Chapter 9
Robust Regression Examples

Chapter Table of Contents

Chapter 9

Robust Regression Examples

Overview

SAS/IML has three subroutines that can be used for outlier detection and robust regression. The Least Median of Squares (LMS) and Least Trimmed Squares (LTS) subroutines perform *robust regression* (sometimes called *resistant regression*). These subroutines are able to detect outliers and perform a least-squares regression on the remaining observations. The Minimum Volume Ellipsoid Estimation (MVE) subroutine can be used to find the minimum volume ellipsoid estimator, which is the location and robust covariance matrix that can be used for constructing confidence regions and for detecting multivariate outliers and leverage points. Moreover, the MVE subroutine provides a table of robust distances and classical Mahalanobis distances. The LMS, LTS, and MVE subroutines and some other robust estimation theories and methods were developed by Rousseeuw (1984) and Rousseeuw and Leroy (1987). Some statistical applications for MVE are described in Rousseeuw and Van Zomeren (1990).

Whereas robust regression methods like L1 or Huber M-estimators reduce the influence of outliers only (compared to least-squares or L2 regression), resistant regression methods like LMS and LTS can completely disregard influential outliers (sometimes called *leverage points*) from the fit of the model. The algorithms used in the LMS and LTS subroutines are based on the PROGRESS program by Rousseeuw and Leroy (1987). Rousseeuw and Hubert (1996) prepared a new version of PROGRESS to facilitate its inclusion in SAS software, and they have incorporated several recent developments. Among other things, the new version of PROGRESS now yields the exact LMS for simple regression, and the program uses a new definition of the robust coefficient of determination (R^2). Therefore, the outputs may differ slightly from those given in Rousseeuw and Leroy (1987) or those obtained from software based on the older version of PROGRESS. The MVE algorithm is based on the algorithm used in the MINVOL program by Rousseeuw (1984).

The three SAS/IML subroutines are designed for

- LMS: minimizing the hth ordered squared residual
- LTS: minimizing the sum of the h smallest squared residuals
- MVE: minimizing the volume of an ellipsoid containing h points

where h is defined in the range

$$\frac{N}{2} + 1 \;\leq\; h \;\leq\; \frac{3N}{4} + \frac{n+1}{4}$$

In the preceding equation, N is the number of observations and n is the number of regressors. * The value of h determines the *breakdown point*, which is "the smallest fraction of contamination that can cause the estimator T to take on values arbitrarily far from $T(Z)$" (Rousseeuw and Leroy 1987, p.10). Here, T denotes an estimator and $T(Z)$ applies T to a sample Z.

For each parameter vector $\mathbf{b} = (b_1, \ldots, b_n)$, the residual of observation i is $r_i = y_i - \mathbf{x}_i\mathbf{b}$. You then denote the ordered, squared residuals as

$$(r^2)_{1:N} \leq \ldots \leq (r^2)_{N:N}$$

The objective functions for the LMS and LTS optimization problems are defined as follows:

- LMS

  $$F_{\text{LMS}} = (r^2)_{h:N} \longrightarrow \min$$

  Note that, for $h = N/2 + 1$, the hth quantile is the median of the squared residuals. The default h in PROGRESS is $h = \left[\frac{N+n+1}{2}\right]$, which yields the breakdown value (where $[k]$ denotes the integer part of k).

- LTS

  $$F_{\text{LTS}} = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(r^2)_{i:N}} \longrightarrow \min$$

- MVE

  The objective function for the MVE optimization problem is based on the hth quantile $d_{h:N}$ of the Mahalanobis-type distances $\mathbf{d} = (d_1, \ldots, d_N)$,

  $$F_{\text{MVE}} = \sqrt{d_{h:N}\det(\mathbf{C})} \longrightarrow \min$$

  subject to $d_{h:N} = \sqrt{\chi^2_{n,0.5}}$, where $\mathbf{C}$ is the scatter matrix estimate, and the Mahalanobis-type distances are computed as

  $$\mathbf{d} = \text{diag}(\sqrt{(\mathbf{X} - T)^T\mathbf{C}^{-1}(\mathbf{X} - T)})$$

  where T is the location estimate.

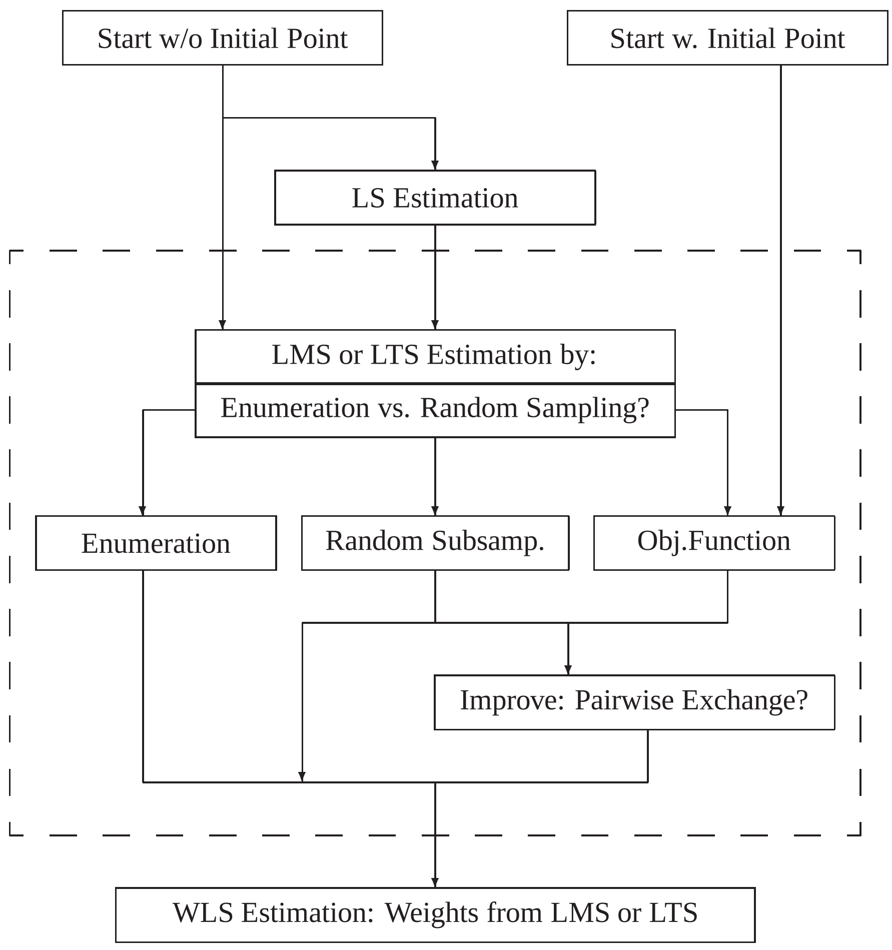
Because of the nonsmooth form of these objective functions, the estimates cannot be obtained with traditional optimization algorithms. For LMS and LTS, the algorithm, as in the PROGRESS program, selects a number of subsets of n observations out of the N given observations, evaluates the objective function, and saves the subset with

*The value of h can be specified (see the "Syntax" section), but in most applications the default value works just fine and the results seem to be quite stable toward different choices of h.

the lowest objective function. As long as the problem size enables you to evaluate all such subsets, the result is a global optimum. If computer time does not permit you to evaluate all the different subsets, a random collection of subsets is evaluated. In such a case, you may not obtain the global optimum.

Note that the LMS, LTS, and MVE subroutines are executed only when the number N of observations is over twice the number n of explanatory variables x_j (including the intercept), that is, if $N > 2n$.

Flow Chart for LMS, LTS, and MVE



Flow Chart Indicating: LS $\longrightarrow$ [LMS or LTS] $\longrightarrow$ WLS

Separate LMS or LTS Part Inside Dashbox Corresponds to MVE

Examples Using LMS and LTS Regression

The following results are based on the updated version of the PROGRESS program by Rousseeuw and Hubert (1996), and they differ slightly from those given in Rousseeuw and Leroy (1987), who use the earlier version of PROGRESS. For space reasons, the output of the tables containing residuals and resistant diagnostics are not included in this document. The macros *prilmts*, *scatlmts*, *primve*, *scatmve*, and *lmsdiap* are used in these examples for printing and plotting the results. See Chapter 18, "Module Library," for more information.

Example 9.1. LMS and LTS with Substantial Leverage Points: Hertzsprung-Russell Star Data

The following data are reported in Rousseeuw and Leroy (1987, p. 27) and are based on Humphrey (1978) and Vansina and De Greve (1982). The 47 observations correspond to the 47 stars of the CYG OB1 cluster in the direction of Cygnus. The regressor variable (column 2) x is the logarithm of the effective temperature at the surface of the star (T_e), and the response variable (column 3) y is the logarithm of its light intensity (L/L_0). The results for LS and LMS on page 28 of Rousseeuw and Leroy (1987) are based on a more precise (five decimal places) version of the data set. This data set is remarkable in that it contains four substantial leverage points (giant stars) corresponding to observations 11, 20, 30, and 34 that greatly affect the results of L_2 and even L_1 regression.

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

a = ab[,2]; b = ab[,3];
```

The following code specifies that most of the output be printed.

```
print "*** Hertzsprung-Russell Star Data: Do LMS ***";
   optn = j(8,1,.);
   optn[2]= 3;    /* ipri */
   optn[3]= 3;    /* ilsq */
   optn[8]= 3;    /* icov */
call lms(sc,coef,wgt,optn,b,a);
```

Output 9.1.1. Some Simple Statistics

```
                    Median and Mean

                          Median            Mean
         VAR1        4.420000000     4.310000000
         Intercep    1.000000000     1.000000000
         Response    5.100000000     5.012127660


             Dispersion and Standard Deviation

                      Dispersion          StdDev
         VAR1       0.1630862440    0.2908234187
         Intercep   0.0000000000    0.0000000000
         Response   0.6671709983    0.5712493409
```

Partial output for LS regression is shown in Output 9.1.2.

Output 9.1.2. Table of Unweighted LS Regression

```
                  ***********************************
                 Unweighted Least-Squares Estimation
                 ***********************************

                      LS Parameter Estimates
-----------------------------------------------------------------
                      Approx                     Lower     Upper
          Estimate  Std Error  T Value   Prob   Wald CI   Wald CI
-----------------------------------------------------------------
VAR1       -0.4133    0.28626  -1.4438  0.156   -0.9744    0.1478
Intercep    6.7935    1.23652   5.4940 175E-8    4.3699    9.2170

                   Sum of Squares = 14.346394626
                      Degrees of Freedom = 45
                  LS Scale Estimate = 0.5646315343

                 COV Matrix of Parameter Estimates

                            VAR1        Intercep
         VAR1        0.081943343    -0.353175807
         Intercep   -0.353175807     1.528970895

                     R-squared = 0.0442737441
                  F(1,45) Statistic = 2.0846120667
                     Probability = 0.1557164396
```

Looking at the column *Best Crit* in the iteration history table, you see that, with complete enumeration, the optimal solution is found very early.

Output 9.1.3. History of Iteration Process

```
 ***********************************************
***      Complete Enumeration for LMS       ***
***********************************************

      Subset Singular        Best Crit   Pct
         271        5 0.39279108982007   25%
         541        8 0.39279108982007   50%
         811       27 0.39279108982007   75%
        1081       45 0.39279108982007  100%
        Minimum Criterion=0.3927910898

    ****************************************
    Least Median of Squares (LMS) Regression
    ****************************************

 Minimizing the 25th Ordered Squared Residual.
  Highest Possible Breakdown Value =  48.94 %
Selection of All 1081 Subsets of 2 Cases Out of 47
      Among 1081 subsets 45 are singular.
```

Output 9.1.4. Results of Optimization

```
          Observations of Best Subset

                       2            29


             Estimated Coefficients

                    VAR1          Intercep
              3.97058824      -12.62794118

      LMS Objective Function = 0.2620588235
       Preliminary LMS Scale = 0.3987301586
         Robust R Squared = 0.5813148789
     Final LMS Scale Estimate = 0.3645644492
```

The output for WLS regression follows. Due to the size of the scaled residuals, six observations (with numbers 7, 9, 11, 20, 30, 34) were assigned zero weights in the following WLS analysis.

Output 9.1.5. Table of Weighted LS Regression

```
                 *********************************
                 Weighted Least-Squares Estimation
                 *********************************

               RLS Parameter Estimates Based on LMS
-----------------------------------------------------------------
                      Approx                         Lower     Upper
          Estimate  Std Error  T Value   Prob   Wald CI   Wald CI
-----------------------------------------------------------------
VAR1        3.0462    0.43734   6.9652  24E-9    2.1890    3.9033
Intercep   -8.5001    1.92631  -4.4126 0.0001  -12.2755   -4.7246

                Weighted Sum of Squares = 4.52819451
                      Degrees of Freedom = 39
                 RLS Scale Estimate = 0.3407455818


                 COV Matrix of Parameter Estimates

                                 VAR1          Intercep
           VAR1           0.191265604      -0.842128459
           Intercep      -0.842128459       3.710661875

                 Weighted R-squared = 0.5543573521
                  F(1,39) Statistic = 48.514065776
                     Probability = 2.3923178E-8
              There are 41 points with nonzero weight.
                   Average Weight = 0.8723404255
```

The LTS regression leads to similar results:

```
print "*** Hertzsprung-Russell Star Data: Do LTS ***";
   optn = j(8,1,.);
   optn[2]= 3;    /* ipri */
   optn[3]= 3;    /* ilsq */
   optn[8]= 3;    /* icov */
call lts(sc,coef,wgt,optn,b,a);
```

Output 9.1.6. History of Iteration Process

```
 ***********************************************
***      Complete Enumeration for LTS       ***
***********************************************

      Subset Singular        Best Crit   Pct
         271        5 0.27426824300686   25%
         541        8 0.27426824300686   50%
         811       27 0.27426824300686   75%
        1081       45 0.27426824300686  100%
         Minimum Criterion=0.274268243

     **************************************
     Least Trimmed Squares (LTS) Regression
     **************************************

 Minimizing Sum of 25 Smallest Squared Residuals.
   Highest Possible Breakdown Value =  48.94 %
Selection of All 1081 Subsets of 2 Cases Out of 47
       Among 1081 subsets 45 are singular.
```

Output 9.1.7. Results of Optimization

```
          Observations of Best Subset

                       2            35

             Estimated Coefficients

                    VAR1          Intercep
              4.24242424      -13.72633939

    LTS Objective Function = 0.1829838175
     Preliminary LTS Scale = 0.4525412929
       Robust R Squared = 0.4038660847
   Final LTS Scale Estimate = 0.3743418666
```

Example 9.2. LMS and LTS: Stackloss Data

This example presents the results for Brownlee's (1965) stackloss data, which is also used for documenting the L1 regression module. The three explanatory variables correspond to measurements for a plant oxidizing ammonia to nitric acid on 21 consecutive days.

- x_1 air flow to the plant
- x_2 cooling water inlet temperature
- x_3 acid concentration

The response variable y_i gives the permillage of ammonia lost (stackloss). These data are also given in Rousseeuw and Leroy (1987, p.76) and Osborne (1985, p.267):

```
print "Stackloss Data";
   aa = { 1  80  27  89  42,
          1  80  27  88  37,
          1  75  25  90  37,
          1  62  24  87  28,
          1  62  22  87  18,
          1  62  23  87  18,
          1  62  24  93  19,
          1  62  24  93  20,
          1  58  23  87  15,
          1  58  18  80  14,
          1  58  18  89  14,
          1  58  17  88  13,
          1  58  18  82  11,
          1  58  19  93  12,
          1  50  18  89   8,
          1  50  18  86   7,
          1  50  19  72   8,
          1  50  19  79   8,
          1  50  20  80   9,
          1  56  20  82  15,
          1  70  20  91  15 };
```

Rousseeuw and Leroy (1987, p.76) cite a large number of papers in which this data set was analyzed before. They state that most researchers “concluded that observations 1, 3, 4, and 21 were outliers” and that some people also reported observation 2 as an outlier.

Consider 2000 Random Subsets

For $N = 21$ and $n = 4$ (three explanatory variables including intercept), you obtain a total of 5985 different subsets of 4 observations out of 21. If you do not specify *optn*[6], the LMS and LTS algorithms draw $N_{rep} = 2000$ random sample subsets. Since there is a large number of subsets with singular linear systems that you do not want to print, you can choose *optn*[2]=2 for reduced printed output.

```
title2 "***Use 2000 Random Subsets***";
      a = aa[,2:4]; b = aa[,5];
      optn = j(8,1,.);
      optn[2]= 2;    /* ipri */
      optn[3]= 3;    /* ilsq */
      optn[8]= 3;    /* icov */

call lms(sc,coef,wgt,optn,b,a);
```

Output 9.2.1. Some Simple Statistics

```
                    Median and Mean

                          Median            Mean
          VAR1       58.00000000     60.42857143
          VAR2       20.00000000     21.09523810
          VAR3       87.00000000     86.28571429
          Intercep    1.00000000      1.00000000
          Response   15.00000000     17.52380952


               Dispersion and Standard Deviation

                      Dispersion          StdDev
          VAR1        5.93040887      9.16826826
          VAR2        2.96520444      3.16077145
          VAR3        4.44780666      5.35857124
          Intercep    0.00000000      0.00000000
          Response    5.93040887     10.17162252
```

The following are the results of LS regression.

Output 9.2.2. Table of Unweighted LS Regression

```
                 ***********************************
                Unweighted Least-Squares Estimation
                ***********************************

                      LS Parameter Estimates
 -----------------------------------------------------------------
                       Approx                       Lower     Upper
           Estimate  Std Error  T Value   Prob   Wald CI   Wald CI
 -----------------------------------------------------------------
 VAR1        0.7156    0.13486   5.3066  58E-6    0.4513    0.9800
 VAR2        1.2953    0.36802   3.5196 0.0026    0.5740    2.0166
 VAR3       -0.1521    0.15629  -0.9733  0.344   -0.4585    0.1542
 Intercep  -39.9197   11.89600  -3.3557 0.0038  -63.2354  -16.6039

                    Sum of Squares = 178.8299616
                      Degrees of Freedom = 17
                  LS Scale Estimate = 3.2433639182


                 COV Matrix of Parameter Estimates

                    VAR1           VAR2           VAR3       Intercep
VAR1           0.0181867     -0.0365107     -0.0071435      0.2875871
VAR2          -0.0365107      0.1354419      0.0000105     -0.6517944
VAR3          -0.0071435      0.0000105      0.0244278     -1.6763208
Intercep       0.2875871     -0.6517944     -1.6763208    141.5147411

                      R-squared = 0.9135769045
                   F(3,17) Statistic = 59.9022259
                     Probability = 3.0163272E-9
```

The following are the LMS results for the 2000 random subsets.

Output 9.2.3. Iteration History and Optimization Results

```
 ***********************************************
***       Random Subsampling for LMS        ***
***********************************************

      Subset Singular        Best Crit   Pct
         500       23  0.1632616086096   25%
        1000       55 0.14051869795752   50%
        1500       79 0.14051869795752   75%
        2000      103 0.12646682816177  100%
        Minimum Criterion=0.1264668282

    ****************************************
    Least Median of Squares (LMS) Regression
    ****************************************

 Minimizing the 13th Ordered Squared Residual.
  Highest Possible Breakdown Value =  42.86 %
        Random Selection of 2103 Subsets
      Among 2103 subsets 103 are singular.


          Observations of Best Subset

        15            11            19            10


             Estimated Coefficients

     VAR1              VAR2              VAR3         Intercep
0.75000000        0.50000000        0.00000000     -39.25000000

         LMS Objective Function = 0.75
      Preliminary LMS Scale = 1.0478510755
         Robust R Squared = 0.96484375
    Final LMS Scale Estimate = 1.2076147288
```

For LMS, observations 1, 3, 4, and 21 have scaled residuals larger than 2.5 (output not shown), and they are considered outliers. The following are the corresponding WLS results.

Output 9.2.4. Table of Weighted LS Regression

```
                 *********************************
                 Weighted Least-Squares Estimation
                 *********************************

               RLS Parameter Estimates Based on LMS
 -----------------------------------------------------------------
                      Approx                           Lower     Upper
           Estimate  Std Error  T Value   Prob   Wald CI   Wald CI
 -----------------------------------------------------------------
 VAR1        0.7977    0.06744  11.8282  25E-9    0.6655    0.9299
 VAR2        0.5773    0.16597   3.4786 0.0041    0.2520    0.9026
 VAR3       -0.0671    0.06160  -1.0886  0.296   -0.1878    0.0537
 Intercep  -37.6525    4.73205  -7.9569 237E-8  -46.9271  -28.3778

               Weighted Sum of Squares = 20.400800254
                      Degrees of Freedom = 13
                 RLS Scale Estimate = 1.2527139846

                 COV Matrix of Parameter Estimates

                   VAR1           VAR2           VAR3       Intercep
VAR1         0.00454803    -0.00792141    -0.00119869     0.00156817
VAR2        -0.00792141     0.02754569    -0.00046339    -0.06501751
VAR3        -0.00119869    -0.00046339     0.00379495    -0.24610225
Intercep     0.00156817    -0.06501751    -0.24610225    22.39230535

                 Weighted R-squared = 0.9750062263
                  F(3,13) Statistic = 169.04317954
                     Probability = 1.158521E-10
              There are 17 points with nonzero weight.
                   Average Weight = 0.8095238095
```

The subroutine, *prilmts()*, which is in *robustmc.sas* file that is contained in the sample library, can be called to print the output summary:

```
call prilmts(3,sc,coef,wgt);
```

Output 9.2.5. First Part of Output Generated by prilmts()

```
             Results of Least Median Squares Estimation

              Quantile. . . . . . . . . . .        13
              Number of Subsets. . . . . .       2103
              Number of Singular Subsets .        103
              Number of Nonzero Weights. .         17
              Objective Function. . . . . .      0.75
              Preliminary Scale Estimate. . 1.0478511
              Final Scale Estimate. . . . . 1.2076147
              Robust R Squared. . . . . . . 0.9648438
              Asymptotic Consistency Factor 1.1413664
               RLS Scale Estimate. . . . .  1.252714
               Weighted Sum of Squares . .   20.4008
               Weighted R-squared. . . . . 0.9750062
               F Statistic . . . . . . . . 169.04318
```

Output 9.2.6. Second Part of Output Generated by prilmts()

```
                    Estimated LMS Coefficients
                  0.75       0.5         0    -39.25

                      Indices of Best Sample
                    15        11        19        10

                    Estimated WLS Coefficients
              0.7976856 0.5773405  -0.06706 -37.65246

                          Standard Errors
              0.0674391 0.1659689 0.0616031 4.7320509

                              T Values
              11.828242 3.4786054 -1.088584 -7.956901

                           Probabilities
              2.4838E-8  0.004078 0.2961071 2.3723E-6

                           Lower Wald CI
              0.6655074 0.2520473   -0.1878 -46.92711

                           Upper Wald CI
              0.9298637 0.9026336 0.0536798 -28.37781
```

Output 9.2.7. Third Part of Output Generated by prilmts()

```
                           LMS Residuals

     6.4176097 2.2772163   6.21059 7.2456884  -0.20702 -0.621059
   :  -0.20702  0.621059 -0.621059  0.621059  0.621059 0.2070197
   : -1.863177 -1.449138  0.621059  -0.20702 0.2070197 0.2070197
   :  0.621059  1.863177 -6.831649

                            Diagnostics
     10.448052 7.9317507        10 11.666667 2.7297297 3.4864865
   : 4.7297297 4.2432432 3.6486486 3.7598351 4.6057675 4.9251688
   : 3.8888889 4.5864209 5.2970297  4.009901  6.679576 4.3053404
   : 4.0199755         3        11


                           WLS Residuals
     4.9634454 0.9185794 5.1312163 6.5250478 -0.535877 -0.996749
   : -0.338162 0.4601047 -0.844485  0.286883 0.7686702 0.3777432
   : -2.000854 -1.074607 1.0731966 0.1143341 -0.297718 0.0770058
   : 0.4679328  1.544002 -6.888934
```

You now want to report the results of LTS for the 2000 random subsets:

```
title2 "***Use 2000 Random Subsets***";
   a = aa[,2:4]; b = aa[,5];
   optn = j(8,1,.);
   optn[2]= 2;    /* ipri */
   optn[3]= 3;    /* ilsq */
   optn[8]= 3;    /* icov */

call lts(sc,coef,wgt,optn,b,a);
```

Output 9.2.8. Iteration History and Optimization Results

```
   ***********************************************
   ***      Random Subsampling for LTS        ***
   ***********************************************


         Subset Singular        Best Crit   Pct
            500       23 0.09950690229748   25%
           1000       55 0.08781379221356   50%
           1500       79 0.08406140720682   75%
           2000      103 0.08406140720682  100%
           Minimum Criterion=0.0840614072

        **************************************
        Least Trimmed Squares (LTS) Regression
        **************************************

   Minimizing Sum of 13 Smallest Squared Residuals.
     Highest Possible Breakdown Value =  42.86 %
           Random Selection of 2103 Subsets
         Among 2103 subsets 103 are singular.

             Observations of Best Subset

           10            11             7            15

                Estimated Coefficients
      VAR1             VAR2             VAR3         Intercep

0.75000000       0.33333333       0.00000000     -35.70512821

        LTS Objective Function = 0.4985185153
         Preliminary LTS Scale = 1.0379336739
           Robust R Squared = 0.9719626168
       Final LTS Scale Estimate = 1.0407755737
```

In addition to observations 1, 3, 4, and 21, which were considered outliers in LMS, observation 2 for LTS has a scaled residual considerably larger than 2.5 (output not shown) and is considered an outlier. Therefore, the WLS results based on LTS are different from those based on LMS.

Output 9.2.9. Table of Weighted LS Regression

```
                  *********************************
                 Weighted Least-Squares Estimation
                 *********************************

               RLS Parameter Estimates Based on LTS
-----------------------------------------------------------------
                      Approx                     Lower     Upper
          Estimate  Std Error  T Value   Prob   Wald CI   Wald CI
-----------------------------------------------------------------
VAR1        0.7569    0.07861   9.6293 108E-8    0.6029    0.9110
VAR2        0.4535    0.13605   3.3335 0.0067    0.1869    0.7202
VAR3       -0.0521    0.05464  -0.9537  0.361   -0.1592    0.0550
Intercep  -34.0575    3.82882  -8.8950 235E-8  -41.5619  -26.5532

               Weighted Sum of Squares = 10.273044977
                      Degrees of Freedom = 11
                 RLS Scale Estimate = 0.9663918355

                 COV Matrix of Parameter Estimates

                   VAR1           VAR2           VAR3       Intercep
VAR1         0.00617916    -0.00577686    -0.00230059    -0.03429007
VAR2        -0.00577686     0.01850969     0.00025825    -0.06974088
VAR3        -0.00230059     0.00025825     0.00298523    -0.13148741
Intercep    -0.03429007    -0.06974088    -0.13148741    14.65985290

                 Weighted R-squared = 0.9622869127
                  F(3,11) Statistic = 93.558645037
                     Probability = 4.1136826E-8
              There are 15 points with nonzero weight.
                   Average Weight = 0.7142857143
```

Consider All 5985 Subsets

You now report the results of LMS for all different subsets:

```
title2 "*** Use All 5985 Subsets***";
   a = aa[,2:4]; b = aa[,5];
   optn = j(8,1,.);
   optn[2]= 2;    /* ipri */
   optn[3]= 3;    /* ilsq */
   optn[6]= -1;   /* nrep: all 5985 subsets */
   optn[8]= 3;    /* icov */

call lms(sc,coef,wgt,optn,b,a);
```

Output 9.2.10. Iteration History and Optimization Results for LMS

```
           ***********************************************
           ***      Complete Enumeration for LMS       ***
           ***********************************************

                 Subset Singular        Best Crit   Pct
                   1497       36 0.18589932664216   25%
                   2993       87 0.15826842822584   50%
                   4489      149 0.14051869795752   75%
                   5985      266 0.12646682816177  100%
                   Minimum Criterion=0.1264668282

               ****************************************
               Least Median of Squares (LMS) Regression
               ****************************************

            Minimizing the 13th Ordered Squared Residual.
             Highest Possible Breakdown Value =  42.86 %
          Selection of All 5985 Subsets of 4 Cases Out of 21
                 Among 5985 subsets 266 are singular.

                     Observations of Best Subset

                    8            10            15            19

                        Estimated Coefficients

              VAR1              VAR2              VAR3         Intercep
        0.75000000        0.50000000        0.00000000     -39.25000000

                    LMS Objective Function = 0.75
                  Preliminary LMS Scale = 1.0478510755
                     Robust R Squared = 0.96484375
               Final LMS Scale Estimate = 1.2076147288
```

Next, report the results of LTS for all different subsets:

```
title2 "*** Use All 5985 Subsets***";
   a = aa[,2:4]; b = aa[,5];
   optn = j(8,1,.);
   optn[2]= 2;    /* ipri */
   optn[3]= 3;    /* ilsq */
   optn[6]= -1;   /* nrep: all 5985 subsets */
   optn[8]= 3;    /* icov */

call lts(sc,coef,wgt,optn,b,a);
```

Output 9.2.11. Iteration History and Optimization Results for LTS

```
          ***********************************************
        ***      Complete Enumeration for LTS       ***
        ***********************************************

              Subset Singular        Best Crit   Pct
                1497       36 0.13544860556893   25%
                2993       87 0.10708384510403   50%
                4489      149 0.08153552986986   75%
                5985      266 0.08153552986986  100%
                Minimum Criterion=0.0815355299

             **************************************
             Least Trimmed Squares (LTS) Regression
             **************************************

        Minimizing Sum of 13 Smallest Squared Residuals.
          Highest Possible Breakdown Value =  42.86 %
       Selection of All 5985 Subsets of 4 Cases Out of 21
              Among 5985 subsets 266 are singular.

                  Observations of Best Subset

                 5            12            17            18

                     Estimated Coefficients

      VAR1              VAR2              VAR3          Intercep
0.72916667        0.41666667        0.00000000      -36.22115385

             LTS Objective Function = 0.4835390299
              Preliminary LTS Scale = 1.0067458407
                Robust R Squared = 0.9736222371
             Final LTS Scale Estimate = 1.009470149
```

Example 9.3. LMS and LTS Univariate (Location) Problem: Barnett and Lewis Data

If you do not specify matrix X of the last input argument, the regression problem is reduced to the estimation of the location parameter a. The following example is described in Rousseeuw and Leroy (1987, p. 175):

```
print "*** Barnett and Lewis (1978) ***";
   b = { 3, 4, 7, 8, 10, 949, 951 };

   optn = j(8,1,.);
   optn[2]= 3;    /* ipri */
   optn[3]= 3;    /* ilsq */
   optn[8]= 3;    /* icov */

call lms(sc,coef,wgt,optn,b);
```

First, show the results of unweighted LS regression.

Output 9.3.1. Table of Unweighted LS Regression

```
          Robust Estimation of Location and Scale
          ***************************************

            ***********************************
            Unweighted Least-Squares Estimation
            ***********************************

         Median = 8   MAD ( * 1.4826) = 5.930408874
        Mean = 276 Standard Deviation = 460.43602523

                       LS Residuals
     -------------------------------------------------
                Observed       Residual        Res / S
     -------------------------------------------------
        1       3.000000    -273.000000      -0.592916
        2       4.000000    -272.000000      -0.590744
        3       7.000000    -269.000000      -0.584229
        4       8.000000    -268.000000      -0.582057
        5      10.000000    -266.000000      -0.577713
        6     949.000000     673.000000       1.461658
        7     951.000000     675.000000       1.466002

                 Distribution of Residuals
  MinRes    1st Qu.     Median       Mean    3rd Qu.     MaxRes
    -273       -272       -268          0       -266        675
```

The output for LMS regression follows.

Output 9.3.2. Table of LMS Results

```
        ************************************
            Least Median of Squares (LMS) Method
            ************************************

          Minimizing 4th Ordered Squared Residual.
        Highest Possible Breakdown Value =  57.14 %
                LMS Objective Function = 2.5
                     LMS Location = 5.5
            Preliminary LMS Scale = 5.4137257125
               Final LMS Scale = 3.0516389039

                      LMS Residuals
     -------------------------------------------------
                Observed       Residual        Res / S
     -------------------------------------------------
        1       3.000000      -2.500000      -0.819232
        2       4.000000      -1.500000      -0.491539
        3       7.000000       1.500000       0.491539
        4       8.000000       2.500000       0.819232
        5      10.000000       4.500000       1.474617
        6     949.000000     943.500000     309.178127
        7     951.000000     945.500000     309.833512

                 Distribution of Residuals
  MinRes    1st Qu.     Median       Mean    3rd Qu.     MaxRes
    -2.5       -1.5        2.5      270.5        4.5      945.5
```

You obtain the LMS location estimate 6.5 compared with the mean 276 (which is the LS estimate of the location parameter) and the median 8. The scale estimate σ^* in the univariate problem is a resistant (high breakdown) estimator for the dispersion of the data (refer to Rousseeuw and Leroy 1987, p. 178).

For weighted LS regression, the last two observations are ignored (given zero weights).

Output 9.3.3. Table of Weighted LS Regression

```
                 *********************************
                 Weighted Least-Squares Estimation
                 *********************************

                        Weighted Mean = 6.4
            Weighted Standard Deviation = 2.8809720582
             There are 5 points with nonzero weight.
                  Average Weight = 0.7142857143

                        Weighted LS Residuals
----------------------------------------------------------------
           Observed       Residual        Res / S         Weight
----------------------------------------------------------------
   1       3.000000      -3.400000      -1.180157       1.000000
   2       4.000000      -2.400000      -0.833052       1.000000
   3       7.000000       0.600000       0.208263       1.000000
   4       8.000000       1.600000       0.555368       1.000000
   5      10.000000       3.600000       1.249578       1.000000
   6     949.000000     942.600000     327.181236              0
   7     951.000000     944.600000     327.875447              0

                     Distribution of Residuals
    MinRes    1st Qu.     Median       Mean    3rd Qu.     MaxRes
      -3.4       -2.4        1.6      269.6        3.6      944.6
```

```
      optn = j(8,1,.);
      optn[2]= 3;    /* ipri */
      optn[3]= 3;    /* ilsq */
      optn[8]= 3;    /* icov */

   call lts(sc,coef,wgt,optn,b);
```

The results for LTS are similar to those reported for LMS in Rousseeuw and Leroy (1987).

Output 9.3.4. Table of LTS Results

```
             *********************************
                 Least Trimmed Squares (LTS) Method
                 *********************************

          Minimizing Sum of 4 Smallest Squared Residuals.
            Highest Possible Breakdown Value =  57.14 %
               LTS Objective Function = 2.0615528128
                         LTS Location = 5.5
                Preliminary LTS Scale = 4.7050421234
                   Final LTS Scale = 3.0516389039

                           LTS Residuals
         -------------------------------------------------
                    Observed       Residual        Res / S
         -------------------------------------------------
            1       3.000000      -2.500000      -0.819232
            2       4.000000      -1.500000      -0.491539
            3       7.000000       1.500000       0.491539
            4       8.000000       2.500000       0.819232
            5      10.000000       4.500000       1.474617
            6     949.000000     943.500000     309.178127
            7     951.000000     945.500000     309.833512

                     Distribution of Residuals
      MinRes    1st Qu.     Median       Mean    3rd Qu.     MaxRes
        -2.5       -1.5        2.5      270.5        4.5      945.5
```

Since nonzero weights are chosen for the same observations as with LMS, the WLS results based on LTS agree with those based on LMS (shown previously).

In summary, you obtain the following estimates for the location parameter:

- LS estimate (unweighted mean) = 276
- Median = 8
- LMS estimate = 5.5
- LTS estimate = 5.5
- WLS estimate (weighted mean based on LMS or LTS) = 6.4

Examples Using MVE Regression

Example 9.4. Brainlog Data

The following data, consisting of the body weights (in kilograms) and brain weights (in grams) of $N = 28$ animals, are reported by Jerison (1973) and can be found also in Rousseeuw and Leroy (1987, p. 57). Instead of the original data, this example uses the logarithms of the measurements of the two variables.

```
title "*** Brainlog Data: Do MVE ***";
   aa={ 1.303338E-001  9.084851E-001 ,
```

```
      2.6674530         2.6263400  ,
      1.5602650         2.0773680  ,
      1.4418520         2.0606980  ,
   1.703332E-002  7.403627E-001    ,
      4.0681860         1.6989700  ,
      3.4060290         3.6630410  ,
      2.2720740         2.6222140  ,
      2.7168380         2.8162410  ,
      1.0000000         2.0606980  ,
   5.185139E-001         1.4082400 ,
      2.7234560         2.8325090  ,
      2.3159700         2.6085260  ,
      1.7923920         3.1205740  ,
      3.8230830         3.7567880  ,
      3.9731280         1.8450980  ,
   8.325089E-001         2.2528530 ,
      1.5440680         1.7481880  ,
  -9.208187E-001          .0000000 ,
     -1.6382720 -3.979400E-001     ,
   3.979400E-001         1.0827850 ,
      1.7442930         2.2430380  ,
      2.0000000         2.1959000  ,
      1.7173380         2.6434530  ,
      4.9395190         2.1889280  ,
  -5.528420E-001  2.787536E-001    ,
  -9.136401E-001  4.771213E-001    ,
      2.2833010         2.2552720  };
```

By default, the MVE subroutine (like the MINVOL subroutine) uses only 1500 randomly selected subsets rather than all subsets. The following specification of the options vector requires that all 3276 subsets of 3 cases out of 28 cases are generated and evaluated:

```
title2 "***MVE for BrainLog Data***";
title3 "*** Use All Subsets***";
   optn = j(8,1,.);
   optn[1]= 3;              /* ipri */
   optn[2]= 1;              /* pcov: print COV */
   optn[3]= 1;              /* pcor: print CORR */
   optn[6]= -1;             /* nrep: all subsets */
call mve(sc,xmve,dist,optn,aa);
```

Specifying *optn*[1]=3, *optn*[2]=1, and *optn*[3]=1 requests that all output be printed. Therefore, the first part of the output shows the classical scatter and correlation matrix.

Output 9.4.1. Some Simple Statistics

```
                 ****************************************
                 Minimum Volume Ellipsoid (MVE) Estimation
                 ****************************************

                  Consider Ellipsoids Containing 15 Cases.

                        Classical Covariance Matrix

                                    VAR1               VAR2
                  VAR1       2.681651236        1.330084693
                  VAR2       1.330084693        1.085753755


                        Classical Correlation Matrix

                                    VAR1               VAR2
                  VAR1       1.000000000        0.779493464
                  VAR2       0.779493464        1.000000000


                               Classical Mean

                             VAR1      1.637857
                             VAR2      1.921947
```

The second part of the output shows the results of the combinatoric optimization (complete subset sampling).

Output 9.4.2. Iteration History for MVE

```
              ***********************************************
              ***      Complete Enumeration for MVE       ***
              ***********************************************

                    Subset Singular        Best Crit   Pct
                       819        0 0.43970910597153   25%
                      1638        0 0.43970910597153   50%
                      2457        0 0.43970910597153   75%
                      3276        0 0.43970910597153  100%
                       Minimum Criterion=0.439709106
                     Among 3276 subsets 0 are singular.


                        Observations of Best Subset

                              1            22            28

                       Initial MVE Location Estimates

                           VAR1       1.385975933
                           VAR2       1.802265033

                         Initial MVE Scatter Matrix

                                    VAR1               VAR2
                  VAR1       4.901852512        3.293713910
                  VAR2       3.293713910        2.340065093
```

The third part of the output shows the optimization results after local improvement.

Output 9.4.3. Table of MVE Results

```
 *********************************************
Final MVE Estimates (Using Local Improvement)
*********************************************

   Number of Points with Nonzero Weight=24

        Robust MVE Location Estimates

            VAR1       1.295282380
            VAR2       1.873372279

          Robust MVE Scatter Matrix

                      VAR1              VAR2
   VAR1       2.056659294       1.529025017
   VAR2       1.529025017       1.204135359


     Eigenvalues of Robust Scatter Matrix


            VAR1       3.217727401
            VAR2       0.043067251

          Robust Correlation Matrix

                      VAR1              VAR2
   VAR1       1.000000000       0.971618466
   VAR2       0.971618466       1.000000000
```

The final output presents a table containing the classical Mahalanobis distances, the robust distances, and the weights identifying the outlier observations.

Output 9.4.4. Mahalanobis and Robust Distances

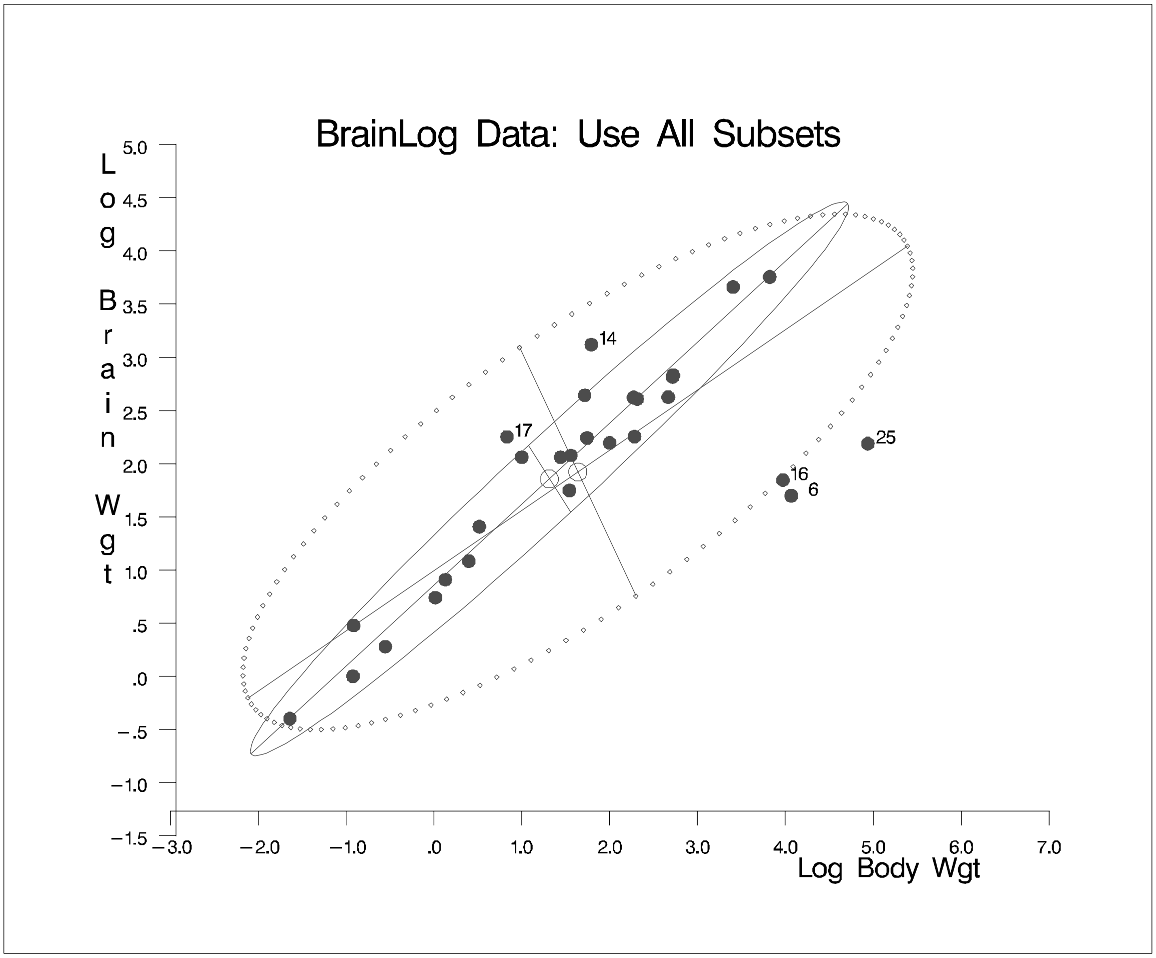
```
                   Classical and Robust Distances
         ------------------------------------------------
                Mahalanobis         Robust
                   Distance       Distance          Weight
         ------------------------------------------------
             1     1.006591       0.897076        1.000000
             2     0.695261       1.405302        1.000000
             3     0.300831       0.186726        1.000000
             4     0.380817       0.318701        1.000000
             5     1.146485       1.135697        1.000000
             6     2.644176       8.828036               0
             7     1.708334       1.699233        1.000000
             8     0.706522       0.686680        1.000000
             9     0.858404       1.084163        1.000000
            10     0.798698       1.580835        1.000000
            11     0.686485       0.693346        1.000000
            12     0.874349       1.071492        1.000000
            13     0.677791       0.717545        1.000000
            14     1.721526       3.398698               0
            15     1.761947       1.762703        1.000000
            16     2.369473       7.999472               0
            17     1.222253       2.805954               0
            18     0.203178       1.207332        1.000000
            19     1.855201       1.773317        1.000000
            20     2.266268       2.074971        1.000000
            21     0.831416       0.785954        1.000000
            22     0.416158       0.342200        1.000000
            23     0.264182       0.918383        1.000000
            24     1.046120       1.782334        1.000000
            25     2.911101       9.565443               0
            26     1.586458       1.543748        1.000000
            27     1.582124       1.808423        1.000000
            28     0.394664       1.523235        1.000000

                  Distribution of Robust Distances
     MinRes    1st Qu.     Median       Mean    3rd Qu.     MaxRes
 0.18672628 0.84151489 1.46426852 2.12846426 1.79537845 9.56544318

                    Cutoff Value = 2.7162030315
The cutoff value is the square root of the 0.975 quantile of the
        chi square distribution with 2 degrees of freedom
 There are 5 points with larger distances receiving zero weights.
                These may include boundary cases.
Only points whose robust distances are substantially larger than
         the cutoff value should be considered outliers.
```

Again, you can call the subroutine *scatmve()*, which is included in the sample library in file *robustmc.sas*, to plot the classical and robust confidence ellipsoids:

```
    optn = j(8,1,.); optn[6]= -1;
    vnam = { "Log Body Wgt","Log Brain Wgt" };
    filn = "brl";
    titl = "BrainLog Data: Use All Subsets";
 call scatmve(2,optn,.9,aa,vnam,titl,1,filn);
```

The output follows.

Output 9.4.5. BrainLog Data: Classical and Robust Ellipsoid



Example 9.5. MVE: Stackloss Data

This example analyzes the three regressors of Brownlee's (1965) stackloss data. By default, the MVE subroutine, like the MINVOL subroutine, tries only 2000 randomly selected subsets in its search. There are, in total, 5985 subsets of 4 cases out of 21 cases.

```
title2 "***MVE for Stackloss Data***";
title3 "*** Use All Subsets***";
   a = aa[,2:4];
   optn = j(8,1,.);
   optn[1]= 2;            /* ipri */
   optn[2]= 1;            /* pcov: print COV */
```

```
      optn[3]= 1;              /* pcor: print CORR */
      optn[6]= -1;             /* nrep: use all subsets */
   call mve(sc,xmve,dist,optn,a);
```

The first part of the output shows the classical scatter and correlation matrix.

Output 9.5.1. Some Simple Statistics

```
          *****************************************
          Minimum Volume Ellipsoid (MVE) Estimation
          *****************************************

           Consider Ellipsoids Containing 12 Cases.

                 Classical Covariance Matrix

                    VAR1              VAR2              VAR3
VAR1         84.05714286       22.65714286       24.57142857
VAR2         22.65714286        9.99047619        6.62142857
VAR3         24.57142857        6.62142857       28.71428571


                 Classical Correlation Matrix

                    VAR1              VAR2              VAR3
VAR1         1.000000000       0.781852333       0.500142875
VAR2         0.781852333       1.000000000       0.390939538
VAR3         0.500142875       0.390939538       1.000000000


                        Classical Mean

                      VAR1      60.42857
                      VAR2      21.09524
                      VAR3      86.28571
```

The second part of the output shows the results of the optimization (complete subset sampling).

Output 9.5.2. Iteration History

```
          ***********************************************
          ***      Complete Enumeration for MVE      ***
          ***********************************************

                Subset Singular        Best Crit   Pct
                  1497       29 253.312430606991   25%
                  2993       64 224.084073229268   50%
                  4489      114  165.83005346003   75%
                  5985      208  165.63436283899  100%
                  Minimum Criterion=165.63436284
                Among 5985 subsets 208 are singular.


                    Observations of Best Subset

                   7            10            14            20


                   Initial MVE Location Estimates

                       VAR1       58.50000000
                       VAR2       20.25000000
                       VAR3       87.00000000


                     Initial MVE Scatter Matrix

                       VAR1              VAR2              VAR3
     VAR1       34.8290147        28.4131436        62.3256053
     VAR2       28.4131436        38.0369503        58.6593933
     VAR3       62.3256053        58.6593933       267.6334818
```

The third part of the output shows the optimization results after local improvement.

Output 9.5.3. Table of MVE Results

```
           *********************************************
           Final MVE Estimates (Using Local Improvement)
           *********************************************

              Number of Points with Nonzero Weight=17

                   Robust MVE Location Estimates

                       VAR1       56.70588235
                       VAR2       20.23529412
                       VAR3       85.52941176

                     Robust MVE Scatter Matrix

                       VAR1              VAR2              VAR3
     VAR1       23.47058824        7.57352941       16.10294118
     VAR2        7.57352941        6.31617647        5.36764706
     VAR3       16.10294118        5.36764706       32.38970588


                Eigenvalues of Robust Scatter Matrix

                       VAR1       46.59743102
                       VAR2       12.15593848
                       VAR3        3.42310109


                     Robust Correlation Matrix

                       VAR1              VAR2              VAR3
     VAR1       1.000000000       0.622026950       0.584036133
     VAR2       0.622026950       1.000000000       0.375278187
     VAR3       0.584036133       0.375278187       1.000000000
```

The final output presents a table containing the classical Mahalanobis distances, the robust distances, and the weights identifying the outlying observations (that is, the leverage points when explaining y with these three regressor variables).

Output 9.5.4. Mahalanobis and Robust Distances

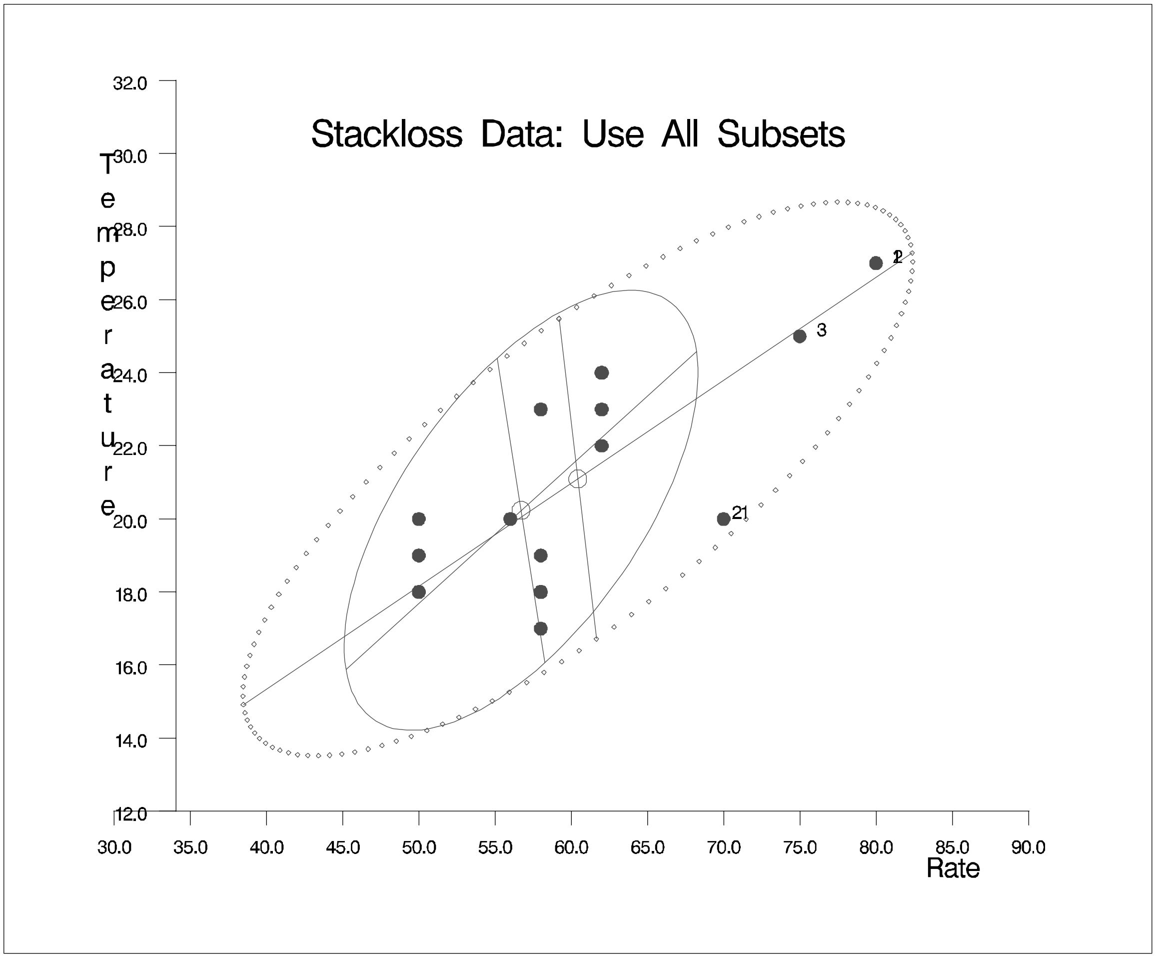
```
                 Classical and Robust Distances
         ------------------------------------------------
                Mahalanobis        Robust
                   Distance       Distance         Weight
         ------------------------------------------------
            1      2.253603       5.528395              0
            2      2.324745       5.637357              0
            3      1.593712       4.197235              0
            4      1.271898       1.588734       1.000000
            5      0.303357       1.189335       1.000000
            6      0.772895       1.308038       1.000000
            7      1.852661       1.715924       1.000000
            8      1.852661       1.715924       1.000000
            9      1.360622       1.226680       1.000000
           10      1.745997       1.936256       1.000000
           11      1.465702       1.493509       1.000000
           12      1.841504       1.913079       1.000000
           13      1.482649       1.659943       1.000000
           14      1.778785       1.689210       1.000000
           15      1.690241       2.230109       1.000000
           16      1.291934       1.767582       1.000000
           17      2.700016       2.431021       1.000000
           18      1.503155       1.523316       1.000000
           19      1.593221       1.710165       1.000000
           20      0.807054       0.675124       1.000000
           21      2.176761       3.657281              0

               Distribution of Robust Distances
    MinRes    1st Qu.     Median       Mean    3rd Qu.     MaxRes
 0.6751245 1.50841208 1.71592421 2.22829602 2.08318267 5.63735735

                 Cutoff Value = 3.0575159206
The cutoff value is the square root of the 0.975 quantile of the
        chi square distribution with 3 degrees of freedom
 There are 4 points with larger distances receiving zero weights.
                These may include boundary cases.
Only points whose robust distances are substantially larger than
         the cutoff value should be considered outliers.
```

The following specification generates three bivariate plots of the classical and robust tolerance ellipsoids, one plot for each pair of variables:

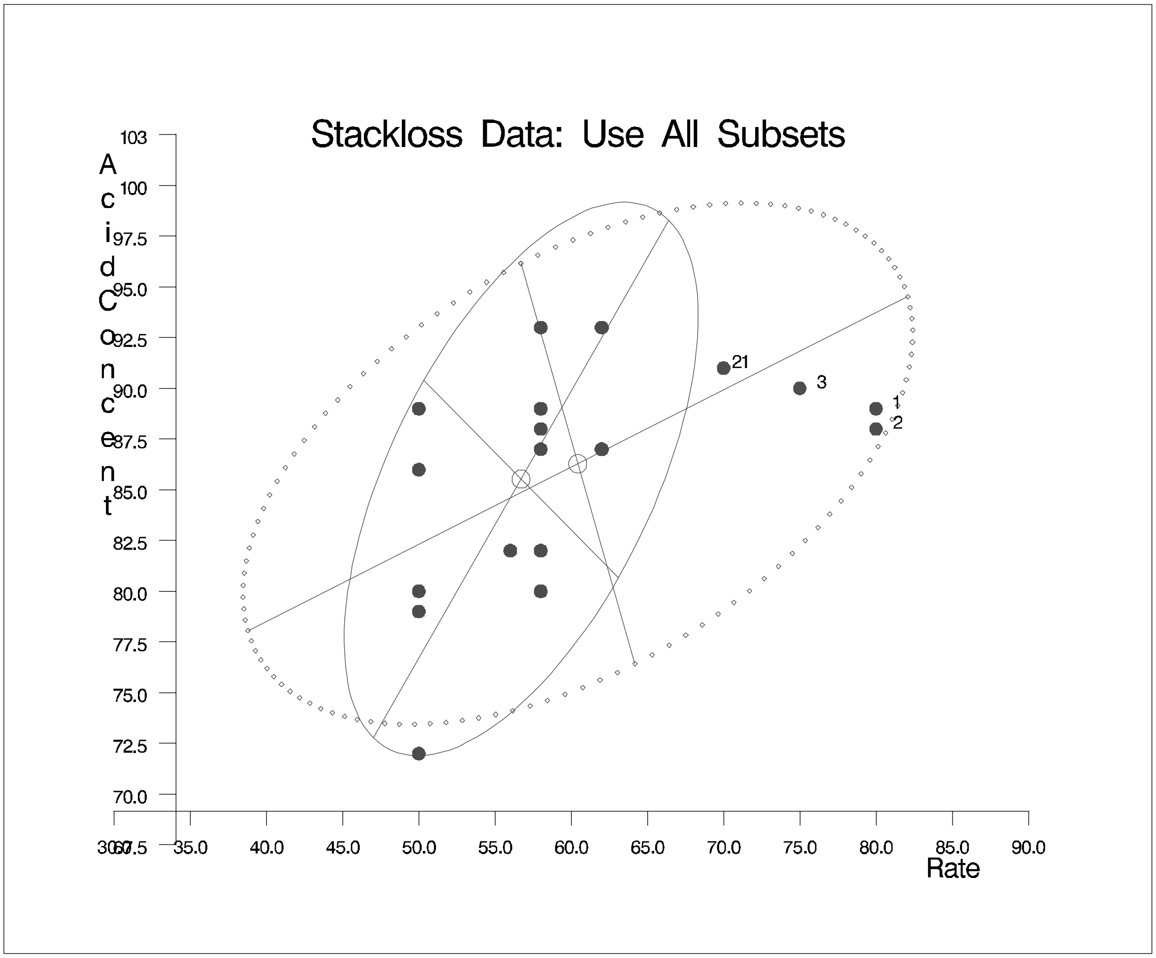
```
    optn = j(8,1,.); optn[6]= -1;
    vnam = { "Rate", "Temperature", "AcidConcent" };
    filn = "stl";
    titl = "Stackloss Data: Use All Subsets";
call scatmve(2,optn,.9,a,vnam,titl,1,filn);
```

The output follows.

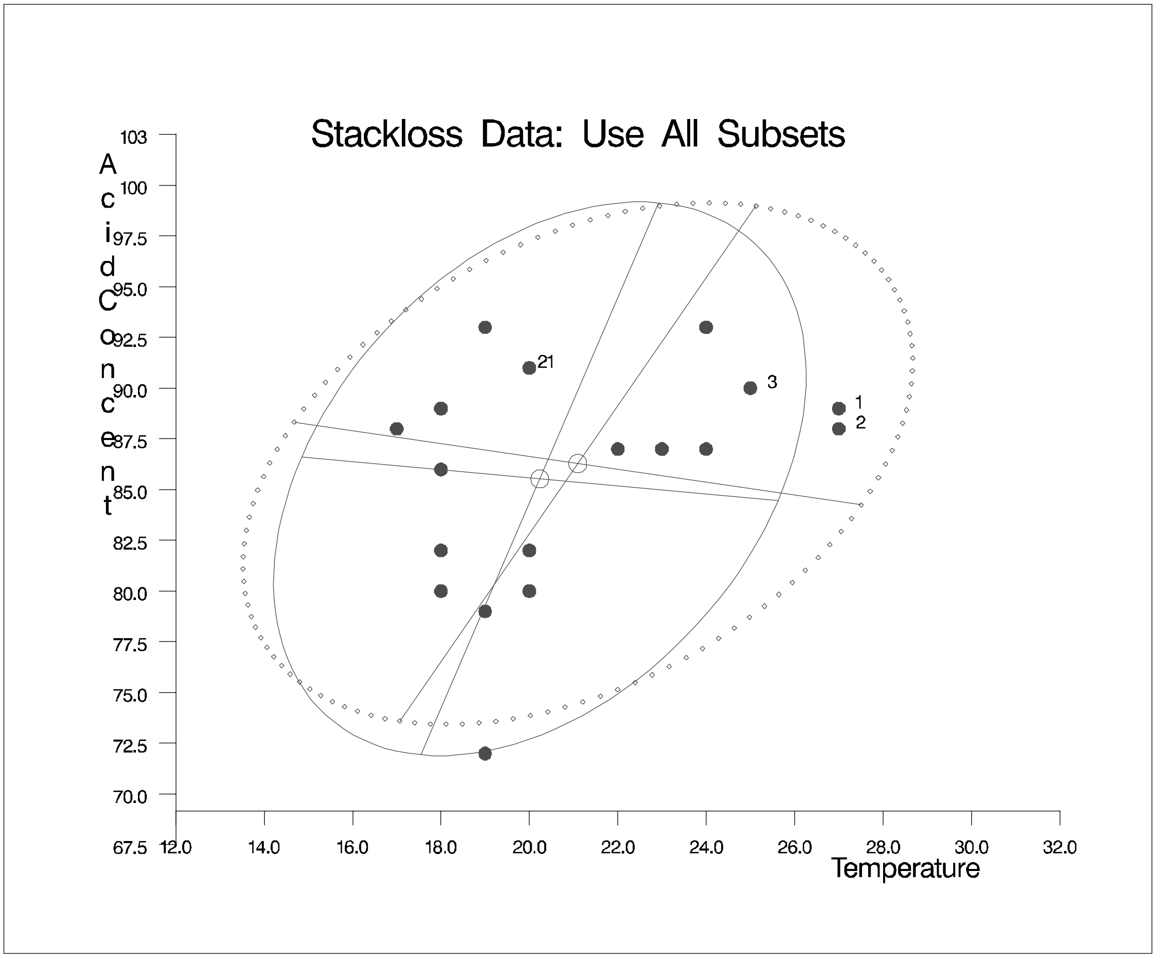
Output 9.5.5. Stackloss Data: Rate vs. Temperature



Output 9.5.6. Stackloss Data: Rate vs. Acid Concent



Output 9.5.7. Stackloss Data: Temperature vs. Acid Concent



Examples Combining Robust Residuals and Robust Distances

This section is based entirely on Rousseeuw and Van Zomeren (1990). Observations $\mathbf{x}_i$, which are far away from most of the other observations, are called *leverage points*. One classical method inspects the Mahalanobis distances MD_i to find outliers $\mathbf{x}_i$:

$$MD_i = \sqrt{(\mathbf{x}_i - \mu)\mathbf{C}^{-1}(\mathbf{x}_i - \mu)^T}$$

where $\mathbf{C}$ is the classical sample covariance matrix.

Note that the MVE subroutine prints the classical Mahalanobis distances MD_i together with the robust distances RD_i. In classical linear regression, the diagonal elements h_{ii} of the *hat* matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

are used to identify leverage points. Rousseeuw and Van Zomeren (1990) report the following monotone relationship between the h_{ii} and MD_i

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

and point out that neither the MD_i nor the h_{ii} are entirely safe for detecting leverage points reliably. Multiple outliers do not necessarily have large MD_i values because of the *masking effect*.

The definition of a *leverage point* is, therefore, based entirely on the outlyingness of $\mathbf{x}_i$ and is not related to the response value y_i. By including the y_i value in the definition, Rousseeuw and Van Zomeren (1990) distinguish between the following:

- *Good leverage points* are points $(\mathbf{x}_i, y_i)$ that are close to the regression plane; that is, good leverage points improve the precision of the regression coefficients.
- *Bad leverage points* are points $(\mathbf{x}_i, y_i)$ that are far from the regression plane; that is, bad leverage points reduce the precision of the regression coefficients.

Rousseeuw and Van Zomeren (1990) propose to plot the standardized residuals of robust regression (LMS or LTS) versus the robust distances RD_i obtained from MVE. Two horizontal lines corresponding to residual values of $+2.5$ and -2.5 are useful to distinguish between small and large residuals, and one vertical line corresponding to the $\sqrt{\chi^2_{n,.975}}$ is used to distinguish between small and large distances.

Example 9.6. Hawkins-Bradu-Kass Data

The first 14 observations of this data set (refer to Hawkins, Bradu, and Kass 1984) are leverage points; however, only observations 12, 13, and 14 have large h_{ii} and only observations 12 and 14 have large MD_i values.

```
title "Hawkins, Bradu, Kass (1984) Data";
   aa = { 1  10.1  19.6  28.3   9.7,
          2   9.5  20.5  28.9  10.1,
          3  10.7  20.2  31.0  10.3,
          4   9.9  21.5  31.7   9.5,
          5  10.3  21.1  31.1  10.0,
          6  10.8  20.4  29.2  10.0,
          7  10.5  20.9  29.1  10.8,
          8   9.9  19.6  28.8  10.3,
```

```
 9   9.7  20.7  31.0   9.6,
10   9.3  19.7  30.3   9.9,
11  11.0  24.0  35.0  -0.2,
12  12.0  23.0  37.0  -0.4,
13  12.0  26.0  34.0   0.7,
14  11.0  34.0  34.0   0.1,
15   3.4   2.9   2.1  -0.4,
16   3.1   2.2   0.3   0.6,
17   0.0   1.6   0.2  -0.2,
18   2.3   1.6   2.0   0.0,
19   0.8   2.9   1.6   0.1,
20   3.1   3.4   2.2   0.4,
21   2.6   2.2   1.9   0.9,
22   0.4   3.2   1.9   0.3,
23   2.0   2.3   0.8  -0.8,
24   1.3   2.3   0.5   0.7,
25   1.0   0.0   0.4  -0.3,
26   0.9   3.3   2.5  -0.8,
27   3.3   2.5   2.9  -0.7,
28   1.8   0.8   2.0   0.3,
29   1.2   0.9   0.8   0.3,
30   1.2   0.7   3.4  -0.3,
31   3.1   1.4   1.0   0.0,
32   0.5   2.4   0.3  -0.4,
33   1.5   3.1   1.5  -0.6,
34   0.4   0.0   0.7  -0.7,
35   3.1   2.4   3.0   0.3,
36   1.1   2.2   2.7  -1.0,
37   0.1   3.0   2.6  -0.6,
38   1.5   1.2   0.2   0.9,
39   2.1   0.0   1.2  -0.7,
40   0.5   2.0   1.2  -0.5,
41   3.4   1.6   2.9  -0.1,
42   0.3   1.0   2.7  -0.7,
43   0.1   3.3   0.9   0.6,
44   1.8   0.5   3.2  -0.7,
45   1.9   0.1   0.6  -0.5,
46   1.8   0.5   3.0  -0.4,
47   3.0   0.1   0.8  -0.9,
48   3.1   1.6   3.0   0.1,
49   3.1   2.5   1.9   0.9,
50   2.1   2.8   2.9  -0.4,
51   2.3   1.5   0.4   0.7,
52   3.3   0.6   1.2  -0.5,
53   0.3   0.4   3.3   0.7,
54   1.1   3.0   0.3   0.7,
55   0.5   2.4   0.9   0.0,

56   1.8   3.2   0.9   0.1,
57   1.8   0.7   0.7   0.7,
58   2.4   3.4   1.5  -0.1,
59   1.6   2.1   3.0  -0.3,
60   0.3   1.5   3.3  -0.9,
61   0.4   3.4   3.0  -0.3,
```

```
            62   0.9   0.1   0.3   0.6,
            63   1.1   2.7   0.2  -0.3,
            64   2.8   3.0   2.9  -0.5,
            65   2.0   0.7   2.7   0.6,
            66   0.2   1.8   0.8  -0.9,
            67   1.6   2.0   1.2  -0.7,
            68   0.1   0.0   1.1   0.6,
            69   2.0   0.6   0.3   0.2,
            70   1.0   2.2   2.9   0.7,
            71   2.2   2.5   2.3   0.2,
            72   0.6   2.0   1.5  -0.2,
            73   0.3   1.7   2.2   0.4,
            74   0.0   2.2   1.6  -0.9,
            75   0.3   0.4   2.6   0.2 };

     a = aa[,2:4]; b = aa[,5];
```

The data are listed also in Rousseeuw and Leroy (1987, p. 94).

The complete enumeration must inspect 1,215,450 subsets.

Output 9.6.1. Iteration History for MVE

```
 ***********************************************
***      Complete Enumeration for MVE       ***
***********************************************

      Subset Singular        Best Crit   Pct
      121545        0 51.1042755960104   10%
      243090        2 51.1042755960104   20%
      364635        4 51.1042755960104   30%
      486180        7 51.1042755960104   40%
      607725        9 51.1042755960104   50%
      729270       22 6.27172477029496   60%
      850815       67 6.27172477029496   70%
      972360      104 5.91230765636768   80%
     1093905      135 5.91230765636768   90%
     1215450      185 5.91230765636768  100%
        Minimum Criterion=5.9123076564
    Among 1215450 subsets 185 are singular.
```

The following output reports the robust parameter estimates for MVE.

Output 9.6.2. Robust Location Estimates

```
                 Robust MVE Location Estimates

                    VAR1       1.513333333
                    VAR2       1.808333333
                    VAR3       1.701666667

                  Robust MVE Scatter Matrix

                    VAR1              VAR2              VAR3
     VAR1       1.114395480       0.093954802       0.141672316
     VAR2       0.093954802       1.123149718       0.117443503
     VAR3       0.141672316       0.117443503       1.074742938
```

Output 9.6.3. MVE Scatter Matrix

```
             Eigenvalues of Robust Scatter Matrix

                    VAR1       1.339637154
                    VAR2       1.028124757
                    VAR3       0.944526224

                  Robust Correlation Matrix

                    VAR1              VAR2              VAR3
     VAR1       1.000000000       0.083980892       0.129453270
     VAR2       0.083980892       1.000000000       0.106895118
     VAR3       0.129453270       0.106895118       1.000000000
```

Output 9.6.4 shows the classical Mahalanobis and robust distances obtained by complete enumeration. The first 14 observations are recognized as outliers (leverage points).

Output 9.6.4. Mahalanobis and Robust Distances

Classical and Robust Distances

	Mahalanobis Distance	Robust Distance	Weight
1	1.916821	29.541649	0
2	1.855757	30.344481	0
3	2.313658	31.985694	0
4	2.229655	33.011768	0
5	2.100114	32.404938	0
6	2.146169	30.683153	0
7	2.010511	30.794838	0
8	1.919277	29.905756	0
9	2.221249	32.092048	0
10	2.333543	31.072200	0
11	2.446542	36.808021	0
12	3.108335	38.071382	0
13	2.662380	37.094539	0
14	6.381624	41.472255	0
15	1.815487	1.994672	1.000000
16	2.151357	2.202278	1.000000
17	1.384915	1.918208	1.000000
18	0.848155	0.819163	1.000000
19	1.148941	1.288387	1.000000
20	1.591431	2.046703	1.000000
21	1.089981	1.068327	1.000000
22	1.548776	1.768905	1.000000
23	1.085421	1.166951	1.000000
24	0.971195	1.304648	1.000000
25	0.799268	2.030417	1.000000
26	1.168373	1.727131	1.000000
27	1.449625	1.983831	1.000000
28	0.867789	1.073856	1.000000
29	0.576399	1.168060	1.000000
30	1.568868	2.091386	1.000000

Output 9.6.4. (continued)

Classical and Robust Distances

	Mahalanobis Distance	Robust Distance	Weight
31	1.838496	1.793386	1.000000
32	1.307230	1.743558	1.000000
33	0.981988	1.264121	1.000000
34	1.175014	2.052641	1.000000
35	1.243636	1.872695	1.000000
36	0.850804	1.136658	1.000000
37	1.832378	2.050041	1.000000
38	0.752061	1.522734	1.000000
39	1.265041	1.885970	1.000000
40	1.112038	1.068841	1.000000
41	1.699757	2.063398	1.000000
42	1.765040	1.785637	1.000000
43	1.870090	2.166100	1.000000
44	1.420448	2.018610	1.000000
45	1.075973	1.944449	1.000000
46	1.344171	1.872483	1.000000
47	1.966328	2.408721	1.000000
48	1.424238	1.892539	1.000000
49	1.569756	1.594109	1.000000
50	0.423972	1.458595	1.000000
51	1.302651	1.569843	1.000000
51	1.302651	1.569843	1.000000
52	2.076055	2.205601	1.000000
53	2.210443	2.492631	1.000000
54	1.414288	1.884937	1.000000
55	1.230455	1.360622	1.000000
56	1.331101	1.626276	1.000000
57	0.832744	1.432408	1.000000
58	1.404401	1.723091	1.000000
59	0.591235	1.263700	1.000000
60	1.889737	2.087849	1.000000

Output 9.6.4. (continued)

```
                   Classical and Robust Distances
         -------------------------------------------------
                Mahalanobis        Robust
                   Distance      Distance         Weight
         -------------------------------------------------
           61      1.674945      2.286045       1.000000
           62      0.759533      2.024702       1.000000
           63      1.292259      1.783035       1.000000
           64      0.973868      1.835207       1.000000
           65      1.148208      1.562278       1.000000
           66      1.296746      1.444491       1.000000
           67      0.629827      0.552899       1.000000
           68      1.549548      2.101580       1.000000
           69      1.070511      1.827919       1.000000
           70      0.997761      1.354151       1.000000
           71      0.642927      0.988770       1.000000
           72      1.053395      0.908316       1.000000
           73      1.472178      1.314779       1.000000
           74      1.646461      1.516083       1.000000
           75      1.899178      2.042560       1.000000

                 Distribution of Robust Distances
     MinRes    1st Qu.     Median       Mean    3rd Qu.     MaxRes
 0.55289874 1.44449066 1.88493749 7.56960939 2.16610046 41.4722551

                   Cutoff Value = 3.0575159206
The cutoff value is the square root of the 0.975 quantile of the
        chi square distribution with 3 degrees of freedom
There are 14 points with larger distances receiving zero weights.
                These may include boundary cases.
Only points whose robust distances are substantially larger than
         the cutoff value should be considered outliers.
```

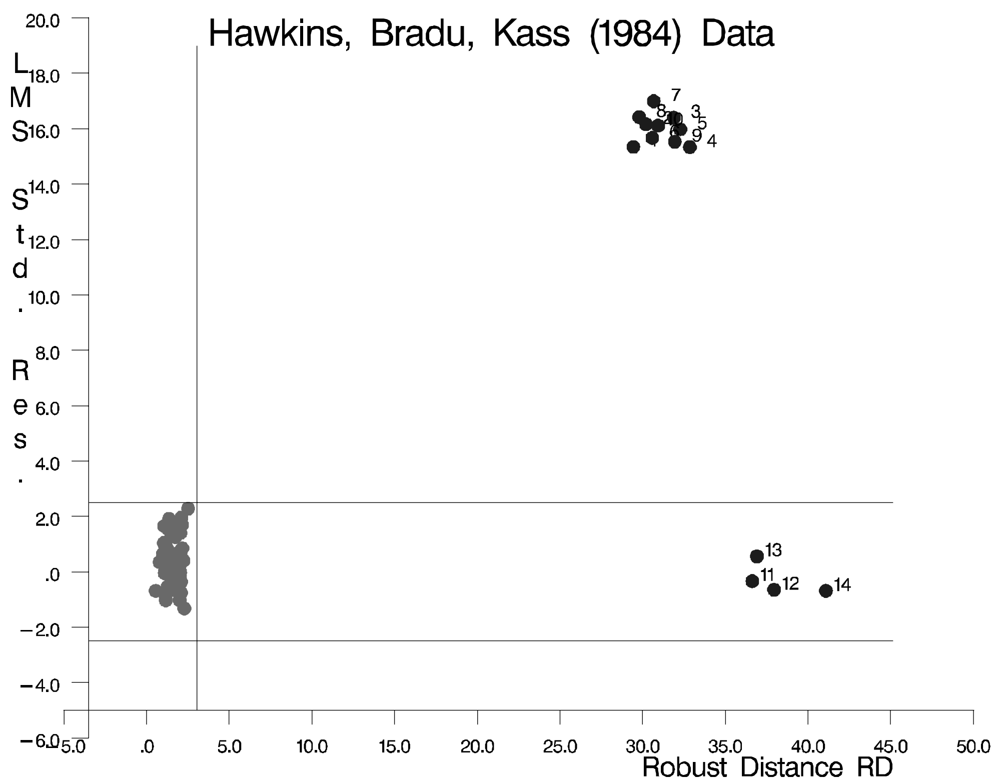
The following two graphs show

- the plot of standardized LMS residuals vs. robust distances RD_i
- the plot of standardized LS residuals vs. Mahalanobis distances MD_i

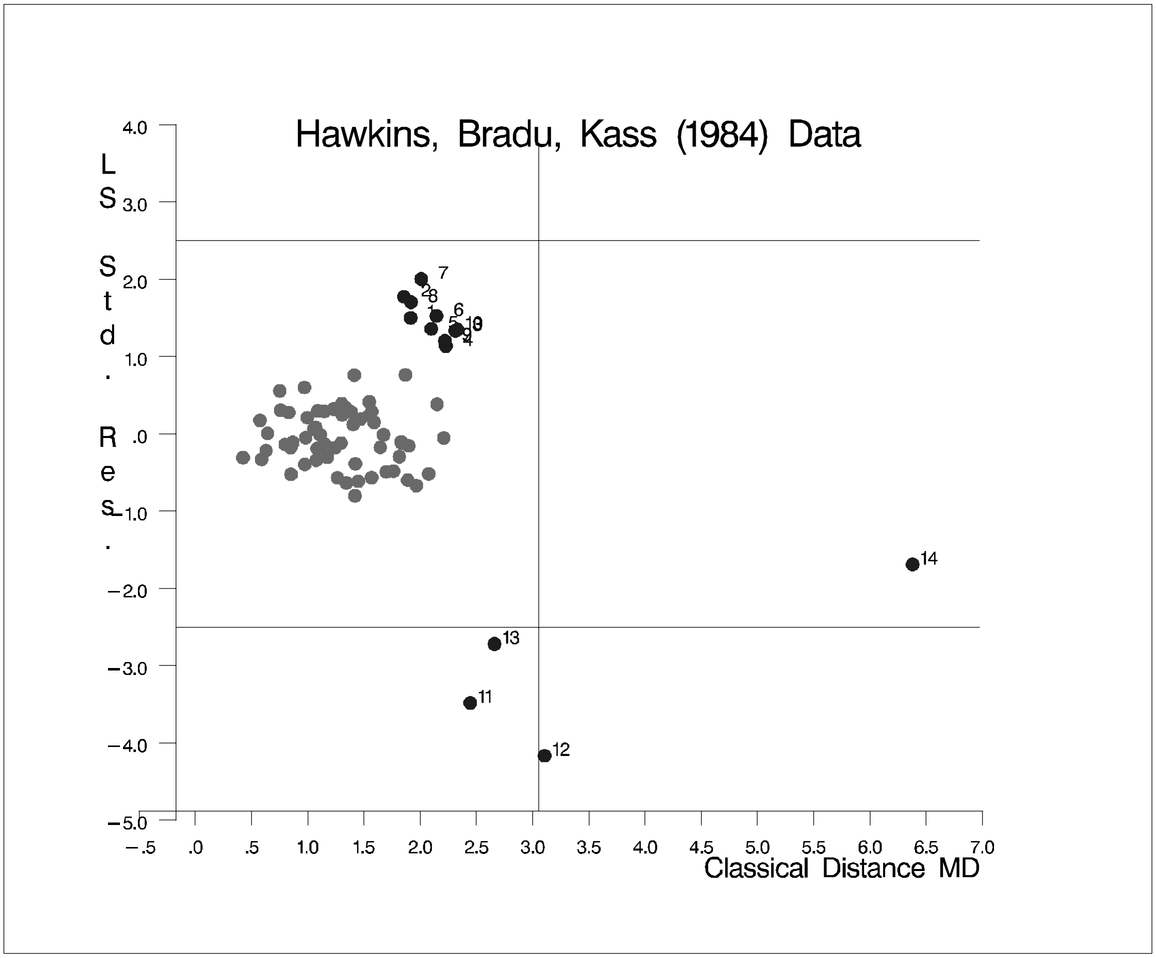
The graph identifies the four good leverage points 11, 12, 13, and 14, which have small standardized LMS residuals but large robust distances, and the 10 bad leverage points $1, \ldots, 10$, which have large standardized LMS residuals and large robust distances.

The output follows.

Output 9.6.5. Hawkins-Bradu-Kass Data: LMS Residuals vs. Robust Distances



Output 9.6.6. Hawkins-Bradu-Kass Data: LS Residuals vs. Mahalanobis Distances



Example 9.7. Stackloss Data

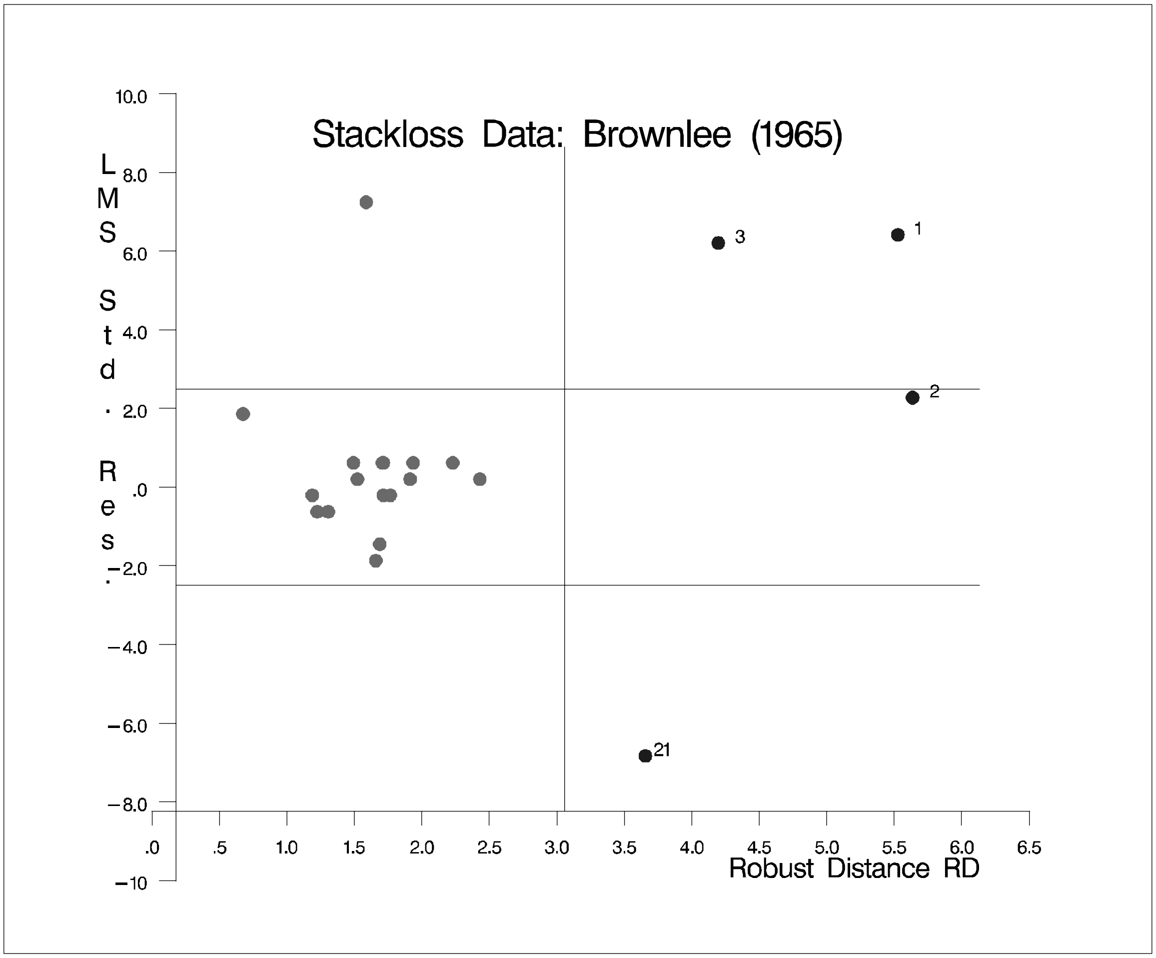
The following two graphs show

- the plot of standardized LMS residuals vs. robust distances RD_i
- the plot of standardized LS residuals vs. Mahalanobis distances MD_i

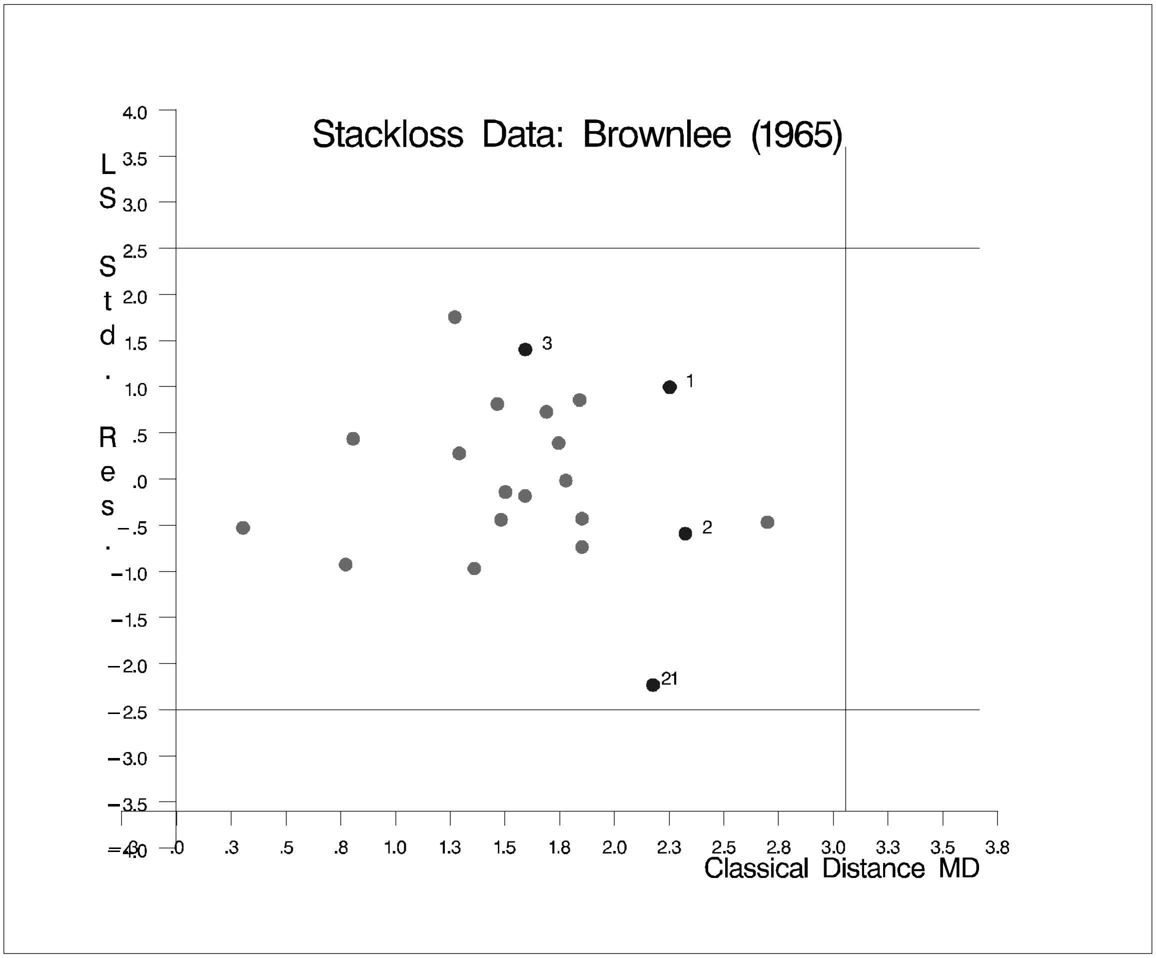
In the first plot, you see that case 4 is a regression outlier but not a leverage point, so it is a vertical outlier. Cases 1, 3, and 21 are bad leverage points, whereas case 2 is a good leverage point. Note that case 21 lies near the boundary line between vertical outliers and bad leverage points and that case 2 is very close to the boundary between good and bad leverage points.

The output follows.

Output 9.7.1. Stackloss Data: LMS Residuals vs. Robust Distances



Output 9.7.2. Stackloss Data: LS Residuals vs. Mahalanobis Distances



References

Afifi, A.A. and Azen, S.P. (1979), *Statistical Analysis, A Computer Oriented Approach*, New York: Academic Press.

Barnett, V. and Lewis, T. (1978), *Outlier in Statistical Data*, New York: John Wiley & Sons, Inc.

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

Ezekiel, M. and Fox, K.A. (1959), *Methods of Correlation and Regression Analysis*, 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.

Jerison, H. J. (1973), *Evolution of the Brain and Intelligence*, New York: Academic Press.

Osborne, M.R. (1985), *Finite Algorithms in Optimization and Data Analysis*, New York: John Wiley & Sons, Inc.

Prescott, P. (1975), "An Approximate Test for Outliers in Linear Models," *Technometrics*, 17, 129–132.

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

Rousseeuw, P.J. (1985), "Multivariate Estimation with High Breakdown Point," in *Mathematical Statistics and Applications*, Dordrecht: Reidel Publishing Company, 283–297.

Rousseeuw, P.J. and Hubert, M. (1996), "Recent Developments in PROGRESS," Technical Report, University of Antwerp.

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

Rousseeuw, P.J. and Van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633 –639.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/IML User's Guide, Version 8,* Cary, NC: SAS Institute Inc., 1999. 846 pp.

SAS/IML User's Guide, Version 8

ISBN 1–58025–553–1

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

1st printing, October 1999