NAG Toolbox

nag_correg_glm_binomial (g02gb)

1 Purpose

nag_correg_glm_binomial (g02gb) fits a generalized linear model with binomial errors.

2 Syntax

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

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 23: *offset* and *weight* were removed from the interface; **v**, **wt**, **tol**, **maxit**, **iprint** and **eps** were made optional.

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 arguments 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.

nag_correg_glm_binomial (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

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

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

nag_correg_glm_binomial (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 $i = 1, 2, \ldots, n$. 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 may be obtained by applying constraints to the arguments. These solutions can be obtained by using nag_correg_glm_constrain (g02gk) after using nag_correg_glm_binomial (g02gb). Only certain linear combinations of the arguments will have unique estimates, these are known as estimable functions and can be estimated and tested using nag_correg_glm_estfunc (g02gn).

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

5.1 Compulsory Input Parameters

1: **link** – CHARACTER(1)

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_p** – CHARACTER(1)

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: $\mathbf{x}(ldx, \mathbf{m})$ – REAL (KIND=nag_wp) array

ldx, the first dimension of the array, must satisfy the constraint $ldx \geq \mathbf{n}$.

$\mathbf{x}(i,j)$ must contain the ith observation for the jth independent variable, for $i = 1, 2, \ldots, \mathbf{n}$ and $j = 1, 2, \ldots, \mathbf{m}$.

4: $\mathbf{isx}(\mathbf{m})$ – INTEGER array

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

$\mathbf{isx}(j) > 0$
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.

5: **ip** – INTEGER

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

Constraint: **ip** > 0.

6: **y**(**n**) – REAL (KIND=nag_wp) array

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

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

7: **t**(**n**) – REAL (KIND=nag_wp) array

t, the binomial denominator.

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

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the arrays **y**, **t** and the first dimension of the arrays **x**, **v**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

Constraint: **n** ≥ 2.

2: **m** – INTEGER

Default: the dimension of the array **isx** and the second dimension of the array **x**. (An error is raised if these dimensions are not equal.)

m, the total number of independent variables.

Constraint: **m** ≥ 1.

3: **wt**(:) – REAL (KIND=nag_wp) array

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 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', **wt**$(i) \geq 0.0$, for $i = 1, 2, \ldots, n$.

4: **v**(**n**, **ip** + **7**) – REAL (KIND=nag_wp) array

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.

5: **tol** – REAL (KIND=nag_wp)

Default: 0

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 function will use $10 \times$ ***machine precision*** instead.

Constraint: **tol** ≥ 0.0.

6: **maxit** – INTEGER

Default: 10

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

If **maxit** $= 0$, a default value of 10 is used.

Constraint: **maxit** ≥ 0.

7: **iprint** – INTEGER

Default: 0

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 nag_file_set_unit_advisory (x04ab)).

8: **eps** – REAL (KIND=nag_wp)

Default: 0

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 function will use ***machine precision*** instead.

Constraint: **eps** ≥ 0.0.

5.3 Output Parameters

1: **dev** – REAL (KIND=nag_wp)

The deviance for the fitted model.

2: **idf** – INTEGER

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

3: **b**(**ip**) – REAL (KIND=nag_wp) 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 ith positive value in the array **isx**.

If **mean** = 'Z', $\mathbf{b}(i)$ will contain the coefficient of the variable contained in column j of **x**, where $\mathbf{isx}(j)$ is the ith positive value in the array **isx**.

4: **irank** – INTEGER

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** × (largest singular value).

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

5: **se**(**ip**) – REAL (KIND=nag_wp) array

The standard errors of the linear parameters.

$\mathbf{se}(i)$ contains the standard error of the parameter estimate in $\mathbf{b}(i)$, for $i = 1, 2, \ldots, \mathbf{ip}$.

6: **covar**($\mathbf{ip} \times (\mathbf{ip}+\mathbf{1})/\mathbf{2}$) – REAL (KIND=nag_wp) 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 $\mathbf{b}(i)$ and the parameter estimate given in $\mathbf{b}(j)$, $j \geq i$, is stored in $\mathbf{cov}((j \times (j-1)/2 + i))$.

7: $\mathbf{v}(\mathbf{n}, \mathbf{ip}+\mathbf{7})$ – REAL (KIND=nag_wp) array

Auxiliary information on the fitted model.

$\mathbf{v}(i,1)$ contains the linear predictor value, η_i, for $i = 1, 2, \ldots, n$.

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

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

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

$\mathbf{v}(i,5)$ contains the deviance residual, r_i, for $i = 1, 2, \ldots, n$.

$\mathbf{v}(i,6)$ contains the leverage, h_i, for $i = 1, 2, \ldots, n$.

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

$\mathbf{v}(i,j)$ for $j = 8, \ldots, \mathbf{ip}+7$, 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** – INTEGER

ifail = 0 unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

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

Errors or warnings detected by the function:

ifail = 1

On entry, **n** < 2,
or **m** < 1,
or ldx < **n**,

or $ldv < \mathbf{n}$,
or $\mathbf{ip} < 1$,
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 $\mathbf{maxit} < 0$,
or $\mathbf{tol} < 0.0$,
or $\mathbf{eps} < 0.0$.

ifail $= 2$

On entry, $weight =$ 'W' and a value of $\mathbf{wt} < 0.0$.

ifail $= 3$

On entry, a value of $\mathbf{isx} < 0$,
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, $\mathbf{t}(i) < 0$ for some $i = 1, 2, \ldots, n$.

ifail $= 5$

On entry, $\mathbf{y}(i) < 0.0$,
or $\mathbf{y}(i) > \mathbf{t}(i)$ 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 $= 9$ (*warning*)

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 $= 10$ (*warning*)

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

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

ifail $= -999$

Dynamic memory allocation failed.

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

```
function g02gb_example

fprintf('g02gb example results\n\n');

x = [  1;      0;     -1];
y = [ 19;     29;     24];
t = [516;    560;    293];

[n,m]   = size(x);
isx = ones(m,1,nag_int_name);
ip  = nag_int(m+1);

link   = 'G';
mean_p = 'M';
eps = 1e-6;
tol = 5e-5;

% Fit generalized linear model with Binomial errors
[dev, idf, b, irank, se, covar, v, ifail] = ...
  g02gb( ...
         link, mean_p, x, isx, ip, y, t, 'eps', eps, 'tol', tol);

%  Display results
fprintf('Deviance           = %12.4e\n', dev);
fprintf('Degrees of freedom = %2d\n', idf);
fprintf('\nVariable   Parameter estimate   Standard error\n\n');
ivar = double([1:ip]');
fprintf('%6d%20.4e%20.4e\n',[ivar b se]');
fprintf('\n       n          y        fv     residual       h\n\n');
for j=1:n
  fprintf('%10.1f%10.1f%10.2f%12.4f%10.3f\n',t(j),y(j),v(j,2),v(j,5),v(j,6));
end
```

9.2 Program Results

```
     g02gb example results

Deviance           =   7.3539e-02
Degrees of freedom =  1

Variable   Parameter estimate   Standard error

     1         -2.8682e+00          1.2173e-01
     2         -4.2637e-01          1.5981e-01
```

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