NAG Toolbox

nag_correg_linregm_fit (g02da)

1 Purpose

nag_correg_linregm_fit (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. nag_correg_linregm_fit (g02da) may be used to perform a weighted regression.

2 Syntax

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

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 $(y - X\beta)^{\mathrm{T}}W(y - X\beta)$.

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

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

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., $P = \begin{pmatrix} P_1 & P_0 \end{pmatrix}$, 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 arguments. These solutions can be obtained by using nag_correg_linregm_constrain (g02dk) after using nag_correg_linregm_fit (g02da). Only certain linear combinations of the arguments 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 nag_correg_linregm_fit (g02da).

The output of nag_correg_linregm_fit (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 nag_correg_linregm_obs_edit (g02dc), adding or deleting a variable using nag_correg_linregm_var_add (g02de) and nag_correg_linregm_var_del (g02df) or estimating and testing an estimable function using nag_correg_linregm_estfunc (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_p** – CHARACTER(1)

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: **x**(ldx, **m**) – REAL (KIND=nag_wp) array

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

3: **isx**(**m**) – INTEGER 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) ≥ 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.

4: **ip** – INTEGER

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**;
otherwise 1 $\leq$ **ip** $\leq$ **n**.

5: **y**(**n**) – REAL (KIND=nag_wp) array

y, the observations on the dependent variable.

5.2 Optional Input Parameters

1: **n** – INTEGER

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

n, the number of observations.

Constraint: **n** ≥ 2.

2: **m** – INTEGER

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

m, the total number of independent variables in the dataset.

Constraint: **m** ≥ 1.

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

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

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

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

4: **tol** – REAL (KIND=nag_wp)

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

Default: 0.000001

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.

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

5.3 Output Parameters

1: **rss** – REAL (KIND=nag_wp)

The residual sum of squares for the regression.

2: **idf** – INTEGER

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

3: $\mathbf{b}(\mathbf{ip})$ – REAL (KIND=nag_wp) 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 $\mathbf{mean} =$ '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 $\mathbf{mean} =$ '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: $\mathbf{se}(\mathbf{ip})$ – REAL (KIND=nag_wp) array

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

5: $\mathbf{covar}(\mathbf{ip} \times (\mathbf{ip}+\mathbf{1})/\mathbf{2})$ – REAL (KIND=nag_wp) 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: $\mathbf{res}(\mathbf{n})$ – REAL (KIND=nag_wp) array

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

7: $\mathbf{h}(\mathbf{n})$ – REAL (KIND=nag_wp) array

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

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

The results of the QR decomposition:

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

the upper triangular part of columns 2 to $\mathbf{ip}+1$ contain the R matrix;

the strictly lower triangular part of columns 2 to **ip** + 1 contain details of the Q matrix.

9: **svd** – LOGICAL

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

10: **irank** – INTEGER

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**) – REAL (KIND=nag_wp) 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 nag_lapack_dgeqrf (f08ae) for details).

If **svd** = *true*, the first **ip** elements of **p** will contain the zeta values for the QR decomposition (see nag_lapack_dgeqrf (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**(**max**(**2**, **5** × (**ip** − **1**) + **ip** × **ip**)) – REAL (KIND=nag_wp) array

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

13: **ifail** – INTEGER

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

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 $= 5$ (*warning*)

The degrees of freedom for the residuals are zero, i.e., the designated number of arguments 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 nag_eigen_real_triang_svd (f02wu). This is an unlikely error.

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 of nag_correg_linregm_fit (g02da) is closely related to the accuracy of nag_eigen_real_triang_svd (f02wu) and nag_lapack_dgeqrf (f08ae). These function documents should be consulted.

8 Further Comments

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

9 Example

Data from an experiment with four treatments and three observations per treatment are read in. The treatments are represented by dummy $(0-1)$ variables. An unweighted model is fitted with a mean included in the model. nag_correg_ssqmat (g02bu) is then called to calculate the total sums of squares and the coefficient of determination (R_2), adjusted R_2 and Akaike's information criteria (AIC) are calculated.

nag_correg_ssqmat (g02bu) is then called to calculate the total sums of squares and the coefficient of determination (R_2), adjusted R_2 and Akaike's information criteria (AIC) are calculated.

9.1 Program Text

```
function g02da_example

fprintf('g02da example results\n\n');

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];
y = [33.63;     39.62;     38.18;     41.46;     38.02;     35.83;
     35.99;     36.58;     42.92;     37.80;     40.43;     37.89];

[n,m]  = size(x);
isx    = ones(m,1,nag_int_name);
mean_p = 'M';
ip     = nag_int(m+1);

% Fit general linear regression model
[rss, idf, b, se, covar, res, h, q, svd, irank, p, wk, ifail] = ...
  g02da(mean_p, x, isx, ip, y);

% Calculate total sums of squares about mean
[sw, wmean, c, ifail] = g02bu(y);

% Effective number of observations
en = double(idf + irank);
% Calculate R-squared, corrected R-squared and AIC
rsq = 1 - rss/c(1);
mult = (en-1)/double(idf);
arsq = 1 - mult*(1-rsq);
aic = en*log(rss/en) + 2*double(irank);

% Display results
if svd
  fprintf('Model not of full rank, rank = %4d\n\n', irank);
end
fprintf('Residual sum of squares = %12.4e\n', rss);
fprintf('Degrees of freedom      = %4d\n', idf);
fprintf('R-squared               = %12.4e\n', rsq);
fprintf('Adjusted R-squared      = %12.4e\n', arsq);
fprintf('AIC                     = %12.4e\n', aic);
fprintf('\nVariable   Parameter estimate   Standard error\n\n');
ivar = double([1:ip]');
fprintf('%6d%20.4e%20.4e\n',[ivar b se]');
fprintf('\n   Obs          Residuals             H\n\n');
ivar = double([1:n]');
fprintf('%6d%20.4e%20.4e\n',[ivar res h]');
```

9.2 Program Results

```
     g02da example results

Model not of full rank, rank =    4

Residual sum of squares =   2.2227e+01
Degrees of freedom      =    8
R-squared               =   7.0042e-01
Adjusted R-squared      =   5.8808e-01
AIC                     =   1.5397e+01

Variable   Parameter estimate   Standard error

     1          3.0557e+01          3.8494e-01
     2          5.4467e+00          8.3896e-01
     3          6.7433e+00          8.3896e-01
     4          1.1047e+01          8.3896e-01
     5          7.3200e+00          8.3896e-01

   Obs          Residuals             H

     1         -2.3733e+00          3.3333e-01
     2          1.7433e+00          3.3333e-01
     3          8.8000e-01          3.3333e-01
     4         -1.4333e-01          3.3333e-01
     5          1.4333e-01          3.3333e-01
     6         -1.4700e+00          3.3333e-01
```

```
     7         -1.8867e+00          3.3333e-01
     8          5.7667e-01          3.3333e-01
     9          1.3167e+00          3.3333e-01
    10          1.7967e+00          3.3333e-01
    11         -1.1733e+00          3.3333e-01
    12          5.9000e-01          3.3333e-01
```