

G02GAF – NAG Fortran Library Routine Document

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

G02GAF fits a generalized linear model with normal errors.

2 Specification

```

SUBROUTINE G02GAF(LINK, MEAN, OFFSET, WEIGHT, N, X, LDX, M, ISX,
1                IP, Y, WT, S, A, RSS, IDF, B, IRANK, SE, COV, V,
2                LDV, TOL, MAXIT, IPRINT, EPS, WK, IFAIL)
  INTEGER        N, LDX, M, ISX(M), IP, IDF, IRANK, LDV, MAXIT,
1                IPRINT, IFAIL
  real          X(LDX,M), Y(N), WT(*), S, A, RSS, 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 Normal errors consists of the following elements:

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

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right),$$

where μ is the mean and σ^2 is the variance.

- (b) X , a set of p independent variables for each observation, x_1, x_2, \dots, 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 , i.e., $\eta = g(\mu)$. 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, the residual sum of squares = $\sum (y_i - \hat{\mu}_i)^2$

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 ,

$$w = \left(\frac{d\eta}{d\mu}\right)^2.$$

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 .

G02GAF 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 a n by p column orthogonal matrix.

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

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

If R is not of full rank a solution is obtained by means of a singular value decomposition (SVD) of R .

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

where D is a k by k diagonal matrix with non-zero diagonal elements, k being the rank of R and $w^{\frac{1}{2}}X$.

This gives the solution

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

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 residual sum of squares.

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 residual sum of squares, in particular comparing the difference in residual sums of squares between nested models, i.e., when one model is a sub-model of the other.

Let RSS_f be the residual sum of squares for the full model with degrees of freedom ν_f and let RSS_s be the residual sum of squares for the sub-model with degrees of freedom ν_s then:

$$F = \frac{(RSS_s - RSS_f)/(\nu_s - \nu_f)}{RSS_f/\nu_f},$$

has, approximately, a F -distribution with $(\nu_s - \nu_f)$, ν_f degrees of freedom.

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

$$C = R^{-1}R^{-T}\sigma^2 \text{ in the full rank case, otherwise}$$

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

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 i th diagonal elements of H , h_i , give a measure of the influence of the i th values of the independent variables on the fitted regression model. These are sometimes known as leverages.

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

G02GAF also computes the residuals, r :

$$r_i = y_i - \hat{\mu}_i$$

An option allows prior weights, ω_i to be used, this gives a model with:

$$\sigma_i^2 = \frac{\sigma^2}{\omega_i}.$$

In many linear regression models the first term is taken as a mean term or an intercept, i.e., $x_{i,1} = 1$, for $i = 1, 2, \dots, n$; 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 variable 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 G02GAF. 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^* :

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

4 References

- [1] Cook R D and Weisberg S (1982) *Residuals and Influence in Regression* Chapman and Hall
- [2] 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.
- If LINK = 'E', then an exponent link is used.
 If LINK = 'I', then an identity link is used. The user is advised not to use G02GAF with an identity link as G02DAF provides a more efficient way of fitting such a model.
 If LINK = 'L', then a log link is used.
 If LINK = 'S', then a square root link is used.
 If LINK = 'R', then 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.
- If MEAN = 'M' (Mean), a mean term, intercept, will be included in the model.
 If MEAN = 'Z' (Zero), 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.
- If OFFSET = 'Y', then an offset is required and the offsets must be supplied in the 7th column of V.
 If 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.
- If WEIGHT = 'U' (Unweighted), no prior weights are used.
 If WEIGHT = 'W' (Weighted), prior weights are used and weights must be supplied in WT.
- Constraint:* WEIGHT = 'U' or 'W'.
- 5: N** — INTEGER *Input*
On entry: the number of observations, n .
- Constraint:* $N \geq 2$.

- 6:** X(LDX,M) — *real* array *Input*
On entry: X(*i*,*j*) must contain the *i*th observation for the *j*th independent variable, for *i* = 1, 2, ..., *n*; *j* = 1, 2, ..., M.
- 7:** LDX — INTEGER *Input*
On entry: the first dimension of the array X as declared in the (sub)program from which G02GAF is called.
Constraint: LDX ≥ N.
- 8:** M — INTEGER *Input*
On entry: the total number of independent variables.
Constraint: M ≥ 1.
- 9:** ISX(M) — INTEGER array *Input*
On entry: indicates which independent variables are to be included in the model.
 If ISX(*j*) > 0, then the variable contained in the *j*th column of X is included in the regression model.
Constraints:
 ISX(*j*) ≥ 0, for *j* = 1, 2, ..., M.
 If MEAN = 'M', then exactly IP – 1 values of ISX must be > 0.
 If MEAN = 'Z', then 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: IP > 0.
- 11:** Y(N) — *real* array *Input*
On entry: the observations on the dependent variable, *y_i*, for *i* = 1, 2, ..., *n*.
- 12:** WT(*) — *real* array *Input*
On entry:
 If WEIGHT = 'W', then WT must contain the weights to be used with the model, ω_i . If WT(*i*) = 0.0, then the *i*th observation is not included in the model, in which case the effective number of observations is the number of observations with non-zero weights.
 If WEIGHT = 'U', then WT is not referenced and the effective number of observations is *n*.
Constraint: if WEIGHT = 'W', WT(*i*) ≥ 0.0, for *i* = 1, 2, ..., *n*.
- 13:** S — *real* *Input/Output*
On entry: the scale parameter for the model, σ^2 .
 If S = 0.0, then the scale parameter is estimated with the routine using the residual mean square.
Constraint: S ≥ 0.0.
On exit:
 if on input S = 0.0, then S contains the estimated value of the scale parameter, $\hat{\sigma}^2$,
 if on input S ≠ 0.0, then S is unchanged on exit.

- 14:** A — *real* *Input*
On entry:
 If LINK = 'E', then A must contain the power of the exponential.
 If LINK \neq 'E', A is not referenced.
Constraint: if LINK = 'E', A \neq 0.0.
- 15:** RSS — *real* *Output*
On exit: the residual sum of squares for the fitted model.
- 16:** IDF — INTEGER *Output*
On exit: the degrees of freedom associated with the residual sum of squares for the fitted model.
- 17:** B(IP) — *real* array *Output*
On exit: the estimates of the parameters of the generalized linear model, $\hat{\beta}$.
 If MEAN = 'M', then 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 i th positive value in the array ISX.
 If MEAN = 'Z', then B(i) will contain the coefficient of the variable contained in column j of X , where ISX(j) is the i th 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 than 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* 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 $i = 1, 2, \dots, \text{IP}$.
- 20:** COV(IP*(IP+1)/2) — *real* array *Output*
On exit: 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($j \times (j - 1)/2 + i$).
- 21:** V(LDV,IP+7) — *real* array *Input/Output*
On entry:
 If OFFSET = 'N', V need not be set.
 If OFFSET = 'Y', V($i, 7$), for $i = 1, 2, \dots, n$ 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 $i = 1, 2, \dots, n$.
 V($i, 2$) contains the fitted value, $\hat{\mu}_i$, for $i = 1, 2, \dots, n$.
 V($i, 3$) is only included for consistency with other routines. V($i, 3$) = 1.0, for $i = 1, 2, \dots, n$.
 V($i, 4$) contains the working weight, w_i , for $i = 1, 2, \dots, n$.
 V($i, 5$) contains the residual, r_i , for $i = 1, 2, \dots, n$.
 V($i, 6$) contains the leverage, h_i , for $i = 1, 2, \dots, n$.

$V(i,7)$ contains the offset, for $i = 1, 2, \dots, n$. If `OFFSET = 'N'`, then all values will be zero.
 $V(i, j)$ for $j = 8, \dots, IP + 7$, contains the results of the *QR* decomposition or the singular value decomposition.

If the model is not of full rank, i.e., $IRANK < IP$, then 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 G02GAF is called.

Constraint: $LDV \geq N$.

23: TOL — *real* *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 interactions is less than $TOL \times (1.0 + \text{current residual sum of squares})$. This is approximately an absolute precision if the residual sum of squares is small and a relative precision if the residual sum of squares is large.

If $0.0 \leq TOL < \textit{machine precision}$, then the routine will use $10 \times \textit{machine precision}$.

Constraint: $TOL \geq 0.0$.

24: MAXIT — INTEGER *Input*

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

If $MAXIT = 0$, then a default value of 10 is used.

Constraint: $MAXIT \geq 0$.

25: IPRINT — INTEGER *Input*

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

If $IPRINT \leq 0$, then there is no printing.

If $IPRINT > 0$, then 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).

26: EPS — *real* *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 $0.0 \leq EPS < \textit{machine precision}$, then the routine will use $\textit{machine precision}$ instead.

Constraint: $EPS \geq 0.0$.

27: WK(($IP * IP + 3 * IP + 22$)/2) — *real* array *Workspace*

28: IFAIL — INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. Users who are unfamiliar with this parameter should refer to Chapter P01 for details.

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

For this routine, because the values of output parameters may be useful even if IFAIL \neq 0 on exit, users are recommended to set IFAIL to -1 before entry. **It is then 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`).

Errors or warnings specified 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`

A fitted value is at a boundary. This will only occur with `LINK = 'L', 'R' or 'E'`. 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 = 5`

The singular value decomposition has failed to converge. This is a unlikely error exit, see `F02WUF`.

`IFAIL = 6`

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 = 7`

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

`IFAIL = 8`

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

7 Accuracy

The accuracy is determined by TOL as described in Section 5. As the residual sum of squares is a function of μ^2 the accuracy of the $\hat{\beta}$'s will depend on the link used and may be of the order $\sqrt{\text{TOL}}$.

8 Further Comments

None.

9 Example

The model:

$$y = \frac{1}{\beta_1 + \beta_2 x} + \epsilon$$

for a sample of 10 observations.

9.1 Program Text

Note. The listing of the example program presented below uses bold italicised terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```

*      G02GAF Example Program Text
*      Mark 14 Release.  NAG Copyright 1989.
*      .. Parameters ..
      INTEGER          NMAX, MMAX
      PARAMETER        (NMAX=5,MMAX=2)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*      .. Local Scalars ..
      real            A, EPS, RSS, S, TOL
      INTEGER          I, IDF, IFAIL, IP, IPRINT, IRANK, J, M, MAXIT, N
      CHARACTER        LINK, MEAN, OFFSET, WEIGHT
*      .. Local Arrays ..
      real            B(MMAX), COV((MMAX*MMAX+MMAX)/2), SE(MMAX),
+                    V(NMAX,7+MMAX), WK((MMAX*MMAX+3*MMAX+22)/2),
+                    WT(NMAX), X(NMAX,MMAX), Y(NMAX)
      INTEGER          ISX(MMAX)
*      .. External Subroutines ..
      EXTERNAL         G02GAF
*      .. Executable Statements ..
      WRITE (NOUT,*) 'G02GAF Example Program Results'
*      Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) LINK, MEAN, OFFSET, WEIGHT, N, M, S, 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), WT(I)
20          CONTINUE
         ELSE
            DO 40 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), Y(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
        IF (LINK.EQ.'E' .OR. LINK.EQ.'e') READ (NIN,*) A
*      Set control parameters
        EPS = 0.000001e0
        TOL = 0.00005e0
        MAXIT = 10
        IFAIL = -1
*
        CALL G02GAF(LINK,MEAN,OFFSET,WEIGHT,N,X,NMAX,M,ISX,IP,Y,WT,S,A,
+              RSS,IDF,B,IRANK,SE,COV,V,NMAX,TOL,MAXIT,IPRINT,EPS,
+              WK,IFAIL)
*
        IF (IFAIL.EQ.0 .OR. IFAIL.GE.6) THEN
          WRITE (NOUT,*)
          WRITE (NOUT,99999) 'Residual sum of squares = ', RSS,
+          ' Degrees of freedom = ', IDF
          WRITE (NOUT,*)
          WRITE (NOUT,*) '      Estimate      Standard error'
          WRITE (NOUT,*)
          DO 80 I = 1, IP
            WRITE (NOUT,99998) B(I), SE(I)
80      CONTINUE
          WRITE (NOUT,*)
          WRITE (NOUT,*) '      Y      FV      Residual      H'
          WRITE (NOUT,*)
          DO 100 I = 1, N
            WRITE (NOUT,99997) Y(I), V(I,2), V(I,5), V(I,6)
100     CONTINUE
          END IF
        END IF
        STOP
*
99999 FORMAT (1X,A,e12.4,A,I2)
99998 FORMAT (1X,2F14.4)
99997 FORMAT (1X,F7.1,F10.2,F12.4,F10.3)
        END

```

9.2 Program Data

G02GAF Example Program Data

```

'R' 'M' 'N' 'U' 5 1 0.0 0
1.0 25.0
2.0 10.0
3.0 6.0
4.0 4.0
5.0 3.0
1

```

9.3 Program Results

G02GAF Example Program Results

Residual sum of squares = 0.3872E+00 Degrees of freedom = 3

	Estimate	Standard error		
	-0.0239	0.0028		
	0.0638	0.0026		
Y	FV	Residual	H	
25.0	25.04	-0.0387	0.995	
10.0	9.64	0.3613	0.458	
6.0	5.97	0.0320	0.268	
4.0	4.32	-0.3221	0.167	
3.0	3.39	-0.3878	0.112	
