NAG Fortran Library Routine Document

G07DBF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G07DBF computes an M-estimate of location with (optional) simultaneous estimation of the scale using Huber's algorithm.

2 Specification

```
      SUBROUTINE G07DBF(ISIGMA, N, X, IPSI, C, H1, H2, H3, DCHI, THETA, SIGMA,
     1                  MAXIT, TOL, RS, NIT, WRK, IFAIL)
      INTEGER           ISIGMA, N, IPSI, MAXIT, NIT, IFAIL
      real              X(N), C, H1, H2, H3, DCHI, THETA, SIGMA, TOL, RS(N),
     1                  WRK(N)
```

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

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

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

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

The following functions are available for ψ and χ in G07DBF.

(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**

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

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

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

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

$\psi_{h_1,h_2,h_3}(t) = t$	$0 \leq t \leq h_1$	$\chi(t) = \dfrac{\lvert t\rvert^2}{2}\ \lvert t\rvert \leq d$
$\psi_{h_1,h_2,h_3}(t) = h_1$	$h_1 \leq t \leq h_2$	
$\psi_{h_1,h_2,h_3}(t) = h_1(h_3 - t)/(h_3 - h_2)$	$h_2 \leq t \leq h_3$	$\chi(t) = \dfrac{d^2}{2}\ \lvert t\rvert > d$
$\psi_{h_1,h_2,h_3}(t) = 0$	$t > h_3$	

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

$\psi(t) = \sin t$	$-\pi \leq t \leq \pi$	$\chi(t) = \dfrac{\lvert t\rvert^2}{2}\ \lvert t\rvert \leq d$
$\psi(t) = 0$	otherwise	$\chi(t) = \dfrac{d^2}{2}\ \lvert t\rvert > d$

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

$\psi(t) = t(1 - t^2)^2$	$\lvert t\rvert \leq 1$	$\chi(t) = \dfrac{\lvert t\rvert^2}{2}\ \lvert t\rvert \leq d$
$\psi(t) = t(1 - t^2)^2 = 0$	otherwise	$\chi(t) = \dfrac{d^2}{2}\ \lvert t\rvert > d$

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

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

The initial values for $\hat{\theta}$ and $\hat{\sigma}$ may either be user-supplied or calculated within G07DBF as the sample median and an estimate of σ based on the median absolute deviation respectively.

G07DBF is based upon subroutine 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

1: ISIGMA – INTEGER *Input*

On entry: 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: N – INTEGER *Input*

On entry: the number of observations, n.

Constraint: N > 1.

3: X(N) – ***real*** array *Input*

On entry: the vector of observations, $x_1, x_2, \ldots, x_n$.

4: IPSI – INTEGER *Input*

On entry: 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.

5: C – ***real*** *Workspace*

If IPSI = 1 on entry, C must specify the parameter, c, of Huber's ψ function. C is not referenced if IPSI ≠ 1.

Constraint: C > 0.0 if IPSI = 1.

6: H1 – ***real*** *Input*
7: H2 – ***real*** *Input*
8: H3 – ***real*** *Input*

If IPSI = 2 on entry, H1, H2, and H3 must specify the parameters, h_1, h_2, and h_3, of Hampel's piecewise linear ψ function. H1, H2, and H3 are not referenced if IPSI ≠ 2.

Constraint: 0 ≤ H1 ≤ H2 ≤ H3 and H3 > 0.0 if IPSI = 2.

9: DCHI – ***real*** *Input*

On entry: the parameter, d, of the χ function. DCHI is not referenced if IPSI = 0.

Constraint: DCHI > 0.0 if IPSI ≠ 0.

10: THETA – ***real*** *Input/Output*

On entry: if SIGMA > 0 then THETA must be set to the required starting value of the estimation of the location parameter $\hat{\theta}$. A reasonable initial value for $\hat{\theta}$ will often be the sample mean or median.

On exit: the M-estimate of the location parameter, $\hat{\theta}$.

11: SIGMA – ***real*** *Input/Output*

On entry: the role of SIGMA depends on the value assigned to ISIGMA (see above) 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 G07DBF 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 G07DBF will determine the value of σ_c as the median absolute deviation adjusted to reduce bias (see G07DAF) 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.

On exit: SIGMA contains the M-estimate of the scale parameter, $\hat{\sigma}$, if ISIGMA was assigned the value 1 on entry, otherwise SIGMA will contain the initial fixed value σ_c.

12: MAXIT – INTEGER *Input*

On entry: the maximum number of iterations that should be used during the estimation.

Suggested value: MAXIT = 50.

Constraint: MAXIT > 0.

13: TOL – ***real*** *Input*

On entry: 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.

14: RS(N) – ***real*** array *Output*

On exit: the Winsorized residuals.

15: NIT – INTEGER *Output*

On exit: the number of iterations that were used during the estimation.

16: WRK(N) – ***real*** array *Output*

On exit: if SIGMA ≤ 0.0 on entry, WRK will contain the n observations in ascending order.

17: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. Users who are unfamiliar with this parameter should refer to Chapter P01 for details.

On exit: IFAIL = 0 unless the routine detects an error (see Section 6).

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, for users not familiar with this parameter the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

6 Error Indicators and Warnings

If on entry IFAIL = 0 or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, $N \leq 1$,
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 standardised residuals $\frac{x_i - \hat{\theta}_k}{\sigma_c}$, 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 routine with a larger value of σ_c or with ISIGMA $= 1$.

7 Accuracy

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

8 Further Comments

When the user supplies 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 standardised 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, Hampels'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) G07DBF determines the starting values, σ is estimated simultaneously.

(b) The user supplies the starting values, σ is estimated simultaneously.

(c) G07DBF determines the starting values, σ is fixed.

(d) The user supplies the starting values, σ is fixed.

9.1 Program Text

Note: the listing of the example program presented below uses ***bold italicised*** terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```
*     G07DBF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX
      PARAMETER        (NMAX=25)
*     .. Local Scalars ..
      real             C, DCHI, H1, H2, H3, SIGMA, SIGSAV, THESAV,
     +                 THETA, TOL
      INTEGER          I, IFAIL, IPSI, ISIGMA, MAXIT, N, NIT
*     .. Local Arrays ..
      real             RS(NMAX), WRK(NMAX), X(NMAX)
*     .. External Subroutines ..
      EXTERNAL         G07DBF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G07DBF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N
      WRITE (NOUT,*)
      IF (N.LE.NMAX) THEN
         READ (NIN,*) (X(I),I=1,N)
         READ (NIN,*) IPSI, H1, H2, H3, DCHI, MAXIT
         WRITE (NOUT,*)
     +     '          Input parameters     Output parameters'
         WRITE (NOUT,*) 'ISIGMA   SIGMA   THETA   TOL    SIGMA  THETA'
   20    READ (NIN,*,END=40) ISIGMA, SIGMA, THETA, TOL
         SIGSAV = SIGMA
         THESAV = THETA
         IFAIL = 0
*
         CALL G07DBF(ISIGMA,N,X,IPSI,C,H1,H2,H3,DCHI,THETA,SIGMA,MAXIT,
     +               TOL,RS,NIT,WRK,IFAIL)
*
         WRITE (NOUT,99999) ISIGMA, SIGSAV, THESAV, TOL, SIGMA, THETA
         GO TO 20
      ELSE
         WRITE (NOUT,99998) 'N is out of range: N =', N
      END IF
   40 STOP
*
99999 FORMAT (1X,I3,3X,2F8.4,F7.4,F9.4,F8.4,I4)
99998 FORMAT (1X,A,I5)
      END
```

9.2 Program Data

```
G07DBF Example Program Data
11                                : NUMBER OF OBSERVATIONS
13.0 11.0 16.0 5.0 3.0 18.0 9.0 8.0 6.0 27.0 7.0  : OBSERVATIONS
 2    1.5  3.0 4.5  1.5  50       :IPSI  H1   H2   H3  DCHI  MAXIT
  1     -1.0    0.0   0.0001      :ISIGMA  SIGMA  THETA   TOL
  1      7.0    2.0   0.0001
  0     -1.0    0.0   0.0001
  0      7.0    2.0   0.0001
```

9.3 Program Results

```
 G07DBF Example Program 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
```