NAG Toolbox for MATLAB

g02de

1 Purpose

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

2 Syntax

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

3 Description

A linear regression model may be built up by adding new independent variables to an existing model. g02de updates the QR decomposition used in the computation of the linear regression model. The QR decomposition may come from g02da or a previous call to 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 parameters,

and ϵ is a vector of length n of unknown random errors such that $\operatorname{var}\epsilon = V\sigma^2$, 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 parameters 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.,

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

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 $i = p+2, p+3, \ldots, n$, is nonzero.

The new parameter estimates, $\hat{\beta}$, can then be obtained by a call to 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: **weight – string**

Indicates if weights are to be used.

weight = 'U' (Unweighted)

Least-squares estimation is used.

weight = 'W' (Weighted)

Weighted least-squares is used and weights must be supplied in array **wt**.

Constraint: **weight** = 'U' or 'W'.

2: **ip – int32 scalar**

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

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

3: **q(ldq,ip + 2) – double array**

ldq, the first dimension of the array, must be at least **n**.

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

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

4: **p(ip + 1) – double 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 f08ae for details).

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

5: **wt(*) – double array**

Note: the dimension of the array **wt** must be at least **n** if **weight** = 'W', and at least 1 otherwise.

If **weight** = 'W', **wt** must contain the weights to be used in the weighted regression.

If **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 **weight** = 'U', **wt** is not referenced and the effective number of observations is n.

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

6: **x(n) – double array**

x, the new independent variable.

5.2 Optional Input Parameters

1: **n – int32 scalar**

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

n, the number of observations.

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

2: **tol – double scalar**

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.

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

Default: 0.000001

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

5.3 Input Parameters Omitted from the MATLAB Interface

ldq

5.4 Output Parameters

1: **q(ldq,ip + 2) – double array**

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

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: **p(ip + 1) – double 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 – double scalar**

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 8.

4: **ifail – int32 scalar**

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

6 Error Indicators and Warnings

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

ifail = 1

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

ifail $= 2$

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

ifail $= 3$

The new independent variable is a linear combination of existing variables. The $(\mathbf{ip}+1)$th column of **q** will therefore be null.

7 Accuracy

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

8 Further Comments

It should be noted that the residual sum of squares produced by 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 g02dd should be used to calculate the residual sum of squares.

9 Example

```
weight = 'U';
ip = int32(0);
q = [4.32, 0;
     5.21, 0;
     6.49, 0;
     7.1,  0;
     7.94, 0;
     8.53, 0;
     8.84, 0;
     9.02, 0;
     9.27, 0;
     9.43, 0;
     9.68, 0;
     9.83, 0];
p = [0];
wt = [0];
x = [1;
     1;
     1;
     1;
     1;
     1;
     1;
     1;
     1;
     1;
     1];
[qOut, pOut, rss, ifail] = g02de(weight, ip, q, p, wt, x)
```

```
qOut =
  -27.6147   -3.4641
   -1.9437    0.2543
   -0.6637    0.2543
   -0.0537    0.2543
    0.7863    0.2543
    1.3763    0.2543
    1.6863    0.2543
    1.8663    0.2543
    2.1163    0.2543
    2.2763    0.2543
    2.5263    0.2543
    2.6763    0.2543
```

```
pOut =
    1.1352
rss =
   36.2666
ifail =
           0
```