NAG C Library Function Document

nag_full_step_regsn (g02efc)

1 Purpose

nag_full_step_regsn (g02efc) calculates a full stepwise selection from p variables by using Clarke's sweep algorithm on the correlation matrix of a design and data matrix, Z. The (weighted) variance-covariance, (weighted) means and sum of weights of Z must be supplied.

2 Specification

```
#include <nag.h>
#include <nagg02.h>

void nag_full_step_regsn (Integer m, Integer n, const double wmean[],
     const double c[], double sw, Integer isx[], double fin, double fout, double tau,
     double b[], double se[], double *rsq, double *rms, Integer *df,
     void (*monfun)(Nag_FullStepwise flag, Integer var, double val,
          Nag_Comm *comm),
     Nag_Comm *comm, NagError *fail)
```

3 Description

The general multiple linear regression model is defined by

$$y = \beta_0 + X\beta + \varepsilon,$$

where

y is a vector of n observations on the dependent variable,

β_0 is an intercept coefficient,

X is a n by p matrix of p explanatory variables,

β is a vector of p unknown coefficients, and

ε is a vector of length n of unknown, normally distributed, random errors.

The function employs a full stepwise regression to select a subset of explanatory variables from the p available variables (the intercept is included in the model) and computes regression coefficients and their standard errors, and various other statistical quantities, by minimizing the sum of squares of residuals. The method applies repeatedly a forward selection step followed by a backward elimination step and halts when neither step updates the current model.

The criterion used to update a current model is the variance ratio of residual sum of squares. Let s_1 and s_2 be the residual sum of squares of the current model and this model after undergoing a single update, with degrees of freedom q_1 and q_2, respectively. Then the condition:

$$\frac{(s_2 - s_1)/(q_2 - q_1)}{s_1/q_1} > f_1,$$

must be satisfied if a variable k will be considered for entry to the current model, and the condition:

$$\frac{(s_1 - s_2)/(q_1 - q_2)}{s_1/q_1} < f_2,$$

must be satisfied if a variable k will be considered for removal from the current model, where f_1 and f_2 are user-supplied values and $f_2 \le f_1$.

In the entry step the entry statistic is computed for each variable not in the current model. If no variable is associated with a test value that exceeds f_1 then this step is terminated; otherwise the variable associated with the largest value for the entry statistic is entered into the model.

In the removal step the removal statistic is computed for each variable in the current model. If no variable is associated with a test value less than f_2 then this step is terminated; otherwise the variable associated with the smallest value for the removal statistic is removed from the model.

The data values X and y are not provided as input to the function. Instead, summary statistics of the design and data matrix $Z = (X \mid y)$ are required.

Explanatory variables are entered into and removed from the current model by using sweep operations on the correlation matrix R of Z, given by:

$$R = \left(\begin{array}{ccc|c} 1 & \dots & r_{1p} & r_{1y} \\ \vdots & \ddots & \vdots & \vdots \\ r_{p1} & \dots & 1 & r_{py} \\ \hline r_{y1} & \dots & r_{yp} & 1 \end{array}\right),$$

where r_{ij} is the correlation between the explanatory variables i and j, for $i,j = 1,2,\dots,p$, and r_{yi} (and r_{iy}) is the correlation between the response variable y and the ith explanatory variable, for $i = 1,2,\dots,p$.

A sweep operation on the kth row and column ($k \leq p$) of R replaces:

r_{kk} by $-1/r_{kk}$;

r_{ik} by $r_{ik}/|r_{kk}|$, for $i = 1,2,\dots,p+1$ $(i \neq k)$;

r_{kj} by $r_{kj}/|r_{kk}|$, for $j = 1,2,\dots,p+1$ $(j \neq k)$;

r_{ij} by $r_{ij} - r_{ik}r_{kj}/|r_{kk}|$, for $i = 1,2,\dots,p+1$ $(i \neq k)$; for $j = 1,2,\dots,p+1$ $(j \neq k)$.

The kth explanatory variable is eligible for entry into the current model if it satisfies the collinearity tests: $r_{kk} > \tau$ and

$$\left(r_{ii} - \frac{r_{ik}r_{ki}}{r_{kk}}\right)\tau \leq 1,$$

for a user-supplied value (> 0) of τ and where the index i runs over explanatory variables in the current model. The sweep operation is its own inverse, therefore pivoting on an explanatory variable k in the current model has the effect of removing it from the model.

Once the stepwise model selection procedure is finished, the function calculates:

(a) the least squares estimate for the ith explanatory variable included in the fitted model;

(b) standard error estimates for each coefficient in the final model;

(c) the square root of the mean square of residuals and its degrees of freedom;

(d) the multiple correlation coefficient.

The function makes use of the symmetry of the sweep operations and correlation matrix which reduces by almost one half the storage and computation required by the sweep algorithm, see Clarke (1981) for details.

4 References

Clarke M R B (1981) Algorithm AS 178: the Gauss–Jordan sweep operator with detection of collinearity *Applied Statistics* **31** 166–169

Dempster A P (1969) *Elements of Continuous Multivariate Analysis* Addison–Wesley

Draper N R and Smith H (1985) *Applied Regression Analysis* (2nd Edition) Wiley

5 Arguments

1: **m** – Integer *Input*

On entry: the number of explanatory variables available in the design matrix, Z.

Constraint: $\mathbf{m} > 1$.

2: **n** – Integer *Input*

On entry: the number of observations used in the calculations.

Constraint: $\mathbf{n} > 1$.

3: $\mathbf{wmean}[\mathbf{m}+\mathbf{1}]$ – const double *Input*

On entry: the mean of the design matrix, Z.

4: $\mathbf{c}[dim]$ – const double *Input*

Note: the dimension, *dim*, of the array **c** must be at least $(\mathbf{m}+1)\times(\mathbf{m}+2)/2$.

On entry: the upper-triangular variance-covariance matrix packed by column for the design matrix, Z. The function computes the correlation matrix R from C.

5: **sw** – double *Input*

On entry: if weights were used to calculate **c** then **sw** is the sum of positive weight values; otherwise **sw** is the number of observations used to calculate **c**.

Constraint: $\mathbf{sw} > 1.0$.

6: $\mathbf{isx}[\mathbf{m}]$ – Integer *Input/Output*

On entry: the value of $\mathbf{isx}[j-1]$ determines the set of variables used to perform full stepwise model selection, for $j = 1, 2, \ldots, \mathbf{m}$.

$\mathbf{isx}[j-1] = -1$

To exclude the variable corresponding to the jth column of X from the final model.

$\mathbf{isx}[j-1] = 1$

To consider the variable corresponding to the jth column of X for selection in the final model.

$\mathbf{isx}[j-1] = 2$

To force the inclusion of the variable corresponding to the jth column of X in the final model.

Constraint: $\mathbf{isx}[j-1] = -1, 1$ or 2, for $j = 1, 2, \ldots, \mathbf{m}$.

On exit: the value of $\mathbf{isx}[j-1]$ indicates the status of the jth explanatory variable in the model.

$\mathbf{isx}[j-1] = -1$

Forced exclusion.

$\mathbf{isx}[j-1] = 0$

Excluded.

$\mathbf{isx}[j-1] = 1$

Selected.

$\mathbf{isx}[j-1] = 2$

Forced selection.

7: **fin** – double *Input*

On entry: the value of the variance ratio which an explanatory variable must exceed to be included in a model.

Suggested value: **fin** $= 4.0$

Constraint: **fin** > 0.0.

8: **fout** – double *Input*

On entry: the explanatory variable in a model with the lowest variance ratio value is removed from the model if its value is less than **fout**. **fout** is usually set equal to the value of **fin**; a value less than **fin** is occasionally preferred.

Suggested value: **fout** = **fin**

Constraint: $0.0 \leq$ **fout** $\leq$ **fin**.

9: **tau** – double *Input*

On entry: the tolerance, τ, for detecting collinearities between variables when adding or removing an explanatory variable from a model. Explanatory variables deemed to be collinear are excluded from the final model.

Suggested value: **tau** $= 1.0 \times 10^{-6}$

Constraint: **tau** > 0.0.

10: **b**[**m** + **1**] – double *Output*

On exit: **b**[0] contains the estimate for the intercept term in the fitted model. If **isx**$[j-1] \neq 0$ then **b**$[j]$ contains the estimate for the jth explanatory variable in the fitted model; otherwise **b**$[j] = 0$.

11: **se**[**m** + **1**] – double *Output*

On exit: **se**$[j-1]$ contains the standard error for the estimate of **b**$[j-1]$, for $j = 1, 2, \ldots,$ **m** $+ 1$

12: **rsq** – double * *Output*

On exit: the R^2-statistic for the fitted regression model.

13: **rms** – double * *Output*

On exit: the mean square of residuals for the fitted regression model.

14: **df** – Integer * *Output*

On exit: the number of degrees of freedom for the sum of squares of residuals.

15: **monfun** – function, supplied by the user *External Function*

You may define your own function or specify the NAG defined default function nag_full_step_regsn_monit (g02ewc). If this facility is not required then the NAG defined null function macroNULLFN can be substituted.

```
void monfun (Nag_FullStepwise flag, Integer var, double val, Nag_Comm *comm)
```

1: **flag** – Nag_FullStepwise *Input*

On entry: the value of **flag** indicates the stage of the stepwise selection of explanatory variables.

flag = **Nag_AddVar**

Variable **var** was added to the current model.

flag = **Nag_BeginBackward**

Beginning the backward elimination step.

flag = **Nag_ColinearVar**

Variable **var** failed the collinearity test and is excluded from the model.

flag = **Nag_DropVar**

Variable **var** was dropped from the current model.

flag = **Nag_BeginForward**

Beginning the forward selection step

flag = **Nag_NoRemoveVar**

Backward elimination did not remove any variables from the current model.

flag = **Nag_BeginStepwise**

Starting stepwise selection procedure.

flag = **Nag_VarianceRatio**

The variance ratio for variable **var** takes the value **val**.

flag = **Nag_FinishStepwise**

Finished stepwise selection procedure.

2: **var** – Integer *Input*

On entry: the index of the explanatory variable in the design matrix Z to which **flag** pertains.

3: **val** – double *Input*

On entry: if **flag** = **Nag_VarianceRatio** then **val** is the variance ratio value for the coefficient associated with explanatory variable index **var**.

4: **comm** – Nag_Comm * *Communication Structure*

Pointer to structure of type **Nag_Comm**; the following members are relevant to **monfun**.

user – double *
iuser – Integer *
p – Pointer

The type Pointer will be `void *`. Before calling nag_full_step_regsn (g02efc) these pointers may be allocated memory by the user and initialized with various quantities for use by **monfun** when called from nag_full_step_regsn (g02efc).

16: **comm** – Nag_Comm * *Communication Structure*

The NAG communication argument (see Section 2.2.1.1 of the Essential Introduction).

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

The NAG error argument (see Section 2.6 of the Essential Introduction).

6 Error Indicators and Warnings

NE_ALLOC_FAIL

Dynamic memory allocation failed.

NE_BAD_PARAM

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

NE_FREE_VARS

No free variables from which to select.

NE_INT

On entry, $\mathbf{m} \leq 1$: $\mathbf{m} = \langle value \rangle$.

On entry, $\mathbf{n} \leq 1$: $\mathbf{n} = \langle value \rangle$.

NE_INT_ARRAY_ELEM_CONS

On entry, invalid value for $\mathbf{isx}[\langle value \rangle] = \langle value \rangle$.

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.

NE_MODEL_INFEASIBLE

All variables are collinear, no model to select.

NE_NOT_POS_DEF

Matrix not positive-definite, results may be inaccurate.

NE_REAL

On entry, $\mathbf{fin} \leq 0.0$: $\mathbf{fin} = \langle value \rangle$.

On entry, $\mathbf{sw} \leq 1$: $\mathbf{sw} = \langle value \rangle$.

On entry, $\mathbf{tau} \leq 0.0$: $\mathbf{tau} = \langle value \rangle$.

NE_REAL_2

On entry, $\mathbf{fout} \leq 0.0$ or $\mathbf{fout} > \mathbf{fin}$: $\mathbf{fout} = \langle value \rangle$; $\mathbf{fin} = \langle value \rangle$.

NE_ZERO_DIAG

On entry at least one diagonal element of $\mathbf{c} \leq 0.0$.

7 Accuracy

nag_full_step_regsn (g02efc) returns a warning if the design and data matrix is not positive-definite.

8 Further Comments

Although the condition for removing or adding a variable to the current model is based on a ratio of variances, these values should not be interpreted as F-statistics with the usual interpretation of significance unless the probability levels are adjusted to account for correlations between variables under consideration and the number of possible updates (see, e.g., Draper and Smith (1985)).

The function allocates internally $\mathcal{O}(4 \times \mathbf{m} + (\mathbf{m} + 1) \times (\mathbf{m} + 2)/2 + 2)$ of double storage.

9 Example

A program that calculates a full stepwise model selection for the Hald data described in Dempster (1969). Means, the upper-triangular variance-covariance matrix and the sum of weights are calculated by nag_sum_sqs (g02buc). An example monitor function is supplied to print information at each step of the model selection process.

9.1 Program Text

```
/* nag_full_stepwise (g02efc) Example Program.
 *
 * Copyright 2004 Numerical Algorithms Group.
 *
 * Mark 8, 2004.
 */

#include <stdio.h>
#include <nag.h>
#include <nag_stdlib.h>
#include <nagg02.h>

int main(void)
{
  /* Scalars */
  double fin, fout, rms, rsq, sw, tau;
  Integer df, exit_status, i, j, m, n, pdx;

  /* Arrays */
  double *b = 0, *c = 0, *se = 0, *wmean = 0, *x = 0;
  Integer *isx = 0;
  /* Nag types */
  Nag_OrderType order;
  Nag_SumSquare mean;
  Nag_Comm comm;
  NagError fail;

#ifdef NAG_COLUMN_ORDER
#define X(I,J) x[(J-1)*pdx + I - 1]
  order = Nag_ColMajor;
#else
#define X(I,J) x[(I-1)*pdx + J - 1]
  order = Nag_RowMajor;
#endif

  INIT_FAIL(fail);
  exit_status = 0;
  mean = Nag_AboutMean;

  Vprintf("nag_full_step_regsn (g02efc) Example Program Results\n\n");

  /* Skip heading in data file */
  Vscanf("%*[^\n]");
  Vscanf("%ld %ld %lf %lf %lf", &n, &m, &fin,
         &fout, &tau);
  Vscanf("%*[^\n]");

  if (n > 1 && m > 1)
    {
      /* Allocate memory */
      if ( !(b = NAG_ALLOC(m+1, double)) ||
           !(c = NAG_ALLOC((m+1)*(m+2)/2, double)) ||
           !(se = NAG_ALLOC(m+1, double)) ||
           !(wmean = NAG_ALLOC(m+1, double)) ||
           !(x = NAG_ALLOC(n * (m+1), double)) ||
           !(isx = NAG_ALLOC(m, Integer)) )
        {
          Vprintf("Allocation failure\n");
          exit_status = -1;
          goto END;
        }
#ifdef NAG_COLUMN_ORDER
      pdx = n;
#else
      pdx = m+1;
```

```
#endif
    }

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

  for (j = 1; j <= m; ++j)
    {
      Vscanf("%"NAG_IFMT, &isx[j-1]);
    }
  Vscanf("%*[^\n]");

  /* Compute upper-triangular correlation matrix */
  /* nag_sum_sqs (g02buc).
   * Computes a weighted sum of squares matrix
   */
  nag_sum_sqs(order, mean, n, m+1, x, pdx, 0, &sw, wmean, c, &fail);
  if (fail.code != NE_NOERROR)
    {
      Vprintf("Error from nag_sum_sqs (g02buc).\n%s\n.",fail.message);
      exit_status = 1;
      goto END;
    }

  /* Perform stepwise selection of variables */
  /* nag_full_step_regsn (g02efc).
   * Stepwise linear regression
   */
  nag_full_step_regsn(m, n, wmean, c, sw, isx, fin, fout, tau, b, se, &rsq,
                      &rms, &df, g02ewc, &comm, &fail);
  if (fail.code != NE_NOERROR)
    {
      Vprintf("Error from nag_full_step_regsn (g02efc).\n%s\n.",fail.message);
      exit_status = 1;
      goto END;
    }

  /* Display summary information for fitted model */
  Vprintf("\n");
  Vprintf("Fitted Model Summary\n");
  Vprintf("%-13s\t%-9s\t%s\n", "Term", "Estimate", "Standard Error");
  Vprintf("%-13s\t%-9.3e\t%-14.3e\n", "Intercept:", b[0], se[0]);
  for (i = 1; i <= m; ++i)
    {
      j = isx[i-1];
      if (j == 1 || j == 2)
        {
          Vprintf("%-10s%3ld\t%9.3e\t%-14.3e\n",
                  "Variable:", i, b[i], se[i]);
        }
    }
  Vprintf("\n");
  Vprintf("RMS: %-12.3e\n\n", rms);

 END:
  if (b) NAG_FREE(b);
  if (c) NAG_FREE(c);
  if (se) NAG_FREE(se);
  if (wmean) NAG_FREE(wmean);
  if (x) NAG_FREE(x);
  if (isx) NAG_FREE(isx);

  return exit_status;
}
```

9.2 Program Data

```
nag_full_step_regsn (g02efc) Example Program Data
13 4 4 2 1.0e-6 1     :  N,M,FIN,FOUT,TAU,MONLEV
 7 26  6 60  78.5
 1 29 15 52  74.3
11 56  8 20 104.3
11 31  8 47  87.6
 7 52  6 33  95.9
11 55  9 22 109.2
 3 71 17  6 102.7
 1 31 22 44  72.5
 2 54 18 22  93.1
21 47  4 26 115.9
 1 40 23 34  83.8
11 66  9 12 113.3
10 68  8 12 109.4     : End of X array of size N by M+1
1 1 1 1               : Array ISX
```

9.3 Program Results

```
nag_full_step_regsn (g02efc) Example Program Results


Starting Stepwise Selection

Forward Selection
Variable 1   Variance ratio = 1.260e+01
Variable 2   Variance ratio = 2.196e+01
Variable 3   Variance ratio = 4.403e+00
Variable 4   Variance ratio = 2.280e+01

Adding variable 4   to model

Backward Selection
Variable 4   Variance ratio = 2.280e+01

Keeping all current variables

Forward Selection
Variable 1   Variance ratio = 1.082e+02
Variable 2   Variance ratio = 1.725e-01
Variable 3   Variance ratio = 4.029e+01

Adding variable 1   to model

Backward Selection
Variable 1   Variance ratio = 1.082e+02
Variable 4   Variance ratio = 1.593e+02

Keeping all current variables

Forward Selection
Variable 2   Variance ratio = 5.026e+00
Variable 3   Variance ratio = 4.236e+00

Adding variable 2   to model

Backward Selection
Variable 1   Variance ratio = 1.540e+02
Variable 2   Variance ratio = 5.026e+00
Variable 4   Variance ratio = 1.863e+00

Dropping variable 4   from model

Forward Selection
Variable 3   Variance ratio = 1.832e+00
Variable 4   Variance ratio = 1.863e+00

Finished Stepwise Selection
```

```
Fitted Model Summary
Term          Estimate  Standard Error
Intercept:    5.258e+01 2.294e+00
Variable:   1 1.468e+00 1.213e-01
Variable:   2 6.623e-01 4.585e-02

RMS: 5.790e+00
```