

NAG Toolbox

nag_surviv_coxmodel (g12ba)

1 Purpose

nag_surviv_coxmodel (g12ba) returns parameter estimates and other statistics that are associated with the Cox proportional hazards model for fixed covariates.

2 Syntax

```
[dev, b, se, sc, cov, res, nd, tp, sur, ifail] = nag_surviv_coxmodel(offset, ns, z, isz, t, ic, omega, isi, b, ndmax, tol, maxit, iprint, 'n', n, 'm', m, 'ip', ip)
```

```
[dev, b, se, sc, cov, res, nd, tp, sur, ifail] = g12ba(offset, ns, z, isz, t, ic, omega, isi, b, ndmax, tol, maxit, iprint, 'n', n, 'm', m, 'ip', ip)
```

3 Description

The proportional hazard model relates the time to an event, usually death or failure, to a number of explanatory variables known as covariates. Some of the observations may be right-censored, that is the exact time to failure is not known, only that it is greater than a known time.

Let t_i , for $i = 1, 2, \dots, n$, be the failure time or censored time for the i th observation with the vector of p covariates z_i . It is assumed that censoring and failure mechanisms are independent. The hazard function, $\lambda(t, z)$, is the probability that an individual with covariates z fails at time t given that the individual survived up to time t . In the Cox proportional hazards model (see Cox (1972)) $\lambda(t, z)$ is of the form:

$$\lambda(t, z) = \lambda_0(t) \exp(z^T \beta + \omega)$$

where λ_0 is the base-line hazard function, an unspecified function of time, β is a vector of unknown arguments and ω is a known offset.

Assuming there are ties in the failure times giving $n_d < n$ distinct failure times, $t_{(1)} < \dots < t_{(n_d)}$ such that d_i individuals fail at $t_{(i)}$, it follows that the marginal likelihood for β is well approximated (see Kalbfleisch and Prentice (1980)) by:

$$L = \prod_{i=1}^{n_d} \frac{\exp(s_i^T \beta + \omega_i)}{\left[\sum_{l \in R(t_{(i)})} \exp(z_l^T \beta + \omega_l) \right]^{d_i}} \quad (1)$$

where s_i is the sum of the covariates of individuals observed to fail at $t_{(i)}$ and $R(t_{(i)})$ is the set of individuals at risk just prior to $t_{(i)}$, that is, it is all individuals that fail or are censored at time $t_{(i)}$ along with all individuals that survive beyond time $t_{(i)}$. The maximum likelihood estimates (MLEs) of β , given by $\hat{\beta}$, are obtained by maximizing (1) using a Newton–Raphson iteration technique that includes step halving and utilizes the first and second partial derivatives of (1) which are given by equations (2) and (3) below:

$$U_j(\beta) = \frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n_d} [s_{ji} - d_i \alpha_{ji}(\beta)] = 0 \quad (2)$$

for $j = 1, 2, \dots, p$, where s_{ji} is the j th element in the vector s_i and

$$\alpha_{ji}(\beta) = \frac{\sum_{l \in R(t_{(i)})} z_{jl} \exp(z_l^T \beta + \omega_l)}{\sum_{l \in R(t_{(i)})} \exp(z_l^T \beta + \omega_l)}$$

Similarly,

$$I_{hj}(\beta) = -\frac{\partial^2 \ln L}{\partial \beta_h \partial \beta_j} = \sum_{i=1}^{n_d} d_i \gamma_{hji} \quad (3)$$

where

$$\gamma_{hji} = \frac{\sum_{l \in R(t_{(i)})} z_{hl} z_{jl} \exp(z_l^T \beta + \omega_l)}{\sum_{l \in R(t_{(i)})} \exp(z_l^T \beta + \omega_l)} - \alpha_{hi}(\beta) \alpha_{ji}(\beta), \quad h, j = 1, \dots, p.$$

$U_j(\beta)$ is the j th component of a score vector and $I_{hj}(\beta)$ is the (h, j) element of the observed information matrix $I(\beta)$ whose inverse $I(\beta)^{-1} = [I_{hj}(\beta)]^{-1}$ gives the variance-covariance matrix of β .

It should be noted that if a covariate or a linear combination of covariates is monotonically increasing or decreasing with time then one or more of the β_j 's will be infinite.

If $\lambda_0(t)$ varies across ν strata, where the number of individuals in the k th stratum is n_k , for $k = 1, 2, \dots, \nu$ with $n = \sum_{k=1}^{\nu} n_k$, then rather than maximizing (1) to obtain $\hat{\beta}$, the following marginal likelihood is maximized:

$$L = \prod_{k=1}^{\nu} L_k, \quad (4)$$

where L_k is the contribution to likelihood for the n_k observations in the k th stratum treated as a single sample in (1). When strata are included the covariate coefficients are constant across strata but there is a different base-line hazard function λ_0 .

The base-line survivor function associated with a failure time $t_{(i)}$, is estimated as $\exp(-\hat{H}(t_{(i)}))$, where

$$\hat{H}(t_{(i)}) = \sum_{t_{(j)} \leq t_{(i)}} \left(\frac{d_i}{\sum_{l \in R(t_{(j)})} \exp(z_l^T \hat{\beta} + \omega_l)} \right), \quad (5)$$

where d_i is the number of failures at time $t_{(i)}$. The residual for the l th observation is computed as:

$$r(t_l) = \hat{H}(t_l) \exp(z_l^T \hat{\beta} + \omega_l)$$

where $\hat{H}(t_l) = \hat{H}(t_{(i)})$, $t_{(i)} \leq t_l < t_{(i+1)}$. The deviance is defined as $-2 \times (\text{logarithm of marginal likelihood})$. There are two ways to test whether individual covariates are significant: the differences between the deviances of nested models can be compared with the appropriate χ^2 -distribution; or, the asymptotic normality of the parameter estimates can be used to form z tests by dividing the estimates by their standard errors or the score function for the model under the null hypothesis can be used to form z tests.

4 References

- Cox D R (1972) Regression models in life tables (with discussion) *J. Roy. Statist. Soc. Ser. B* **34** 187–220
- Gross A J and Clark V A (1975) *Survival Distributions: Reliability Applications in the Biomedical Sciences* Wiley
- Kalbfleisch J D and Prentice R L (1980) *The Statistical Analysis of Failure Time Data* Wiley

5 Parameters

5.1 Compulsory Input Parameters

1: **offset** – CHARACTER(1)

Indicates if an offset is to be used.

offset = 'Y'

An offset must be included in **omega**.

offset = 'N'

No offset is included in the model.

Constraint: **offset** = 'Y' or 'N'.

2: **ns** – INTEGER

The number of strata. If **ns** > 0 then the stratum for each observation must be supplied in **isi**.

Constraint: **ns** ≥ 0.

3: **z**(*ldz*, **m**) – REAL (KIND=nag_wp) array

ldz, the first dimension of the array, must satisfy the constraint *ldz* ≥ **n**.

The *i*th row must contain the covariates which are associated with the *i*th failure time given in **t**.

4: **isz**(**m**) – INTEGER array

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

isz(*j*) ≥ 1

The *j*th covariate is included in the model.

isz(*j*) = 0

The *j*th covariate is excluded from the model and not referenced.

Constraint: **isz**(*j*) ≥ 0 and at least one and at most $n_0 - 1$ elements of **isz** must be nonzero where n_0 is the number of observations excluding any with zero value of **isi**.

5: **t**(**n**) – REAL (KIND=nag_wp) array

The vector of *n* failure censoring times.

6: **ic**(**n**) – INTEGER array

The status of the individual at time *t* given in **t**.

ic(*i*) = 0

The *i*th individual has failed at time **t**(*i*).

ic(*i*) = 1

The *i*th individual has been censored at time **t**(*i*).

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

7: **omega**(:) – REAL (KIND=nag_wp) array

The dimension of the array **omega** must be at least **n** if **offset** = 'Y', and at least 1 otherwise

If **offset** = 'Y', the offset, ω_i , for $i = 1, 2, \dots, \mathbf{n}$. Otherwise **omega** is not referenced.

8: **isi**(:) – INTEGER array

The dimension of the array **isi** must be at least **n** if **ns** > 0, and at least 1 otherwise

If **ns** > 0, the stratum indicators which also allow data points to be excluded from the analysis.

If $\mathbf{ns} = 0$, \mathbf{isi} is not referenced.

$$\mathbf{isi}(i) = k$$

The i th data point is in the k th stratum, where $k = 1, 2, \dots, \mathbf{ns}$.

$$\mathbf{isi}(i) = 0$$

The i th data point is omitted from the analysis.

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

9: $\mathbf{b}(\mathbf{ip})$ – REAL (KIND=nag_wp) array

Suggested value: in many cases an initial value of zero for $\mathbf{b}(j)$ may be used. For other suggestions see Section 9.

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

10: \mathbf{ndmax} – INTEGER

The dimension of the array \mathbf{tp} and the first dimension of the array \mathbf{sur} .

Constraint: $\mathbf{ndmax} \geq$ the number of distinct failure times. This is returned in \mathbf{nd} .

11: \mathbf{tol} – REAL (KIND=nag_wp)

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

Constraint: $\mathbf{tol} \geq 10 \times \text{machine precision}$.

12: \mathbf{maxit} – INTEGER

The maximum number of iterations to be used for computing the estimates. If \mathbf{maxit} is set to 0 then the standard errors, score functions, variance-covariance matrix and the survival function are computed for the input value of β in \mathbf{b} but β is not updated.

Constraint: $\mathbf{maxit} \geq 0$.

13: \mathbf{iprint} – INTEGER

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

$$\mathbf{iprint} \leq 0$$

No printing.

$$\mathbf{iprint} \geq 1$$

The deviance and the current estimates are printed every \mathbf{iprint} iterations. When printing occurs the output is directed to the current advisory message unit (see `nag_file_set_unit_advisory` (x04ab)).

5.2 Optional Input Parameters

1: \mathbf{n} – INTEGER

Default: the dimension of the arrays \mathbf{t} , \mathbf{ic} and the first dimension of the array \mathbf{z} . (An error is raised if these dimensions are not equal.)

n , the number of data points.

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

2: \mathbf{m} – INTEGER

Default: the dimension of the array \mathbf{isz} and the second dimension of the array \mathbf{z} . (An error is raised if these dimensions are not equal.)

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

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

3: **ip** – INTEGER

Default: the dimension of the array **b**.

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

Constraints:

$\mathbf{ip} \geq 1$;

\mathbf{ip} = number of nonzero values of **isz**.

5.3 Output Parameters

1: **dev** – REAL (KIND=nag_wp)

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

2: **b(ip)** – REAL (KIND=nag_wp) array

Suggested value: in many cases an initial value of zero for **b(j)** may be used. For other suggestions see Section 9.

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

3: **se(ip)** – REAL (KIND=nag_wp) array

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

4: **sc(ip)** – REAL (KIND=nag_wp) array

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

5: **covar(ip × (ip + 1)/2)** – REAL (KIND=nag_wp) 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 **b(i)** and **b(j)**, $j \geq i$, is stored in **cov(j(j - 1)/2 + i)**.

6: **res(n)** – REAL (KIND=nag_wp) array

The residuals, $r(t_l)$, for $l = 1, 2, \dots, \mathbf{n}$.

7: **nd** – INTEGER

The number of distinct failure times.

8: **tp(ndmax)** – REAL (KIND=nag_wp) array

tp(i) contains the i th distinct failure time, for $i = 1, 2, \dots, \mathbf{nd}$.

9: **sur(ndmax, :)** – REAL (KIND=nag_wp) array

The second dimension of the array **sur** will be $\max(\mathbf{ns}, 1)$.

If $\mathbf{ns} = 0$, **sur(i, 1)** contains the estimated survival function for the i th distinct failure time.

If $\mathbf{ns} > 0$, **sur(i, k)** contains the estimated survival function for the i th distinct failure time in the k th stratum.

10: **ifail** – INTEGER

ifail = 0 unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail = 1

On entry, **offset** \neq 'Y' or 'N',
 or **m** < 1,
 or **n** < 2,
 or **ns** < 0,
 or **ip** < 1,
 or **ldz** < **n**,
 or **tol** < $10 \times$ *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 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,
 or all observations are censored, i.e., **ic**(*i*) = 1 for all *i*,
 or **ndmax** is too small.

ifail = 3

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

ifail = 4

Overflow has been detected. Try using different starting values.

ifail = 5 (*warning*)

Convergence has not been achieved in **maxit** iterations. The progress toward 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.

ifail = 6 (*warning*)

In the current iteration 10 step halvings have been performed without decreasing the deviance from the previous iteration. Convergence is assumed.

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 is specified by `tol`.

8 Further Comments

`nag_surviv_coxmodel` (g12ba) 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.

If the initial estimates are poor then there may be a problem with overflow in calculating $\exp(\beta^T z_i)$ or there may be non-convergence. Reasonable estimates can often be obtained by fitting an exponential model using `nag_correg_glm_poisson` (g02gc).

9 Example

The data are the remission times for two groups of leukemia patients (see page 242 of Gross and Clark (1975)). A dummy variable indicates which group they come from. An initial estimate is computed using the exponential model and then the Cox proportional hazard model is fitted and parameter estimates and the survival function are printed.

9.1 Program Text

```
function g12ba_example

fprintf('g12ba example results\n\n');

% Data
n = 42;
m = 1;
z      = zeros(n,m);
ic     = zeros(n,1,nag_int_name);
ic(31:end) = 1;
z(22:end) = 1;
isz    = [nag_int(1)];

t = [ 1;    1;    2;    2;    3;    4;    4;
      5;    5;    8;    8;    8;    8;   11;
     11;   12;   12;   15;   17;   22;   23;
      6;    6;    6;    7;   10;   13;   16;
     22;   23;    6;    9;   10;   11;   17;
     19;   20;   25;   32;   32;   34;   35];

% Parameters for Cox proportional hazards model
offset = 'No-offset';
ns     = nag_int(0);
omega  = [0];
isi    = [nag_int(0)];
b(1)   = [0];
ndmax  = nag_int(n);
tol    = 5e-05;
maxit  = nag_int(20);
iprint = nag_int(0);

% Fit Cox proportional hazards model
[dev, b, se, sc, covar, res, nd, tp, sur, ifail] = ...
    g12ba( ...
        offset, ns, z, isz, t, ic, omega, isi, b, ndmax, tol, maxit, iprint);

% Display results
ns = max(ns,1);
fprintf(' Parameter          Estimate          Standard Error\n\n');
ivar = [1:m]';
```

```

fprintf('%6d%18.4f%18.4f\n',[ivar b se]');
fprintf('\n Deviance = %13.3e\n\n', dev);
fprintf('      Time      Survivor Function\n\n');
results = [tp(1:nd) sur(1:nd,1:ns)];
fprintf('%7.0f%16.4f\n', results');

```

9.2 Program Results

g12ba example results

| Parameter | Estimate | Standard Error |
|-----------|----------|----------------|
| 1 | -1.5092 | 0.4096 |

Deviance = 1.728e+02

| Time | Survivor Function |
|------|-------------------|
| 1 | 0.9640 |
| 2 | 0.9264 |
| 3 | 0.9065 |
| 4 | 0.8661 |
| 5 | 0.8235 |
| 6 | 0.7566 |
| 7 | 0.7344 |
| 8 | 0.6506 |
| 10 | 0.6241 |
| 11 | 0.5724 |
| 12 | 0.5135 |
| 13 | 0.4785 |
| 15 | 0.4447 |
| 16 | 0.4078 |
| 17 | 0.3727 |
| 22 | 0.2859 |
| 23 | 0.1908 |
