamar, R.H. (1997), "Optimal Locally Robust M Estimate of Regression," *Journal of Statist. Planning and Inference*, 64, 309–323.

Zaman, A., Rousseeuw, P.J., Orhan, M. (2001), "Econometric Applications of High-Breakdown Robust Regression Techniques," *Econometrics Letters*, 71, 1–8.

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

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

SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

support.sas.com/publishing

SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

support.sas.com/spn

sas | THE POWER TO KNOW®