NAG Library Routine Document

G11CAF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G11CAF returns parameter estimates for the conditional logistic analysis of stratified data, for example, data from case-control studies and survival analyses.

2 Specification

```
SUBROUTINE G11CAF (N, M, NS, Z, LDZ, ISZ, IP, IC, ISI, DEV, B, SE, SC, COV,     &
                   NCA, NCT, TOL, MAXIT, IPRINT, WK, LWK, IFAIL)

INTEGER            N, M, NS, LDZ, ISZ(M), IP, IC(N), ISI(N), NCA(NS),           &
                   NCT(NS), MAXIT, IPRINT, LWK, IFAIL
REAL (KIND=nag_wp) Z(LDZ,M), DEV, B(IP), SE(IP), SC(IP), COV(IP*(IP+1)/2),      &
                   TOL, WK(LWK)
```

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

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

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 G12BAF) 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 G12BAF uses an approximation. For an exact estimate, the data can be expanded using G12ZAF to create the risk sets/strata and G11CAF used.

4 References

Cox D R (1972) 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

1: N – INTEGER *Input*

On entry: n, the number of observations.

Constraint: $\mathrm{N} \geq 2$.

2: M – INTEGER *Input*

On entry: the number of covariates in array Z.

Constraint: $\mathrm{M} \geq 1$.

3: NS – INTEGER *Input*

On entry: the number of strata, n_s.

Constraint: $\mathrm{NS} \geq 1$.

4: Z(LDZ,M) – REAL (KIND=nag_wp) array *Input*

On entry: the ith row must contain the covariates which are associated with the ith observation.

5: LDZ – INTEGER *Input*

On entry: the first dimension of the array Z as declared in the (sub)program from which G11CAF is called.

Constraint: $\mathrm{LDZ} \geq \mathrm{N}$.

6: ISZ(M) – INTEGER array *Input*

On entry: indicates which subset of covariates are to be included in the model.

If $\mathrm{ISZ}(j) \geq 1$, the jth covariate is included in the model.

If $\mathrm{ISZ}(j) = 0$, the jth covariate is excluded from the model and not referenced.

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

7: IP – INTEGER *Input*

On entry: p, the number of covariates included in the model as indicated by ISZ.

Constraint: IP ≥ 1 and IP = number of nonzero values of ISZ .

8: IC(N) – INTEGER array *Input*

On entry: indicates whether the ith observation is a case or a control.

If IC$(i) = 0$, indicates that the ith observation is a case.

If IC$(i) = 1$, indicates that the ith observation is a control.

Constraint: IC$(i) = 0$ or 1, for $i = 1, 2, \ldots,$ N.

9: ISI(N) – INTEGER array *Input*

On entry: stratum indicators which also allow data points to be excluded from the analysis.

If ISI$(i) = k$, indicates that the ith observation is from the kth stratum, where $k = 1, 2, \ldots,$ NS.

If ISI$(i) = 0$, indicates that the ith observation is to be omitted from the analysis.

Constraint: $0 \leq$ ISI$(i) \leq$ NS and more than IP values of ISI$(i) > 0$, for $i = 1, 2, \ldots,$ N.

10: DEV – REAL (KIND=nag_wp) *Output*

On exit: the deviance, that is, $-2 \times$, (maximized log marginal likelihood).

11: B(IP) – REAL (KIND=nag_wp) array *Input/Output*

On entry: initial estimates of the covariate coefficient parameters β. B(j) must contain the initial estimate of the coefficent of the covariate in Z corresponding to the jth 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.

On exit: 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.

12: SE(IP) – REAL (KIND=nag_wp) array *Output*

On exit: SE(j) is the asymptotic standard error of the estimate contained in B(j) and score function in SC(j), for $j = 1, 2, \ldots,$ IP.

13: SC(IP) – REAL (KIND=nag_wp) array *Output*

On exit: SC(j) is the value of the score function $U_j(\beta)$ for the estimate contained in B(j).

14: COV(IP $\times$ (IP + 1)/2) – REAL (KIND=nag_wp) array *Output*

On exit: 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 B(i) and B(j), $j \geq i$, is given in COV$(j(j-1)/2 + i)$.

15: NCA(NS) – INTEGER array *Output*

On exit: NCA(i) contains the number of cases in the ith stratum, for $i = 1, 2, \ldots,$ NS.

16: NCT(NS) – INTEGER array *Output*

On exit: NCT(i) contains the number of controls in the ith stratum, for $i = 1, 2, \ldots,$ NS.

17: TOL – REAL (KIND=nag_wp) *Input*

On entry: indicates the accuracy required for the estimation. Convergence is assumed when the decrease in deviance is less than $\mathrm{TOL} \times (1.0 + \mathrm{CurrentDeviance})$. This corresponds approximately to an absolute accuracy if the deviance is small and a relative accuracy if the deviance is large.

Constraint: $\mathrm{TOL} \geq 10 \times$ ***machine precision***.

18: MAXIT – INTEGER *Input*

On entry: 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: $\mathrm{MAXIT} \geq 0$.

19: IPRINT – INTEGER *Input*

On entry: indicates if the printing of information on the iterations is required.

$\mathrm{IPRINT} \leq 0$
No printing.

$\mathrm{IPRINT} \geq 1$
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 X04ABF).

Suggested value: $\mathrm{IPRINT} = 0$.

20: WK(LWK) – REAL (KIND=nag_wp) array *Workspace*
21: LWK – INTEGER *Input*

On entry: the dimension of the array WK as declared in the (sub)program from which G11CAF is called.

Constraint: $\mathrm{LWK} \geq pn_0 + (c_{\mathrm{m}} + 1)(p + 1)(p + 2)/2 + c_{\mathrm{m}}$, where n_0 is the number of observations included in the model, i.e., the number of observations for which $\mathrm{ISI}(i) \neq 0$ and c_{m} is the maximum number of observations in any stratum.

22: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this parameter, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

On exit: $\mathrm{IFAIL} = 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

6 Error Indicators and Warnings

If on entry $\mathrm{IFAIL} = 0$ or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

$\mathrm{IFAIL} = 1$

On entry, $\mathrm{M} < 1$,
or $\mathrm{N} < 2$,
or $\mathrm{NS} < 1$,
or $\mathrm{IP} < 1$,

or LDZ $<$ N,
or TOL $<$ 10 × ***machine precision***,
or MAXIT $<$ 0.

IFAIL $= 2$

On entry, ISZ$(i) < 0$, for some i,
or the value of IP is incompatible with ISZ,
or IC$(i) \neq 1$ or 0.
or ISI$(i) < 0$ or ISI$(i) >$ NS,
or the number of values of ISZ$(i) > 0$ is greater than or equal to n_0, the number of observations excluding any with 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 routines G02GBF and G02GCF.

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 G02GCF.

G11CAF 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

The data was used for illustrative purposes by Smith *et al.* (1981) and consists of two strata and two covariates. The data is input, the model is fitted and the results are printed.

9.1 Program Text

```
    Program g11cafe

!     G11CAF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g11caf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: dev, tol
      Integer                          :: cm, i, ifail, ip, iprint, ldz, lwk,  &
                                          m, maxit, n, n0, ns
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: b(:), cov(:), sc(:), se(:), wk(:),   &
                                          z(:,:)
      Integer, Allocatable             :: cnt(:), ic(:), isi(:), isz(:),       &
                                          nca(:), nct(:)
!     .. Intrinsic Procedures ..
      Intrinsic                        :: count, maxval, sum
!     .. Executable Statements ..
      Write (nout,*) 'G11CAF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in problem size and control parameters
      Read (nin,*) n, m, ns, maxit, iprint, tol

      ldz = n
      Allocate (z(ldz,m),isz(m),ic(n),isi(n),nca(ns),nct(ns),cnt(ns))

!     Read in data
      Read (nin,*)(isi(i),ic(i),z(i,1:m),i=1,n)

!     Read in variable inclusion flags
      Read (nin,*) isz(1:m)

!     Calculate IP
      ip = count(isz(1:m)>0)

!     Calculate number of observations in the model and maximum number of
!     observations in any stratum
      cnt(1:ns) = 0
      Do i = 1, n
        If (isi(i)>0 .And. isi(i)<=ns) Then
          cnt(isi(i)) = cnt(isi(i)) + 1
        End If
      End Do
      cm = maxval(cnt(1:ns))
      n0 = sum(cnt(1:ns))

      lwk = ip*n0 + (cm+1)*(ip+1)*(ip+2)/2 + cm
      Allocate (b(ip),se(ip),sc(ip),cov(ip*(ip+1)/2),wk(lwk))

!     Read in initial estimate for B
      Read (nin,*) b(1:ip)

!     Calculate parameter estimates
      ifail = 0
      Call g11caf(n,m,ns,z,ldz,isz,ip,ic,isi,dev,b,se,sc,cov,nca,nct,tol, &
        maxit,iprint,wk,lwk,ifail)

!     Display results
      Write (nout,99999) ' Deviance = ', dev
      Write (nout,*)
      Write (nout,*) ' Strata     No. Cases   No. Controls'
      Write (nout,*)
      Write (nout,99998)(i,nca(i),nct(i),i=1,ns)
      Write (nout,*)
      Write (nout,*) ' Parameter      Estimate', '       Standard Error'
      Write (nout,*)
      Write (nout,99997)(i,b(i),se(i),i=1,ip)

99999 Format (A,E13.4)
99998 Format (3X,I3,10X,I2,10X,I2)
99997 Format (I6,10X,F8.4,10X,F8.4)
    End Program g11cafe
```

9.2 Program Data

```
G11CAF Example Program Data
7 2 2 10 0 1.0E-5 :: N,M,NS,MAXIT,IPRINT,TOL
1 0 0 1
1 0 1 2
1 1 0 1
1 1 1 3
2 0 0 1
2 1 1 0
2 1 0 2           :: End of ISI,IC,Z
1 1               :: ISZ
0.0 0.0           :: B
```

9.3 Program Results

```
 G11CAF Example Program Results

 Deviance =    0.5475E+01

  Strata     No. Cases   No. Controls

     1           2           2
     2           1           2

  Parameter      Estimate       Standard Error

     1           -0.5223            1.3901
     2           -0.2674            0.8473
```