NAG Library Routine Document

G02GDF

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

G02GDF fits a generalized linear model with gamma errors.

2 Specification

```
SUBROUTINE G02GDF (LINK, MEAN, OFFSET, WEIGHT, N, X, LDX, M, ISX, IP, Y,        &
                   WT, S, A, DEV, IDF, B, IRANK, SE, COV, V, LDV, TOL,          &
                   MAXIT, IPRINT, EPS, WK, IFAIL)
INTEGER            N, LDX, M, ISX(M), IP, IDF, IRANK, LDV, MAXIT, IPRINT,       &
                   IFAIL
REAL (KIND=nag_wp) X(LDX,M), Y(N), WT(*), S, A, DEV, B(IP), SE(IP),             &
                   COV(IP*(IP+1)/2), V(LDV,IP+7), TOL, EPS,                     &
                   WK((IP*IP+3*IP+22)/2)
CHARACTER(1)       LINK, MEAN, OFFSET, WEIGHT
```

3 Description

A generalized linear model with gamma errors consists of the following elements:

(a) a set of n observations, y_i, from a gamma distribution with probability density function:

$$\frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu}\exp\left(-\frac{\nu y}{\mu}\right)\frac{1}{y}$$

ν being constant for the sample.

(b) X, a set of p independent variables for each observation, $x_1, x_2, \ldots, x_p$.

(c) a linear model:

$$\eta = \sum \beta_j x_j.$$

(d) a link between the linear predictor, η, and the mean of the distribution, μ, $\eta = g(\mu)$. The possible link functions are:

(i) exponent link: $\eta = \mu^a$, for a constant a,

(ii) identity link: $\eta = \mu$,

(iii) log link: $\eta = \log\mu$,

(iv) square root link: $\eta = \sqrt{\mu}$,

(v) reciprocal link: $\eta = \frac{1}{\mu}$.

(e) a measure of fit, an adjusted deviance. This is a function related to the deviance, but defined for $y = 0$:

$$\sum_{i=1}^{n}\mathrm{dev}^*(y_i, \hat{\mu}_i) = \sum_{i=1}^{n}2\left(\log(\hat{\mu}_i) + \left(\frac{y_i}{\hat{\mu}_i}\right)\right).$$

The linear parameters are estimated by iterative weighted least squares. An adjusted dependent variable, z, is formed:

$$z = \eta + (y - \mu)\frac{d\eta}{d\mu}$$

and a working weight, w,

$$w = \left(\tau\frac{d\eta}{d\mu}\right)^2, \qquad \text{where} \qquad \tau = \frac{1}{\mu}.$$

At each iteration an approximation to the estimate of β, $\hat{\beta}$ is found by the weighted least squares regression of z on X with weights w.

G02GDF finds a QR decomposition of $w^{\frac{1}{2}}X$, i.e.,

$w^{\frac{1}{2}}X = QR$ where R is a p by p triangular matrix and Q is an n by p column orthogonal matrix.

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

$R\hat{\beta} = Q^{\mathrm{T}} w^{\frac{1}{2}} z$

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 $w^{\frac{1}{2}}X$.

This gives the solution

$$\hat{\beta} = P_1 D^{-1} \begin{pmatrix} Q_* & 0 \\ 0 & I \end{pmatrix} Q^{\mathrm{T}} w^{\frac{1}{2}} z,$$

where P_1 is the first k columns of P, i.e., $P = (P_1 P_0)$.

The iterations are continued until there is only a small change in the deviance.

The initial values for the algorithm are obtained by taking

$$\hat{\eta} = g(y).$$

The scale parameter, ν^{-1} is estimated by a moment estimator:

$$\hat{\nu}^{-1} = \sum_{i=1}^{n} \frac{[(y_i - \hat{\mu}_i)/\hat{\mu}]^2}{(n-k)}.$$

The fit of the model can be assessed by examining and testing the deviance, in particular, by comparing the difference in deviance between nested models, i.e., when one model is a sub-model of the other. The difference in deviance or adjusted deviance between two nested models with known ν has, asymptotically, a χ^2-distribution with degrees of freedom given by the difference in the degrees of freedom associated with the two deviances.

The parameters estimates, $\hat{\beta}$, are asymptotically Normally distributed with variance-covariance matrix:

$C = R^{-1} {R^{-1}}^{\mathrm{T}} \nu^{-1}$ in the full rank case, otherwise

$C = P_1 D^{-2} P_1^{\mathrm{T}} \nu^{-1}$.

The residuals and influence statistics can also be examined.

The estimated linear predictor $\hat{\eta} = X\hat{\beta}$, can be written as $Hw^{\frac{1}{2}}z$ 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. These are known as leverages.

The fitted values are given by $\hat{\mu} = g^{-1}(\hat{\eta})$.

G02GDF also computes the Anscombe residuals, r:

$$r_i = \frac{3\left(y_i^{\frac{1}{3}} - \hat{\mu}_i^{\frac{1}{3}}\right)}{\hat{\mu}_i^{\frac{1}{3}}}.$$

An option allows the use of prior weights, ω_i. This gives a model with:

$$\nu_i = \nu\omega_i.$$

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.

Often only some of the possible independent variables are included in a model, the facility to select variables to be included in the model is provided.

If part of the linear predictor can be represented by a variables with a known coefficient then this can be included in the model by using an offset, o:

$$\eta = o + \sum \beta_j x_j.$$

If the model is not of full rank the solution given 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 G02GKF after using G02GDF. Only certain linear combinations of the parameters will have unique estimates, these are known as estimable functions, and can be estimated and tested using G02GNF.

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

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

4 References

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

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

5 Parameters

1: LINK – CHARACTER(1) *Input*

On entry: indicates which link function is to be used.

LINK = 'E'
An exponential link is used.

LINK = 'I'
An identity link is used.

LINK = 'L'
A log link is used.

LINK = 'S'
A square root link is used.

LINK = 'R'
A reciprocal link is used.

Constraint: LINK = 'E', 'I', 'L', 'S' or 'R'.

2: 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'.

3: OFFSET – CHARACTER(1) *Input*

On entry: indicates if an offset is required.

OFFSET = 'Y'
An offset is required and the offsets must be supplied in the seventh column of V.

OFFSET = 'N'
No offset is required.

Constraint: OFFSET = 'N' or 'Y'.

4: WEIGHT – CHARACTER(1) *Input*

On entry: indicates if prior weights are to be used.

WEIGHT = 'U'
No prior weights are used.

WEIGHT = 'W'
Prior weights are used and weights must be supplied in WT.

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

5: N – INTEGER *Input*

On entry: n, the number of observations.

Constraint: $\mathrm{N} \geq 2$.

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

On entry: $\mathrm{X}(i,j)$ must contain the ith observation for the jth independent variable, for $i = 1, 2, \ldots, \mathrm{N}$ and $j = 1, 2, \ldots, \mathrm{M}$.

7: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02GDF is called.

Constraint: $\mathrm{LDX} \geq \mathrm{N}$.

8: M – INTEGER *Input*

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

Constraint: $\mathrm{M} \geq 1$.

9: ISX(M) – INTEGER array *Input*

On entry: indicates which independent variables are to be included in the model.

If $\mathrm{ISX}(j) > 0$, the variable contained in the jth column of X is included in the regression model.

Constraints:

$\mathrm{ISX}(j) \geq 0$, for $i = 1, 2, \ldots, \mathrm{M}$;
if MEAN = 'M', exactly $\mathrm{IP} - 1$ values of ISX must be > 0;
if MEAN = 'Z', exactly IP values of ISX must be > 0.

10: IP – INTEGER *Input*

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

Constraint: $\mathrm{IP} > 0$.

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

On entry: y, the dependent variable.

Constraint: $\mathrm{Y}(i) \geq 0.0$, for $i = 1, 2, \ldots, n$.

12: 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 $\mathrm{WT}(i) = 0.0$, 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: if WEIGHT = 'W', $\mathrm{WT}(i) \geq 0.0$, for $i = 1, 2, \ldots, n$.

13: S – REAL (KIND=nag_wp) *Input/Output*

On entry: the scale parameter for the gamma model, ν^{-1}.

S = 0.0
The scale parameter is estimated with the routine using the formula described in Section 3.

Constraint: $\mathrm{S} \geq 0.0$.

On exit: if on input S = 0.0, S contains the estimated value of the scale parameter, $\hat{\nu}^{-1}$.

If on input $\mathrm{S} \neq 0.0$, S is unchanged on exit.

14: A – REAL (KIND=nag_wp) *Input*

On entry: if LINK = 'E', A must contain the power of the exponential.

If LINK $\neq$ 'E', A is not referenced.

Constraint: if LINK = 'E', $\mathrm{A} \neq 0.0$.

15: DEV – REAL (KIND=nag_wp) *Output*

On exit: the adjusted deviance for the fitted model.

16: IDF – INTEGER *Output*

On exit: the degrees of freedom asociated with the deviance for the fitted model.

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

On exit: the estimates of the parameters of the generalized linear model, $\hat{\beta}$.

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

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

18: IRANK – INTEGER *Output*

On exit: the rank of the independent variables.

If the model is of full rank then IRANK = IP.

If the model is not of full rank then IRANK is an estimate of the rank of the independent variables. IRANK is calculated as the number of singular values greater that EPS $\times$ (largest singular value). It is possible for the SVD to be carried out but for IRANK to be returned as IP.

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

On exit: the standard errors of the linear parameters.

SE(i) contains the standard error of the parameter estimate in B(i), for $i = 1, 2, \ldots,$ IP.

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

On exit: the upper triangular part of the variance-covariance matrix of the IP parameter estimates given in B. They are stored in packed form 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($(j \times (j-1)/2 + i)$).

21: V(LDV,IP $+ 7$) – REAL (KIND=nag_wp) array *Input/Output*

On entry: if OFFSET = 'N', V need not be set.

If OFFSET = 'Y', V($i, 7$), for $i = 1, 2, \ldots, n$, must contain the offset values o_i. All other values need not be set.

On exit: auxiliary information on the fitted model.

V($i, 1$) contains the linear predictor value, η_i, for $i = 1, 2, \ldots, n$.

V($i, 2$) contains the fitted value, $\hat{\mu}_i$, for $i = 1, 2, \ldots, n$.

V($i, 3$) contains the variance standardization, $\frac{1}{\tau_i}$, for $i = 1, 2, \ldots, n$.

V($i, 4$) contains the square root of the working weight, $w_i^{\frac{1}{2}}$, for $i = 1, 2, \ldots, n$.

V($i, 5$) contains the Anscombe residual, r_i, for $i = 1, 2, \ldots, n$.

V($i, 6$) contains the leverage, h_i, for $i = 1, 2, \ldots, n$.

V($i, 7$) contains the offset, o_i, for $i = 1, 2, \ldots, n$. If OFFSET = 'N', all values will be zero.

V(i, j), for $j = 8, \ldots,$ IP $+ 7$, contains the results of the QR decomposition or the singular value decomposition.

If the model is not of full rank, i.e., IRANK $<$ IP, the first IP rows of columns 8 to IP $+ 7$ contain the P^* matrix.

22: LDV – INTEGER *Input*

On entry: the first dimension of the array V as declared in the (sub)program from which G02GDF is called.

Constraint: LDV $\geq$ N.

23: TOL – REAL (KIND=nag_wp) *Input*

On entry: indicates the accuracy required for the fit of the model.

The iterative weighted least squares procedure is deemed to have converged if the absolute change in deviance between iterations is less than TOL $\times$ (1.0 + Current Deviance). This is approximately an absolute precision if the deviance is small and a relative precision if the deviance is large.

If 0.0 $\leq$ TOL $<$ ***machine precision*** then the routine will use 10 $\times$ ***machine precision*** instead.

Constraint: TOL $\geq$ 0.0.

24: MAXIT – INTEGER *Input*

On entry: the maximum number of iterations for the iterative weighted least squares.

MAXIT $= 0$
A default value of 10 is used.

Constraint: MAXIT ≥ 0.

25: IPRINT – INTEGER *Input*

On entry: indicates if the printing of information on the iterations is required.

IPRINT ≤ 0
There is no printing.

IPRINT > 0
Every IPRINT iteration, the following are printed:

the deviance;

the current estimates;

and if the weighted least squares equations are singular then this is indicated.

When printing occurs the output is directed to the current advisory message unit (see X04ABF).

26: EPS – REAL (KIND=nag_wp) *Input*

On entry: the value of EPS 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 EPS the stricter the criterion for selecting the singular value decomposition.

If $0.0 \leq$ EPS $<$ ***machine precision*** then the routine will use ***machine precision*** instead.

Constraint: EPS ≥ 0.0.

27: WK$((\mathrm{IP} \times \mathrm{IP} + 3 \times \mathrm{IP} + 22)/2)$ – REAL (KIND=nag_wp) array *Workspace*

28: 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, because for this routine the values of the output parameters may be useful even if IFAIL $\neq 0$ on exit, the recommended value is -1. **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).

Note: G02GDF may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, N < 2,
or M < 1,
or LDX $<$ N,

or LDV < N,
or IP < 1,
or LINK ≠ 'E','I','L','S' or 'R',
or S < 0.0,
or LINK = 'E' and A = 0.0,
or MEAN ≠ 'M' or 'Z',
or WEIGHT ≠ 'U' or 'W',
or OFFSET ≠ 'N' or 'Y',
or MAXIT < 0,
or TOL < 0.0,
or EPS < 0.0.

IFAIL = 2

On entry, WEIGHT = 'W' and a value of WT < 0.0.

IFAIL = 3

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 = 4

On entry, Y(i) < 0.0 for some $i = 1, 2, \ldots, n$.

IFAIL = 5

A fitted value is at the boundary, i.e., $\hat{\mu} = 0.0$. This may occur if there are small values of y and the model is not suitable for the data. The model should be reformulated with, perhaps, some observations dropped.

IFAIL = 6

The singular value decomposition has failed to converge. This is an unlikely error exit.

IFAIL = 7

The iterative weighted least squares has failed to converge in MAXIT (or default 10) iterations. The value of MAXIT could be increased but it may be advantageous to examine the convergence using the IPRINT option. This may indicate that the convergence is slow because the solution is at a boundary in which case it may be better to reformulate the model.

IFAIL = 8

The rank of the model has changed during the weighted least squares iterations. The estimate for β returned may be reasonable, but you should check how the deviance has changed during iterations.

IFAIL = 9

The degrees of freedom for error are 0. A saturated model has been fitted.

7 Accuracy

The accuracy depends on TOL as described in Section 5. As the adjusted deviance is a function of $\log \mu$, the accuracy of the $\hat{\beta}$s will be a function of TOL, so TOL should be set to a smaller value than the accuracy required for $\hat{\beta}$.

8 Further Comments

None.

9 Example

A set of 10 observations from two groups is input and a model for the two groups is fitted.

9.1 Program Text

```
    Program g02gdfe

!     G02GDF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g02gdf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter                 :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)                 :: a, dev, eps, s, tol
      Integer                            :: i, idf, ifail, ip, iprint, irank,    &
                                            ldv, ldx, lwk, lwt, m, maxit, n
      Character (1)                      :: link, mean, offset, weight
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable    :: b(:), cov(:), se(:), v(:,:), wk(:),  &
                                            wt(:), x(:,:), y(:)
      Integer, Allocatable               :: isx(:)
!     .. Intrinsic Procedures ..
      Intrinsic                          :: count
!     .. Executable Statements ..
      Write (nout,*) 'G02GDF Example Program Results'
      Write (nout,*)

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

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

      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

!     Read in power for exponential link
      If (link=='E' .Or. link=='e') Then
        Read (nin,*) a
      End If

      ldv = n
      lwk = (ip*ip+3*ip+22)/2
```

```
      Allocate (b(ip),se(ip),cov(ip*(ip+1)/2),v(ldv,ip+7),wk(lwk))

!     Read in the offset
      If (offset=='Y' .Or. offset=='y') Then
        Read (nin,*) v(1:n,7)
      End If

!     Read in control parameters
      Read (nin,*) iprint, eps, tol, maxit

!     Fit generalized linear model with Gamma errors
      ifail = -1
      Call g02gdf(link,mean,offset,weight,n,x,ldx,m,isx,ip,y,wt,s,a,dev,idf,b, &
        irank,se,cov,v,ldv,tol,maxit,iprint,eps,wk,ifail)
      If (ifail/=0) Then
        If (ifail<7) Then
          Go To 100
        End If
      End If

!     Display results
      Write (nout,99999) 'Deviance = ', dev
      Write (nout,99998) 'Degrees of freedom = ', idf
      Write (nout,*)
      Write (nout,*) '     Estimate     Standard error'
      Write (nout,*)
      Write (nout,99997)(b(i),se(i),i=1,ip)
      Write (nout,*)
      Write (nout,*) '     Y        FV     Residual        H'
      Write (nout,*)
      Write (nout,99996)(y(i),v(i,2),v(i,5),v(i,6),i=1,n)

100   Continue

99999 Format (1X,A,E12.4)
99998 Format (1X,A,I0)
99997 Format (1X,2F14.4)
99996 Format (1X,F7.1,F10.2,F12.4,F10.3)
    End Program g02gdfe
```

9.2 Program Data

```
G02GDF Example Program Data
'R' 'M' 'N' 'U' 10 1 0.0 :: LINK,MEAN,OFFSET,WEIGHT,N,M,S
1.0  1.0
1.0  0.3
1.0 10.5
1.0  9.7
1.0 10.9
0.0 0.62
0.0 0.12
0.0 0.09
0.0 0.50
0.0 2.14                 :: End of X,Y
 1                       :: ISX
0  1.0E-6  5.0E-5  0     :: IPRINT,EPS,TOL,MAXIT
```

9.3 Program Results

```
 G02GDF Example Program Results

 Deviance =   0.3503E+02
 Degrees of freedom = 8

      Estimate     Standard error

         1.4408        0.6678
        -1.2865        0.6717

      Y        FV     Residual        H
```

```
     1.0       6.48      -1.3909     0.200
     0.3       6.48      -1.9228     0.200
    10.5       6.48       0.5236     0.200
     9.7       6.48       0.4318     0.200
    10.9       6.48       0.5678     0.200
     0.6       0.69      -0.1107     0.200
     0.1       0.69      -1.3287     0.200
     0.1       0.69      -1.4815     0.200
     0.5       0.69      -0.3106     0.200
     2.1       0.69       1.3665     0.200
```