NAG C Library Function Document

nag_step_regsn (g02eec)

1 Purpose

nag_step_regsn (g02eec) carries out one step of a forward selection procedure in order to enable the 'best' linear regression model to be found.

2 Specification

```
void nag_step_regsn (Nag_OrderType order, Integer *istep, Nag_IncludeMean mean,
      Integer n, Integer m, const double x[], Integer pdx, const char *var_names[],
      const Integer sx[], Integer maxip, const double y[], const double wt[],
      double fin, Boolean *addvar, const char *newvar[], double *chrss, double *f,
      const char *model[], Integer *nterm, double *rss, Integer *idf, Integer *ifr,
      const char *free_vars[], double exss[], double q[], Integer pdq, double p[],
      NagError *fail)
```

3 Description

One method of selecting a linear regression model from a given set of independent variables is by forward selection. The following procedure is used:

(i) Select the best fitting independent variable, i.e., the independent variable which gives the smallest residual sum of squares. If the F-test for this variable is greater than a chosen critical value, F_c, then include the variable in the model, else stop.

(ii) Find the independent variable that leads to the greatest reduction in the residual sum of squares when added to the current model.

(iii) If the F-test for this variable is greater than a chosen critical value, F_c, then include the variable in the model and go to (b), otherwise stop.

At any step the variables not in the model are known as the free terms.

nag_step_regsn (g02eec) allows the user to specify some independent variables that must be in the model, these are known as forced variables.

The computational procedure involves the use of QR decompositions, the R and the Q matrices being updated as each new variable is added to the model. In addition the matrix $Q^T X_{free}$, where X_{free} is the matrix of variables not included in the model, is updated.

nag_step_regsn (g02eec) computes one step of the forward selection procedure at a call. The results produced at each step may be printed or used as inputs to nag_regsn_mult_linear_upd_model (g02ddc), in order to compute the regression coefficients for the model fitted at that step. Repeated calls to nag_step_regsn (g02eec) should be made until $F < F_c$ is indicated.

4 References

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

Weisberg S (1985) *Applied Linear Regression* Wiley

5 Parameters

Note: after the initial call to nag_step_regsn (g02eec) with **istep** $= 0$ all parameters except **fin** must not be changed by the user between calls.

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: **istep** – Integer * *Input/Output*

On entry: indicates which step in the forward selection process is to be carried out.

If **istep** = 0, then the process is initialised.

Constraint: **istep** ≥ 0.

On exit: **istep** is incremented by 1.

3: **mean** – Nag_IncludeMean *Input*

On entry: indicates if a mean term is to be included.

If **mean** = **Nag_MeanInclude**, a mean term, intercept, will be included in the model.

If **mean** = **Nag_MeanZero**, the model will pass through the origin, zero-point.

Constraint: **mean** = **Nag_MeanInclude** or **Nag_MeanZero**.

4: **n** – Integer *Input*

On entry: the number of observations.

Constraint: **n** ≥ 2.

5: **m** – Integer *Input*

On entry: the total number of independent variables in the data set.

Constraint: **m** ≥ 1.

6: **x**[dim] – const double *Input*

Note: the dimension, dim, of the array **x** must be at least $\max(1, \mathbf{pdx} \times \mathbf{m})$ when **order** = **Nag_ColMajor** and at least $\max(1, \mathbf{pdx} \times \mathbf{n})$ when **order** = **Nag_RowMajor**.

Where $\mathbf{X}(i, j)$ appears in this document, it refers to the array element

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

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

On entry: $\mathbf{X}(i, j)$ must contain the ith observation for the jth independent variable, for $i = 1, 2, \ldots, \mathbf{n}$; $j = 1, 2, \ldots, \mathbf{m}$.

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**, **pdx** $\geq$ **n**;
if **order** = **Nag_RowMajor**, **pdx** $\geq$ **m**.

8: **var_names**[**m**] – char * *Input*

On entry: **var_names**$[i-1]$ must contain the name of the independent variable in row i of **x**, for $i = 1, 2, \ldots, \mathbf{m}$.

9: **sx**[**m**] – const Integer *Input*

On entry: indicates which independent variables could be considered for inclusion in the regression.

If **sx**$[j-1] \geq 2$, then the variable contained in the jth column of **x** is automatically included in the regression model, for $j = 1, 2, \ldots,$ **m**.

If **sx**$[j-1] = 1$, then the variable contained in the jth column of **x** is considered for inclusion in the regression model, for $j = 1, 2, \ldots,$ **m**.

If **sx**$[j-1] = 0$, the variable in the jth column is not considered for inclusion in the model, for $j = 1, 2, \ldots,$ **m**.

Constraint: **sx**$[j-1] \geq 0$ and at least one value of **sx**$[j-1] = 1$, for $j = 1, 2, \ldots,$ **m**.

10: **maxip** – Integer *Input*

On entry: the maximum number of independent variables to be included in the model.

Constraints:

if **mean** = **Nag_MeanInclude**, **maxip** $\geq 1 +$ number of values of **sx** > 0;
if **mean** = **Nag_MeanZero**, **maxip** $\geq$ number of values of **sx** > 0.

11: **y**[**n**] – const double *Input*

On entry: the dependent variable.

12: **wt**[dim] – const double *Input*

Note: the dimension, dim, of the array **wt** must be at least **n**.

On entry: **wt** must contain the weights to be used in the weighted regression, W.

If **wt**$[i-1] = 0.0$, then the ith observation is not included in the model, in which case the effective number of observations is the number of observations with non-zero weights.

If weights are not provided then **wt** must be set to the **NULL** pointer, i.e., `(double *)0`, and the effective number of observations is **n**.

Constraint: if **wt** is not **NULL**, **wt**$[i] \geq 0.0$ for $i = 0, 1, \ldots,$ **n** $- 1$.

13: **fin** – double *Input*

On entry: the critical value of the F statistic for the term to be included in the model, F_{c}.

Suggested value: 2.0 is a commonly used value in exploratory modelling.

Constraint: **fin** ≥ 0.0.

14: **addvar** – Boolean * *Output*

On exit: indicates if a variable has been added to the model.

If **addvar** = **TRUE**, then a variable has been added to the model.

If **addvar** = **FALSE**, then no variable had an F value greater than F_{c} and none were added to the model.

15: **newvar**[1] – char * *Output*

On exit: if **addvar** = **TRUE**, then **newvar** contains the name of the variable added to the model.

16: **chrss** – double * *Output*

On exit: if **addvar** = **TRUE**, then **chrss** contains the change in the residual sum of squares due to adding variable **newvar**.

17: **f** – double * *Output*

On exit: if **addvar** = **TRUE**, then **f** contains the F statistic for the inclusion of the variable in **newvar**.

18: **model**[**maxip**] – char * *Input/Output*

On entry: if **istep** = 0, then need not be set.

If **istep** $\neq 0$, then must contain the values returned by the previous call to nag_step_regsn (g02eec).

On exit: the names of the variables in the current model.

19: **nterm** – Integer * *Input/Output*

On entry: if **istep** = 0, then **nterm** need not be set.

If **istep** $\neq 0$, then **nterm** must contain the value returned by the previous call to nag_step_regsn (g02eec).

On exit: the number of independent variables in the current model, not including the mean, if any.

20: **rss** – double * *Input/Output*

On entry: if **istep** = 0, then **rss** need not be set.

If **istep** $\neq 0$, then **rss** must contain the value returned by the previous call to nag_step_regsn (g02eec).

On exit: the residual sums of squares for the current model.

21: **idf** – Integer * *Input/Output*

On entry: if **istep** = 0, then **idf** need not be set.

If **istep** $\neq 0$, then **idf** must contain the value returned by the previous call to nag_step_regsn (g02eec).

On exit: the degrees of freedom for the residual sum of squares for the current model.

22: **ifr** – Integer * *Input/Output*

On entry: if **istep** = 0, then **ifr** need not be set.

If **istep** $\neq 0$, then **ifr** must contain the value returned by the previous call to nag_step_regsn (g02eec).

On exit: the number of free independent variables, i.e., the number of variables not in the model that are still being considered for selection.

23: **free_vars**[**maxip**] – char * *Input/Output*

On entry: if **istep** = 0, then **free_vars** need not be set.

If **istep** $\neq 0$, then **free_vars** must contain the values returned by the previous call to nag_step_regsn (g02eec).

On exit: the first **ifr** values of **free_vars** contain the names of the free variables.

24: **exss**[**maxip**] – double *Output*

On exit: the first **ifr** values of **exss** contain what would be the change in regression sum of squares if the free variables had been added to the model, i.e., the extra sum of squares for the free variables. **exss**$[i-1]$ contains what would be the change in regression sum of squares if the variable **free_vars**$[i-1]$ had been added to the model.

25: **q**[dim] – double *Input/Output*

Note: the dimension, dim, of the array **q** must be at least $\max(1, \textbf{pdq} \times \textbf{maxip} + 2)$ when **order** = **Nag_ColMajor** and at least $\max(1, \textbf{pdq} \times \textbf{n})$ when **order** = **Nag_RowMajor**.

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

On entry: if **istep** = 0, then **q** need not be set.

If **istep** $\neq 0$, then **q** must contain the values returned by the previous call to nag_step_regsn (g02eec).

On exit: the results of the QR decomposition for the current model:

the first column of **q** contains $c = Q^T y$ (or $Q^T W^{\frac{1}{2}} y$ where W is the vector of weights if used);
the upper triangular part of columns 2 to $ip + 1$ contain the R matrix;
the strictly lower triangular part of columns 2 to $ip + 1$ contain details of the Q matrix;
the remaining $ip + 1$ to $ip +$ **ifr** columns of contain $Q^T X_{free}$ (or $Q^T W^{\frac{1}{2}} X_{free}$),

where $ip =$ **nterm**, or $ip =$ **nterm** $+ 1$ if **mean** = **Nag_MeanInclude**.

26: **pdq** – Integer *Input*

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

Constraints:

if **order** = **Nag_ColMajor**, **pdq** $\geq$ **n**;
if **order** = **Nag_RowMajor**, **pdq** $\geq$ **maxip** $+ 2$.

27: **p**[**maxip** $+ 1$] – double *Input/Output*

On entry: if **istep** = 0, then **p** need not be set.

If **istep** $\neq 0$, then **p** must contain the values returned by the previous call to nag_step_regsn (g02eec).

On exit: the first ip elements of **p** contain details of the QR decomposition, where $ip =$ **nterm**, or $ip =$ **nterm** $+ 1$ if **mean** = **Nag_MeanInclude**.

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

The NAG error parameter (see the Essential Introduction).

6 Error Indicators and Warnings

NE_INT

On entry, **istep** = ⟨*value*⟩.
Constraint: **istep** ≥ 0.

On entry, **n** = ⟨*value*⟩.
Constraint: **n** ≥ 2.

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

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

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

NE_INT_2

On entry, **istep** = ⟨*value*⟩, **nterm** = ⟨*value*⟩.
Constraint: if **istep** $\neq 0$, **nterm** > 0.

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

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

On entry, **pdq** = ⟨*value*⟩, **n** = ⟨*value*⟩.
Constraint: **pdq** $\geq$ **n**.

On entry, **istep** and **nterm** are inconsistent: **istep** = ⟨*value*⟩, **nterm** = ⟨*value*⟩.

NE_INT_REAL

On entry, **istep** = ⟨*value*⟩, **rss** = ⟨*value*⟩.
Constraint: if **istep** $\neq 0$, **rss** > 0.

NE_DENOM_ZERO

Denominator of **f** statistic is ≤ 0.0.

NE_FREE_VARS

There are no free variables in the regression.

NE_FULL_RANK

Forced variables not of full rank.

NE_INT_ARRAY_ELEM_CONS

On entry, **sx**[⟨*value*⟩] < 0.

NE_REAL

On entry, **fin** = ⟨*value*⟩.
Constraint: $FIN \geq 0.0$.

On entry, with non-zero **istep**, **rss** ≤ 0.0: **rss** = ⟨*value*⟩.

NE_REAL_ARRAY_ELEM_CONS

On entry, **wt**[⟨*value*⟩] < 0.0.

NE_ZERO_DF

Degrees of freedom for error will equal 0 if new variable is added.

On entry, number of forced variables $\geq$ **n**, i.e., **idf** would be zero.

NE_ZERO_VARS

Maximum number of variables to be included is 0.

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

As nag_step_regsn (g02eec) uses a QR transformation the results will often be more accurate than traditional algorithms using methods based on the cross-products of the dependent and independent variables.

8 Further Comments

None.

9 Example

The data, from an oxygen uptake experiment, is given by Weisberg (1985). The names of the variables are as given in Weisberg (1985). The independent and dependent variables are read and nag_step_regsn (g02eec) is repeatedly called until **addvar** = **FALSE**. At each step the F statistic, the free variables and their extra sum of squares are printed; also, except for when **addvar** = **FALSE**, the new variable, the change in the residual sum of squares and the terms in the model are printed.

9.1 Program Text

```
/* nag_step_regsn (g02eec) Example Program.
 *
 * Copyright 2002 Numerical Algorithms Group.
 *
 * Mark 7, 2002.
 */

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

int main(void)
{

  /* Scalars */
  Boolean addvar;
  double chrss, f, fin, rss;
  Integer exit_status, i, idf, ifr, istep, j, m,
    maxip, n, nterm, pdq, pdx;
  NagError fail;
  Nag_OrderType order;
  Nag_IncludeMean mean_enum;
  char mean, weight;

  /* Arrays */
  const char *newvar;
  double *exss=0, *p=0, *q=0, *wt=0, *x=0, *y=0, *wtptr=0;
  Integer *sx=0;

  const char **free_vars, **model;

  const char *vname[] = { "DAY", "BOD", "TKN", "TS", "TVS", "COD"};

#ifdef NAG_COLUMN_MAJOR
#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;
  Vprintf("g02eec Example Program Results\n");

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

  Vscanf("%ld%ld ' %c ' ' %c '%*[^\n] ", &n, &m, &mean, &weight);
  maxip = m;

  /* Allocate memory */
  if ( !(exss = NAG_ALLOC(maxip, double)) ||
       !(p = NAG_ALLOC(maxip+1, double)) ||
       !(q = NAG_ALLOC(n * (maxip+2), double)) ||
       !(wt = NAG_ALLOC(n, double)) ||
       !(x = NAG_ALLOC(n * m, double)) ||
       !(y = NAG_ALLOC(n, double)) ||
       !(sx = NAG_ALLOC(m, Integer)) ||
       !(free_vars =  NAG_ALLOC(maxip, const char *)) ||
       !(model =  NAG_ALLOC(maxip, const char *))
 )
    {
      Vprintf("Allocation failure\n");
      exit_status = -1;
      goto END;
    }

#ifdef NAG_COLUMN_MAJOR
  pdx = n;
  pdq = n;
#else
  pdx = m;
  pdq = maxip+2;
#endif

  if (weight == 'W' || weight == 'w')
    {
      for (i = 1; i <= n; ++i)
        {
          for (j = 1; j <= m; ++j)
            Vscanf("%lf", &X(i,j));
          Vscanf("%lf%lf%*[^\n] ", &y[i - 1], &wt[i - 1]);
          wtptr = wt;
        }
    }
  else
    {
      for (i = 1; i <= n; ++i)
        {
          for (j = 1; j <= m; ++j)
            Vscanf("%lf", &X(i,j));
          Vscanf("%lf%*[^\n] ", &y[i - 1]);
        }
    }

  for (j = 1; j <= m; ++j)
    Vscanf("%ld", &sx[j - 1]);
  Vscanf("%*[^\n] ");

  Vscanf("%lf%*[^\n] ", &fin);

  if (mean == 'M')
    mean_enum = Nag_MeanInclude;
  else if (mean == 'Z')
    mean_enum = Nag_MeanZero;
  else
    {
      Vprintf("Incorrect value for mean: '%c'\n", mean);
```

```
      exit_status = -1;
      goto END;
    }

  Vprintf("\n");

  istep = 0;
  for (i = 1; i <= m; ++i)
    {
      g02eec(order, &istep, mean_enum, n, m, x, pdx, vname, sx, maxip,
             y, wtptr, fin, &addvar, &newvar, &chrss, &f, model, &nterm,
             &rss, &idf, &ifr, free_vars, exss, q, pdq, p, &fail);

      if (fail.code != NE_NOERROR)
        {
          Vprintf("Error from g02eec.\n%s\n", fail.message);
          exit_status = 1;
          goto END;
        }

      Vprintf("Step %ld\n", istep);
      if (!addvar)
        {
          Vprintf("No further variables added maximum F =%7.2f\n", f);
          Vprintf("Free variables:        ");

          for (j = 1; j <= ifr; ++j)
            Vprintf("%3.3s %s", free_vars[j-1], j%6 == 0 || j == ifr ?"\n":" ");

          Vprintf("\n");
          Vprintf("Change in residual sums of squares for free variables:\n");

          Vprintf("                 ");
          for (j = 1; j <= ifr; ++j)
            {
              Vprintf("%9.4f", exss[j - 1]);
              Vprintf("%s", j%6 == 0 || j == ifr ?"\n":" ");
            }
          goto END;
        }
      else
        {
          Vprintf("Added variable is %3s\n", newvar);
          Vprintf("Change in residual sum of squares =%13.4e\n", chrss);
          Vprintf("F Statistic = %7.2f\n\n", f);
          Vprintf("Variables in model: ");

          for (j = 1; j <= nterm; ++j)
            Vprintf("%3s %s", model[j-1], j%6 == 0 || j == nterm ?"\n":" ");
          Vprintf("Residual sum of squares = %13.4e\n", rss);
          Vprintf("Degrees of freedom = %ld\n\n", idf);
          if (ifr == 0)
            {
              Vprintf("No free variables remaining\n");
              goto END;
            }

          Vprintf("%s%6s", "Free variables:  ", "");
          for (j = 1; j <= ifr; ++j)
            {
              Vprintf("%3.3s ", free_vars[j-1]);
              Vprintf(j%6 == 0 || j == ifr ?"\n":" ");
            }
          Vprintf("Change in residual sums of squares for free variables:\n");
          Vprintf("                 ");

          for (j = 1; j <= ifr; ++j)
            Vprintf("%9.4f%s", exss[j - 1], j%6 == 0 || j == ifr ?"\n":" ");
          Vprintf("\n");
        }
    }
```

```
 END:
  if (model) NAG_FREE(model);
  if (free_vars) NAG_FREE(free_vars);
  if (exss) NAG_FREE(exss);
  if (p) NAG_FREE(p);
  if (q) NAG_FREE(q);
  if (wt) NAG_FREE(wt);
  if (x) NAG_FREE(x);
  if (y) NAG_FREE(y);
  if (sx) NAG_FREE(sx);

  return exit_status;
}
```

9.2 Program Data

```
g02eec Example Program Data
  20 6 'M' 'U'
   0. 1125.0 232.0 7160.0 85.9 8905.0  1.5563
   7.  920.0 268.0 8804.0 86.5 7388.0  0.8976
  15.  835.0 271.0 8108.0 85.2 5348.0  0.7482
  22. 1000.0 237.0 6370.0 83.8 8056.0  0.7160
  29. 1150.0 192.0 6441.0 82.1 6960.0  0.3010
  37.  990.0 202.0 5154.0 79.2 5690.0  0.3617
  44.  840.0 184.0 5896.0 81.2 6932.0  0.1139
  58.  650.0 200.0 5336.0 80.6 5400.0  0.1139
  65.  640.0 180.0 5041.0 78.4 3177.0 -0.2218
  72.  583.0 165.0 5012.0 79.3 4461.0 -0.1549
  80.  570.0 151.0 4825.0 78.7 3901.0  0.0000
  86.  570.0 171.0 4391.0 78.0 5002.0  0.0000
  93.  510.0 243.0 4320.0 72.3 4665.0 -0.0969
 100.  555.0 147.0 3709.0 74.9 4642.0 -0.2218
 107.  460.0 286.0 3969.0 74.4 4840.0 -0.3979
 122.  275.0 198.0 3558.0 72.5 4479.0 -0.1549
 129.  510.0 196.0 4361.0 57.7 4200.0 -0.2218
 151.  165.0 210.0 3301.0 71.8 3410.0 -0.3979
 171.  244.0 327.0 2964.0 72.5 3360.0 -0.5229
 220.   79.0 334.0 2777.0 71.9 2599.0 -0.0458
  0     1    1      1     1    2
 2.0
```

9.3 Program Results

```
g02eec Example Program Results

Step 1
Added variable is  TS
Change in residual sum of squares =   4.7126e-01
F Statistic =    7.38

Variables in model: COD   TS
Residual sum of squares =    1.0850e+00
Degrees of freedom = 17

Free variables:        TKN  BOD  TVS
Change in residual sums of squares for free variables:
                    0.1175    0.0600    0.2276

Step 2
No further variables added maximum F =   1.59
Free variables:        TKN  BOD  TVS

Change in residual sums of squares for free variables:
                    0.0979    0.0207    0.0217
```