NAG Toolbox for MATLAB

g02da

1 Purpose

g02da performs a general multiple linear regression when the independent variables may be linearly dependent. Parameter estimates, standard errors, residuals and influence statistics are computed. g02da may be used to perform a weighted regression.

2 Syntax

```
[rss, idf, b, se, cov, res, h, q, svd, irank, p, wk, ifail] =
g02da(mean, weight, x, isx, ip, y, wt, 'n', n, 'm', m, 'tol', tol)
```

3 Description

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 $\mathrm{var}\,\epsilon = V\sigma^2$, 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 parameters β minimize $(y - X\beta)^{\mathrm{T}}(y - X\beta)$ while the weighted least-squares estimates minimize $(y - X\beta)^{\mathrm{T}}W(y - X\beta)$.

g02da finds a QR decomposition of X (or $W^{1/2}X$ in weighted case), i.e.,

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

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 $Q^{\mathrm{T}}W^{1/2}y$) 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,

$$R = Q_* \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} P^{\mathrm{T}},$$

where D is a k by k diagonal matrix with nonzero diagonal elements, k being the rank of R, and Q_* and P are p by p orthogonal matrices. This gives the solution

$$\hat{\beta} = P_1 D^{-1} Q_{*_1}^{\mathrm{T}} c_1,$$

P_1 being the first k columns of P, i.e., $P = (\,P_1 \quad P_0\,)$, and Q_{*_1} being the first k columns of Q_*.

Details of the **svd**, are made available, in the form of the matrix P^*:

$$P^* = \begin{pmatrix} D^{-1}P_1^{\mathrm{T}} \\ P_0^{\mathrm{T}} \end{pmatrix}.$$

This will be only one of the possible solutions. Other estimates may be obtained by applying constraints to the parameters. These solutions can be obtained by using g02dk after using g02da. Only certain linear combinations of the parameters will have unique estimates; these are known as estimable functions.

The fit of the model can be examined by considering the residuals, $r_i = y_i - \hat{y}$, where $\hat{y} = X\hat{\beta}$ are the fitted values. The fitted values can be written as Hy for an n by n matrix H. The ith diagonal elements of H, h_i, give a measure of the influence of the ith values of the independent variables on the fitted regression model. The values h_i are sometimes known as leverages. Both r_i and h_i are provided by g02da.

The output of g02da also includes $\hat{\beta}$, the residual sum of squares and associated degrees of freedom, $(n-k)$, the standard errors of the parameter estimates and the variance-covariance matrix of the parameter estimates.

In many linear regression models the first term is taken as a mean term or an intercept, i.e., $X_{i,1} = 1$, for $i = 1, 2, \ldots, n$. This is provided as an option. Also only some of the possible independent variables are required to be included in a model, a facility to select variables to be included in the model is provided.

Details of the QR decomposition and, if used, the **svd**, are made available. These allow the regression to be updated by adding or deleting an observation using g02dc, adding or deleting a variable using g02de and g02df or estimating and testing an estimable function using g02dn.

4 References

Cook R D and Weisberg S 1982 *Residuals and Influence in Regression* Chapman and Hall

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: **mean – string**

Indicates if a mean term is to be included.

mean = 'M'

A mean term, intercept, will be included in the model.

mean = 'Z'

The model will pass through the origin, zero-point.

Constraint: **mean** = 'M' or 'Z'.

2: **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'.

3: **x(ldx,m) – double array**

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

x(i,j) must contain the ith observation for the jth independent variable, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

4: **isx(m) – int32 array**

Indicates which independent variables are to be included in the model.

isx$(j) > 0$

The variable contained in the jth column of **x** is included in the regression model.

Constraints:

isx$(j) \geq 0$, for $j = 1, 2, \ldots, m$;
if **mean** = 'M', exactly **ip** $- 1$ values of **isx** must be > 0;
if **mean** = 'Z', exactly **ip** values of **isx** must be > 0.

5: **ip – int32 scalar**

the number of independent variables in the model, including the mean or intercept if present.

Constraints:

if **mean** = 'M', $1 \leq$ **ip** $\leq$ **m** $+ 1$;
if **mean** = 'Z', $1 \leq$ **ip** $\leq$ **m**.

6: **y(n) – double array**

y, observations on the dependent variable.

7: **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. The values of **res** and **h** will be set to zero for observations with zero 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 $i = 1, 2, \ldots, n$.

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the arrays **y**, **res**, **h** and the dimension of the arrays **y**, **rss**, **h**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

Constraint: **n** ≥ 2.

2: **m – int32 scalar**

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

m, the total number of independent variables in the data set.

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

3: **tol – double scalar**

The value of **tol** is used to decide if the independent variables are of full rank and if not what is the rank of the independent variables. The smaller the value of **tol** the stricter the criterion for selecting the singular value decomposition. If $\mathbf{tol} = 0.0$, the singular value decomposition will never be used; this may cause run time errors or inaccurate results if the independent variables are not of full rank.

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

Default: 0.000001

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

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldq

5.4 Output Parameters

1: **rss – double scalar**

The residual sum of squares for the regression.

2: **idf – int32 scalar**

The degrees of freedom associated with the residual sum of squares.

3: **b(ip) – double array**

$\mathbf{b}(i)$, $i = 1, 2, \ldots, \mathbf{ip}$ contains the least-squares estimates of the parameters of the regression model, $\hat{\beta}$.

If $\mathbf{mean} = \text{'M'}$, $\mathbf{b}(1)$ will contain the estimate of the mean parameter and $\mathbf{b}(i+1)$ will contain the coefficient of the variable contained in column j of **x**, where $\mathbf{isx}(j)$ is the ith positive value in the array **isx**.

If $\mathbf{mean} = \text{'Z'}$, $\mathbf{b}(i)$ will contain the coefficient of the variable contained in column j of **x**, where $\mathbf{isx}(j)$ is the ith positive value in the array **isx**.

4: **se(ip) – double array**

$\mathbf{se}(i)$, $i = 1, 2, \ldots, \mathbf{ip}$ contains the standard errors of the **ip** parameter estimates given in **b**.

5: **cov(ip × (ip + 1)/2) – double array**

The first $\mathbf{ip} \times (\mathbf{ip}+1)/2$ elements of **cov** contain the upper triangular part of the variance-covariance matrix of the **ip** parameter estimates given in **b**. They are stored packed by column, i.e., the covariance between the parameter estimate given in $\mathbf{b}(i)$ and the parameter estimate given in $\mathbf{b}(j)$, $j \geq i$, is stored in $\mathbf{cov}(j \times (j-1)/2 + i)$.

6: **res(n) – double array**

The (weighted) residuals, r_i, for $i = 1, 2, \ldots, n$.

7: **h(n) – double array**

The diagonal elements of H, h_i, for $i = 1, 2, \ldots, n$.

8: **q(ldq,ip + 1) – double array**

The results of the *QR* decomposition:

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

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.

9: **svd – logical scalar**

If a singular value decomposition has been performed then **svd** will be **true**, otherwise **svd** will be **false**.

10: **irank – int32 scalar**

The rank of the independent variables.

If **svd** = **false**, **irank** = **ip**.

If **svd** = **true**, **irank** is an estimate of the rank of the independent variables.

irank is calculated as the number of singular values greater that **tol** × (largest singular value). It is possible for the **svd** to be carried out but **irank** to be returned as **ip**.

11: **p(2 × ip + ip × ip) – double array**

Details of the *QR* decomposition and **svd** if used.

If **svd** = **false**, only the first **ip** elements of **p** are used these will contain the zeta values for the *QR* decomposition (see f08ae for details).

If **svd** = **true**, the first **ip** elements of **p** will contain the zeta values for the *QR* decomposition (see f08ae for details) and the next **ip** elements of **p** contain singular values. The following **ip** by **ip** elements contain the matrix P^* stored by columns.

12: **wk(5 × (ip − 1) + ip × ip) – double array**

If on exit **svd** = **true**, **wk** contains information which is needed by g02dg; otherwise **wk** is used as workspace.

13: **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, **n** < 2,
or **m** < 1,
or **ldx** < **n**,
or **ldq** < **n**,
or **tol** < 0.0,
or **ip** ≤ 0,
or **ip** > **n**.

ifail = 2

On entry, **mean** ≠ 'M' or 'Z',
or **weight** ≠ 'W' or 'U'.

ifail $= 3$

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

ifail $= 4$

On entry, a value of **isx** < 0,
or the value of **ip** is incompatible with the values of **mean** and **isx**,
or **ip** is greater than the effective number of observations.

ifail $= 5$

The degrees of freedom for the residuals are zero, i.e., the designated number of parameters is equal to the effective number of observations. In this case the parameter estimates will be returned along with the diagonal elements of H, but neither standard errors nor the variance-covariance matrix will be calculated.

ifail $= 6$

The singular value decomposition has failed to converge, see f02wu. This is an unlikely error.

7 Accuracy

The accuracy of g02da is closely related to the accuracy of f08ae and f02wu. These function documents should be consulted.

8 Further Comments

Standardized residuals and further measures of influence can be computed using g02fa. g02da requires, in particular, the results stored in **res** and **h**.

9 Example

```
mean = 'M';
weight = 'U';
x = [1, 0, 0, 0;
     0, 0, 0, 1;
     0, 1, 0, 0;
     0, 0, 1, 0;
     0, 0, 0, 1;
     0, 1, 0, 0;
     0, 0, 0, 1;
     1, 0, 0, 0;
     0, 0, 1, 0;
     1, 0, 0, 0;
     0, 0, 1, 0;
     0, 1, 0, 0];
isx = [int32(1);
     int32(1);
     int32(1);
     int32(1)];
ip = int32(5);
y = [33.63;
     39.62;
     38.18;
     41.46;
     38.02;
     35.83;
     35.99;
     36.58;
     42.92;
     37.8;
     40.43;
```

```
     37.89];
wt = [];
[rss, idf, b, se, cov, res, h, q, svd, irank, p, wk, ifail] = g02da(mean,
weight, x, isx, ip, y, wt)
```

```
rss =
   22.2268
idf =
           8
b =
   30.5567
    5.4467
    6.7433
   11.0467
    7.3200
se =
    0.3849
    0.8390
    0.8390
    0.8390
    0.8390
cov =
    0.1482
    0.0370
    0.7038
    0.0370
   -0.2223
    0.7038
    0.0370
   -0.2223
   -0.2223
    0.7038
    0.0370
   -0.2223
   -0.2223
   -0.2223
    0.7038
res =
   -2.3733
    1.7433
    0.8800
   -0.1433
    0.1433
   -1.4700
   -1.8867
    0.5767
    1.3167
    1.7967
   -1.1733
    0.5900
h =
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
    0.3333
q =
 -132.3142   -3.4641   -0.8660   -0.8660   -0.8660   -0.8660
   -4.3850    0.2543    1.5000   -0.5000   -0.5000   -0.5000
    3.4507    0.2543    0.2464   -1.4142    0.7071    0.7071
   -4.5642    0.2543    0.2464   -0.1494   -1.2247    1.2247
   -2.0473    0.2543    0.2464   -0.1494   -0.2512   -0.0000
   -1.6798    0.2543    0.2464    0.4236    0.0476   -0.0351
```

```
   -2.4286    0.2543    0.2464   -0.1494   -0.2512    0.3511
    0.9458    0.2543   -0.3431   -0.0580   -0.0975    0.4564
    1.7380    0.2543    0.2464   -0.1494    0.4137   -0.1404
    2.1658    0.2543   -0.3431   -0.0580   -0.0975    0.4564
   -0.7520    0.2543    0.2464   -0.1494    0.4137   -0.1404
    0.3802    0.2543    0.2464    0.4236    0.0476   -0.0351
svd =
     1
irank =
          4
p =
    1.1352
    1.1308
    1.2340
    1.2280
    1.1908
    3.8730
    1.7321
    1.7321
    1.7321
    0.0000
    0.2309
   -0.0000
         0
   -0.0000
   -0.4472
    0.0577
   -0.4644
   -0.0817
   -0.1664
    0.4472
    0.0577
    0.2625
   -0.4249
   -0.0232
    0.4472
    0.0577
   -0.0182
    0.1590
    0.4737
    0.4472
    0.0577
    0.2201
    0.3476
   -0.2841
    0.4472
wk =
   -1.0000
   -0.0000
    0.0000
   -0.0000
   -0.0000
    0.0000
   -0.9288
   -0.2285
    0.2919
   -0.0000
         0
   -0.1634
    0.9591
    0.2310
   -0.0000
    0.0000
   -0.3327
    0.1669
   -0.9281
    0.0000
    0.0000
    0.0000
   -0.0000
```

```
   -0.0000
   -1.0000
    0.2887
    0.2625
   -0.4249
   -0.0232
    4.0000
    3.2500
    5.2500
    4.2500
    5.2500
    0.0000
    0.8075
   -0.7071
    0.9995
    1.0000
    0.5898
         0
    0.0312
    0.0000
   -0.8075
         0
ifail =
           0
```