NAG Library Routine Document

G02DAF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G02DAF performs a general multiple linear regression when the independent variables may be linearly dependent. Parameter estimates, standard errors, residuals and influence statistics are computed. G02DAF may be used to perform a weighted regression.

2 Specification

```
SUBROUTINE G02DAF (MEAN, WEIGHT, N, X, LDX, M, ISX, IP, Y, WT, RSS, IDF, B,      &
                   SE, COV, RES, H, Q, LDQ, SVD, IRANK, P, TOL, WK, IFAIL)

INTEGER            N, LDX, M, ISX(M), IP, IDF, LDQ, IRANK, IFAIL
REAL (KIND=nag_wp) X(LDX,M), Y(N), WT(*), RSS, B(IP), SE(IP),                   &
                   COV(IP*(IP+1)/2), RES(N), H(N), Q(LDQ,IP+1),                  &
                   P(2*IP+IP*IP), TOL, WK(max(2,5*(IP-1)+IP*IP))
LOGICAL            SVD
CHARACTER(1)       MEAN, WEIGHT
```

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 $\operatorname{var}\epsilon = V\sigma^2$, 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 $(y - X\beta)^{\mathrm{T}}W(y - X\beta)$.

G02DAF 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., $P = \begin{pmatrix} P_1 & P_0 \end{pmatrix}$, 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 parameters. These solutions can be obtained by using G02DKF after using G02DAF. 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 G02DAF.

The output of G02DAF 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 G02DCF, adding or deleting a variable using G02DEF and G02DFF or estimating and testing an estimable function using G02DNF.

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

1: MEAN – CHARACTER(1) *Input*

On entry: 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 – CHARACTER(1) *Input*

On entry: indicates if weights are to be used.

WEIGHT = 'U'
Least squares estimation is used.

WEIGHT = 'W'
Weighted least squares is used and weights must be supplied in array WT.

Constraint: WEIGHT = 'U' or 'W'.

3: N – INTEGER *Input*

On entry: n, the number of observations.

Constraint: N ≥ 2.

4: X(LDX,M) – REAL (KIND=nag_wp) array *Input*

On entry: 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.

5: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02DAF is called.

Constraint: LDX $\geq$ N.

6: M – INTEGER *Input*

On entry: m, the total number of independent variables in the dataset.

Constraint: M ≥ 1.

7: ISX(M) – INTEGER array *Input*

On entry: 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.

8: IP – INTEGER *Input*

On entry: 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.

9: Y(N) – REAL (KIND=nag_wp) array *Input*

On entry: y, observations on the dependent variable.

10: WT($*$) – REAL (KIND=nag_wp) array *Input*

Note: the dimension of the array WT must be at least N if WEIGHT = 'W', and at least 1 otherwise.

On entry: 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: if WEIGHT = 'W', WT(i) $\geq$ 0.0, for $i = 1, 2, \ldots, n$.

11: RSS – REAL (KIND=nag_wp) *Output*

On exit: the residual sum of squares for the regression.

12: IDF – INTEGER *Output*

On exit: the degrees of freedom associated with the residual sum of squares.

13: B(IP) – REAL (KIND=nag_wp) array *Output*

On exit: B(i), $i = 1, 2, \ldots,$ IP contains the least squares estimates of the parameters of the regression model, $\hat{\beta}$.

If MEAN = 'M', B(1) will contain the estimate of the mean parameter and B($i + 1$) will contain the coefficient of the variable contained in column j of X, where ISX(j) is the ith positive value in the array ISX.

If MEAN = 'Z', B(i) will contain the coefficient of the variable contained in column j of X, where ISX(j) is the ith positive value in the array ISX.

14: SE(IP) – REAL (KIND=nag_wp) array *Output*

On exit: SE(i), $i = 1, 2, \ldots,$ IP contains the standard errors of the IP parameter estimates given in B.

15: COV(IP $\times$ (IP $+ 1)/2$) – REAL (KIND=nag_wp) array *Output*

On exit: the first IP $\times$ (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 B(i) and the parameter estimate given in B(j), $j \geq i$, is stored in COV($j \times (j - 1)/2 + i$).

16: RES(N) – REAL (KIND=nag_wp) array *Output*

On exit: the (weighted) residuals, r_i, for $i = 1, 2, \ldots, n$.

17: H(N) – REAL (KIND=nag_wp) array *Output*

On exit: the diagonal elements of H, h_i, for $i = 1, 2, \ldots, n$.

18: Q(LDQ,IP $+ 1$) – REAL (KIND=nag_wp) array *Output*

On exit: 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.

19: LDQ – INTEGER *Input*

On entry: the first dimension of the array Q as declared in the (sub)program from which G02DAF is called.

Constraint: LDQ $\geq$ N.

20: SVD – LOGICAL *Output*

On exit: if a singular value decomposition has been performed then SVD will be .TRUE., otherwise SVD will be .FALSE..

21: IRANK – INTEGER *Output*

On exit: 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.

22: P($2 \times \mathrm{IP} + \mathrm{IP} \times \mathrm{IP}$) – REAL (KIND=nag_wp) array *Output*

On exit: 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 F08AEF (DGEQRF) for details).

If SVD = .TRUE., the first IP elements of P will contain the zeta values for the QR decomposition (see F08AEF (DGEQRF) 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.

23: TOL – REAL (KIND=nag_wp) *Input*

On entry: 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 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: TOL = 0.000001.

Constraint: TOL $\geq$ 0.0.

24: WK($\max(2, 5 \times (\mathrm{IP} - 1) + \mathrm{IP} \times \mathrm{IP})$) – REAL (KIND=nag_wp) array *Output*

On exit: if on exit SVD = .TRUE., WK contains information which is needed by G02DGF; otherwise WK is used as workspace.

25: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this parameter, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

On exit: IFAIL = 0 unless the routine detects an error or a warning has been flagged (see Section 6).

6 Error Indicators and Warnings

If on entry IFAIL $= 0$ or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

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 $\neq$ 'M' or 'Z',
or WEIGHT $\neq$ '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 F02WUF. This is an unlikely error.

7 Accuracy

The accuracy of G02DAF is closely related to the accuracy of F02WUF and F08AEF (DGEQRF). These routine documents should be consulted.

8 Further Comments

Standardized residuals and further measures of influence can be computed using G02FAF. G02FAF 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. G02BUF 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.

G02BUF 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

```
    Program g02dafe

!     G02DAF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g02buf, g02daf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: aic, arsq, en, mult, rsq, rss, sw, tol
      Integer                          :: i, idf, ifail, ip, irank, ldq, ldx,  &
                                          lwt, m, n
      Logical                          :: svd
      Character (1)                    :: mean, weight
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: b(:), cov(:), h(:), p(:), q(:,:),    &
                                          res(:), se(:), wk(:), wt(:), x(:,:), &
                                          y(:)
      Real (Kind=nag_wp)               :: c(1), wmean(1)
      Integer, Allocatable             :: isx(:)
!     .. Intrinsic Procedures ..
      Intrinsic                        :: count, log, real
!     .. Executable Statements ..
      Write (nout,*) 'G02DAF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in the problem size
      Read (nin,*) n, m, weight, mean

      If (weight=='W' .Or. weight=='w') Then
        lwt = n
      Else
        lwt = 0
      End If
      ldx = n
      Allocate (x(ldx,m),y(n),wt(lwt),isx(m))

!     Read in data
      If (lwt>0) Then
        Read (nin,*)(x(i,1:m),y(i),wt(i),i=1,n)
      Else
        Read (nin,*)(x(i,1:m),y(i),i=1,n)
      End If

!     Read in variable inclusion flags
      Read (nin,*) isx(1:m)

!     Calculate IP
      ip = count(isx(1:m)>0)
      If (mean=='M' .Or. mean=='m') Then
        ip = ip + 1
      End If

      ldq = n
      Allocate (b(ip),cov((ip*ip+ip)/2),h(n),p(ip*(ip+ &
        2)),q(ldq,ip+1),res(n),se(ip),wk(ip*ip+5*(ip-1)))

!     Use suggested value for tolerance
      tol = 0.000001E0_nag_wp

!     Fit general linear regression model
      ifail = -1
      Call g02daf(mean,weight,n,x,ldx,m,isx,ip,y,wt,rss,idf,b,se,cov,res,h,q, &
        ldq,svd,irank,p,tol,wk,ifail)
      If (ifail/=0) Then
        If (ifail/=5) Then
          Go To 100
        End If
      End If

!     Calculate (weighted) total sums of squares, adjusted for mean if required
!     If in G02DAF, an intercept is added to the regression by including a
!     column of 1’s in X, rather than by using the MEAN argument then
!     MEAN = ’M’ should be used in this call to G02BUF.
      ifail = 0
      Call g02buf(mean,weight,n,1,y,n,wt,sw,wmean,c,ifail)

!     Get effective number of observations (=N if there are no zero weights)
      en = real(idf+irank,kind=nag_wp)

!     Calculate R-squared, corrected R-squared and AIC
      rsq = 1.0_nag_wp - rss/c(1)
      If (mean==’M’ .Or. mean==’m’) Then
        mult = (en-1.0E0_nag_wp)/(en-real(irank,kind=nag_wp))
      Else
        mult = en/(en-real(irank,kind=nag_wp))
      End If
      arsq = 1.0_nag_wp - mult*(1.0_nag_wp-rsq)
      aic = en*log(rss/en) + 2.0_nag_wp*real(irank,kind=nag_wp)

!     Display results
      If (svd) Then
        Write (nout,99999) ’Model not of full rank, rank = ’, irank
        Write (nout,*)
      End If
      Write (nout,99998) ’Residual sum of squares = ’, rss
      Write (nout,99999) ’Degrees of freedom      = ’, idf
      Write (nout,99998) ’R-squared               = ’, rsq
      Write (nout,99998) ’Adjusted R-squared      = ’, arsq
      Write (nout,99998) ’AIC                     = ’, aic
      Write (nout,*)
      Write (nout,*) ’Variable   Parameter estimate   ’, ’Standard error’
      Write (nout,*)
      If (ifail==0) Then
        Write (nout,99997)(i,b(i),se(i),i=1,ip)
      Else
        Write (nout,99996)(i,b(i),i=1,ip)
      End If
      Write (nout,*)
      Write (nout,*) ’   Obs          Residuals              H’
      Write (nout,*)
      Write (nout,99997)(i,res(i),h(i),i=1,n)

100   Continue

99999 Format (1X,A,I4)
99998 Format (1X,A,E12.4)
99997 Format (1X,I6,2E20.4)
99996 Format (1X,I6,E20.4)
    End Program g02dafe
```

9.2 Program Data

```
G02DAF Example Program Data
 12 4 ’U’ ’M’
1.0 0.0 0.0 0.0 33.63
0.0 0.0 0.0 1.0 39.62
0.0 1.0 0.0 0.0 38.18
0.0 0.0 1.0 0.0 41.46
0.0 0.0 0.0 1.0 38.02
0.0 1.0 0.0 0.0 35.83
0.0 0.0 0.0 1.0 35.99
```

```
1.0 0.0 0.0 0.0 36.58
0.0 0.0 1.0 0.0 42.92
1.0 0.0 0.0 0.0 37.80
0.0 0.0 1.0 0.0 40.43
0.0 1.0 0.0 0.0 37.89
 1   1   1   1
```

9.3 Program Results

```
 G02DAF Example Program Results

 Model not of full rank, rank =    4

 Residual sum of squares =   0.2223E+02
 Degrees of freedom      =    8
 R-squared               =   0.7004E+00
 Adjusted R-squared      =   0.5881E+00
 AIC                     =   0.1540E+02

 Variable   Parameter estimate   Standard error

      1          0.3056E+02          0.3849E+00
      2          0.5447E+01          0.8390E+00
      3          0.6743E+01          0.8390E+00
      4          0.1105E+02          0.8390E+00
      5          0.7320E+01          0.8390E+00

    Obs          Residuals              H

      1         -0.2373E+01          0.3333E+00
      2          0.1743E+01          0.3333E+00
      3          0.8800E+00          0.3333E+00
      4         -0.1433E+00          0.3333E+00
      5          0.1433E+00          0.3333E+00
      6         -0.1470E+01          0.3333E+00
      7         -0.1887E+01          0.3333E+00
      8          0.5767E+00          0.3333E+00
      9          0.1317E+01          0.3333E+00
     10          0.1797E+01          0.3333E+00
     11         -0.1173E+01          0.3333E+00
     12          0.5900E+00          0.3333E+00
```