

NAG Library Function Document

nag_glm_binomial (g02gbc)

1 Purpose

nag_glm_binomial (g02gbc) fits a generalized linear model with binomial errors.

2 Specification

```
#include <nag.h>
#include <nagg02.h>

void nag_glm_binomial (Nag_Link link, Nag_IncludeMean mean, Integer n,
    const double x[], Integer tdx, Integer m, const Integer sx[],
    Integer ip, const double y[], const double binom_t[], const double wt[],
    const double offset[], double *dev, double *df, double b[],
    Integer *rank, double se[], double cov[], double v[], Integer tdv,
    double tol, Integer max_iter, Integer print_iter, const char *outfile,
    double eps, NagError *fail)
```

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 function $\eta = g(\mu)$, linking the linear predictor, η and the mean of the distribution, $\mu = \pi t$.

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:

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

The linear 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 ,

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

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_glm_binomial (g02gbc) finds a QR decomposition of $w^{\frac{1}{2}}X$, i.e.,

$$w^{\frac{1}{2}}X = QR \text{ where } R \text{ is a } p \text{ by } p \text{ triangular matrix and } Q \text{ is an } n \text{ by } p \text{ column orthogonal matrix.}$$

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 nonzero 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 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:

$C = R^{-1}R^{-T}$ 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 $Hw^{\frac{1}{2}}z$ 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 known as leverages.

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

nag_glm_binomial (g02gbc) 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 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 arguments. These solutions can be obtained by using `nag_glm_tran_model` (g02gkc) after using `nag_glm_binomial` (g02gbc).

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_glm_est_func` (g02gnc).

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

5 Arguments

- 1: **link** – Nag_Link *Input*
On entry: indicates which link function is to be used.
link = Nag_Logistic
 A logistic link is used.
link = Nag_Probit
 A probit link is used.
link = Nag_Compl
 A complementary log-log link is used.
Constraint: **link** = Nag_Logistic, Nag_Probit or Nag_Compl.
- 2: **mean** – Nag_IncludeMean *Input*
On entry: indicates if a mean term is to be included.
mean = Nag_MeanInclude
 A mean term, (intercept), will be included in the model.
mean = Nag_MeanZero
 The model will pass through the origin, zero point.
Constraint: **mean** = Nag_MeanInclude or Nag_MeanZero.
- 3: **n** – Integer *Input*
On entry: the number of observations, n .
Constraint: $n \geq 2$.
- 4: **x**[**n** × **tdx**] – const double *Input*
On entry: **x**[($i - 1$) × **tdx** + $j - 1$] must contain the i th observation for the j th independent variable, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, \mathbf{m}$.
- 5: **tdx** – Integer *Input*
On entry: the stride separating matrix column elements in the array **x**.
Constraint: **tdx** \geq **m**.

- 6: **m** – Integer *Input*
On entry: the total number of independent variables.
Constraint: $\mathbf{m} \geq 1$.
- 7: **sx[m]** – const Integer *Input*
On entry: indicates which independent variables are to be included in the model. If $\mathbf{sx}[j - 1] > 0$, then the variable contained in the j th column of **x** is included in the regression model.
Constraints:
 $\mathbf{sx}[j - 1] \geq 0$, for $j = 1, 2, \dots, \mathbf{m}$;
 if **mean** = Nag_MeanInclude, then exactly **ip** – 1 values of **sx** must be > 0 ;
 if **mean** = Nag_MeanZero, then exactly **ip** values of **sx** must be > 0 .
- 8: **ip** – Integer *Input*
On entry: the number p of independent variables in the model, including the mean or intercept if present.
Constraint: $\mathbf{ip} > 0$.
- 9: **y[n]** – const double *Input*
On entry: observations on the dependent variable, y_i , for $i = 1, 2, \dots, n$.
Constraint: $0.0 \leq \mathbf{y}[i - 1] \leq \mathbf{binom_t}[i - 1]$, for $i = 1, 2, \dots, n$.
- 10: **binom_t[n]** – const double *Input*
On entry: the binomial denominator, t .
Constraint: $\mathbf{binom_t}[i] \geq 0.0$, for $i = 1, 2, \dots, n$.
- 11: **wt[n]** – const double *Input*
On entry: if weighted estimates are required, then **wt** must contain the weights to be used. Otherwise **wt** need not be defined and may be set to **NULL**.
 If $\mathbf{wt}[i - 1] = 0.0$, 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 positive weights.
 If **wt** is **NULL**, then the effective number of observations is n .
Constraint: **wt** is **NULL** or $\mathbf{wt}[i - 1] \geq 0.0$, for $i = 1, 2, \dots, n$.
- 12: **offset[n]** – const double *Input*
On entry: if an offset is required then **offset** must contain the values of the offset o . Otherwise **offset** must be supplied as **NULL**.
- 13: **dev** – double * *Output*
On exit: the deviance for the fitted model.
- 14: **df** – double * *Output*
On exit: the degrees of freedom associated with the deviance for the fitted model.
- 15: **b[ip]** – double *Output*
On exit: $\mathbf{b}[i - 1]$, $i = 1, \dots, \mathbf{ip}$ contains the estimates of the arguments of the generalized linear model, $\hat{\beta}$.

If **mean** = Nag_MeanInclude, then **b**[0] will contain the estimate of the mean argument and **b**[*i*] will contain the coefficient of the variable contained in column *j* of **x**, where **sx**[*j* – 1] is the *i*th positive value in the array **sx**.

If **mean** = Nag_MeanZero, then **b**[*i* – 1] will contain the coefficient of the variable contained in column *j* of **x**, where **sx**[*j* – 1] is the *i*th positive value in the array **sx**.

16: **rank** – Integer * Output

On exit: the rank of the independent variables.

If the model is of full rank, then **rank** = **ip**.

If the model is not of full rank, then **rank** is an estimate of the rank of the independent variables. **rank** 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 **rank** to be returned as **ip**.

17: **se**[**ip**] – double Output

On exit: the standard errors of the linear arguments.

se[*i* – 1] contains the standard error of the parameter estimate in **b**[*i* – 1], for *i* = 1, 2, ..., **ip**.

18: **cov**[**ip** × (**ip** + 1)/2] – double Output

On exit: the **ip** × (**ip** + 1)/2 elements of **cov** contain 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* ≥ *i*, is stored in **cov**[*j*(*j* + 1)/2 + *i*], for *i* = 0, 1, ..., **ip** – 1 and *j* = *i*, ..., **ip** – 1.

19: **v**[**n** × **tdv**] – double Output

On exit: auxiliary information on the fitted model.

v[(*i* – 1) × **tdv**], contains the linear predictor value, η_i , for *i* = 1, 2, ..., *n*.

v[(*i* – 1) × **tdv** + 1], contains the fitted value, $\hat{\mu}_i$, for *i* = 1, 2, ..., *n*.

v[(*i* – 1) × **tdv** + 2], contains the variance standardization, τ_i , for *i* = 1, 2, ..., *n*.

v[(*i* – 1) × **tdv** + 3], contains the working weight, w_i , for *i* = 1, 2, ..., *n*.

v[(*i* – 1) × **tdv** + 4], contains the deviance residual, r_i , for *i* = 1, 2, ..., *n*.

v[(*i* – 1) × **tdv** + 5], contains the leverage, h_i , for *i* = 1, 2, ..., *n*.

v[(*i* – 1) × **tdv** + *j* – 1], for *j* = 7, 8, ..., **ip** + 6, contains the results of the *QR* decomposition or the singular value decomposition.

If the model is not of full rank, i.e., **rank** < **ip**, then the first **ip** rows of columns 7 to **ip** + 6 contain the P^* matrix.

20: **tdv** – Integer Input

On entry: the stride separating matrix column elements in the array **v**.

Constraint: **tdv** ≥ **ip** + 6.

21: **tol** – double 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**× (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 $0.0 \leq \mathbf{tol} < \mathbf{machine\ precision}$, then the function will use $10 \times \mathbf{machine\ precision}$.

Constraint: **tol** ≥ 0.0.

- 22: **max_iter** – Integer *Input*
On entry: the maximum number of iterations for the iterative weighted least squares.
 If **max_iter** = 0, then a default value of 10 is used.
Constraint: **max_iter** ≥ 0.
- 23: **print_iter** – Integer *Input*
On entry: indicates if the printing of information on the iterations is required and the rate at which printing is produced.
print_iter ≤ 0
 There is no printing.
print_iter > 0
 The following items are printed every **print_iter** iterations:
 (i) the deviance,
 (ii) the current estimates, and
 (iii) if the weighted least squares equations are singular then this is indicated.
- 24: **outfile** – const char * *Input*
On entry: a null terminated character string giving the name of the file to which results should be printed. If **outfile** is **NULL** or an empty string then the `stdout` stream is used. Note that the file will be opened in the append mode.
- 25: **eps** – double *Input*
On entry: the value of **eps** is used to decide if the independent variables are of full rank and, if not, what the rank of the independent variables is. The smaller the value of **eps** the stricter the criterion for selecting the singular value decomposition.
 If $0.0 \leq \mathbf{eps} < \mathit{machine\ precision}$, then the function will use *machine precision* instead.
Constraint: **eps** ≥ 0.0.
- 26: **fail** – NagError * *Input/Output*
 The NAG error argument (see Section 3.6 in the Essential Introduction).

6 Error Indicators and Warnings

NE_2_INT_ARG_LT

On entry, **tdv** = $\langle value \rangle$ while **ip** = $\langle value \rangle$. These arguments must satisfy **tdv** ≥ **ip** + 6.

On entry, **tdx** = $\langle value \rangle$ while **m** = $\langle value \rangle$. These arguments must satisfy **tdx** ≥ **m**.

NE_2_REAL_ARG_GT

On entry, $\mathbf{y}[\langle value \rangle] = \langle value \rangle$ while $\mathbf{binom_t}[\langle value \rangle] = \langle value \rangle$. These arguments must satisfy $\mathbf{y}[\langle value \rangle] \leq \mathbf{binom_t}[\langle value \rangle]$.

NE_ALLOC_FAIL

Dynamic memory allocation failed.

NE_BAD_PARAM

On entry, argument **link** had an illegal value.

On entry, argument **mean** had an illegal value.

NE_INT_ARG_LT

On entry, **ip** = $\langle value \rangle$.

Constraint: **ip** ≥ 1 .

On entry, **m** = $\langle value \rangle$.

Constraint: **m** ≥ 1 .

On entry, **max_iter** must not be less than 0: **max_iter** = $\langle value \rangle$.

On entry, **n** = $\langle value \rangle$.

Constraint: **n** ≥ 2 .

On entry, **sx**[$\langle value \rangle$] must not be less than 0: **sx**[$\langle value \rangle$] = $\langle value \rangle$.

NE_IP_GT_OBSERV

Argument **ip** is greater than the effective number of observations.

NE_IP_INCOMP_SX

Argument **ip** is incompatible with arguments **mean** and **sx**.

NE_LSQ_ITER_NOT_CONV

The iterative weighted least squares has failed to converge in **max_iter** = $\langle value \rangle$ iterations. The value of **max_iter** could be increased but it may be advantageous to examine the convergence using the **print_iter** 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.

NE_NOT_APPEND_FILE

Cannot open file $\langle string \rangle$ for appending.

NE_NOT_CLOSE_FILE

Cannot close file $\langle string \rangle$.

NE_RANK_CHANGED

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.

NE_REAL_ARG_LT

On entry, **binom_t**[$\langle value \rangle$] must not be less than 0.0: **binom_t**[$\langle value \rangle$] = $\langle value \rangle$.

On entry, **eps** must not be less than 0.0: **eps** = $\langle value \rangle$.

On entry, **tol** must not be less than 0.0: **tol** = $\langle value \rangle$.

On entry, **wt**[$\langle value \rangle$] must not be less than 0.0: **wt**[$\langle value \rangle$] = $\langle value \rangle$.

On entry, **y**[$\langle value \rangle$] must not be less than 0.0: **y**[$\langle value \rangle$] = $\langle value \rangle$.

NE_SVD_NOT_CONV

The singular value decomposition has failed to converge.

NE_VALUE_AT_BOUNDARY_B

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

NE_ZERO_DOF_ERROR

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 adjusted deviance is a function of $\log \mu$ the accuracy of the $\hat{\beta}$'s will be a function of **tol**. **tol** should therefore be set to a smaller value than the accuracy required for $\hat{\beta}$.

8 Parallelism and Performance

Not applicable.

9 Further Comments

None.

10 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).

10.1 Program Text

```

/* nag_glm_binomial (g02gbc) Example Program.
 *
 * Copyright 1996 Numerical Algorithms Group.
 *
 * Mark 4, 1996.
 * Mark 8 revised, 2004.
 *
 */

#include <nag.h>
#include <stdio.h>
#include <nag_stdlib.h>
#include <ctype.h>
#include <nagg02.h>

#define X(I, J) x[(I) *tdx + J]
#define V(I, J) v[(I) *tdv + J]

int main(void)
{
    Integer          exit_status = 0, i, *ivar = 0, j, m, max_iter, n, nvar,
                    print_iter;
    Integer          rank, tdv, tdx;
    Nag_IncludeMean mean;
    Nag_Link         link;
    Nag_Boolean      weight;
    char             nag_enum_arg[40];
    double           dev, df, eps, tol;
    double           *beta = 0, *binom = 0, *cov = 0, *offsetptr = 0, *se = 0;
    double           *v = 0, *wt = 0, *wtptr, *x = 0, *y = 0;
    NagError         fail;

    INIT_FAIL(fail);

    printf("nag_glm_binomial (g02gbc) Example Program Results\n");
    /* Skip heading in data file */
    scanf("%*[\n]");
    scanf("%39s", nag_enum_arg);
    /* nag_enum_name_to_value (x04nac).
     * Converts NAG enum member name to value

```

```

*/
link = (Nag_Link) nag_enum_name_to_value(nag_enum_arg);
scanf("%39s", nag_enum_arg);
mean = (Nag_IncludeMean) nag_enum_name_to_value(nag_enum_arg);
scanf("%39s", nag_enum_arg);
weight = (Nag_Boolean) nag_enum_name_to_value(nag_enum_arg);
scanf("%ld %ld %ld", &n, &m, &print_iter);

if (n >= 2 && m >= 1)
{
    if (!(binom = NAG_ALLOC(n, double)) ||
        !(wt = NAG_ALLOC(n, double)) ||
        !(x = NAG_ALLOC(n*m, double)) ||
        !(y = NAG_ALLOC(n, double)) ||
        !(ivar = NAG_ALLOC(m, Integer)))
    {
        printf("Allocation failure\n");
        exit_status = -1;
        goto END;
    }
    tdx = m;
}
else
{
    printf("Invalid n or m.\n");
    exit_status = 1;
    return exit_status;
}
if (weight)
{
    wtptr = wt;
    for (i = 0; i < n; i++)
    {
        for (j = 0; j < m; j++)
            scanf("%lf", &X(i, j));
        scanf("%lf%lf%lf", &y[i], &binom[i], &wt[i]);
    }
}
else
{
    wtptr = (double *) 0;
    for (i = 0; i < n; i++)
    {
        for (j = 0; j < m; j++)
            scanf("%lf", &X(i, j));
        scanf("%lf%lf", &y[i], &binom[i]);
    }
}
for (j = 0; j < m; j++)
    scanf("%ld", &ivar[j]);

/* Calculate nvar */
nvar = 0;
for (i = 0; i < m; i++)
    if (ivar[i] > 0) nvar += 1;
if (mean == Nag_MeanInclude)
    nvar += 1;

if (!(beta = NAG_ALLOC(nvar, double)) ||
    !(v = NAG_ALLOC((n)*(nvar+6), double)) ||
    !(se = NAG_ALLOC(nvar, double)) ||
    !(cov = NAG_ALLOC(nvar*(nvar+1)/2, double)))
{
    printf("Allocation failure\n");
    exit_status = -1;
    goto END;
}
tdv = nvar+6;

/* Set other control parameters */

```

```

max_iter = 10;
tol = 5e-5;
eps = 1e-6;

/* nag_glm_binomial (g02gbc).
 * Fits a generalized linear model with binomial errors
 */
nag_glm_binomial(link, mean, n, x, tdx, m, ivar, nvar, y, binom, wtptr,
                 offsetptr, &dev, &df, beta, &rank,
                 se, cov, v, tdv, tol, max_iter, print_iter, "",
                 eps, &fail);

if (fail.code == NE_NOERROR || fail.code == NE_SVD_NOT_CONV ||
    fail.code == NE_LSQ_ITER_NOT_CONV ||
    fail.code == NE_RANK_CHANGED || fail.code == NE_ZERO_DOF_ERROR)
{
    if (fail.code != NE_NOERROR) {
        printf("Error from nag_glm_binomial (g02gbc).\n%s\n",
              fail.message);
    }
    printf("\nDeviance = %13.4e\n", dev);
    printf("Degrees of freedom = %3.1f\n\n", df);
    printf("      Estimate      Standard error\n\n");
    for (i = 0; i < nvar; i++)
        printf("%14.4f%14.4f\n", beta[i], se[i]);
    printf("\n");
    printf("      binom      y      fitted value      Residual      Leverage\n\n");
    for (i = 0; i < n; ++i)
    {
        printf("%10.1f%7.1f%10.2f%12.4f%10.3f\n", binom[i], y[i],
              V(i, 1), V(i, 4), V(i, 5));
    }
}
else
{
    printf("Error from nag_glm_binomial (g02gbc).\n%s\n",
          fail.message);
    exit_status = 1;
    goto END;
}
END:
NAG_FREE(binom);
NAG_FREE(wt);
NAG_FREE(x);
NAG_FREE(y);
NAG_FREE(beta);
NAG_FREE(v);
NAG_FREE(se);
NAG_FREE(cov);
NAG_FREE(ivar);

return exit_status;
}

```

10.2 Program Data

```

nag_glm_binomial (g02gbc) Example Program Data
Nag_Logistic Nag_MeanInclude Nag_FALSE 3 1 0
1.0 19. 516.
0.0 29. 560.
-1.0 24. 293.
1

```

10.3 Program Results

nag_glm_binomial (g02gbc) Example Program Results

Deviance = 7.3539e-02
Degrees of freedom = 1.0

Estimate	Standard error
-2.8682	0.1217
-0.4264	0.1598

binom	y	fitted value	Residual	Leverage
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
