NAG Toolbox for MATLAB

g11ca

1 Purpose

g11ca returns parameter estimates for the conditional logistic analysis of stratified data, for example, data from case-control studies and survival analyses.

2 Syntax

```
[dev, b, se, sc, cov, nca, nct, ifail] = g11ca(ns, z, isz, ic, isi, b,
tol, maxit, 'n', n, 'm', m, 'ip', ip, 'iprint', iprint)
```

3 Description

In the analysis of binary data, the logistic model is commonly used. This relates the probability of one of the outcomes, say $y = 1$, to p explanatory variates or covariates by

$$\mathrm{Prob}(y = 1) = \frac{\exp\left(\alpha + z^{\mathrm{T}}\beta\right)}{1 + \exp(\alpha + z^{\mathrm{T}}\beta)},$$

where β is a vector of unknown coefficients for the covariates z and α is a constant term. If the observations come from different strata or groups, α would vary from strata to strata. If the observed outcomes are independent then the ys follow a Bernoulli distribution, i.e., a binomial distribution with sample size one and the model can be fitted as a generalized linear model with binomial errors.

In some situations the number of observations for which $y = 1$ may not be independent. For example, in epidemiological research, case-control studies are widely used in which one or more observed cases are matched with one or more controls. The matching is based on fixed characteristics such as age and sex, and is designed to eliminate the effect of such characteristics in order to more accurately determine the effect of other variables. Each case-control group can be considered as a stratum. In this type of study the binomial model is not appropriate, except if the strata are large, and a conditional logistic model is used. This considers the probability of the cases having the observed vectors of covariates given the set of vectors of covariates in the strata. In the situation of one case per stratum, the conditional likelihood for n_{s} strata can be written as

$$L = \prod_{i=1}^{n_s} \frac{\exp\left(z_i^{\mathrm{T}}\beta\right)}{\left[\sum_{l \in S_i} \exp\left(z_l^{\mathrm{T}}\beta\right)\right]}, \tag{1}$$

where S_i is the set of observations in the ith stratum, with associated vectors of covariates z_l, $l \in S_i$, and z_i is the vector of covariates of the case in the ith stratum. In the general case of c_i cases per strata then the full conditional likelihood is

$$L = \prod_{i=1}^{n_s} \frac{\exp\left(s_i^{\mathrm{T}}\beta\right)}{\left[\sum_{l \in C_i} \exp\left(s_l^{\mathrm{T}}\beta\right)\right]}, \tag{2}$$

where s_i is the sum of the vectors of covariates for the cases in the ith stratum and s_l, $l \in C_i$ refer to the sum of vectors of covariates for all distinct sets of c_i observations drawn from the ith stratum. The conditional likelihood can be maximized by a Newton–Raphson procedure. The covariances of the parameter estimates can be estimated from the inverse of the matrix of second derivatives of the logarithm of the conditional likelihood, while the first derivatives provide the score function, $U_j(\beta)$, for $j = 1, 2, \ldots, p$, which can be used for testing the significance of parameters.

If the strata are not small, C_i can be large so to improve the speed of computation, the algorithm in Howard 1972 and described by Krailo and Pike 1984 is used.

A second situation in which the above conditional likelihood arises is in fitting Cox's proportional hazard model (see g12ba) in which the strata refer to the risk sets for each failure time and where the failures are cases. When ties are present in the data g12ba uses an approximation. For an exact estimate, the data can be expanded using g12za to create the risk sets/strata and g11ca used.

4 References

Cox D R 1972b Regression models in life tables (with discussion) *J. Roy. Statist. Soc. Ser. B* **34** 187–220

Cox D R and Hinkley D V 1974 *Theoretical Statistics* Chapman and Hall

Howard S 1972 Remark on the paper by Cox,D.R. (1972): Regression methods *J. R. Statist. Soc.* **B 34** and life tables 187–220

Krailo M D and Pike M C 1984 Algorithm AS 196. Conditional multivariate logistic analysis of stratified case-control studies *Appl. Statist.* **33** 95–103

Smith P G, Pike M C, Hill P, Breslow N E and Day N E 1981 Algorithm AS 162. Multivariate conditional logistic analysis of stratum-matched case-control studies *Appl. Statist.* **30** 190–197

5 Parameters

5.1 Compulsory Input Parameters

1: **ns – int32 scalar**

the number of strata, n_s.

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

2: **z(ldz,m) – double array**

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

The *i*th row must contain the covariates which are associated with the *i*th observation.

3: **isz(m) – int32 array**

Indicates which subset of covariates are to be included in the model.

If $\mathbf{isz}(j) \geq 1$, the *j*th covariate is included in the model.

If $\mathbf{isz}(j) = 0$, the *j*th covariate is excluded from the model and not referenced.

Constraint: $\mathbf{isz}(j) \geq 0$ and at least one value must be nonzero.

4: **ic(n) – int32 array**

Indicates whether the *i*th observation is a case or a control.

If $\mathbf{ic}(i) = 0$, indicates that the *i*th observation is a case.

If $\mathbf{ic}(i) = 1$, indicates that the *i*th observation is a control.

Constraint: $\mathbf{ic}(i) = 0$ or 1,, for $i = 1, 2, \ldots, \mathbf{n}$.

5: **isi(n) – int32 array**

Stratum indicators which also allow data points to be excluded from the analysis.

If $\mathbf{isi}(i) = k$, indicates that the *i*th observation is from the *k*th stratum, where $k = 1, 2, \ldots, \mathbf{ns}$.

If $\mathbf{isi}(i) = 0$, indicates that the *i*th observation is to be omitted from the analysis.

Constraint: $0 \leq \mathbf{isi}(i) \leq \mathbf{ns}$ and more than **ip** values of $\mathbf{isi}(i) > 0$,, for $i = 1, 2, \ldots, \mathbf{n}$.

6: **b(ip) – double array**

Initial estimates of the covariate coefficient parameters β. $\mathbf{b}(j)$ must contain the initial estimate of the coefficent of the covariate in **z** corresponding to the *j*th nonzero value of **isz**.

Suggested value: in many cases an initial value of zero for **b**(j) may be used. For another suggestion see Section 8.

7: **tol – double scalar**

Indicates the accuracy required for the estimation. Convergence is assumed when the decrease in deviance is less than **tol** $\times$ (1.0 + CurrentDeviance). This corresponds approximately to an absolute accuracy if the deviance is small and a relative accuracy if the deviance is large.

Constraint: **tol** $\geq 10 \times$ ***machine precision***.

8: **maxit – int32 scalar**

The maximum number of iterations required for computing the estimates. If **maxit** is set to 0 then the standard errors, the score functions and the variance-covariance matrix are computed for the input value of β in **b** but β is not updated.

Constraint: **maxit** ≥ 0.

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the arrays **ic**, **isi**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

Constraint: **n** ≥ 2.

2: **m – int32 scalar**

Default: The dimension of the arrays **z**, **isz**. (An error is raised if these dimensions are not equal.)

the number of covariates in array **z**.

Constraint: **m** ≥ 1.

3: **ip – int32 scalar**

Default: The dimension of the arrays **b**, **se**, **sc**. (An error is raised if these dimensions are not equal.)

p, the number of covariates included in the model as indicated by **isz**.

Constraint: **ip** $\geq$ 1and **ip** = number of nonzero values of **isz**

4: **iprint – int32 scalar**

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

iprint ≤ 0

No printing.

iprint ≥ 1

The deviance and the current estimates are printed every **iprint** iterations. When printing occurs the output is directed to the current advisory message unit (see x04ab).

Suggested value: **iprint** = 0

Default: 0

5.3 Input Parameters Omitted from the MATLAB Interface

ldz, wk, lwk

5.4 Output Parameters

1: **dev – double scalar**

The deviance, that is, $-2 \times$,(maximized log marginal likelihood).

2: **b(ip) – double array**

Suggested value: in many cases an initial value of zero for $\mathbf{b}(j)$ may be used. For another suggestion see Section 8.

$\mathbf{b}(j)$ contains the estimate $\hat{\beta}_i$ of the coefficient of the covariate stored in the ith column of **z** where i is the jth nonzero value in the array **isz**.

3: **se(ip) – double array**

$\mathbf{se}(j)$ is the asymptotic standard error of the estimate contained in $\mathbf{b}(j)$ and score function in $\mathbf{sc}(j)$, for $j = 1, 2, \ldots, \mathbf{ip}$.

4: **sc(ip) – double array**

$\mathbf{sc}(j)$ is the value of the score function $U_j(\beta)$ for the estimate contained in $\mathbf{b}(j)$.

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

The variance-covariance matrix of the parameter estimates in **b** stored in packed form by column, i.e., the covariance between the parameter estimates given in $\mathbf{b}(i)$ and $\mathbf{b}(j)$, $j \geq i$, is given in $\mathbf{cov}(j(j-1)/2 + i)$.

6: **nca(ns) – int32 array**

$\mathbf{nca}(i)$ contains the number of cases in the ith stratum, for $i = 1, 2, \ldots, \mathbf{ns}$.

7: **nct(ns) – int32 array**

$\mathbf{nct}(i)$ contains the number of controls in the ith stratum, for $i = 1, 2, \ldots, \mathbf{ns}$.

8: **ifail – int32 scalar**

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

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail = 1

On entry, $\mathbf{m} < 1$,
or $\mathbf{n} < 2$,
or $\mathbf{ns} < 1$,
or $\mathbf{ip} < 1$,
or $\mathbf{ldz} < \mathbf{n}$,
or $\mathbf{tol} < 10 \times$ ***machine precision***,
or $\mathbf{maxit} < 0$.

ifail = 2

On entry, $\mathbf{isz}(i) < 0$, for some i,
or the value of **ip** is incompatible with **isz**,
or $\mathbf{ic}(i) \neq 1$ or 0.
or $\mathbf{isi}(i) < 0$ or $\mathbf{isi}(i) > \mathbf{ns}$,
or the number of values of $\mathbf{isz}(i) > 0$ is greater than or equal to n_0, the number of observations excluding any with $\mathbf{isi}(i) = 0$.

ifail = 3

The value of **lwk** is too small.

ifail = 4

Overflow has been detected. Try using different starting values.

ifail = 5

The matrix of second partial derivatives is singular. Try different starting values or include fewer covariates.

ifail = 6

Convergence has not been achieved in **maxit** iterations. The progress towards convergence can be examined by using a nonzero value of **iprint**. Any non-convergence may be due to a linear combination of covariates being monotonic with time.

Full results are returned.

7 Accuracy

The accuracy is specified by **tol**.

8 Further Comments

The other models described in Section 3 can be fitted using the generalized linear modelling functions g02gb and g02gc.

The case with one case per stratum can be analysed by having a dummy response variable y such that $y = 1$ for a case and $y = 0$ for a control, and fitting a Poisson generalized linear model with a log link and including a factor with a level for each strata. These models can be fitted by using g02gc.

g11ca uses mean centering, which involves subtracting the means from the covariables prior to computation of any statistics. This helps to minimize the effect of outlying observations and accelerates convergence. In order to reduce the risk of the sums computed by Howard's algorithm becoming too large, the scaling factor described in Krailo and Pike 1984 is used.

If the initial estimates are poor then there may be a problem with overflow in calculating $\exp\left(\beta^{\mathrm{T}} z_i\right)$ or there may be non-convergence. Reasonable estimates can often be obtained by fitting an unconditional model.

9 Example

```
ns = int32(2);
z = [0, 1;
     1, 2;
     0, 1;
     1, 3;
     0, 1;
     1, 0;
     0, 2];
isz = [int32(1);
     int32(1)];
ic = [int32(0);
     int32(0);
     int32(1);
     int32(1);
     int32(0);
     int32(1);
     int32(1)];
isi = [int32(1);
```

```
     int32(1);
     int32(1);
     int32(1);
     int32(2);
     int32(2);
     int32(2)];
b = [0;
     0];
tol = 1e-05;
maxit = int32(10);
[dev, bOut, se, sc, cov, nca, nct, ifail] = g11ca(ns, z, isz, ic, isi, b,
tol, maxit)
```

```
dev =
    5.4749
bOut =
   -0.5223
   -0.2674
se =
    1.3901
    0.8473
sc =
   1.0e-05 *
   -0.4794
   -0.7901
cov =
    1.9325
   -0.2317
    0.7180
nca =
           2
           1
nct =
           2
           2
ifail =
           0
```