NAG Library Routine Document

G08RBF

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

G08RBF 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 when some of the observations may be right-censored.

2 Specification

```
SUBROUTINE G08RBF (NS, NV, NSUM, Y, IP, X, LDX, ICEN, GAMMA, NMAX, TOL,     &
                   PRVR, LDPRVR, IRANK, ZIN, ETA, VAPVEC, PAREST, WORK,     &
                   LWORK, IWA, IFAIL)
INTEGER            NS, NV(NS), NSUM, IP, LDX, ICEN(NSUM), NMAX, LDPRVR,     &
                   IRANK(NMAX), LWORK, IWA(4*NMAX), IFAIL
REAL (KIND=nag_wp) Y(NSUM), X(LDX,IP), GAMMA, TOL, PRVR(LDPRVR,IP),         &
                   ZIN(NMAX), ETA(NMAX), VAPVEC(NMAX*(NMAX+1)/2),           &
                   PAREST(4*IP+1), WORK(LWORK)
```

3 Description

Analysis of data can be made by replacing observations by their ranks. The analysis produces inference for the regression model where the location parameters of the observations, θ_i, for $i = 1, 2, \ldots, n$, are related by $\theta = X\beta$. Here X is an n by p matrix of explanatory variables and β is a vector of p unknown regression parameters. The observations are replaced by their ranks and an approximation, based on Taylor's series expansion, made to the rank marginal likelihood. For details of the approximation see Pettitt (1982).

An observation is said to be right-censored if we can only observe Y_j^* with $Y_j^* \leq Y_j$. We rank censored and uncensored observations as follows. Suppose we can observe Y_j, for $j = 1, 2, \ldots, n$, directly but Y_j^*, for $j = n+1, \ldots, q$ and $n \leq q$, are censored on the right. We define the rank r_j of Y_j, for $j = 1, 2, \ldots, n$, in the usual way; r_j equals i if and only if Y_j is the ith smallest amongst the $Y_1, Y_2, \ldots, Y_n$. The right-censored Y_j^*, for $j = n+1, n+2, \ldots, q$, has rank r_j if and only if Y_j^* lies in the interval $\left[Y_{(r_j)}, Y_{(r_j+1)}\right]$, with $Y_0 = -\infty$, $Y_{(n+1)} = +\infty$ and $Y_{(1)} < \cdots < Y_{(n)}$ the ordered Y_j, for $j = 1, 2, \ldots, n$.

The distribution of the Y is assumed to be of the following form. Let $F_L(y) = e^y/(1+e^y)$, the logistic distribution function, and consider the distribution function $F_\gamma(y)$ defined by $1 - F_\gamma = [1 - F_L(y)]^{1/\gamma}$. This distribution function can be thought of as either the distribution function of the minimum, $X_{1,\gamma}$, of a random sample of size γ^{-1} from the logistic distribution, or as the $F_\gamma(y - \log\gamma)$ being the distribution function of a random variable having the F-distribution with 2 and $2\gamma^{-1}$ degrees of freedom. This family of generalized logistic distribution functions $\left[F_\gamma(.);0 \leq \gamma < \infty\right]$ naturally links the symmetric logistic distribution ($\gamma = 1$) with the skew extreme value distribution ($\lim\gamma \to 0$) and with the limiting negative exponential distribution ($\lim\gamma \to \infty$). For this family explicit results are available for right-censored data. See Pettitt (1983) for details.

Let l_R denote the logarithm of the rank marginal likelihood of the observations and define the $q \times 1$ vector a by $a = l_R'(\theta = 0)$, and let the q by q diagonal matrix B and q by q symmetric matrix A be given by $B - A = -l_R''(\theta = 0)$. Then various statistics can be found from the analysis.

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

(b) The estimated variance-covariance matrix of the score statistic in (a).

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

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

(e) The χ^2 statistic $Q = \hat{\beta}_R M^{-1}\ \hat{\beta}_r = a^{\mathrm{T}}X\left(X^{\mathrm{T}}(B - A)X\right)^{-1}X^{\mathrm{T}}a$, 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., $Z_i = \hat{\beta}_{R_i}/se\left(\hat{\beta}_{R_i}\right)$ 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,

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

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

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

and

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

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

Kalbfleisch J D and Prentice R L (1980) *The Statistical Analysis of Failure Time Data* Wiley

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

Pettitt A N (1983) Approximate methods using ranks for regression with censored data *Biometrika* **70** 121–132

5 Parameters

1: NS – INTEGER *Input*

On entry: the number of samples.

Constraint: $\mathrm{NS} \geq 1$.

2: NV(NS) – INTEGER array *Input*

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

Constraint: $\mathrm{NV}(i) \geq 1$, 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 (KIND=nag_wp) array *Input*

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

5: IP – INTEGER *Input*

On entry: the number of parameters to be fitted.

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

6: X(LDX, IP) – REAL (KIND=nag_wp) array *Input*

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

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

7: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G08RBF is called.

Constraint: $\mathrm{LDX} \geq \mathrm{NSUM}$.

8: ICEN(NSUM) – INTEGER array *Input*

On entry: defines the censoring variable for the observations in Y.

$\mathrm{ICEN}(i) = 0$
If $\mathrm{Y}(i)$ is uncensored.

$\mathrm{ICEN}(i) = 1$
If $\mathrm{Y}(i)$ is censored.

Constraint: $\mathrm{ICEN}(i) = 0$ or 1, for $i = 1,2,\ldots,\mathrm{NSUM}$.

9: GAMMA – REAL (KIND=nag_wp) *Input*

On entry: the value of the parameter defining the generalized logistic distribution. For $\mathrm{GAMMA} \leq 0.0001$, the limiting extreme value distribution is assumed.

Constraint: $\mathrm{GAMMA} \geq 0.0$.

10: NMAX – INTEGER *Input*

On entry: the value of the largest sample size.

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

11: TOL – REAL (KIND=nag_wp) *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 $\left|Y_i - Y_j\right| < \mathrm{TOL}$.

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

12: PRVR(LDPRVR, IP) – REAL (KIND=nag_wp) 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 $1 \leq i \leq j \leq \mathrm{IP}$, $\mathrm{PRVR}(i,j)$ contains an estimate of the covariance between the ith and jth score statistics. For $1 \leq j \leq i \leq \mathrm{IP} - 1$, PRVR$(i+1, j)$ contains an estimate of the covariance between the ith and jth parameter estimates.

13: LDPRVR – INTEGER *Input*

On entry: the first dimension of the array PRVR as declared in the (sub)program from which G08RBF is called.

Constraint: LDPRVR $\geq$ IP $+ 1$.

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

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

15: ZIN(NMAX) – REAL (KIND=nag_wp) array *Output*

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

16: ETA(NMAX) – REAL (KIND=nag_wp) array *Output*

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

17: VAPVEC(NMAX $\times$ (NMAX $+ 1)/2$) – REAL (KIND=nag_wp) 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.

18: PAREST($4 \times$ IP $+ 1$) – REAL (KIND=nag_wp) array *Output*

On exit: the statistics calculated by the routine.

The first IP components of PAREST contain the score statistics.

The next IP elements contain the parameter estimates.

PAREST($2 \times$ 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.

19: WORK(LWORK) – REAL (KIND=nag_wp) array *Workspace*
20: LWORK – INTEGER *Input*

On entry: the dimension of the array WORK as declared in the (sub)program from which G08RBF is called.

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

21: IWA($4 \times$ NMAX) – INTEGER array *Workspace*

22: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

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, if you are 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.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

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 ≤ 0.0,
or NMAX $\leq$ IP,
or LDPRVR $<$ IP $+ 1$,
or LDX $<$ NSUM,
or NMAX $\neq \max_{1 \leq i \leq \mathrm{NS}}(\mathrm{NV}(i))$,
or NV$(i) \leq 0$ for some i, $i = 1, 2, \ldots,$ NS,
or NSUM $\neq \sum_{i=1}^{\mathrm{NS}} \mathrm{NV}(i)$,
or IP < 1,
or GAMMA < 0.0,
or LWORK $<$ NMAX $\times$ (IP $+ 1$).

IFAIL $= 2$

On entry, ICEN$(i) \neq 0$ or 1, for some $1 \leq i \leq$ NSUM.

IFAIL $= 3$

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

IFAIL $= 4$

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

IFAIL $= 5$

On entry, at least one column of the matrix X has all its elements equal.

IFAIL $= -99$

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

See Section 3.8 in the Essential Introduction for further information.

IFAIL $= -399$

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

See Section 3.7 in the Essential Introduction for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.6 in the Essential Introduction for further information.

7 Accuracy

The computations are believed to be stable.

8 Parallelism and Performance

G08RBF is threaded by NAG for parallel execution in multithreaded implementations of the NAG Library.

G08RBF makes calls to BLAS and/or LAPACK routines, which may be threaded within the vendor library used by this implementation. Consult the documentation for the vendor library for further information.

Please consult the X06 Chapter Introduction for information on how to control and interrogate the OpenMP environment used within this routine. Please also consult the Users' Note for your implementation for any additional implementation-specific information.

9 Further Comments

The time taken by G08RBF 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.

10 Example

This example fits a regression model to a single sample of 40 observations using just one explanatory variable.

10.1 Program Text

```
    Program g08rbfe

!     G08RBF Example Program Text

!     Mark 25 Release. NAG Copyright 2014.

!     .. Use Statements ..
      Use nag_library, Only: g08rbf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: gamma, tol
      Integer                          :: i, ifail, ip, j, ldprvr, ldx, liwa,  &
                                          lparest, lvapvec, lwork, nmax, ns,   &
                                          nsum
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: eta(:), parest(:), prvr(:,:),        &
                                          vapvec(:), work(:), x(:,:), y(:),    &
                                          zin(:)
      Integer, Allocatable             :: icen(:), irank(:), iwa(:), nv(:)
!     .. Intrinsic Procedures ..
      Intrinsic                        :: maxval, sum
!     .. Executable Statements ..
      Write (nout,*) 'G08RBF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read number of samples, number of parameters to be fitted,
!     distribution power parameter and tolerance criterion for ties.
      Read (nin,*) ns, ip, gamma, tol

      Allocate (nv(ns))

!     Read the number of observations in each sample
```

```
      Read (nin,*) nv(1:ns)

!     Calculate NSUM, NMAX and various array lengths
      nsum = sum(nv(1:ns))
      nmax = maxval(nv(1:ns))
      ldx = nsum
      ldprvr = ip + 1
      lvapvec = nmax*(nmax+1)/2
      lparest = 4*ip + 1
      lwork = nmax*(ip+1)
      liwa = 4*nmax

      Allocate (y(nsum),x(ldx,ip),icen(nsum),prvr(ldprvr,ip),irank(nmax), &
        zin(nmax),eta(nmax),vapvec(lvapvec),parest(lparest),work(lwork), &
        iwa(liwa))

!     Read in observations, design matrix and censoring variable
      Read (nin,*)(y(i),x(i,1:ip),icen(i),i=1,nsum)

!     Display input information
      Write (nout,99999) 'Number of samples =', ns
      Write (nout,99999) 'Number of parameters fitted =', ip
      Write (nout,99998) 'Distribution power parameter =', gamma
      Write (nout,99998) 'Tolerance for ties =', tol

      ifail = 0
      Call g08rbf(ns,nv,nsum,y,ip,x,ldx,icen,gamma,nmax,tol,prvr,ldprvr,irank, &
        zin,eta,vapvec,parest,work,lwork,iwa,ifail)

!     Display results
      Write (nout,*)
      Write (nout,*) 'Score statistic'
      Write (nout,99997) parest(1:ip)
      Write (nout,*)
      Write (nout,*) 'Covariance matrix of score statistic'
      Do j = 1, ip
        Write (nout,99997) prvr(1:j,j)
      End Do
      Write (nout,*)
      Write (nout,*) 'Parameter estimates'
      Write (nout,99997) parest((ip+1):(2*ip))
      Write (nout,*)
      Write (nout,*) 'Covariance matrix of parameter estimates'
      Do i = 1, ip
        Write (nout,99997) prvr(i+1,1:i)
      End Do
      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)

99999 Format (1X,A,I2)
99998 Format (1X,A,F10.5)
99997 Format (1X,F9.3)
99996 Format (1X,A,F9.3,A,I2,A)
99995 Format (1X,F9.3,F14.3)
    End Program g08rbfe
```

10.2 Program Data

```
G08RBF Example Program Data
 1 1 0.00001 0.00001
 40
 143.0 0.0 0 164.0 0.0 0 188.0 0.0 0 188.0 0.0 0 190.0 0.0 0
 192.0 0.0 0 206.0 0.0 0 209.0 0.0 0 213.0 0.0 0 216.0 0.0 0
 220.0 0.0 0 227.0 0.0 0 230.0 0.0 0 234.0 0.0 0 246.0 0.0 0
```

```
265.0 0.0 0 304.0 0.0 0 216.0 0.0 1 244.0 0.0 1 142.0 1.0 0
156.0 1.0 0 163.0 1.0 0 198.0 1.0 0 205.0 1.0 0 232.0 1.0 0
232.0 1.0 0 233.0 1.0 0 233.0 1.0 0 233.0 1.0 0 233.0 1.0 0
239.0 1.0 0 240.0 1.0 0 261.0 1.0 0 280.0 1.0 0 280.0 1.0 0
296.0 1.0 0 296.0 1.0 0 323.0 1.0 0 204.0 1.0 1 344.0 1.0 1
```

10.3 Program Results

```
G08RBF Example Program Results

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

Score statistic
    4.584

Covariance matrix of score statistic
    7.653

Parameter estimates
    0.599

Covariance matrix of parameter estimates
    0.131

Chi-squared statistic =    2.746 with 1 d.f.

Standard errors of estimates and
approximate z-statistics
    0.361         1.657
```