NAG Toolbox

nag_correg_linregm_var_add (g02de)

1 Purpose

nag_correg_linregm_var_add (g02de) adds a new independent variable to a general linear regression model.

2 Syntax

```
[q, p, rss, ifail] = nag_correg_linregm_var_add(ip, q, p, x, 'n', n, 'wt', wt, 'tol', tol)
[q, p, rss, ifail] = g02de(ip, q, p, x, 'n', n, 'wt', wt, 'tol', tol)
```

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 23: *weight* was removed from the interface; **wt** was made optional.

3 Description

A linear regression model may be built up by adding new independent variables to an existing model. nag_correg_linregm_var_add (g02de) updates the QR decomposition used in the computation of the linear regression model. The QR decomposition may come from nag_correg_linregm_fit (g02da) or a previous call to nag_correg_linregm_var_add (g02de). The general linear regression model is defined by

$$y = X\beta + \epsilon,$$

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

X is an n by p matrix of the independent variables of column rank k,

β is a vector of length p of unknown arguments,

and ϵ is a vector of length n of unknown random errors such that $\mathrm{var}\,\epsilon = V\sigma^2$, where V is a known diagonal matrix.

If $V = I$, the identity matrix, then least squares estimation is used. If $V \neq I$, then for a given weight matrix $W \propto V^{-1}$, weighted least squares estimation is used.

The least squares estimates, $\hat{\beta}$ of the arguments β minimize $(y - X\beta)^{\mathrm{T}}(y - X\beta)$ while the weighted least squares estimates, minimize $(y - X\beta)^{\mathrm{T}}W(y - X\beta)$.

The parameter estimates may be found by computing a QR decomposition of X (or $W^{\frac{1}{2}}X$ in the weighted case), i.e.,

$$X = QR^* \quad \left(\text{or} \quad W^{\frac{1}{2}}X = QR^*\right),$$

where $R^* = \begin{pmatrix} R \\ 0 \end{pmatrix}$ and R is a p by p upper triangular matrix and Q is an n by n orthogonal matrix.

If R is of full rank, then $\hat{\beta}$ is the solution to

$$R\hat{\beta} = c_1,$$

where $c = Q^{\mathrm{T}}y$ (or $Q^{\mathrm{T}}W^{\frac{1}{2}}y$) and c_1 is the first p elements of c.

If R is not of full rank a solution is obtained by means of a singular value decomposition (SVD) of R.

To add a new independent variable, x_{p+1}, R and c have to be updated. The matrix Q_{p+1} is found such that $Q_{p+1}^{\mathrm{T}}\left[R : Q^{\mathrm{T}}x_{p+1}\right]$ (or $Q_{p+1}^{\mathrm{T}}\left[R : Q^{\mathrm{T}}W^{\frac{1}{2}}x_{p+1}\right]$) is upper triangular. The vector c is then updated by multiplying by Q_{p+1}^{T}.

The new independent variable is tested to see if it is linearly related to the existing independent variables by checking that at least one of the values $\left(Q^{\mathrm{T}}x_{p+1}\right)_i$, for $i = p+2, \ldots, n$, is nonzero.

The new parameter estimates, $\hat{\beta}$, can then be obtained by a call to nag_correg_linregm_update (g02dd).

The function can be used with $p = 0$, in which case R and c are initialized.

4 References

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

Golub G H and Van Loan C F (1996) *Matrix Computations* (3rd Edition) Johns Hopkins University Press, Baltimore

Hammarling S (1985) The singular value decomposition in multivariate statistics *SIGNUM Newsl.* **20(3)** 2–25

McCullagh P and Nelder J A (1983) *Generalized Linear Models* Chapman and Hall

Searle S R (1971) *Linear Models* Wiley

5 Parameters

5.1 Compulsory Input Parameters

1: **ip** – INTEGER

p, the number of independent variables already in the model.

Constraint: **ip** ≥ 0 and **ip** $<$ **n**.

2: **q**$(ldq, \mathbf{ip}+\mathbf{2})$ – REAL (KIND=nag_wp) array

ldq, the first dimension of the array, must satisfy the constraint $ldq \geq$ **n**.

If **ip** $\neq 0$, **q** must contain the results of the QR decomposition for the model with p arguments as returned by nag_correg_linregm_fit (g02da) or a previous call to nag_correg_linregm_var_add (g02de).

If **ip** $= 0$, the first column of **q** should contain the n values of the dependent variable, y.

3: **p**$(\mathbf{ip}+\mathbf{1})$ – REAL (KIND=nag_wp) array

Contains further details of the QR decomposition used. The first **ip** elements of **p** must contain the zeta values for the QR decomposition (see nag_lapack_dgeqrf (f08ae) for details).

The first **ip** elements of array **p** are provided by nag_correg_linregm_fit (g02da) or by previous calls to nag_correg_linregm_var_add (g02de).

4: **x**$(\mathbf{n})$ – REAL (KIND=nag_wp) array

x, the new independent variable.

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the array **x** and the first dimension of the array **q**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

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

2: **wt**(:) – REAL (KIND=nag_wp) array

The dimension of the array **wt** must be at least **n** if $weight = \text{'W'}$, and at least 1 otherwise

If provided, **wt** must contain the weights to be used.

If $\mathbf{wt}(i) = 0.0$, the ith observation is not included in the model, in which case the effective number of observations is the number of observations with nonzero weights.

If **wt** is not provided the effective number of observations is n.

Constraint: if $weight = \text{'W'}$, $\mathbf{wt}(i) \geq 0.0$, for $i = 1, 2, \ldots, n$.

3: **tol** – REAL (KIND=nag_wp)

Suggested value: $\mathbf{tol} = 0.000001$.

Default: 0.000001

The value of **tol** is used to decide if the new independent variable is linearly related to independent variables already included in the model. If the new variable is linearly related then c is not updated. The smaller the value of **tol** the stricter the criterion for deciding if there is a linear relationship.

Constraint: $\mathbf{tol} > 0.0$.

5.3 Output Parameters

1: $\mathbf{q}(ldq, \mathbf{ip}+\mathbf{2})$ – REAL (KIND=nag_wp) array

The results of the QR decomposition for the model with $p+1$ arguments:

the first column of **q** contains the updated value of c;

the columns 2 to $\mathbf{ip}+1$ are unchanged;

the first $\mathbf{ip}+1$ elements of column $\mathbf{ip}+2$ contain the new column of R, while the remaining $\mathbf{n}-\mathbf{ip}-1$ elements contain details of the matrix Q_{p+1}.

2: $\mathbf{p}(\mathbf{ip}+\mathbf{1})$ – REAL (KIND=nag_wp) array

The first **ip** elements of **p** are unchanged and the $(\mathbf{ip}+1)$th element contains the zeta value for Q_{p+1}.

3: **rss** – REAL (KIND=nag_wp)

The residual sum of squares for the new fitted model.

Note: this will only be valid if the model is of full rank, see Section 9.

4: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Note: nag_correg_linregm_var_add (g02de) may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected by the function:

ifail $= 1$

On entry, **n** < 1,
or **ip** < 0,
or **ip** $\geq$ **n**,
or $ldq <$ **n**,
or **tol** ≤ 0.0,
or $weight \neq$ 'U' or 'W'.

ifail $= 2$

On entry, $weight =$ 'W' and a value of **wt** < 0.0.

ifail $= 3$ (*warning*)

The new independent variable is a linear combination of existing variables. The $(\mathbf{ip}+2)$th column of **q** will therefore be null.

ifail $= -99$

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

ifail $= -399$

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

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

The accuracy is closely related to the accuracy of nag_lapack_dormqr (f08ag) which should be consulted for further details.

8 Further Comments

It should be noted that the residual sum of squares produced by nag_correg_linregm_var_add (g02de) may not be correct if the model to which the new independent variable is added is not of full rank. In such a case nag_correg_linregm_update (g02dd) should be used to calculate the residual sum of squares.

9 Example

A dataset consisting of 12 observations is read in. The four independent variables are stored in the array **x** while the dependent variable is read into the first column of **q**. If the character variable $mean$ indicates that a mean should be included in the model a variable taking the value 1.0 for all observations is set up and fitted. Subsequently, one variable at a time is selected to enter the model as indicated by the input value of $indx$. After the variable has been added the parameter estimates are calculated by nag_correg_linregm_update (g02dd) and the results printed. This is repeated until the input value of $indx$ is 0.

9.1 Program Text

```
     function g02de_example

fprintf('g02de example results\n\n');

x = [1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0;
     1.0  1.5  2.0  2.5  3.0  3.5  4.0  4.5  5.0  5.5  6.0  6.5;
     0.0  0.0  0.0  0.0  0.0  0.0  1.0  1.0  1.0  1.0  1.0  1.0;
     0.0  0.0  0.0  0.0  0.0  0.0  4.0  4.5  5.0  5.5  6.0  6.5;
     1.4  2.2  4.5  6.1  7.1  7.7  8.3  8.6  8.8  9.0  9.3  9.2];
[m,n] = size(x);
y = [4.32; 5.21; 6.49; 7.10; 7.94; 8.53;
     8.84; 9.02; 9.27; 9.43; 9.68; 9.83];

q = zeros(n,m+1);
q(:,1) = y;
p = zeros(m*(m+2),1);;
ip = nag_int(0);

% Add variables to model one at a time
for j = 1:m
  [q, p, rss, ifail] = g02de( ...
                              ip, q, p, x(j,1:n));
  ip = ip + 1;
  fprintf('\nVariable %4d added\n',ip);

  % Calculate parameter estimates
  rsst = 0;
  [rsst, idf, b, se, covar, svd, irank, p2, ifail] = ...
  g02dd(nag_int(n), ip, q, rsst);

  if svd
    fprintf('Model not of full rank\n\n');
  end
  fprintf('Residual sum of squares = %12.4e\n', rsst);
  fprintf('Degrees of freedom      = %4d\n', idf);
  fprintf('\nVariable   Parameter estimate   Standard error\n\n');
  ivar = double([1:ip]');
  fprintf('%6d%20.4e%20.4e\n',[ivar b se]');
end
```

9.2 Program Results

```
     g02de example results

Variable    1 added
Residual sum of squares =   3.6267e+01
Degrees of freedom      =   11

Variable   Parameter estimate   Standard error

     1          7.9717e+00          5.2416e-01

Variable    2 added
Residual sum of squares =   4.0164e+00
Degrees of freedom      =   10

Variable   Parameter estimate   Standard error

     1          4.4100e+00          4.3756e-01
     2          9.4979e-01          1.0599e-01

Variable    3 added
Residual sum of squares =   3.8872e+00
Degrees of freedom      =    9

Variable   Parameter estimate   Standard error

     1          4.2236e+00          5.6734e-01
```

```
     2          1.0554e+00          2.2217e-01
     3         -4.1962e-01          7.6695e-01

Variable    4 added
Residual sum of squares =   1.8702e-01
Degrees of freedom      =    8

Variable   Parameter estimate   Standard error

     1          2.7605e+00          1.7592e-01
     2          1.7057e+00          7.3100e-02
     3          4.4575e+00          4.2676e-01
     4         -1.3006e+00          1.0338e-01

Variable    5 added
Residual sum of squares =   8.4066e-02
Degrees of freedom      =    7

Variable   Parameter estimate   Standard error

     1          3.1440e+00          1.8181e-01
     2          9.0748e-01          2.7761e-01
     3          2.0790e+00          8.6804e-01
     4         -6.1589e-01          2.4530e-01
     5          2.9224e-01          9.9810e-02
```