NAG Toolbox

nag_univar_robust_1var_mestim (g07db)

1 Purpose

nag_univar_robust_1var_mestim (g07db) computes an M-estimate of location with (optional) simultaneous estimation of the scale using Huber's algorithm.

2 Syntax

```
[theta, sigma, rs, nit, wrk, ifail] = nag_univar_robust_1var_mestim(isigma, x, ipsi, c, h1, h2, h3, dchi, theta, sigma, tol, 'n', n, 'maxit', maxit)
[theta, sigma, rs, nit, wrk, ifail] = g07db(isigma, x, ipsi, c, h1, h2, h3, dchi, theta, sigma, tol, 'n', n, 'maxit', maxit)
```

3 Description

The data consists of a sample of size n, denoted by $x_1, x_2, \ldots, x_n$, drawn from a random variable X. The x_i are assumed to be independent with an unknown distribution function of the form

$$F((x_i - \theta)/\sigma)$$

where θ is a location argument, and σ is a scale argument. M-estimators of θ and σ are given by the solution to the following system of equations:

$$\sum_{i=1}^{n} \psi\left(\left(x_i - \hat{\theta}\right)/\hat{\sigma}\right) = 0 \tag{1}$$

$$\sum_{i=1}^{n} \chi\left(\left(x_i - \hat{\theta}\right)/\hat{\sigma}\right) = (n-1)\beta \tag{2}$$

where ψ and χ are given functions, and β is a constant, such that $\hat{\sigma}$ is an unbiased estimator when x_i, for $i = 1, 2, \ldots, n$ has a Normal distribution. Optionally, the second equation can be omitted and the first equation is solved for $\hat{\theta}$ using an assigned value of $\sigma = \sigma_c$.

The values of $\psi\left(\dfrac{x_i - \hat{\theta}}{\hat{\sigma}}\right)\hat{\sigma}$ are known as the Winsorized residuals.

The following functions are available for ψ and χ in nag_univar_robust_1var_mestim (g07db).

(a) **Null Weights**

$$\psi(t) = t \qquad \chi(t) = \frac{t^2}{2}$$

Use of these null functions leads to the mean and standard deviation of the data.

(b) **Huber's Function**

$$\psi(t) = \max(-c, \min(c, t)) \qquad \chi(t) = \frac{\|t\|^2}{2} \|t\| \leq d$$

$$\chi(t) = \frac{d^2}{2} \|t\| > d$$

(c) **Hampel's Piecewise Linear Function**

$$\psi_{h_1,h_2,h_3}(t) = -\psi_{h_1,h_2,h_3}(-t)$$

$$\psi_{h_1,h_2,h_3}(t) = t \qquad 0 \le t \le h_1 \qquad \chi(t) = \frac{|t|^2}{2} |t| \le d$$

$$\psi_{h_1,h_2,h_3}(t) = h_1 \qquad h_1 \le t \le h_2$$

$$\psi_{h_1,h_2,h_3}(t) = h_1(h_3 - t)/(h_3 - h_2) \qquad h_2 \le t \le h_3 \qquad \chi(t) = \frac{d^2}{2} |t| > d$$

$$\psi_{h_1,h_2,h_3}(t) = 0 \qquad t > h_3$$

(d) **Andrew's Sine Wave Function**

$$\psi(t) = \sin t \qquad -\pi \le t \le \pi \qquad \chi(t) = \frac{|t|^2}{2} |t| \le d$$

$$\psi(t) = 0 \qquad \text{otherwise} \qquad \chi(t) = \frac{d^2}{2} |t| > d$$

(e) **Tukey's Bi-weight**

$$\psi(t) = t\left(1 - t^2\right)^2 \qquad |t| \le 1 \qquad \chi(t) = \frac{|t|^2}{2} |t| \le d$$

$$\psi(t) = t\left(1 - t^2\right)^2 = 0 \qquad \text{otherwise} \qquad \chi(t) = \frac{d^2}{2} |t| > d$$

where c, h_1, h_2, h_3 and d are constants.

Equations (1) and (2) are solved by a simple iterative procedure suggested by Huber:

$$\hat{\sigma}_k = \sqrt{\frac{1}{\beta(n-1)} \left(\sum_{i=1}^{n} \chi\left(\frac{x_i - \hat{\theta}_{k-1}}{\hat{\sigma}_{k-1}} \right) \right) \hat{\sigma}_{k-1}^2}$$

and

$$\hat{\theta}_k = \hat{\theta}_{k-1} + \tfrac{1}{n} \sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}_{k-1}}{\hat{\sigma}_k} \right) \hat{\sigma}_k$$

or

$$\hat{\sigma}_k = \sigma_c, \qquad \text{if} \quad \sigma \quad \text{is fixed.}$$

The initial values for $\hat{\theta}$ and $\hat{\sigma}$ may either be user-supplied or calculated within nag_univar_robust_1var_mestim (g07db) as the sample median and an estimate of σ based on the median absolute deviation respectively.

nag_univar_robust_1var_mestim (g07db) is based upon function LYHALG within the ROBETH library, see Marazzi (1987).

4 References

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

Huber P J (1981) *Robust Statistics* Wiley

Marazzi A (1987) Subroutines for robust estimation of location and scale in ROBETH *Cah. Rech. Doc. IUMSP, No. 3 ROB 1* Institut Universitaire de Médecine Sociale et Préventive, Lausanne

5 Parameters

5.1 Compulsory Input Parameters

1: **isigma** – INTEGER

The value assigned to **isigma** determines whether $\hat{\sigma}$ is to be simultaneously estimated.

isigma $= 0$
The estimation of $\hat{\sigma}$ is bypassed and **sigma** is set equal to σ_c.

isigma $= 1$
$\hat{\sigma}$ is estimated simultaneously.

2: **x**(**n**) – REAL (KIND=nag_wp) array

The vector of observations, $x_1, x_2, \ldots, x_n$.

3: **ipsi** – INTEGER

Which ψ function is to be used.

ipsi $= 0$
$\psi(t) = t$.

ipsi $= 1$
Huber's function.

ipsi $= 2$
Hampel's piecewise linear function.

ipsi $= 3$
Andrew's sine wave,

ipsi $= 4$
Tukey's bi-weight.

4: **c** – REAL (KIND=nag_wp)

If **ipsi** $= 1$, **c** must specify the argument, c, of Huber's ψ function. **c** is not referenced if **ipsi** $\neq 1$.

Constraint: if **ipsi** $= 1$, **c** > 0.0.

5: **h1** – REAL (KIND=nag_wp)
6: **h2** – REAL (KIND=nag_wp)
7: **h3** – REAL (KIND=nag_wp)

If **ipsi** $= 2$, **h1**, **h2** and **h3** must specify the arguments, h_1, h_2, and h_3, of Hampel's piecewise linear ψ function. **h1**, **h2** and **h3** are not referenced if **ipsi** $\neq 2$.

Constraint: $0 \leq$ **h1** $\leq$ **h2** $\leq$ **h3** and **h3** > 0.0 if **ipsi** $= 2$.

8: **dchi** – REAL (KIND=nag_wp)

d, the argument of the χ function. **dchi** is not referenced if **ipsi** $= 0$.

Constraint: if **ipsi** $\neq 0$, **dchi** > 0.0.

9: **theta** – REAL (KIND=nag_wp)

If **sigma** > 0 then **theta** must be set to the required starting value of the estimation of the location argument $\hat{\theta}$. A reasonable initial value for $\hat{\theta}$ will often be the sample mean or median.

10: **sigma** – REAL (KIND=nag_wp)

The role of **sigma** depends on the value assigned to **isigma**, as follows:

if **isigma** $= 1$, **sigma** must be assigned a value which determines the values of the starting points for the calculations of $\hat{\theta}$ and $\hat{\sigma}$. If **sigma** ≤ 0.0 then nag_univar_robust_1var_mestim (g07db) will determine the starting points of $\hat{\theta}$ and $\hat{\sigma}$. Otherwise the value assigned to **sigma** will be taken as the starting point for $\hat{\sigma}$, and **theta** must be assigned a value before entry, see above;

if **isigma** $= 0$, **sigma** must be assigned a value which determines the value of σ_c, which is held fixed during the iterations, and the starting value for the calculation of $\hat{\theta}$. If **sigma** ≤ 0, then nag_univar_robust_1var_mestim (g07db) will determine the value of σ_c as the median absolute deviation adjusted to reduce bias (see nag_univar_robust_1var_median (g07da)) and the starting point for $\hat{\theta}$. Otherwise, the value assigned to **sigma** will be taken as the value of σ_c and **theta** must be assigned a relevant value before entry, see above.

11: **tol** – REAL (KIND=nag_wp)

The relative precision for the final estimates. Convergence is assumed when the increments for **theta**, and **sigma** are less than **tol** $\times \max(1.0, \sigma_{k-1})$.

Constraint: **tol** > 0.0.

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the array **x**.

n, the number of observations.

Constraint: **n** > 1.

2: **maxit** – INTEGER

Suggested value: **maxit** $= 50$.

Default: 50

The maximum number of iterations that should be used during the estimation.

Constraint: **maxit** > 0.

5.3 Output Parameters

1: **theta** – REAL (KIND=nag_wp)

The M-estimate of the location argument, $\hat{\theta}$.

2: **sigma** – REAL (KIND=nag_wp)

Contains the M-estimate of the scale argument, $\hat{\sigma}$, if **isigma** was assigned the value 1 on entry, otherwise **sigma** will contain the initial fixed value σ_c.

3: **rs**(**n**) – REAL (KIND=nag_wp) array

The Winsorized residuals.

4: **nit** – INTEGER

The number of iterations that were used during the estimation.

5: **wrk**(**n**) – REAL (KIND=nag_wp) array

If **sigma** ≤ 0.0 on entry, **wrk** will contain the n observations in ascending order.

6: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, **n** ≤ 1,
or **maxit** ≤ 0,
or **tol** ≤ 0.0,
or **isigma** $\neq 0$ or 1,
or **ipsi** < 0,
or **ipsi** > 4.

ifail $= 2$

On entry, **c** ≤ 0.0 and **ipsi** $= 1$,
or **h1** < 0.0 and **ipsi** $= 2$,
or **h1** $=$ **h2** $=$ **h3** $= 0.0$ and **ipsi** $= 2$,
or **h1** $>$ **h2** and **ipsi** $= 2$,
or **h1** $>$ **h3** and **ipsi** $= 2$,
or **h2** $>$ **h3** and **ipsi** $= 2$,
or **dchi** ≤ 0.0 and **ipsi** $\neq 0$.

ifail $= 3$

On entry, all elements of the input array **x** are equal.

ifail $= 4$

sigma, the current estimate of σ, is zero or negative. This error exit is very unlikely, although it may be caused by too large an initial value of **sigma**.

ifail $= 5$

The number of iterations required exceeds **maxit**.

ifail $= 6$

On completion of the iterations, the Winsorized residuals were all zero. This may occur when using the **isigma** $= 0$ option with a redescending ψ function, i.e., Hampel's piecewise linear function, Andrew's sine wave, and Tukey's biweight.

If the given value of σ is too small, then the standardized residuals $\dfrac{x_i - \hat{\theta}_k}{\sigma_c}$, will be large and all the residuals may fall into the region for which $\psi(t) = 0$. This may incorrectly terminate the iterations thus making **theta** and **sigma** invalid.

Re-enter the function with a larger value of σ_c or with **isigma** $= 1$.

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

On successful exit the accuracy of the results is related to the value of **tol**, see Section 5.

8 Further Comments

When you supply the initial values, care has to be taken over the choice of the initial value of σ. If too small a value of σ is chosen then initial values of the standardized residuals $\frac{x_i - \hat{\theta}_k}{\sigma}$ will be large. If the redescending ψ functions are used, i.e., Hampel's piecewise linear function, Andrew's sine wave, or Tukey's bi-weight, then these large values of the standardized residuals are Winsorized as zero. If a sufficient number of the residuals fall into this category then a false solution may be returned, see page 152 of Hampel *et al.* (1986).

9 Example

The following program reads in a set of data consisting of eleven observations of a variable X.

For this example, Hampel's Piecewise Linear Function is used (**ipsi** $= 2$), values for h_1, h_2 and h_3 along with d for the χ function, being read from the data file.

Using the following starting values various estimates of θ and σ are calculated and printed along with the number of iterations used:

(a) nag_univar_robust_1var_mestim (g07db) determines the starting values, σ is estimated simultaneously.

(b) You must supply the starting values, σ is estimated simultaneously.

(c) nag_univar_robust_1var_mestim (g07db) determines the starting values, σ is fixed.

(d) You must supply the starting values, σ is fixed.

9.1 Program Text

```
function g07db_example

fprintf('g07db example results\n\n');

x = [13; 11; 16;  5;  3; 18;  9;  8;  6; 27;  7];

ipsi = nag_int(2);
c  = 0;
h1 = 1.5;
h2 = 3;
h3 = 4.5;
dchi = 1.5;
tol = 0.0001;

% Loop over input values for isigma sigma and theta
isigma = nag_int([ 1  1  0  0]);
sigma  =         [-1  7 -1  7];
theta  =         [ 0  2  0  2];

fprintf('           Input parameters     Output parameters\n');
fprintf(' isigma   sigma   theta   tol    sigma  theta\n');
for j = 1:numel(theta)

  fprintf('%3d   %8.4f%8.4f%8.4f', isigma(j), sigma(j), theta(j), tol);

  [thetaOut, sigmaOut, rs, nit, wrk, ifail] = ...
  g07db( ...
         isigma(j), x, ipsi, c, h1, h2, h3, dchi, theta(j), sigma(j), tol);

  fprintf(' %8.4f%8.4f\n', sigmaOut, thetaOut);
end
```

9.2 Program Results

```
     g07db example results

           Input parameters     Output parameters
 isigma   sigma   theta   tol    sigma  theta
  1    -1.0000  0.0000  0.0001   6.3247 10.5487
  1     7.0000  2.0000  0.0001   6.3249 10.5487
  0    -1.0000  0.0000  0.0001   5.9304 10.4896
  0     7.0000  2.0000  0.0001   7.0000 10.6500
```