

NAG Toolbox for MATLAB

g02gb

1 Purpose

g02gb fits a generalized linear model with binomial errors.

2 Syntax

```
[dev, idf, b, irank, se, cov, v, ifail] = g02gb(link, mean, offset,
weight, x, isx, ip, y, t, wt, v, tol, maxit, iprint, eps, 'n', n, 'm',
m)
```

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, \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, $\mu = \pi t$, 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: $\log(-\log(1 - \frac{\mu}{t}))$.

(e) a measure of fit, the deviance:

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

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 ,

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

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 .

g02gb 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

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

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

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

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

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

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

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

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

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 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 $\hat{\mu} = g^{-1}(\hat{\eta})$.

g02gb also computes the deviance residuals, r :

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

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 $i = 1, 2, \dots, n$. 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 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 g02gk after using g02gb. Only certain linear combinations of the parameters will have unique estimates, these are known as estimable functions and can be estimated and tested using g02gn.

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

- 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

5.1 Compulsory Input Parameters

1: **link** – string

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** – string

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** – string

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** – string

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: **x(ldx,m) – double array**

ldx, the first dimension of the array, must be at least **n**.

$x(i,j)$ must contain the i th observation for the j th independent variable, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$.

6: **isx(m) – int32 array**

Indicates which independent variables are to be included in the model.

$isx(j) > 0$

The variable contained in the j th column of **x** is included in the regression model.

Constraints:

$isx(j) \geq 0$, for $j = 1, 2, \dots, m$;
 if **mean** = 'M', exactly **ip** – 1 values of **isx** must be > 0 ;
 if **mean** = 'Z', exactly **ip** values of **isx** must be > 0 .

7: **ip – int32 scalar**

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

Constraint: **ip** > 0 .

8: **y(n) – double array**

The observations on the dependent variable, y_i , for $i = 1, 2, \dots, n$.

Constraint: $0.0 \leq y(i) \leq t(i)$, for $i = 1, 2, \dots, n$.

9: **t(n) – double array**

t , the binomial denominator.

Constraint: $t(i) \geq 0.0$, for $i = 1, 2, \dots, n$.

10: **wt(*) – double array**

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

If **weight** = 'W', **wt** must contain the weights to be used in the weighted regression. If $wt(i) = 0.0$, the i th 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: $wt(i) \geq 0.0$ if **weight** = 'W', for $i = 1, 2, \dots, n$.

11: **v(ldv,ip + 7) – double array**

ldv, the first dimension of the array, must be at least **n**.

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.

12: **tol – double scalar**

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 $tol \times (1.0 + \text{Current Deviance})$. This is approximately an absolute precision if the deviance is small and a relative precision if the deviance is large.

If $0.0 \leq \mathbf{tol} < \mathit{machine\ precision}$, the function will use $10 \times \mathit{machine\ precision}$ instead.

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

13: **maxit – int32 scalar**

The maximum number of iterations for the iterative weighted least-squares.

maxit = 0

A default value of 10 is used.

Constraint: $\mathbf{maxit} \geq 0$.

14: **iprint – int32 scalar**

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 x04ab).

15: **eps – double scalar**

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 $0.0 \leq \mathbf{eps} < \mathit{machine\ precision}$, the function will use $\mathit{machine\ precision}$ instead.

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

5.2 Optional Input Parameters

1: **n – int32 scalar**

n , the number of observations.

Constraint: $\mathbf{n} \geq 2$.

2: **m – int32 scalar**

Default: The dimension of the arrays \mathbf{x} , \mathbf{isx} . (An error is raised if these dimensions are not equal.)

m , the total number of independent variables.

Constraint: $\mathbf{m} \geq 1$.

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldv, wk

5.4 Output Parameters

1: **dev – double scalar**

The deviance for the fitted model.

2: **idf – int32 scalar**

The degrees of freedom associated with the deviance for the fitted model.

3: **b(ip) – double array**

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 *i*th 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 *i*th positive value in the array **isx**.

4: **irank – int32 scalar**

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 than **eps** × (largest singular value).

It is possible for the SVD to be carried out but for **irank** to be returned as **ip**.

5: **se(ip) – double array**

The standard errors of the linear parameters.

se(*i*) contains the standard error of the parameter estimate in **b**(*i*), for *i* = 1, 2, ..., **ip**.

6: **cov(ip × (ip + 1)/2) – double array**

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* ≥ *i*, is stored in **cov**(*j* × (*j* - 1)/2 + *i*).

7: **v(ldv, ip + 7) – double array**

Auxiliary information on the fitted model.

v(*i*, 1) contains the linear predictor value, η_i , for *i* = 1, 2, ..., *n*.

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

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

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

v(*i*, 5) contains the deviance residual, r_i , for *i* = 1, 2, ..., *n*.

v(*i*, 6) contains the leverage, h_i , for *i* = 1, 2, ..., *n*.

v(*i*, 7) contains the offset, o_i , for *i* = 1, 2, ..., *n*. If **offset** = 'N', all values will be zero.

v(*i*, *j*) for *j* = 8, ..., **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.

8: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Note: g02gb may return useful information for one or more of the following detected errors or warnings.

ifail = 1

On entry, **n** < 2,
 or **m** < 1,
 or **ldx** < **n**,
 or **ldv** < **n**,
 or **ip** < 1,
 or **link** ≠ 'G', 'P' or 'C'.
 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, **t**(*i*) < 0 for some $i = 1, 2, \dots, n$.

ifail = 5

On entry, **y**(*i*) < 0.0,
 or **y**(*i*) > **t**(*i*) for some $i = 1, 2, \dots, n$.

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

```

link = 'G';
mean = 'M';
offset = 'N';
weight = 'U';
x = [1;
     0;
     -1];
isx = [int32(1)];
ip = int32(2);
y = [19;
     29;
     24];
t = [516;
     560;
     293];
wt = [0];
v = zeros(3, 9);
tol = 5e-05;
maxit = int32(10);
iprint = int32(0);
eps = 1e-06;
[dev, idf, b, irank, se, cov, vOut, ifail] = ...
    g02gb(link, mean, offset, weight, x, isx, ip, y, t, wt, v, tol,
maxit, iprint, eps)

```

```

dev =
    0.0735
idf =
     1
b =
   -2.8682
   -0.4264
irank =
     2
se =
    0.1217
    0.1598
cov =
    0.0148
    0.0014
    0.0255
vOut =
  Columns 1 through 7
   -3.2946   18.4508    0.2371    4.2179    0.1296    0.7687    0
   -2.8682   30.0985    0.1874    5.3367   -0.2070    0.4220    0
   -2.4418   23.4508    0.2153    4.6448    0.1178    0.8093    0
  Columns 8 through 9
   -8.2368    0.4593
    0.5269    6.2573
    0.4586    0.8617
ifail =
     0

```

