NAG Toolbox for MATLAB

g08ra

1 Purpose

g08ra 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 Syntax

```
[prvr, irank, zin, eta, vapvec, parest, ifail] = g08ra(nv, y, x, idist,
nmax, tol, 'ns', ns, 'ip', ip)
```

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

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

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 $\hat{\beta} = MX^{\mathrm{T}}a$ 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 $a_i = E\left(Z_{r_i}\right)$. To define M we need $M^{-1} = X^{\mathrm{T}}(B - A)X$, where B is an n by n diagonal matrix with $B_{ii} = E\left(g'\left(W_{r_i}\right)\right)$ and A is a symmetric matrix with $A_{ij} = \operatorname{cov}\left(Z_{r_i}, Z_{r_j}\right)$. In the case of the Normal distribution, the $Z_1 < \cdots < Z_n$ are standard Normal order statistics and $E\left(g'(W_i)\right) = 1$, for $i = 1, 2, \ldots, n$.

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

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 $X^{\mathrm{T}}(B - A)X$ of the score statistic in (a).

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

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

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

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

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

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

and

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

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

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

5.1 Compulsory Input Parameters

1: **nv(ns) – int32 array**

The number of observations in the ith sample, for $i = 1, 2, \ldots, \mathbf{ns}$.

Constraint: $\mathbf{nv}(i) \geq 1$, for $i = 1, 2, \ldots, \mathbf{ns}$.

2: **y(nsum) – double array**

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

3: **x(ldx,ip) – double array**

ldx, the first dimension of the array, must be at least **nsum**.

The design matrices for each sample. Specifically, $\mathbf{x}\left(\sum_{k=1}^{i-1} \mathbf{nv}(k) + j, l\right)$ 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.

4: **idist – int32 scalar**

The error distribution to be used in the analysis.

idist $= 1$

Normal.

idist $= 2$

Logistic.

idist $= 3$

Extreme value.

idist $= 4$

Double-exponential.

Constraint: $1 \leq$ **idist** ≤ 4.

5: **nmax – int32 scalar**

the value of the largest sample size.

Constraint: **nmax** $= \max_{1 \leq i \leq \mathbf{ns}} (\mathbf{nv}(i))$ and **nmax** $>$ **ip**.

6: **tol – double scalar**

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.

5.2 Optional Input Parameters

1: **ns – int32 scalar**

Default: The dimension of the array **nv**.

the number of samples.

Constraint: **ns** ≥ 1.

2: **ip – int32 scalar**

Default: The dimension of the arrays **x**, **prvr**. (An error is raised if these dimensions are not equal.)

the number of parameters to be fitted.

Constraint: **ip** ≥ 1.

5.3 Input Parameters Omitted from the MATLAB Interface

nsum, ldx, ldprvr, work, lwork, iwa

5.4 Output Parameters

1: **prvr(ldprvr,ip) – double array**

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 $1 \leq i \leq j \leq$ **ip**, **prvr**(i,j) contains an estimate of the covariance between the ith and jth score statistics. For $1 \leq j \leq i \leq$ **ip** $- 1$, **prvr**$(i+1,j)$ contains an estimate of the covariance between the ith and jth parameter estimates.

2: **irank(nmax) – int32 array**

For the one sample case, **irank** contains the ranks of the observations.

3: **zin(nmax) – double array**

For the one sample case, **zin** contains the expected values of the function $g(.)$ of the order statistics.

4: **eta(nmax) – double array**

For the one sample case, **eta** contains the expected values of the function $g\prime(.)$ of the order statistics.

5: **vapvec(nmax × (nmax + 1)/2) – double array**

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.

6: **parest(4 × ip + 1) – double array**

The statistics calculated by the function as follows. 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.

7: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, **ns** < 1,
or **tol** ≤ 0.0,
or **nmax** $\leq$ **ip**,
or **ldprvr** $<$ **ip** $+ 1$,
or **ldx** $<$ **nsum**,
or **nmax** $\neq \max_{1 \leq i \leq \mathbf{ns}}(\mathbf{nv}(i))$,
or **nv**$(i) \leq 0$, for some i, **nv**(i),
or **nsum** $\neq \sum_{i=1}^{\mathbf{ns}} \mathbf{nv}(i)$,
or **ip** < 1,
or **lwork** $<$ **nmax** $\times$ (**ip** $+ 1$).

ifail $= 2$

On entry, **idist** < 1,
or **idist** > 4.

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 $= 5$

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 g08ra 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

```
nv = [int32(20)];
y = [1;
     1;
     3;
     4;
     2;
     4;
     1;
     5;
     4;
     4;
     4;
     4;
     4;
     1;
     4;
     5;
     5;
     4;
     4;
     3];
x = [1, 23;
     1, 32;
     1, 37;
     1, 41;
     1, 41;
     1, 48;
     1, 48;
     1, 55;
     1, 55;
     0, 56;
     1, 57;
     1, 57;
     1, 57;
     0, 58;
     1, 59;
     0, 59;
     0, 60;
     1, 61;
     1, 62;
     1, 62];
idist = int32(2);
nmax = int32(20);
tol = 1e-05;
[parvar, irank, zin, eta, vapvec, parest, ifail] = g08ra(nv, y, x, idist,
nmax, tol)
```

```
parvar =
    0.6733   -4.1587
    1.5604  533.6696
    0.0122    0.0020
```

```
irank =
           1
           2
           6
           8
           5
           9
           3
          18
          10
          11
          12
          13
          14
           4
          15
          19
          20
          16
          17
           7
zin =
   -0.7619
   -0.7619
   -0.7619
   -0.7619
   -0.5238
   -0.3810
   -0.3810
    0.1905
    0.1905
    0.1905
    0.1905
    0.1905
    0.1905
    0.1905
    0.1905
    0.1905
    0.1905
    0.8095
    0.8095
    0.8095
eta =
    0.1948
    0.1948
    0.1948
    0.1948
    0.3463
    0.4069
    0.4069
    0.4242
    0.4242
    0.4242
    0.4242
    0.4242
    0.4242
    0.4242
    0.4242
    0.4242
    0.4242
    0.1616
    0.1616
    0.1616
vapvec =
     array elided
parest =
   -1.0476
   64.3333
   -0.8524
    0.1139
    8.2210
    1.2492
    0.0444
   -0.6824
    2.5673
ifail =
           0
```