NAG C Library Function Document

nag_rank_regsn_censored (g08rbc)

1 Purpose

nag_rank_regsn_censored (g08rbc) 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

```
void nag_rank_regsn_censored (Nag_OrderType order, Integer ns, const Integer nv[],
      const double y[], Integer p, const double x[], Integer pdx,
      const Integer icen[], double gamma, Integer nmax, double tol, double parvar[],
      Integer pdparvar, Integer irank[], double zin[], double eta[], double vapvec[],
      double parest[], NagError *fail)
```

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, $i = 1, 2, \ldots, n$, 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 $j = n+1, n+2, \ldots, q$; $n \leq q$, 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 $[Y_{(r_j)}, Y_{(r_j+1)}]$, 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 $[F_\gamma(.); 0 \leq \gamma < \infty]$ 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 $X^T a$. 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 $\hat{\beta}_R = MX^T a$.

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

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

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

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

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

and

$$d = \sum_{k=1}^{\mathbf{ns}} X_k^T a_k,$$

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

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: **order** – Nag_OrderType *Input*

On entry: the **order** parameter specifies the two-dimensional storage scheme being used, i.e., row-major ordering or column-major ordering. C language defined storage is specified by **order** = **Nag_RowMajor**. See Section 2.2.1.4 of the Essential Introduction for a more detailed explanation of the use of this parameter.

Constraint: **order** = **Nag_RowMajor** or **Nag_ColMajor**.

2: **ns** – Integer *Input*

On entry: the number of samples.

Constraint: **ns** ≥ 1.

3: **nv**[**ns**] – const Integer *Input*

On entry: the number of observations in the ith sample, for $i = 1, 2, \ldots,$ **ns**.

Constraint: **nv**$[i] \geq 1$ for $i = 0, 1, \ldots,$ **ns** $- 1$.

4: **y**[dim] – const double *Input*

Note: the dimension, dim, of the array **y** must be at least $\sum_{i=0}^{\mathbf{ns}-1} \mathbf{nv}[i]$.

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

5: **p** – Integer *Input*

On entry: the number of parameters to be fitted.

Constraint: $\mathbf{p} \geq 1$.

6: **x**[dim] – const double *Input*

Note: the dimension, dim, of the array **x** must be at least $\max(1, \mathbf{pdx} \times \mathbf{p})$ when **order** = **Nag_ColMajor** and at least $\max(1, \mathbf{pdx} \times \sum_{i=0}^{\mathbf{ns}-1} \mathbf{nv}[i])$ when **order** = **Nag_RowMajor**.

If **order** = **Nag_ColMajor**, the (i,j)th element of the matrix X is stored in $\mathbf{x}[(j-1) \times \mathbf{pdx} + i - 1]$ and if **order** = **Nag_RowMajor**, the (i,j)th element of the matrix X is stored in $\mathbf{x}[(i-1) \times \mathbf{pdx} + j - 1]$.

On entry: the design matrices for each sample. Specifically, $\mathbf{x}\left(\sum_{k=1}^{i-1} \mathbf{nv}[k-1]+j, l\right)$ 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: **pdx** – Integer *Input*

On entry: the stride separating matrix row or column elements (depending on the value of **order**) in the array **x**.

Constraints:

if **order** = **Nag_ColMajor**, $\mathbf{pdx} \geq \sum_{i=0}^{\mathbf{ns}-1} \mathbf{nv}[i]$;
if **order** = **Nag_RowMajor**, $\mathbf{pdx} \geq \mathbf{p}$.

8: **icen**[dim] – const Integer *Input*

Note: the dimension, dim, of the array **icen** must be at least $\sum_{i=0}^{\mathbf{ns}-1} \mathbf{nv}[i]$.

On entry: defines the censoring variable for the observations in **y** as follows:

icen$[i-1] = 0$

If $\mathbf{y}[i-1]$ is uncensored.

icen$[i-1] = 1$

If $\mathbf{y}[i-1]$ is censored.

Constraint: **icen**$[i-1] = 0$ or 1, for $i = 1, 2, \ldots, \sum_{i=0}^{\mathbf{ns}-1} \mathbf{nv}[i]$.

9: **gamma** – double *Input*

On entry: the value of the parameter defining the generalized logistic distribution. For **gamma** ≤ 0.0001, the limiting extreme value distribution is assumed.

Constraint: **gamma** > 0.0.

10: **nmax** – Integer *Input*

On entry: the value of the largest sample size.

Constraint: $\mathbf{nmax} = \max_{1 \leq i \leq \mathbf{ns}} (\mathbf{nv}[i-1])$ and $\mathbf{nmax} > \mathbf{p}$.

11: **tol** – double *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 $|Y_i - Y_j| <$ **tol**.

Constraint: **tol** > 0.0.

12: **parvar**[dim] – double *Output*

Note: the dimension, dim, of the array **parvar** must be at least $\max(1, \mathbf{pdparvar} \times \mathbf{p})$ when **order** = **Nag_ColMajor** and at least $\max(1, \mathbf{pdparvar} \times \mathbf{p} + 1)$ when **order** = **Nag_RowMajor**.

Where **PARVAR**(i, j) appears in this document, it refers to the array element

if **order** = **Nag_ColMajor**, **parvar**$[(j-1) \times \mathbf{pdparvar} + i - 1]$;

if **order** = **Nag_RowMajor**, **parvar**$[(i-1) \times \mathbf{pdparvar} + j - 1]$.

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 $1 \le i \le j \le \mathbf{p}$, **PARVAR**$(i, j)$ contains an estimate of the covariance between the ith and jth score statistics. For $1 \le j \le i \le \mathbf{p} - 1$, **PARVAR**$(i + 1, j)$ contains an estimate of the covariance between the ith and jth parameter estimates.

13: **pdparvar** – Integer *Input*

On entry: the stride separating matrix row or column elements (depending on the value of **order**) in the array **parvar**.

Constraints:

if **order** = **Nag_ColMajor**, **pdparvar** $\ge \mathbf{p} + 1$;
if **order** = **Nag_RowMajor**, **pdparvar** $\ge \mathbf{p}$.

14: **irank**[**nmax**] – Integer *Output*

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

15: **zin**[**nmax**] – double *Output*

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

16: **eta**[**nmax**] – double *Output*

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

17: **vapvec**[dim] – double *Output*

Note: the dimension, dim, of the array **vapvec** must be at least **nmax** $\times$ (**nmax** $+ 1)/2$.

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**[dim] – double *Output*

Note: the dimension, dim, of the array **parest** must be at least $4 \times \mathbf{p} + 1$.

On exit: the statistics calculated by the routine as follows. The first **p** components of **parest** contain the score statistics. The next **p** elements contain the parameter estimates. **parest**$[2 \times \mathbf{p}]$ contains the value of the χ^2 statistic. The next **p** elements of **parest** contain the standard errors of the parameter estimates. Finally, the remaining **p** elements of **parest** contain the z-statistics.

19: **fail** – NagError * *Input/Output*

The NAG error parameter (see the Essential Introduction).

6 Error Indicators and Warnings

NE_INT

On entry, **ns** = ⟨*value*⟩.
Constraint: **ns** ≥ 1.

On entry, **p** = ⟨*value*⟩.
Constraint: **p** ≥ 1.

On entry, **pdx** = ⟨*value*⟩.
Constraint: **pdx** > 0.

On entry, **pdparvar** = ⟨*value*⟩.
Constraint: **pdparvar** > 0.

NE_INT_2

On entry, **pdx** = ⟨*value*⟩, **p** = ⟨*value*⟩.
Constraint: **pdx** $\geq$ **p**.

On entry, **pdparvar** = ⟨*value*⟩, **p** = ⟨*value*⟩.
Constraint: **pdparvar** $\geq$ **p** $+ 1$.

On entry, **pdparvar** = ⟨*value*⟩, **p** = ⟨*value*⟩.
Constraint: **pdparvar** $\geq$ **p**.

On entry, **nmax** $\leq$ **p**: **nmax** = ⟨*value*⟩, **p** = ⟨*value*⟩.

On entry, **pdx** $<$ the sum of **nv**$[i]$: **pdx** = ⟨*value*⟩, sum **nv**$[i]$ = ⟨*value*⟩.

NE_INT_ARRAY

On entry, **nv**$[i]$ = ⟨*value*⟩.
Constraint: **nv**$[i] \geq 1$ for $i = 0, \ldots,$ **ns** $- 1$.

NE_INT_ARRAY_ELEM_CONS

M elements of array **icen** are not equal to 0 or 1: M = ⟨*value*⟩.

M elements of array **nv** are less than or equal to zero: M = ⟨*value*⟩.

NE_MAT_ILL_DEFINED

The matrix $X^T(B - A)X$ is either singular or non-positive-definite.

NE_OBSERVATIONS

All the observations were adjudged to be tied.

NE_REAL

On entry, **tol** = ⟨*value*⟩.
Constraint: **tol** > 0.0.

On entry, **gamma** = ⟨*value*⟩.
Constraint: **gamma** ≥ 0.0.

NE_REAL_ARRAY_ELEM_CONS

On entry, all elements in column ⟨*value*⟩ of **x** are equal to ⟨*value*⟩.

NE_SAMPLE

The largest sample size is ⟨*value*⟩ which is not equal to **nmax**, **nmax** = ⟨*value*⟩.

NE_ALLOC_FAIL

Memory allocation failed.

NE_BAD_PARAM

On entry, parameter ⟨*value*⟩ had an illegal value.

NE_INTERNAL_ERROR

An internal error has occurred in this function. Check the function call and any array sizes. If the call is correct then please consult NAG for assistance.

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 40 observations using just one explanatory variable.

9.1 Program Text

```
/* nag_rank_regsn_censored (g08rbc) Example Program.
 *
 * Copyright 2001 Numerical Algorithms Group.
 *
 * Mark 7, 2001.
 */

#include <stdio.h>
#include <nag.h>
#include <nag_stdlib.h>
#include <nagg08.h>

int main(void)
{

  /* Scalars */
  double gamma, tol;
  Integer exit_status, i, p, j, nmax, ns, nsum;
  Integer pdx, pdparvar;
  NagError fail;
  Nag_OrderType order;

  /* Arrays */
  double *eta=0, *parest=0, *parvar=0, *vapvec=0, *x=0, *y=0, *zin=0;
  Integer *icen=0, *irank=0, *iwa=0, *nv=0;

#ifdef NAG_COLUMN_MAJOR
#define X(I,J) x[(J-1)*pdx + I - 1]
#define PARVAR(I,J) parvar[(J-1)*pdparvar + I - 1]
  order = Nag_ColMajor;
#else
#define X(I,J) x[(I-1)*pdx + J - 1]
#define PARVAR(I,J) parvar[(I-1)*pdparvar + J - 1]
  order = Nag_RowMajor;
```

```
#endif

  INIT_FAIL(fail);
  exit_status = 0;
  Vprintf("g08rbc Example Program Results\n");

  /* Skip heading in data file */
  Vscanf("%*[^\n] ");

  /* Read number of samples, number of parameters to be fitted, */
  /* distribution power parameter and tolerance criterion for ties. */

  Vscanf("%ld%ld%lf%lf%*[^\n] ", &ns, &p, &gamma, &tol);
  Vprintf("\n");

  /* Allocate memory to nv only */
  if ( !(nv = NAG_ALLOC(ns, Integer)) )
    {
      Vprintf("Allocation failure\n");
      exit_status = -1;
      goto END;
    }

  Vprintf("Number of samples =%2ld\n", ns);
  Vprintf("Number of parameters fitted =%2ld\n", p);
  Vprintf("Distribution power parameter =%10.5f\n", gamma);

  Vprintf("Tolerance for ties =%10.5f\n", tol);

  Vprintf("\n");
  /* Read the number of observations in each sample */

  for (i = 1; i <= ns; ++i)
    Vscanf("%ld", &nv[i - 1]);
  Vscanf("%*[^\n] ");

  nmax = 0;
  nsum = 0;
  for (i = 1; i <= ns; ++i)
    {
      nsum += nv[i - 1];
      nmax = MAX(nmax,nv[i - 1]);
    }

  /* Allocate memory */
  if ( !(eta = NAG_ALLOC(nmax, double)) ||
       !(parest = NAG_ALLOC(4*p+1, double)) ||
       !(parvar = NAG_ALLOC(7 * 6, double)) ||
       !(vapvec = NAG_ALLOC(nmax*(nmax+1)/2, double)) ||
       !(x = NAG_ALLOC(nsum * p, double)) ||
       !(y = NAG_ALLOC(nsum, double)) ||
       !(zin = NAG_ALLOC(nmax, double)) ||
       !(icen = NAG_ALLOC(nsum, Integer)) ||
       !(irank = NAG_ALLOC(nmax, Integer)) ||
       !(iwa = NAG_ALLOC(400, Integer)) )
    {
      Vprintf("Allocation failure\n");
      exit_status = -1;
      goto END;
    }
#ifdef NAG_COLUMN_MAJOR
  pdx = nsum;
  pdparvar = p+1 ;
#else
  pdx = p;
  pdparvar = p;
#endif

  /* Read in observations, design matrix and censoring variable */

  for (i = 1; i <= nsum; ++i)
```

```
    {
      Vscanf("%lf", &y[i - 1]);

      for (j = 1; j <= p; ++j)
        {
          Vscanf("%lf", &X(i,j));
        }
      Vscanf("%ld", &icen[i - 1]);
    }
  Vscanf("%*[^\n] ");

  g08rbc(order, ns, nv, y, p, x, pdx, icen, gamma,
         nmax, tol, parvar, pdparvar, irank, zin, eta, vapvec,
         parest, &fail);
  if (fail.code != NE_NOERROR)
    {
      Vprintf("Error from g08rbc.\n%s\n", fail.message);
      exit_status = 1;
      goto END;
    }

  Vprintf("Score statistic\n");

  for (i = 1; i <= p; ++i)
    Vprintf("%9.3f\n", parest[i - 1]);
  Vprintf("\n");

  Vprintf("Covariance matrix of score statistic\n");
  for (j = 1; j <= p; ++j)
    {
      for (i = 1; i <= j; ++i)
        Vprintf("%9.3f\n", PARVAR(i,j));
      Vprintf("\n");
    }

  Vprintf("Parameter estimates\n");
  for (i = 1; i <= p; ++i)
    Vprintf("%9.3f\n", parest[p + i - 1]);
  Vprintf("\n");
  Vprintf("Covariance matrix of parameter estimates\n");
  for (i = 1; i <= p; ++i)
    {
      for (j = 1; j <= i; ++j)
        Vprintf("%9.3f\n", PARVAR(i + 1,j));
      Vprintf("\n");
    }

  Vprintf("Chi-squared statistic =%9.3f with%2ld d.f.\n", parest[p * 2], p);
  Vprintf("\n");

  Vprintf("Standard errors of estimates and\n");
  Vprintf("approximate z-statistics\n");

  for (i = 1; i <= p; ++i)
     Vprintf("%9.3f%14.3f\n", parest[2*p + 1 + i - 1], parest[p * 3 + 1 + i -
1]);
 END:
  if (eta) NAG_FREE(eta);
  if (parest) NAG_FREE(parest);
  if (parvar) NAG_FREE(parvar);
  if (vapvec) NAG_FREE(vapvec);
  if (x) NAG_FREE(x);
  if (y) NAG_FREE(y);
  if (zin) NAG_FREE(zin);
  if (icen) NAG_FREE(icen);
  if (irank) NAG_FREE(irank);
  if (iwa) NAG_FREE(iwa);
  if (nv) NAG_FREE(nv);

  return exit_status;
}
```

9.2 Program Data

```
g08rbc 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
```

9.3 Program Results

```
g08rbc 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
```