NAG Toolbox for MATLAB

g02ga

1 Purpose

g02ga fits a generalized linear model with normal errors.

2 Syntax

```
[s, rss, idf, b, irank, se, cov, v, ifail] = g02ga(link, mean, offset, weight, x, isx, ip, y, wt, s, a, v, tol, maxit, iprint, eps, 'n', n, 'm', m)
```

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, $x_1, x_2, \ldots, x_p$.

(c) a linear model:

$$\eta = \sum \beta_j x_j.$$

(d) a link between the linear predictor, η, and the mean of the distribution, 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.

g02ga 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 an n by p column orthogonal matrix.

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

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

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

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

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

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

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 $\mathbf{rss}_f$ be the residual sum of squares for the full model with degrees of freedom ν_f and let $\mathbf{rss}_s$ be the residual sum of squares for the sub-model with degrees of freedom ν_s then:

$$F = \frac{\left(\mathrm{RSS}_s - \mathrm{RSS}_f\right) / \left(\nu_s - \nu_f\right)}{\mathrm{RSS}_f / \nu_f},$$

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:

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

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

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

g02ga 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 $i = 1, 2, \ldots, n$; 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 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 be may be obtained by applying constraints to the parameters. These solutions can be obtained by using g02gk after using g02ga. 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^*:

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

4 References

Cook R D and Weisberg S 1982 *Residuals and Influence in Regression* Chapman and Hall

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 = 'E'

An exponent link is used.

link = 'I'

An identity link is used. You are advised not to use g02ga with an identity link as g02da provides a more efficient way of fitting such a model.

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 – 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**.

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

6: **isx(m) – int32 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:

$\mathbf{isx}(j) \geq 0$, for $i = 1,2,\ldots,\mathbf{m}$;
if **mean** = 'M', exactly **ip** $- 1$ values of **isx** must be > 0;
if **mean** = 'Z', exactly **ip** values of **isx** must be > 0.

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 $i = 1,2,\ldots,n$.

9: **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 with the model, ω_i. If $\mathbf{wt}(i) = 0.0$, 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: $\mathbf{wt}(i) \geq 0.0$ if **weight** = 'W', for $i = 1,2,\ldots,n$.

10: **s – double scalar**

The scale parameter for the model, σ^2.

s $= 0.0$

The scale parameter is estimated with the function using the residual mean square.

Constraint: **s** ≥ 0.0.

11: **a – double scalar**

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

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

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

12: **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', $\mathbf{v}(i, 7)$, for $i = 1, 2, \ldots, n$, must contain the offset values o_i. All other values need not be set.

13: **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 interactions is less than **tol** $\times$ (1.0 + 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** $<$ ***machine precision***, g02ga will use $10 \times$ ***machine precision***.

Constraint: **tol** ≥ 0.0.

14: **maxit – int32 scalar**

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

maxit $= 0$

A default value of 10 is used.

Constraint: **maxit** ≥ 0.

15: **iprint – int32 scalar**

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

iprint ≤ 0

There is no printing.

iprint > 0

Every **iprint** iteration, the following is 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 x04ab).

16: **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 $0.0 \leq \mathbf{eps} <$ ***machine precision***, the function will use ***machine precision*** instead.

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

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the arrays **y**, **wt**, **offset**, **v**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

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

2: **m – int32 scalar**

Default: The dimension of the arrays **x**, **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: **s – double scalar**

If on input $\mathbf{s} = 0.0$, **s** contains the estimated value of the scale parameter, $\hat{\sigma}^2$.

If on input $\mathbf{s} \neq 0.0$, **s** is unchanged on exit.

2: **rss – double scalar**

The residual sum of squares for the fitted model.

3: **idf – int32 scalar**

The degrees of freedom associated with the residual sum of squares for the fitted model.

4: **b(ip) – double array**

The estimates of the parameters of the generalized linear model, $\hat{\beta}$.

If $\mathbf{mean} = \text{'M'}$, $\mathbf{b}(1)$ will contain the estimate of the mean parameter and $\mathbf{b}(i+1)$ 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**.

If $\mathbf{mean} = \text{'Z'}$, $\mathbf{b}(i)$ 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**.

5: **irank – int32 scalar**

The rank of the independent variables.

If the model is of full rank, $\mathbf{irank} = \mathbf{ip}$.

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** $\times$ (largest singular value). It is possible for the SVD to be carried out but for **irank** to be returned as **ip**.

6: **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 $i = 1, 2, \ldots,$ **ip**.

7: **cov(ip** $\times$ **(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 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**$(j \times (j-1)/2 + i)$.

8: **v(ldv,ip** $+$ **7) – double array**

Auxiliary information on the fitted model.

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

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

v$(i, 3)$ is only included for consistency with other functions. **v**$(i, 3) = 1.0$, for $i = 1, 2, \ldots, n$.

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

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

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

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

v(i,j), for $j = 8, \ldots,$ **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.

9: **ifail – int32 scalar**

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

6 Error Indicators and Warnings

Note: g02ga 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** $\neq$ 'E','I','L','S' or 'R',
or **s** < 0.0,
or **link** $=$ 'E' and **a** $= 0.0$,
or **mean** $\neq$ 'M' or 'Z',
or **weight** $\neq$ 'U' or 'W',
or **offset** $\neq$ '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 an unlikely error exit, see f02wu.

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 you 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 $\hat{\beta}$ will depend on the link used and may be of the order $\sqrt{\textbf{tol}}$.

8 Further Comments

None.

9 Example

```
link = 'R';
mean = 'M';
offset = 'N';
weight = 'U';
x = [1;
     2;
     3;
     4;
     5];
isx = [int32(1)];
ip = int32(2);
y = [25;
     10;
     6;
     4;
```

```
     3];
wt = [0];
s = 0;
a = 0;
v = zeros(5, 9);
tol = 5e-05;
maxit = int32(10);
iprint = int32(0);
eps = 1e-06;
[sOut, rss, idf, b, irank, se, cov, vOut, ifail] = ...
      g02ga(link, mean, offset, weight, x, isx, ip, y, wt, s, a, v, tol,
maxit, iprint, eps)
```

```
sOut =
    0.1291
rss =
    0.3872
idf =
           3
b =
   -0.0239
    0.0638
irank =
           2
se =
    0.0028
    0.0026
cov =
   1.0e-05 *
    0.7723
   -0.7177
    0.6957
vOut =
  Columns 1 through 7
    0.0399   25.0387    1.0000 -626.9347   -0.0387    0.9954         0
    0.1037    9.6387    1.0000  -92.9036    0.3613    0.4577         0
    0.1676    5.9680    1.0000  -35.6173    0.0320    0.2681         0
    0.2314    4.3221    1.0000  -18.6803   -0.3221    0.1666         0
    0.2952    3.3878    1.0000  -11.4769   -0.3878    0.1121         0
  Columns 8 through 9
  635.1594  655.2205
    0.1038  136.2024
    0.0398    0.4013
    0.0209    0.3166
    0.0128    0.2597
ifail =
           0
```