NAG Fortran Library Routine Document

G08RAF

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

G08RAF calculates the parameter estimates, score statistics and their variance-covariance matrices for the linear model using a likelihood based on the ranks of the observations.

2 Specification

```
      SUBROUTINE G08RAF(NS, NV, NSUM, Y, IP, X, NX, IDIST, NMAX, TOL, PARVAR,
     1                  NPVAR, IRANK, ZIN, ETA, VAPVEC, PAREST, WORK, LWORK,
     2                  IWA, IFAIL)
      INTEGER           NS, NV(NS), NSUM, IP, NX, IDIST, NMAX, NPVAR,
     1                  IRANK(NMAX), LWORK, IWA(NMAX), IFAIL
      real              Y(NSUM), X(NX,IP), TOL, PARVAR(NPVAR,IP), ZIN(NMAX),
     1                  ETA(NMAX), VAPVEC(NMAX*(NMAX+1)/2), PAREST(4*IP+1),
     2                  WORK(LWORK)
```

3 Description

Analysis of data can be made by replacing observations by their ranks. The analysis produces inference for regression parameters arising from the following model.

For random variables $Y_1, Y_2, \ldots, Y_n$ we assume that, after an arbitrary monotone increasing differentiable transformation, $h(.)$, the model

$$h(Y_i) = x_i^T \beta + \epsilon_i \tag{1}$$

holds, where x_i is a known vector of explanatory variables and β is a vector of p unknown regression coefficients. The ϵ_i are random variables assumed to be independent and identically distributed with a completely known distribution which can be one of the following: Normal, logistic, extreme value or double-exponential. In Pettitt (1982) an estimate for β is proposed as $\hat{\beta} = MX^T a$ with estimated variance-covariance matrix M. The statistics a and M depend on the ranks r_i of the observations Y_i and the density chosen for ϵ_i.

The matrix X is the n by p matrix of explanatory variables. It is assumed that X is of rank p and that a column or a linear combination of columns of X is not equal to the column vector of 1 or a multiple of it. This means that a constant term cannot be included in the model (1). The statistics a and M are found as follows. Let ϵ_i have pdf $f(\epsilon)$ and let $g = -f'/f$. Let $W_1, W_2, \ldots, W_n$ be order statistics for a random sample of size n with the density $f(.)$. Define $Z_i = g(W_i)$, then $a_i = E(Z_{r_i})$. To define M we need $M^{-1} = X^T(B - A)X$, where B is an n by n diagonal matrix with $B_{ii} = E(g'(W_{r_i}))$ and A is a symmetric matrix with $A_{ij} = \text{cov}(Z_{r_i}, Z_{r_j})$. In the case of the Normal distribution, the $Z_1 < \cdots < Z_n$ are standard Normal order statistics and $E(g'(W_i)) = 1$, for $i = 1, 2, \ldots, n$.

The analysis can also deal with ties in the data. Two observations are adjudged to be tied if $|Y_i - Y_j| <$ TOL, where TOL is a user-supplied tolerance level.

Various statistics can be found from the analysis:

(a) The score statistic $X^T a$. This statistic is used to test the hypothesis $H_0 : \beta = 0$, see (e).

(b) The estimated variance-covariance matrix $X^T(B - A)X$ of the score statistic in (a).

(c) The estimate $\hat{\beta} = MX^T a$.

(d) The estimated variance-covariance matrix $M = (X^T(B - A)X)^{-1}$ of the estimate $\hat{\beta}$.

(e) The χ^2 statistic $Q = \hat{\beta}^T M^{-1} \hat{\beta} = a^T X(X^T(B-A)X)^{-1} X^T a$ used to test $H_0 : \beta = 0$. Under H_0, Q has an approximate χ^2 distribution with p degrees of freedom.

(f) The standard errors $M_{ii}^{1/2}$ of the estimates given in (c).

(g) Approximate z-statistics, i.e., $Z_i = \hat{\beta}_i / se(\hat{\beta}_i)$ for testing $H_0 : \beta_i = 0$. For $i = 1, 2, \ldots, n$, Z_i has an approximate $N(0,1)$ distribution.

In many situations, more than one sample of observations will be available. In this case we assume the model

$$h_k(Y_k) = X_k^T \beta + e_k, \quad k = 1, 2, \ldots, \mathrm{NS},$$

where NS is the number of samples. In an obvious manner, Y_k and X_k are the vector of observations and the design matrix for the kth sample respectively. Note that the arbitrary transformation h_k can be assumed different for each sample since observations are ranked within the sample.

The earlier analysis can be extended to give a combined estimate of β as $\hat{\beta} = Dd$, where

$$D^{-1} = \sum_{k=1}^{\mathrm{NS}} X_k^T (B_k - A_k) X_k$$

and

$$d = \sum_{k=1}^{\mathrm{NS}} X_k^T a_k,$$

with a_k, B_k and A_k defined as a, B and A above but for the kth sample.

The remaining statistics are calculated as for the one sample case.

4 References

Pettitt A N (1982) Inference for the linear model using a likelihood based on ranks *J. Roy. Statist. Soc. Ser. B* **44** 234–243

5 Parameters

1: NS – INTEGER *Input*

On entry: the number of samples.

Constraint: NS ≥ 1.

2: NV(NS) – INTEGER array *Input/Output*

On entry: the number of observations in the ith sample, for $i = 1, 2, \ldots, \mathrm{NS}$.

On exit: used as internal workspace prior to being restored and hence is unchanged.

Constraint: NV(i) ≥ 1, for $i = 1, 2, \ldots, \mathrm{NS}$.

3: NSUM – INTEGER *Input*

On entry: the total number of observations.

Constraint: $\mathrm{NSUM} = \sum_{i=1}^{\mathrm{NS}} \mathrm{NV}(i)$.

4: Y(NSUM) – ***real*** array *Input*

On entry: the observations in each sample. Specifically, $\mathrm{Y}\left(\sum_{k=1}^{i-1} \mathrm{NV}(k) + j\right)$ must contain the jth observation in the ith sample.

5: IP – INTEGER *Input*

On entry: the number of parameters to be fitted.

Constraint: $\text{IP} \geq 1$.

6: X(NX,IP) – ***real*** array *Input*

On entry: the design matrices for each sample. Specifically, $\mathrm{X}\left(\sum_{k=1}^{i-1}\mathrm{NV}(k)+j,l\right)$ must contain the value of the lth explanatory variable for the jth observation in the ith sample.

Constraint: X must not contain a column with all elements equal.

7: NX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G08RAF is called.

Constraint: $\text{NX} \geq \text{NSUM}$.

8: IDIST – INTEGER *Input*

On entry: the error distribution to be used in the analysis as follows:

$\text{IDIST} = 1$

Normal.

$\text{IDIST} = 2$

Logistic.

$\text{IDIST} = 3$

Extreme value.

$\text{IDIST} = 4$

Double-exponential.

Constraint: IDIST = 1, 2, 3 or 4.

9: NMAX – INTEGER *Input*

On entry: the value of the largest sample size.

Constraint: $\text{NMAX} = \max_{1\leq i\leq \text{NS}}(\text{NV}(i))$ and $\text{NMAX} > \text{IP}$.

10: TOL – ***real*** *Input*

On entry: the tolerance for judging whether two observations are tied. Thus, observations Y_i and Y_j are adjudged to be tied if $|Y_i - Y_j| < \text{TOL}$.

Constraint: $\text{TOL} > 0.0$.

11: PARVAR(NPVAR,IP) – ***real*** array *Output*

On exit: the variance-covariance matrices of the score statistics and the parameter estimates, the former being stored in the upper triangle and the latter in the lower triangle. Thus for $1 \leq i \leq j \leq \text{IP}$, $\text{PARVAR}(i,j)$ contains an estimate of the covariance between the ith and jth score statistics. For $1 \leq j \leq i \leq \text{IP}-1$, $\text{PARVAR}(i+1,j)$ contains an estimate of the covariance between the ith and jth parameter estimates.

12: NPVAR – INTEGER *Input*

On entry: the first dimension of the array PARVAR as declared in the (sub)program from which G08RAF is called.

Constraint: NPVAR $\geq$ IP+ 1.

13: IRANK(NMAX) – INTEGER array *Output*

On exit: for the one sample case, IRANK contains the ranks of the observations.

14: ZIN(NMAX) – ***real*** array *Output*

On exit: for the one sample case, ZIN contains the expected values of the function $g(.)$ of the order statistics.

15: ETA(NMAX) – ***real*** array *Output*

On exit: for the one sample case, ETA contains the expected values of the function $g'(.)$ of the order statistics.

16: VAPVEC(NMAX*(NMAX+1)/2) – ***real*** array *Output*

On exit: for the one sample case, VAPVEC contains the upper triangle of the variance-covariance matrix of the function $g(.)$ of the order statistics stored column-wise.

17: PAREST(4*IP+1) – ***real*** array *Output*

On exit: the statistics calculated by the routine as follows. The first IP components of PAREST contain the score statistics. The next IP elements contain the parameter estimates. PAREST$(2 \times \mathrm{IP} + 1)$ contains the value of the χ^2 statistic. The next IP elements of PAREST contain the standard errors of the parameter estimates. Finally, the remaining IP elements of PAREST contain the z-statistics.

18: WORK(LWORK) – ***real*** array *Workspace*
19: LWORK – INTEGER *Input*

On entry: the dimension of the array WORK as declared in the (sub)program from which G08RAF is called.

Constraint: LWORK $\geq$ NMAX$\times$ (IP+1).

20: IWA(NMAX) – INTEGER array *Workspace*

21: 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, NS < 1,
or TOL $\leq$ 0.0,
or NMAX $\leq$ IP,
or NPVAR < IP + 1,
or NX < NSUM,
or NMAX $\neq \max\limits_{1\leq i\leq \mathrm{NS}} (\mathrm{NV}(i))$,
or NV(i) $\leq$ 0, for some i, NV(i),
or NSUM $\neq \sum\limits_{i=1}^{\mathrm{NS}} \mathrm{NV}(i)$,
or IP < 1,
or LWORK < NMAX × (IP + 1).

IFAIL = 2

On entry, IDIST < 1,
or IDIST > 4.

IFAIL = 3

On entry, all the observations are adjudged to be tied. The user is advised to check the value supplied for TOL.

IFAIL = 4

The matrix $X^T(B-A)X$ is either ill-conditioned or not positive-definite. This error should only occur with extreme rankings of the data.

IFAIL = 5

The matrix X has at least one of its columns with all elements equal.

7 Accuracy

The computations are believed to be stable.

8 Further Comments

The time taken by the routine depends on the number of samples, the total number of observations and the number of parameters fitted.

In extreme cases the parameter estimates for certain models can be infinite, although this is unlikely to occur in practice. See Pettitt (1982) for further details.

9 Example

A program to fit a regression model to a single sample of 20 observations using two explanatory variables. The error distribution will be taken to be logistic.

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.

```
*     G08RAF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NSMAX, NX, NMXMAX, NSMMAX, IPMAX, NPVAR, LWORK
      PARAMETER        (NSMAX=5,NX=100,NMXMAX=NX,NSMMAX=NX,IPMAX=6,
     +                 NPVAR=IPMAX+1,LWORK=NMXMAX*(IPMAX+1))
*     .. Local Scalars ..
      real             TOL
      INTEGER          I, IDIST, IFAIL, IP, J, NMAX, NS, NSUM
*     .. Local Arrays ..
      real             ETA(NMXMAX), PAREST(4*IPMAX+1),
     +                 PARVAR(NPVAR,IPMAX), VAPVEC(NMXMAX*(NMXMAX+1)/2),
     +                 WORK(LWORK), X(NX,IPMAX), Y(NSMMAX), ZIN(NMXMAX)
      INTEGER          IRANK(NMXMAX), IWA(NMXMAX), NV(NSMAX)
*     .. External Subroutines ..
      EXTERNAL         G08RAF
*     .. Intrinsic Functions ..
      INTRINSIC        MAX
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G08RAF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
*     Read number of samples, number of parameters to be fitted,
*     error distribution parameter and tolerance criterion for ties.
      READ (NIN,*) NS, IP, IDIST, TOL
      WRITE (NOUT,*)
      IF (NS.GT.0 .AND. NS.LE.NSMAX .AND. IP.GT.0 .AND. IP.LE.IPMAX)
     +    THEN
         WRITE (NOUT,99999) 'Number of samples =', NS
         WRITE (NOUT,99999) 'Number of parameters fitted =', IP
         WRITE (NOUT,99999) 'Distribution =', IDIST
         WRITE (NOUT,99998) 'Tolerance for ties =', TOL
*        Read the number of observations in each sample.
         READ (NIN,*) (NV(I),I=1,NS)
         NMAX = 0
         NSUM = 0
         DO 20 I = 1, NS
            NSUM = NSUM + NV(I)
            NMAX = MAX(NMAX,NV(I))
   20    CONTINUE
         IF (NMAX.GT.0 .AND. NMAX.LE.NMXMAX .AND. NSUM.GT.0 .AND.
     +       NSUM.LE.NSMMAX) THEN
*           Read in observations and design matrices for each sample.
            READ (NIN,*) (Y(I),(X(I,J),J=1,IP),I=1,NSUM)
            IFAIL = 0
*
            CALL G08RAF(NS,NV,NSUM,Y,IP,X,NX,IDIST,NMAX,TOL,PARVAR,
     +                  NPVAR,IRANK,ZIN,ETA,VAPVEC,PAREST,WORK,LWORK,
     +                  IWA,IFAIL)
*
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Score statistic'
            WRITE (NOUT,99997) (PAREST(I),I=1,IP)
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Covariance matrix of score statistic'
            DO 40 J = 1, IP
               WRITE (NOUT,99997) (PARVAR(I,J),I=1,J)
   40       CONTINUE
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Parameter estimates'
            WRITE (NOUT,99997) (PAREST(IP+I),I=1,IP)
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Covariance matrix of parameter estimates'
            DO 60 I = 1, IP
               WRITE (NOUT,99997) (PARVAR(I+1,J),J=1,I)
   60       CONTINUE
            WRITE (NOUT,*)
            WRITE (NOUT,99996) 'Chi-squared statistic =',
     +        PAREST(2*IP+1), ' with', IP, ' d.f.'
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Standard errors of estimates and'
            WRITE (NOUT,*) 'approximate z-statistics'
            WRITE (NOUT,99995) (PAREST(2*IP+1+I),PAREST(3*IP+1+I),I=1,
     +        IP)
         END IF
      END IF
      STOP
*
99999 FORMAT (1X,A,I2)
99998 FORMAT (1X,A,F8.5)
99997 FORMAT (1X,2F9.3)
99996 FORMAT (1X,A,F9.3,A,I2,A)
99995 FORMAT (1X,F9.3,F14.3)
      END
```

9.2 Program Data

```
G08RAF Example Program Data
 1 2 2 0.00001
 20
 1.0 1.0 23.0
 1.0 1.0 32.0
 3.0 1.0 37.0
 4.0 1.0 41.0
 2.0 1.0 41.0
 4.0 1.0 48.0
 1.0 1.0 48.0
 5.0 1.0 55.0
 4.0 1.0 55.0
 4.0 0.0 56.0
 4.0 1.0 57.0
 4.0 1.0 57.0
 4.0 1.0 57.0
 1.0 0.0 58.0
 4.0 1.0 59.0
 5.0 0.0 59.0
 5.0 0.0 60.0
 4.0 1.0 61.0
 4.0 1.0 62.0
 3.0 1.0 62.0
```

9.3 Program Results

```
 G08RAF Example Program Results

 Number of samples = 1
 Number of parameters fitted = 2
 Distribution = 2
 Tolerance for ties = 0.00001

 Score statistic
    -1.048   64.333

 Covariance matrix of score statistic
     0.673
    -4.159  533.670

 Parameter estimates
    -0.852    0.114

 Covariance matrix of parameter estimates
     1.560
     0.012    0.002

 Chi-squared statistic =    8.221 with 2 d.f.

 Standard errors of estimates and
 approximate z-statistics
     1.249        -0.682
     0.044         2.567
```