Paper 265-27

Robust Regression and Outlier Detection with the ROBUSTREG Procedure

Colin Chen, SAS Institute Inc., Cary, NC

Abstract

Robust regression is an important tool for analyzing data that are contaminated with outliers. It can be used to detect outliers and to provide resistant (stable) results in the presence of outliers. This paper introduces the ROBUSTREG procedure, which is experimental in SAS/STAT® Version 9. The ROBUSTREG procedure implements the most commonly used robust regression techniques. These include M estimation (Huber, 1973), LTS estimation (Rousseeuw, 1984), S estimation (Rousseeuw and Yohai, 1984), and MM estimation (Yohai, 1987). The paper will provide an overview of robust regression methods, describe the syntax of PROC ROBUSTREG, and illustrate the use of the procedure to fit regression models and display outliers and leverage points. This paper will also discuss scalability of the ROBUSTREG procedure for applications in data cleansing and data mining.

Introduction

The main purpose of robust regression is to 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 covariate space (i.e. outliers in the x-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 for 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 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 a procedure can withstand and still maintain its robustness. The performance of this method was improved by the FAST-LTS algorithm of Rousseeuw and Van Driessen (1998).
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.

The following example introduces the basic usage of the ROBUSTREG procedure.

Growth Study

Zaman, Rousseeuw, and Orhan (2001) used the following example to show how these robust 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 using OLS.

```
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 GDP growth per worker (GDP) and the regressors are the constant term, labor force growth (LFG), relative GDP gap (GAP), equipment investment (EQP), and non-equipment investment (NEQ).

The following statements invoke the REG procedure for the OLS analysis:

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

The OLS analysis of Figure 1 indicates that GAP and EQP have a significant influence on GDP at the 5% level.

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

Figure 1. OLS Estimates

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

```
proc robustreg data=growth;
    model GDP  = LFG GAP EQP NEQ /
                 diagnostics leverage;
    output out=robout r=resid sr=stdres;
run;
```

Figure 2 displays model information and summary statistics for variables in the model. Figure 3 displays the M estimates. Besides GAP and EQP, the robust analysis also indicates NEQ has significant impact on GDP. This new finding is explained by Figure 4, which shows that Zambia, the sixtieth country in the data, is an outlier. Figure 4 also displays leverage points; however, there are no serious high leverage points.

The ROBUSTREG Procedure

Model Information

Data Set	MYLIB.GROWTH
Dependent Variable	GDP
Number of Covariates	4
Number of Observations	61
Name of Method	M-Estimation

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation
LFG	0.0118	0.0239	0.02805	0.02113	0.009794
GAP	0.57955	0.8015	0.88625	0.725777	0.21807
EQP	0.0265	0.0433	0.072	0.052325	0.029622
NEQ	0.09555	0.1356	0.1812	0.139856	0.056966
GDP	0.01205	0.0231	0.03095	0.022384	0.015516

Summary Statistics

Variable	MAD
LFG	0.009489
GAP	0.177764
EQP	0.032469
NEQ	0.062418
GDP	0.014974

Figure 2. Model Fitting Information and Summary Statistics

The ROBUSTREG Procedure

Parameter Estimates

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

Parameter Estimates

Parameter	Pr > ChiSq
Intercept	0.0106
LFG	0.5775
GAP	0.0038
EQP	<.0001
NEQ	0.0069
Scale	

Figure 3. M estimates

The ROBUSTREG Procedure

Diagnostics

Obs	Mahalanobis Distance	Robust MCD Distance	Leverage	Robust Residual	Outlier
1	2.6083	4.0639	*	-0.9424	
5	3.4351	6.7391	*	1.4200	
8	3.1876	4.6843	*	-0.1972	
9	3.6752	5.0599	*	-1.8784	
17	2.6024	3.8186	*	-1.7971	
23	2.1225	3.8238	*	1.7161	
27	2.6461	5.0336	*	0.0909	
31	2.9179	4.7140	*	0.0216	
53	2.2600	4.3193	*	-1.8082	
57	3.8701	5.4874	*	0.1448	
58	2.5953	3.9671	*	-0.0978	
59	2.9239	4.1663	*	0.3573	
60	1.8562	2.7135		-4.9798	*
61	1.9634	3.9128	*	-2.5959	

Diagnostics Profile

Name	Percentage	Cutoff
Outlier	0.0164	3.0000
Leverage	0.2131	3.3382

Figure 4. Diagnostics

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

```
proc robustreg method=lts(h=33) fwls
               data=growth;
    model GDP  = LFG GAP EQP NEQ /
                diagnostics leverage ;
    output out=robout r=resid sr=stdres;
run;
```

Figure 5 displays the LTS estimates.

The ROBUSTREG Procedure

LTS Profile

Total Number of Observations	61
Number of Squares Minimized	33
Number of Coefficients	5
Highest Possible Breakdown Value	0.4590

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		0.0076
WScale		0.0109

Figure 5. LTS estimates

Figure 6 displays outlier and leverage point diagnostics based on the LTS estimates. Figure 7 displays the final weighted least square estimates, which are identical to those reported in Zaman, Rousseeuw, and Orhan (2001).

The ROBUSTREG Procedure

Diagnostics

Obs	Mahalanobis Distance	Robust MCD Distance	Leverage	Robust Residual	Outlier
1	2.6083	4.0639	*	-1.0715	
5	3.4351	6.7391	*	1.6574	
8	3.1876	4.6843	*	-0.2324	
9	3.6752	5.0599	*	-2.0896	
17	2.6024	3.8186	*	-1.6367	
23	2.1225	3.8238	*	1.7570	
27	2.6461	5.0336	*	0.2334	
31	2.9179	4.7140	*	0.0971	
53	2.2600	4.3193	*	-1.2978	
57	3.8701	5.4874	*	0.0605	
58	2.5953	3.9671	*	-0.0857	
59	2.9239	4.1663	*	0.4113	
60	1.8562	2.7135		-4.4984	*
61	1.9634	3.9128	*	-2.1201	

Diagnostics Profile

Name	Percentage	Cutoff
Outlier	0.0164	3.0000
Leverage	0.2131	3.3382

Rsquare for LTS-estimation

Rsquare	0.7417678684

Figure 6. Diagnostics and LTS-Rsquare

The ROBUSTREG Procedure

Parameter Estimates for Final Weighted LS

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square
Intercept	1	-0.0222	0.0093	-0.0403	-0.0041	5.75
LFG	1	0.0446	0.1755	-0.2995	0.3886	0.06
GAP	1	0.0245	0.0081	0.0085	0.0404	9.05
EQP	1	0.2824	0.0576	0.1695	0.3953	24.03
NEQ	1	0.0849	0.0311	0.0239	0.1460	7.43
Scale		0.0115				

Parameter Estimates for Final Weighted LS

Parameter	Pr > ChiSq
Intercept	0.0165
LFG	0.7995
GAP	0.0026
EQP	<.0001
NEQ	0.0064
Scale	

Figure 7. Final Weighted LS estimates

The following section provides some theoretical background for robust estimates.

Robust Estimates

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

$$y = X\theta + e$$

where $e = (e_1, ..., e_n)^T$ is an n-vector of unknown errors. It is assumed that (for 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 $L(\cdot/\sigma) = \Phi(\cdot)$, the standard normal distribution with density $\phi(s) = (1/\sqrt{2}\pi)exp(-s^2/2)$. $r = (r_1, ..., r_n)^T$ denotes the n-vector of residuals for a given value of θ and by x_i^T the i-th row of the matrix X.

The Ordinary Least Squares (OLS) estimate $\hat{\theta}_{LS}$ of θ is obtained as the solution of the problem

$$\min_\theta Q_{LS}(\theta)$$

where $Q_{LS}(\theta) = \frac{1}{2}\sum_{i=1}^n r_i^2$.

Taking the partial derivatives of Q_{LS} with respect to the components of θ and setting them equal to zero yields the normal equations

$$XX^T\theta = X^Ty$$

If the rank(X) is equal to p, the solution for θ is

$$\hat{\theta}_{LS} = (X^TX)^{-1}X^Ty$$

The least squares estimate is the maximum likelihood estimate when $L(\cdot/\sigma) = \Phi(\cdot)$. In this case the usual estimate of the scale parameter σ is

$$\hat{\sigma}_{LS} = \sqrt{\frac{1}{(n-p)}Q_{LS}(\hat{\theta})}$$

As shown in the growth study, the OLS estimate can be significantly influenced by a single outlier. To bound the influence of outliers, Huber (1973) introduced the M estimate.

Huber-type Estimates

Instead of minimizing a sum of squares, 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 OLS estimate, ρ is the quadratic function.

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

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

where $\psi = \rho'$. If ρ is convex, $\hat{\theta}_M$ is the unique solution.

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

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

PROC ROBUSTREG provides ten kinds of weight functions (corresponding to ten ρ-functions) through the WEIGHTFUNCTION= option in the MODEL statement. The scale parameter σ can be specified using the SCALE= option in the PROC statement.

If σ is unknown, then the function

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

is minimized with $a > 0$ over θ and σ by alternately improving $\hat{\theta}$ in a location step and $\hat{\sigma}$ in a scale step.

For the scale step, three options can be used to estimate σ:

1. METHOD=M(SCALE=HUBER<(D=d)>) This option obtains $\hat{\sigma}$ 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 $h = \frac{n-p}{n}(d^2 + (1-d^2)\Phi(d) - .5 - d\sqrt{2\pi}e^{-\frac{1}{2}d^2})$ is the Huber constant (refer to Huber 1981, p. 179). You can specify d with the D= option. By default, $d = 2.5$.

2. METHOD=M(SCALE=TUKEY<(D=d)>) This option obtains $\hat{\sigma}$ by solving the supplementary equation

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

   where

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

   $\chi_d^{'}$ being Tukey's Biweight 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 by the D= option. By default, $d = 2.5$.

3. METHOD=M(SCALE=MED) This option obtains $\hat{\sigma}$ by the iteration

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

   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).

SCALE = MED is the default.

High Breakdown Value Estimates

If the data are contaminated in the x-space, M estimation does not do well. This can be shown using a data set created by Hawkins, Bradu, and Kass (1984).

```
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
;
```

Both OLS estimation and M estimation suggest that observations 11 to 14 are serious outliers. However, these four observations were generated from the underlying model and observations 1 to 10 were contaminated. The reason that OLS estimation and M estimation do not pick up the bad observations is that they cannot distinguish good leverage points (observations 11 to 14) from bad leverage points (observations 1 to 10). In such cases, high breakdown value estimates are needed.

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

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

$r_{(1)}^2 \le r_{(2)}^2 \le ... \le r_{(n)}^2$ are the ordered squared residuals $r_i^2 = (y_i - x_i^T\theta)^2$, $i = 1, ..., n$, 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 option H= in the PROC statement. By default, $h = [(3n+p+1)/4]$. The breakdown value is $\frac{n-h}{n}$ for the LTS estimate.

LMS estimate

The *least median of squares (LMS) estimate* is defined as the p-vector

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

where

$$Q_{LMS}(\theta) = r^2_{(h)}$$

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

The breakdown value for the LMS estimate is also $\frac{n-h}{n}$. However the LTS estimate has several advantages over the LMS estimate. Its objective function is smoother, making the LTS estimate less "jumpy" (i.e. sensitive to local effects) than the LMS estimate. Its statistical efficiency is better, because the LTS estimate is asymptotically normal whereas the LMS estimate has a lower convergence rate (Rousseeuw and Leroy (1987)). Another important advantage is that, using the FAST-LTS algorithm by Rousseeuw and Van Driessen (1998), the LTS estimate takes less computing time and is more accurate.

The ROBUSTREG procedure computes LTS estimates. The estimates are mainly used to detect outliers in the data, which are then downweighted in the resulting weighted least square regression.

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

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

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

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

PROC ROBUSTREG provides two kinds of functions for χ:

Tukey: Specified with the option CHIF=TUKEY.

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

The turning constant k_0 controls the breakdown value and efficiency of the S estimate. By specifying the efficiency 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%.

Yohai: Specified with the option CHIF=YOHAI.

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

where $b_0 = 1.792$, $b_1 = -0.972$, $b_2 = 0.432$, $b_3 = -0.052$, and $b_4 = 0.002$. By specifying the efficiency 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%.

The following statements invoke the ROBUSTREG procedure with the LTS estimation method for the *hbk* data.

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

Figure 8 displays the model fitting information and summary statistics for the response variable and independent covariates.

Figure 9 displays information about the LTS fit, which includes the breakdown value of the LTS estimate. In this example, the LTS estimate minimizes the sum of 40 smallest squares of residuals, thus it can still pick up the right model if the remaining 35 observations are contaminated.

The ROBUSTREG Procedure

Model Information

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

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation
x1	0.8	1.8	3.1	3.206667	3.652631
x2	1	2.2	3.3	5.597333	8.239112
x3	0.9	2.1	3	7.230667	11.74031
y	-0.5	0.1	0.7	1.278667	3.492822

Summary Statistics

Variable	MAD
x1	1.927383
x2	1.630862
x3	1.779123
y	0.889561

Figure 8. Model Fitting Information and Summary Statistics

The ROBUSTREG Procedure

LTS Profile

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

Figure 9. LTS Profile

Figure 10 displays parameter estimates for covariates and scale. Two robust estimates of the scale parameter are displayed. The weighted scale estimate (Wscale) is a more efficient estimate of the scale parameter.

The ROBUSTREG Procedure

LTS Parameter Estimates

Parameter	DF	Estimate
Intercept	1	-0.3431
x1	1	0.0901
x2	1	0.0703
x3	1	-0.0731
Scale		0.7451
WScale		0.5749

Figure 10. LTS Parameter Estimates

Figure 11 displays outlier and leverage point diagnostics. The ID variable *index* is used to identify the observations. The first ten observations are identified as outliers and observations 11 to 14 are identified as good leverage points.

The ROBUSTREG Procedure

Diagnostics

Obs	index	Mahalanobis Distance	Robust MCD Distance	Leverage	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	

Diagnostics Profile

Name	Percentage	Cutoff
Outlier	0.1333	3.0000
Leverage	0.1867	3.0575

Figure 11. Diagnostics Profile

Figure 12 displays the final weighted LS estimates. These estimates are OLS estimates computed after deleting the detected outliers.

The ROBUSTREG Procedure

Parameter Estimates for Final Weighted LS

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square
Intercept	1	-0.1805	0.0968	-0.3702	0.0093	3.47
x1	1	0.0814	0.0618	-0.0397	0.2025	1.73
x2	1	0.0399	0.0375	-0.0336	0.1134	1.13
x3	1	-0.0517	0.0328	-0.1159	0.0126	2.48
Scale		0.5165				

Parameter Estimates for Final Weighted LS

Parameter	Pr > ChiSq
Intercept	0.0623
x1	0.1879
x2	0.2875
x3	0.1150
Scale	

Figure 12. Final Weighted LS Estimates

MM estimate

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

1. Compute an initial (consistent) high breakdown value estimate $\hat{\theta}'$. PROC ROBUSTREG provides two kinds of estimates as the initial estimate, the LTS estimate and the S estimate. By default, PROC ROBUSTREG uses the LTS

estimate because of its speed, efficiency, 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 CHI function. To use the S estimate as the initial estimate, you need to 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 initial estimates, you can also specify the initial estimate using the INEST= option in the PROC statement.

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 $\beta = \int \chi(s)d\Phi(s)$.
PROC ROBUSTREG provides two kinds of functions for χ:

Tukey: Specified with the option CHIF=TUKEY.

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

where k_0 can be specified by 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: Specified with the option CHIF=YOHAI.

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

where $b_0 = 1.792$, $b_1 = -0.972$, $b_2 = 0.432$, $b_3 = -0.052$, and $b_4 = 0.002$. k_0 can be specified with the K0= option. The default $k_0 = .7405$ 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 estimate is used here. PROC ROBUSTREG provides two kinds of functions for ρ corresponding to the two kinds of χ functions, respectively.

Tukey: With the option CHIF=TUKEY,

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

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

Yohai: With the option CHIF=YOHAI,

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

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

In the following sections, robust diagnostic and inference are introduced.

Resistant Diagnostic and Outlier Detection

Robust Distance

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 location and scatter matrix for the multivariates. PROC ROBUSTREG implements the FAST-MCD algorithm of Rousseeuw and Van Driessen (1999) for computing these robust multivariate estimates.

High Leverage Points

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

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

Outliers

Residuals $r_i, i = 1, ..., n$ based on the described robust estimates are used to detect outliers in the response direction. The variable OUTLIER is defined as

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

An ODS table called DIAGNOSTICS provides the summary of these two variables if you specify the DIAGNOSTICS and LEVERAGE options in the MODEL statement. As an example, review the syntax for the MODEL statement used by the growth study:

```
model GDP  = LFG GAP EQP NEQ /
             diagnostics leverage;
```

If you do not specify the LEVERAGE option, only the OUTLIER variable is included in the ODS table. However, the DIAGNOSTICS option is required if you specify the LEVERAGE option.

Robust Inference

Robust Measure of Goodness-of-Fit and Model Selection

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 ρ is the objective function for the robust estimate, $\hat{\mu}$ is the robust location estimator, and $\hat{s}$ is the robust scale estimator in the full model.

The Information Criterion is a powerful tool for model selection. The counterpart of the Akaike (1974) AIC criterion for robust regression 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 some robust estimate of σ, and $\hat{\theta}$ is the robust estimator of θ with a p-dimensional design matrix.

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

The robust version of the Schwarz information criteria (BIC) is defined as

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

For the growth study, PROC ROBUSTREG produces the following goodness-of-fit table:

```
                The ROBUSTREG Procedure

                    Goodness-of-Fit
                    Statistics for
                     M-estimation

              Statistic            Value

              Rsquare       0.3177714766
              AICR          80.213370744
              BICR           91.50951378
              Deviance      0.0070081124
```

Figure 13. Goodness-of-Fit

Asymptotic Covariance and Confidence Intervals

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

$$\text{H1: } K^2 \frac{[1/(n-p)] \sum (\psi(r_i))^2}{[(1/n) \sum (\psi'(r_i))]^2} (X^T X)^{-1}$$

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

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

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

Linear Tests

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

$$\mathcal{H}_0 : \quad \theta_j = 0, \quad j = q + 1, ..., p$$

They are briefly described as follows. Refer to Hampel et al. (1986, Chapter 7) for details.

ρ-test:
The robust estimators in the full and reduced model are $\hat{\theta}_0 \in \Omega_0$ and $\hat{\theta}_1 \in \Omega_1$, respectively. Let

$$Q_0 = Q(\hat{\theta}_0) = \min\{Q(\theta)|\theta \in \Omega_0\},$$
$$Q_1 = Q(\hat{\theta}_1) = \min\{Q(\theta)|\theta \in \Omega_1\},$$

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

The robust F test is based on the test statistic

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

Asymptotically $S_n^2 \sim \lambda\chi^2_{p-q}$ under $\mathcal{H}_0$, where the standardization factor is $\lambda = \int \psi^2(s)d\Phi(s) / \int \psi'(s)d\Phi(s)$ and Φ is the c.d.f. of the standard normal distribution. Large values of S_n^2 are significant. This robust F test is a special case of the general τ-test of Hampel et al. (1986, Section 7.2).

R_n^2-test:
The test statistic for the R_n^2-test is defined as

$$R_n^2 = n(\hat{\theta}_{q+1}, ..., \hat{\theta}_p)H_{22}^{-1}(\hat{\theta}_{q+1}, ..., \hat{\theta}_p)^T$$

where H_{22} is the $(p-q) \times (p-q)$ lower right block of the asymptotic covariance matrix of the M estimate $\hat{\theta}_M$ of θ in a p-parameter linear model.

Under $\mathcal{H}_0$, R_n^2 has an asymptotic χ^2 distribution with $p-q$ degrees of freedom. Large absolute values of R_n^2 are significant.

Robust ANOVA

The classical analysis of variance (ANOVA) technique based on least squares is safe to use if the underlying experimental errors are normally distributed. However, data often contain outliers due to recording or other errors. In other cases, extreme responses might be produced by setting control variables in the experiments to extremes. It is important to distinguish these extreme points and determine whether they are outliers or important extreme cases. The ROBUSTREG procedure can be used for robust analysis of variance based on M estimation. Since the independent variables are well designed in the experiments, there are no high leverage points and M estimation is suitable.

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

In an experiment studying the effects of two successive treatments (T1, T2) on the recovery time of mice with certain disease, 16 mice were randomly assigned into four groups for the four different combinations of the treatments. The recovery times (in hours) were recorded.

```
data recover;
   input id T1 $ T2 $ time;
   datalines;
1  0 0 20.2  9  0 1 25.9
2  0 0 23.9  10 0 1 34.5
3  0 0 21.9  11 0 1 25.1
4  0 0 42.4  12 0 1 34.2
5  1 0 27.2  13 1 1 35.0
6  1 0 34.0  14 1 1 33.9
7  1 0 27.4  15 1 1 38.3
8  1 0 28.5  16 1 1 39.9
;
```

The following statements invoke the GLM procedure for a standard ANOVA.

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

The results in Figure 14 indicate that neither treatment is significant at the 10% level.

```
                              The GLM Procedure

Dependent Variable: time

 Source                     DF        Type I SS      Mean Square

 T1                          1       81.4506250       81.4506250
 T2                          1      106.6056250      106.6056250
 T1*T2                       1       21.8556250       21.8556250

        Source                      F Value    Pr > F

        T1                             2.16    0.1671
        T2                             2.83    0.1183
        T1*T2                          0.58    0.4609


 Source                     DF      Type III SS      Mean Square

 T1                          1       81.4506250       81.4506250
 T2                          1      106.6056250      106.6056250
 T1*T2                       1       21.8556250       21.8556250

        Source                      F Value    Pr > F

        T1                             2.16    0.1671
        T2                             2.83    0.1183
        T1*T2                          0.58    0.4609
```

Figure 14. Model ANOVA

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;
run;
```

The parameter estimates in Figure 15 indicate strong significance of both treatments.

The ROBUSTREG Procedure

Parameter Estimates

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

Parameter Estimates

Parameter			Pr > ChiSq
Intercept			<.0001
T1	0		0.0184
T1	1		.
T2	0		0.0081
T2	1		.
T1*T2	0	0	0.9490
T1*T2	0	1	.
T1*T2	1	0	.
T1*T2	1	1	.
Scale			

Figure 15. Model Parameter Estimates

The reason for the difference between the traditional ANOVA and the robust ANOVA is explained by Figure 16, which shows that the fourth observation is an obvious outlier. Further investigation shows that the original value 24.4 for the fourth observation was recorded incorrectly.

The ROBUSTREG Procedure

Diagnostics

Obs	Robust Residual	Outlier
4	5.7722	*

Diagnostics Profile

Name	Percentage	Cutoff
Outlier	0.0625	3.0000

Figure 16. Diagnostics

Figure 17 displays the robust test results. The interaction between the two treatments is not significant.

The ROBUSTREG Procedure

Robust Linear Tests

T1_T2

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

Figure 17. Test of Significance

Graphical Displays

Two particularly useful plots for revealing outliers and leverage points are a scatter plot of the robust residuals against the robust distances (RDPLOT) and a scatter plot of the robust distances against the classical Mahalanobis distances (DDPLOT). You can create these two displays using the data in the ODS table named DIAGNOSTICS.

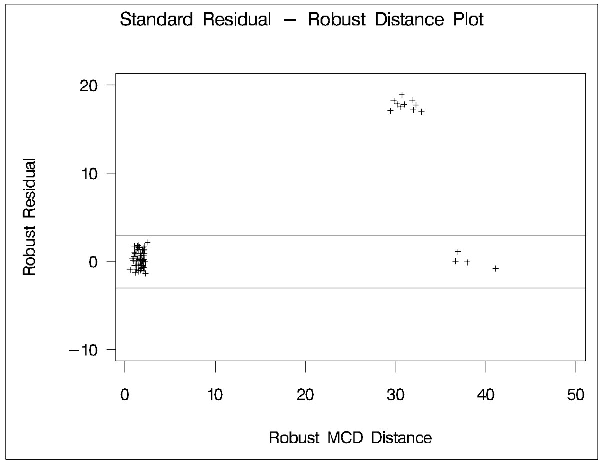

Figure 18. RDPLOT

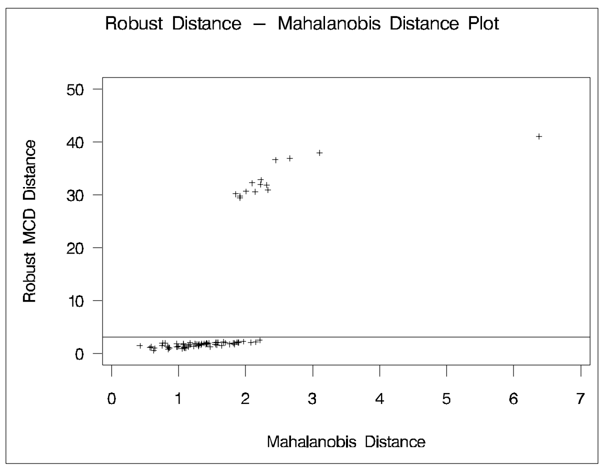

Figure 19. DDPLOT

For the *hbk* data, the following statements create a SAS data set named DIAGNOSTICS and produce the RDPLOT in Figure 18 and the DDPLOT in Figure 19:

```
ods output Diagnostics=Diagnostics;
ods select Diagnostics;

proc robustreg data=hbk method=lts;
    model y = x1 x2 x3 /
```

```
          diagnostics(all) leverage;
    id index;
run;

title "Standard Residual -
       Robust Distance Plot";
symbol v=plus h=2.5 pct;
proc gplot data=Diagnostics;
    plot RResidual * Robustdis /
         hminor = 0
         vminor = 0
         vaxis  = axis1
         vref   = -3 3
         frame;
         axis1 label = ( r=0 a=90 );
run;

title "Robust Distance -
       Mahalanobis Distance Plot";
symbol v=plus h=2.5 pct;
proc gplot data=Diagnostics;
    plot Robustdis * Mahalanobis /
         hminor = 0
         vminor = 0
         vaxis  = axis1
         vref   = 3.0575
         frame;
         axis1 label = ( r=0 a=90 );
run;
```

These plots are helpful in identifying outliers, good, and bad leverage points.

Scalability

The ROBUSTREG procedure implements parallel algorithms for LTS- and S estimation. You can use the global SAS option CPUCOUNTS to specify the number of threads to use in the computation:

OPTIONS CPUCOUNTS=1-256|ACTUAL ;

More details about multithreading in SAS Version 9 can be found in Cohen (2002).

The following table contains some empirical results for LTS estimation we got from using a single processor and multiple processors (with 8 processors) on a SUN multi-processor workstation (time in seconds):

RobustReg Timing and Speedup Results for Narrow Data (10 regressors)

numObs	num Vars	Time 8 threads	Unthreaded time
50000	10	7.78	5.90
100000	10	10.70	23.54
200000	10	23.49	80.30
300000	10	41.41	171.03
400000	10	63.20	296.30
500000	10	93.00	457.00
750000	10	173.00	1003.00
1000000	10	305.00	1770.00

RobustReg Timing and Speedup Results for Narrow Data (10 regressors)

numObs	Scalable speedup	Scalable speedup with intercept adjustment
50000	0.75835	1.26137
100000	2.20000	2.24629
200000	3.41848	3.44375
300000	4.13016	4.01907
400000	4.68829	4.28268
500000	4.91398	4.33051
750000	5.79769	4.94009
1000000	5.80328	4.30432

RobustReg Timing and Speedup Results for Wide Data (5000 observations)

num Vars	num Obs	Time 8 threads	Unthreaded time	Scalable speedup
50	5000	17.69	24.06	1.36009
100	5000	40.29	76.45	1.89749
200	5000	128.14	319.80	2.49571
250	5000	207.21	520.15	2.51026

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. John Wiley & Sons, New York.

Cohen, R. A. (2002), "SAS® Meets Big Iron: High Performance Computing in SAS Analytic Procedures", *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

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., 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*, John Wiley & Sons, New York.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984), "Location of several outliers in multiple regression data using elemental sets," *Technometrics* , **26**, 197-208.

Holland, P. and Welsch, R. (1977), “Robust regression using interatively 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*. John Wiley & Sons, New York.

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

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,* Wiley-Interscience, New York (Series in Applied Probability and Statistics), 329 pages. ISBN 0-471-85233-3.

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. (1998), “Computing LTS Regression for Large Data Sets,” Technical Report, University of Antwerp, submitted.

Rousseeuw, P.J. and Yohai, V. (1984), “Robust Regression by Means of S estimators”, in *Robust and Nonlinear Time Series Analysis*, edited by J. Franke, W. Härdle, and R.D. Martin, Lecture Notes in Statistics **26**, Springer Verlag, New York, 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,* Springer-Verlag, New Work.

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.

Acknowledgments

The author would like to thank Dr. Robert Rodriguez, Professor Peter Rousseeuw, Victor Yohai, Zamma Ruben, Alfio Marazzi, Xuming He and Roger Koenker for reviewing, comments, and discussions.

Contact Information

Lin (Colin) Chen, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919) 531-6388, FAX (919) 677-4444, Email Lin.Chen@sas.com.