NAG Library Routine Document

G02GBF

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

G02GBF fits a generalized linear model with binomial errors.

2 Specification

```
 SUBROUTINE G02GBF(LINK, MEAN, OFFSET, WEIGHT, N, X, LDX, M, ISX, IP, Y,
1                  T, WT, DEV, IDF, B, IRANK, SE, COV, V, LDV, TOL,
2                  MAXIT, IPRINT, EPS, WK, IFAIL)
 INTEGER           N, LDX, M, ISX(M), IP, IDF, IRANK, LDV, MAXIT, IPRINT,
1                  IFAIL
 double precision  X(LDX,M), Y(N), T(N), WT(*), DEV, B(IP), SE(IP),
1                  COV(IP*(IP+1)/2), V(LDV,IP+7), TOL, EPS,
2                  WK((IP*IP+3*IP+22)/2)
 CHARACTER*1       LINK, MEAN, OFFSET, WEIGHT
```

3 Description

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

(a) a set of n observations, y_i, from a binomial distribution:

$$\binom{t}{y}\pi^y(1-\pi)^{t-y}.$$

(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, $\mu = \pi t$, the link function, $\eta = g(\mu)$. The possible link functions are:

(i) logistic link: $\eta = \log\left(\frac{\mu}{t-\mu}\right)$,

(ii) probit link: $\eta = \Phi^{-1}\left(\frac{\mu}{t}\right)$,

(iii) complementary log-log link: $\log\left(-\log\left(1-\frac{\mu}{t}\right)\right)$.

(e) a measure of fit, the deviance:

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

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,

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

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.

G02GBF finds a QR decomposition of $w^{1/2}X$, i.e., $w^{1/2}X = QR$ 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

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

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 $w^{1/2}X$.

This gives the solution

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

P_1 being 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 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 between two nested models 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

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

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

The residuals and influence statistics can also be examined.

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

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

G02GBF also computes the deviance residuals, r:

$$r_i = \operatorname{sign}(y_i - \hat{\mu}_i)\sqrt{\operatorname{dev}(y_i, \hat{\mu}_i)}.$$

An option allows the use of prior weights in the model.

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 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 be may be obtained by applying constraints to the parameters. These solutions can be obtained by using G02GKF after using G02GBF. 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

Cox D R (1983) *Analysis of Binary Data* 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 = 'G'
A logistic link is used.

LINK = 'P'
A probit link is used.

LINK = 'C'
A complementary log-log link is used.

Constraint: LINK = 'G', 'P' or 'C'.

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) – ***double precision*** array *Input*

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

7: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02GBF 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.

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

Constraints:

$\mathrm{ISX}(j) \geq 0$, for $j = 1, 2, \ldots, \mathrm{M}$;
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) – ***double precision*** array *Input*

On entry: the observations on the dependent variable, y_i, for $i = 1, 2, \ldots, n$.

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

12: T(N) – ***double precision*** array *Input*

On entry: t, the binomial denominator.

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

13: WT($*$) – ***double precision*** 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 $i = 1, 2, \ldots, n$.

14: DEV – ***double precision*** *Output*

On exit: the deviance for the fitted model.

15: IDF – INTEGER *Output*

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

16: B(IP) – ***double precision*** 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 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.

17: IRANK – INTEGER *Output*

On exit: the rank of the independent variables.

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

If the model is not of full rank, 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.

18: SE(IP) – ***double precision*** 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.

19: COV(IP $\times$ (IP $+ 1)/2$) – ***double precision*** 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$(j \times (j-1)/2 + i)$.

20: V(LDV,IP $+ 7$) – ***double precision*** 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 deviance 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.

21: LDV – INTEGER *Input*

On entry: the first dimension of the array V as declared in the (sub)program from which G02GBF is called.

Constraint: LDV $\geq$ N.

22: TOL – ***double precision*** *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***, the routine will use $10 \times$ ***machine precision*** instead.

Constraint: TOL ≥ 0.0.

23: 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.

24: IPRINT – INTEGER *Input*

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

IPRINT ≤ 0
  There is no printing.

IPRINT > 0
  The following is printed every IPRINT iterations:

    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).

25: EPS – ***double precision*** *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***, the routine will use ***machine precision*** instead.

Constraint: EPS ≥ 0.0.

26: WK((IP $\times$ IP + 3 $\times$ IP + 22)/2) – ***double precision*** array *Workspace*

27: 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.

On exit: IFAIL = 0 unless the routine detects an error (see Section 6).

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.**

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: G02GBF 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 $\neq$ 'G', 'P' or 'C'.
or MEAN $\neq$ 'M' or 'Z'.
or WEIGHT $\neq$ 'U' or 'W'.
or OFFSET $\neq$ '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, T$(i) < 0$ for some $i = 1, 2, \ldots, n$.

IFAIL $= 5$

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

IFAIL $= 6$

A fitted value is at the boundary, i.e., 0.0 or 1.0. This may occur if there are y values of 0.0 or t and the model is too complex for the data. The model should be reformulated with, perhaps, some observations dropped.

IFAIL $= 7$

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

IFAIL $= 8$

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 = 9

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 = 10

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

7 Accuracy

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

8 Further Comments

None.

9 Example

A linear trend $(x = -1, 0, 1)$ is fitted to data relating the incidence of carriers of Streptococcus pyogenes to size of tonsils. The data is described in Cox (1983).

9.1 Program Text

```
*     G02GBF Example Program Text
*     Mark 14 Release. NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NMAX, MMAX, LDX, LDV
      PARAMETER        (NMAX=3,MMAX=2,LDX=NMAX,LDV=NMAX)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      DOUBLE PRECISION DEV, EPS, TOL
      INTEGER          I, IDF, IFAIL, IP, IPRINT, IRANK, J, M, MAXIT, N
      CHARACTER        LINK, MEAN, OFFSET, WEIGHT
*     .. Local Arrays ..
      DOUBLE PRECISION B(MMAX), COV((MMAX*MMAX+MMAX)/2), SE(MMAX),
     +                 T(NMAX), V(LDV,7+MMAX),
     +                 WK((MMAX*MMAX+3*MMAX+22)/2), WT(NMAX),
     +                 X(LDX,MMAX), Y(NMAX)
      INTEGER          ISX(MMAX)
*     .. External Subroutines ..
      EXTERNAL         G02GBF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02GBF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) LINK, MEAN, OFFSET, WEIGHT, N, M, IPRINT
      IF (N.LE.NMAX .AND. M.LT.MMAX) THEN
         IF (WEIGHT.EQ.'W' .OR. WEIGHT.EQ.'w') THEN
            DO 20 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), Y(I), T(I), WT(I)
   20       CONTINUE
         ELSE
            DO 40 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), Y(I), T(I)
   40       CONTINUE
         END IF
         READ (NIN,*) (ISX(J),J=1,M)
*        Calculate IP
```

```
         IP = 0
         DO 60 J = 1, M
            IF (ISX(J).GT.0) IP = IP + 1
   60    CONTINUE
         IF (MEAN.EQ.'M' .OR. MEAN.EQ.'m') IP = IP + 1
*        Set control parameters
         EPS = 0.000001D0
         TOL = 0.00005D0
         MAXIT = 10
         IFAIL = -1
*
         CALL G02GBF(LINK,MEAN,OFFSET,WEIGHT,N,X,LDX,M,ISX,IP,Y,T,WT,
     +               DEV,IDF,B,IRANK,SE,COV,V,LDV,TOL,MAXIT,IPRINT,EPS,
     +               WK,IFAIL)
*
         IF (IFAIL.EQ.0 .OR. IFAIL.GE.7) THEN
            WRITE (NOUT,*)
            WRITE (NOUT,99999) 'Deviance = ', DEV
            WRITE (NOUT,99998) 'Degrees of freedom = ', IDF
            WRITE (NOUT,*)
            WRITE (NOUT,*) '       Estimate     Standard error'
            WRITE (NOUT,*)
            DO 80 I = 1, IP
               WRITE (NOUT,99997) B(I), SE(I)
   80       CONTINUE
            WRITE (NOUT,*)
            WRITE (NOUT,*)
     +        '       N           Y         FV      Residual      H'
            WRITE (NOUT,*)
            DO 100 I = 1, N
               WRITE (NOUT,99996) T(I), Y(I), V(I,2), V(I,5), V(I,6)
  100       CONTINUE
         END IF
      END IF
*
99999 FORMAT (1X,A,E12.4)
99998 FORMAT (1X,A,I2)
99997 FORMAT (1X,2F14.4)
99996 FORMAT (1X,2F10.1,F10.2,F12.4,F10.3)
      END
```

9.2 Program Data

```
G02GBF Example Program Data
'G' 'M' 'N' 'U' 3 1 0
 1.0 19. 516.
 0.0 29. 560.
-1.0 24. 293.
  1
```

9.3 Program Results

```
 G02GBF Example Program Results

 Deviance =   0.7354E-01
 Degrees of freedom =  1

        Estimate     Standard error

        -2.8682        0.1217
        -0.4264        0.1598

        N           Y         FV      Residual      H

      516.0      19.0      18.45      0.1296      0.769
      560.0      29.0      30.10     -0.2070      0.422
      293.0      24.0      23.45      0.1178      0.809
```