

NAG Toolbox

nag_correg_glm_poisson (g02gc)

1 Purpose

nag_correg_glm_poisson (g02gc) fits a generalized linear model with Poisson errors.

2 Syntax

```
[dev, idf, b, irank, se, cov, v, ifail] = nag_correg_glm_poisson(link, mean, x, isx, ip, y, 'n', n, 'm', m, 'wt', wt, 'a', a, 'v', v, 'tol', tol, 'maxit', maxit, 'iprint', iprint, 'eps', eps)
```

```
[dev, idf, b, irank, se, cov, v, ifail] = g02gc(link, mean, x, isx, ip, y, 'n', n, 'm', m, 'wt', wt, 'a', a, '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**, **eps** and **a** were made optional.

3 Description

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

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

$$\frac{\mu^y e^{-\mu}}{y!}.$$

(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, 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, the deviance:

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

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 ,

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

where $\tau = \sqrt{\mu}$.

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_poisson` (g02gc) 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 arguments estimates, $\hat{\beta}$, are asymptotically Normally distributed with variance-covariance matrix

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

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

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 known as leverages.

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

`nag_correg_glm_poisson` (g02gc) 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 prior weights to be used with 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 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 arguments. These solutions can be obtained by using `nag_correg_glm_constrain` (g02gk) after using `nag_correg_glm_poisson` (g02gc). Only certain linear combinations of the arguments will have unique estimates, these are known as estimable functions, these 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^* :

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

The generalized linear model with Poisson errors can be used to model contingency table data; see Cook and Weisberg (1982) and McCullagh and Nelder (1983).

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

Plackett R L (1974) *The Analysis of Categorical Data* Griffin

5 Parameters

5.1 Compulsory Input Parameters

1: **link** – CHARACTER(1)

Indicates which link function is to be used.

link = 'E'

An exponent 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_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: **x(ldx, m)** – REAL (KIND=nag_wp) array

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

The matrix of all possible independent variables. $\mathbf{x}(i,j)$ must contain the i th element of \mathbf{x} , for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, \mathbf{m}$.

4: **isx(m)** – INTEGER array

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

isx(j) > 0

The variable contained in the j th column of \mathbf{x} is included in the regression model.

Constraints:

isx(j) ≥ 0, for $j = 1, 2, \dots, \mathbf{m}$;

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

y , observations on the dependent variable.

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

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the array **y** 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 provided, **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 **wt** is not provided the effective number of observations is n .

Constraint: if *weight* = 'W', **wt(i)** ≥ 0.0, for $i = 1, 2, \dots, n$.

4: **a** – REAL (KIND=nag_wp)

Default: 0

If **link** = 'E', **a** must contain the power of the exponential.

If **link** \neq 'E', **a** is not referenced.

Constraint: if **a** \neq 0.0, **link** = 'E'.

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

If *offset* = 'N', **v** need not be set.

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

6: **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 $0.0 \leq \mathbf{tol} < \mathbf{machine\ precision}$, the function will use $10 \times \mathbf{machine\ precision}$ instead.

Constraint: **tol** \geq 0.0.

7: **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** \geq 0.

8: **iprint** – INTEGER

Default: 0

Indicates if the printing of information on the iterations is required.

iprint \leq 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 nag_file_set_unit_advisory (x04ab)).

9: **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 $0.0 \leq \mathbf{eps} < \mathbf{machine\ precision}$, the function will use **machine precision** instead.

Constraint: **eps** \geq 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 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** – 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 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)** – REAL (KIND=nag_wp) array
The standard errors of the linear parameters.
se(i) contains the standard error of the parameter estimate in **b**(i), for $i = 1, 2, \dots, \mathbf{ip}$.
- 6: **covar(ip × (ip + 1)/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 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**(($j \times (j - 1)/2 + i$)).
- 7: **v(n, ip + 7)** – REAL (KIND=nag_wp) array
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($i, 3$) contains the variance standardization, $\frac{1}{\tau_i}$, for $i = 1, 2, \dots, n$.
v($i, 4$) contains the square root of the working weight, $w_i^{\frac{1}{2}}$, for $i = 1, 2, \dots, n$.
v($i, 5$) contains the deviance 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, o_i , for $i = 1, 2, \dots, n$. If **offset** = 'N', all values will be zero.
v(i, j) for $j = 8, \dots, \mathbf{ip} + 7$, 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_poisson` (g02gc) 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* < **n**,
 or **ip** < 1,
 or **link** ≠ 'E', 'I', 'L', 'S' or 'R',
 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 $i = 1, 2, \dots, n$.

ifail = 5

A fitted value is at the boundary, i.e., $\hat{\mu} = 0.0$. This may occur if there are y values of 0.0 and the model is too complex 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 (*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 = 9 (*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 depends 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 only be a function of **tol**. **tol** should therefore be set smaller than the accuracy required for $\hat{\beta}$.

8 Further Comments

None.

9 Example

A 3 by 5 contingency table given by Plackett (1974) is analysed by fitting terms for rows and columns. The table is:

141	67	114	79	39
131	66	143	72	35
36	14	38	28	16

9.1 Program Text

```
function g02gc_example
fprintf('g02gc example results\n\n');
x = [1, 0, 0, 1, 0, 0, 0, 0;
     1, 0, 0, 0, 1, 0, 0, 0;
     1, 0, 0, 0, 0, 1, 0, 0;
     1, 0, 0, 0, 0, 0, 1, 0;
     1, 0, 0, 0, 0, 0, 0, 1;
     0, 1, 0, 1, 0, 0, 0, 0;
     0, 1, 0, 0, 1, 0, 0, 0;
     0, 1, 0, 0, 0, 1, 0, 0;
     0, 1, 0, 0, 0, 0, 1, 0;
     0, 1, 0, 0, 0, 0, 0, 1;
     0, 0, 1, 1, 0, 0, 0, 0;
     0, 0, 1, 0, 1, 0, 0, 0;
     0, 0, 1, 0, 0, 1, 0, 0;
     0, 0, 1, 0, 0, 0, 1, 0;
     0, 0, 1, 0, 0, 0, 0, 1];
y = [141; 67; 114; 79; 39;
     131; 66; 143; 72; 35;
     36; 14; 38; 28; 16];
[n,m] = size(x);
isx = ones(m,1,nag_int_name);
ip = nag_int(m+1);
```



```

link = 'L';
mean_p = 'M';
eps = 1e-6;
maxit = nag_int(20);

% Fit generalized linear model with Poisson errors
[dev, idf, b, irank, se, covar, v, ifail] = ...
g02gc( ...
    link, mean_p, x, isx, ip, y, 'eps', eps, 'maxit', maxit);

% 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%16.4f%20.4f\n',[ivar b se]');
fprintf('\n      y          fv      residual          h\n\n');
for j=1:n
    fprintf('%7.1f%10.2f%12.4f%10.3f\n',y(j),v(j,2),v(j,5),v(j,6));
end

```

9.2 Program Results

g02gc example results

```

Deviance                = 9.0379e+00
Degrees of freedom = 8

```

Variable	Parameter estimate	Standard error
----------	--------------------	----------------

1	2.5977	0.0258
2	1.2619	0.0438
3	1.2777	0.0436
4	0.0580	0.0668
5	1.0307	0.0551
6	0.2910	0.0732
7	0.9876	0.0559
8	0.4880	0.0675
9	-0.1996	0.0904

y	fv	residual	h
141.0	132.99	0.6875	0.604
67.0	63.47	0.4386	0.514
114.0	127.38	-1.2072	0.596
79.0	77.29	0.1936	0.532
39.0	38.86	0.0222	0.482
131.0	135.11	-0.3553	0.608
66.0	64.48	0.1881	0.520
143.0	129.41	1.1749	0.601
72.0	78.52	-0.7465	0.537
35.0	39.48	-0.7271	0.488
36.0	39.90	-0.6276	0.393
14.0	19.04	-1.2131	0.255
38.0	38.21	-0.0346	0.382
28.0	23.19	0.9675	0.282
16.0	11.66	1.2028	0.206
