

## **NAG Library Chapter Introduction**

### **g12 – Survival Analysis**

#### **Contents**

<b>1</b>	<b>Scope of the Chapter</b> .....	<b>2</b>
<b>2</b>	<b>Background to the Problems</b> .....	<b>2</b>
2.1	Introduction to Terminology .....	2
2.2	Rank Statistics .....	2
2.3	Estimating the Survivor Function and Hazard Plotting .....	3
2.4	Proportional Hazard Models .....	4
2.5	Cox’s Proportional Hazard Model .....	4
<b>3</b>	<b>Recommendations on Choice and Use of Available Functions</b> .....	<b>4</b>
<b>4</b>	<b>Functionality Index</b> .....	<b>5</b>
<b>5</b>	<b>Auxiliary Functions Associated with Library Function Arguments</b> .....	<b>5</b>
<b>6</b>	<b>Functions Withdrawn or Scheduled for Withdrawal</b> .....	<b>5</b>
<b>7</b>	<b>References</b> .....	<b>5</b>

## 1 Scope of the Chapter

This chapter is concerned with statistical techniques used in the analysis of survival/reliability/failure time data.

Other chapters contain functions which are also used to analyse this type of data. Chapter g02 contains generalized linear models, Chapter g07 contains functions to fit distribution models, and Chapter g08 contains rank based methods.

## 2 Background to the Problems

### 2.1 Introduction to Terminology

This chapter is concerned with the analysis on the time,  $t$ , to a single event. This type of analysis occurs commonly in two areas. In medical research it is known as survival analysis and is often the time from the start of treatment to the occurrence of a particular condition or of death. In engineering it is concerned with reliability and the analysis of failure times, that is how long a component can be used until it fails. In this chapter the time  $t$  will be referred to as the **failure time**.

Let the probability density function of the failure time be  $f(t)$ , then the **survivor function**,  $S(t)$ , which is the probability of surviving to at least time  $t$ , is given by

$$S(t) = \int_t^{\infty} f(\tau) d\tau = 1 - F(t)$$

where  $F(t)$  is the cumulative density function. The **hazard function**,  $\lambda(t)$ , is the probability that failure occurs at time  $t$  given that the individual survived up to time  $t$ , and is given by

$$\lambda(t) = f(t)/S(t).$$

The **cumulative hazard rate** is defined as

$$\Lambda(t) = \int_0^t \lambda(\tau) d\tau,$$

hence  $S(t) = e^{-\Lambda(t)}$ .

It is common in survival analysis for some of the data to be **right-censored**. That is, the exact failure time is not known, only that failure occurred after a known time. This may be due to the experiment being terminated before all the individuals have failed, or an individual being removed from the experiment for a reason not connected with effects being tested in the experiment. The presence of censored data leads to complications in the analysis.

### 2.2 Rank Statistics

There are a number of different rank statistics described in the literature, the most common being the logrank statistic. All of these statistics are designed to test the null hypothesis

$$H_0 : S_1(t) = S_2(t) = \dots = S_g(t), t \leq \tau$$

where  $S_j$  is the survivor function for group  $j$ ,  $g$  is the number of groups being tested and  $\tau$  is the largest observed time, against the alternative hypothesis

$$H_1 : \text{at least one of the } S_j(t) \text{ differ, for some } t \leq \tau.$$

A rank statistics  $T$  is calculated as follows:

Let  $t_i$ , for  $i = 1, 2, \dots, n_d$ , denote the list of distinct failure times across all  $g$  groups and  $w_i$  a series of  $n_d$  weights.

Let  $d_{ij}$  denote the number of failures at time  $t_i$  in group  $j$  and  $n_{ij}$  denote the number of observations in the group  $j$  that are known to have not failed prior to time  $t_i$ , i.e., the size of the risk set for group  $j$  at time  $t_i$ . If a censored observation occurs at time  $t_i$  then that observation is treated as if the censoring had occurred slightly after  $t_i$  and therefore the observation is counted as being part of the risk set at time  $t_i$ .

Finally let

$$d_i = \sum_{j=1}^g d_{ij} \quad \text{and} \quad n_i = \sum_{j=1}^g n_{ij}.$$

The (weighted) number of observed failures in the  $j$ th group,  $O_j$ , is therefore given by

$$O_j = \sum_{i=1}^{n_d} w_i d_{ij}$$

and the (weighted) number of expected failures in the  $j$ th group,  $E_j$ , by

$$E_j = \sum_{i=1}^{n_d} w_i \frac{n_{ij} d_i}{n_i}$$

and if  $x$  denote the vector of differences  $x = (O_1 - E_1, O_2 - E_2, \dots, O_g - E_g)$

$$V_{jk} = \sum_{i=1}^{n_d} w_i^2 \left( \frac{d_i(n_i - d_i)(n_i n_{ik} I_{jk} - n_{ij} n_{ik})}{n_i^2 (n_i - 1)} \right)$$

where  $I_{jk} = 1$  if  $j = k$  and 0 otherwise, then the rank statistic,  $T$ , is calculated as

$$T = xV^{-}x^T$$

where  $V^{-}$  denotes a generalized inverse of the matrix  $V$ .

Under the null hypothesis,  $T \sim \chi_{\nu}^2$  where the degrees of freedom,  $\nu$ , is taken as the rank of the matrix  $V$ .

The different rank statistics are defined by using different weights in the above calculations, for example

logrank statistic  $w_i = 1$

Wilcoxon rank statistic  $w_i = n_i$

Tarone–Ware rank statistic  $w_i = \sqrt{n_i}$

Peto–Peto rank statistic  $w_i = \tilde{S}(t_i)$  where  $\tilde{S}(t_i) = \prod_{t_j \leq t_i} \frac{n_j - d_j + 1}{n_j + 1}$

### 2.3 Estimating the Survivor Function and Hazard Plotting

The most common estimate of the survivor function for censored data is the **Kaplan–Meier** or **product-limit** estimate,

$$\hat{S}(t) = \prod_{j=1}^i \left( \frac{n_j - d_j}{n_j} \right), \quad t_i \leq t < t_{i+1}$$

where  $d_j$  is the number of failures occurring at time  $t_j$  out of  $n_j$  surviving to  $t_j$ . This is a step function with steps at each failure time but not at censored times.

As  $S(t) = e^{-\Lambda(t)}$  the cumulative hazard rate can be estimated by

$$\hat{\Lambda}(t) = -\log(\hat{S}(t)).$$

A plot of  $\hat{\Lambda}(t)$  or  $\log(\hat{\Lambda}(t))$  against  $t$  or  $\log(t)$  is often useful in identifying a suitable parametric model for the survivor times. The following relationships can be used in the identification.

- (a) Exponential distribution:  $\Lambda(t) = \lambda t$ .
- (b) Weibull distribution:  $\log(\Lambda(t)) = \log \lambda + \gamma \log(t)$ .
- (c) Gompertz distribution:  $\log(\Lambda(t)) = \log \lambda + \gamma t$ .
- (d) Extreme value (smallest) distribution:  $\log(\Lambda(t)) = \lambda(t - \gamma)$ .

## 2.4 Proportional Hazard Models

Often in the analysis of survival data the relationship between the hazard function and the number of explanatory variables or covariates is modelled. The covariates may be, for example, group or treatment indicators or measures of the state of the individual at the start of the observational period. There are two types of covariate time independent covariates such as those described above which do not change value during the observational period and time dependent covariates. The latter can be classified as either external covariates, in which case they are not directly involved with the failure mechanism, or as internal covariates which are time dependent measurements taken on the individual.

The most common function relating the covariates to the hazard function is the proportional hazard function

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

where  $\lambda_0(t)$  is a baseline hazard function,  $z$  is a vector of covariates and  $\beta$  is a vector of unknown parameters. The assumption is that the covariates have a multiplicative effect on the hazard.

The form of  $\lambda_0(t)$  can be one of the distributions considered above or a nonparametric function. In the case of the exponential, Weibull and extreme value distributions the proportional hazard model can be fitted to censored data using the method described by Aitkin and Clayton (1980) which uses a generalized linear model with Poisson errors. Other possible models are the gamma distribution and the log-normal distribution.

## 2.5 Cox's Proportional Hazard Model

Rather than using a specified form for the hazard function, Cox (1972) considered the case when  $\lambda_0(t)$  was an unspecified function of time. To fit such a model assuming fixed covariates a marginal likelihood is used. For each of the times at which a failure occurred,  $t_i$ , the set of those who were still in the study is considered this includes any that were censored at  $t_i$ . This set is known as the risk set for time  $t_i$  and denoted by  $R(t_i)$ . Given the risk set the probability that out of all possible sets of  $d_i$  subjects that could have failed the actual observed  $d_i$  cases failed can be written as

$$\frac{\exp(s_i^T \beta)}{\sum \exp(z_l^T \beta)} \quad (1)$$

where  $s_i$  is the sum of the covariates of the  $d_i$  individuals observed to fail at  $t_i$  and the summation is over all distinct sets of  $n_i$  individuals drawn from  $R(t_i)$ . This leads to a complex likelihood. If there are no ties in failure times the likelihood reduces to

$$L = \prod_{i=1}^{n_d} \frac{\exp(z_i^T \beta)}{\left[ \sum_{l \in R(t_i)} \exp(z_l^T \beta) \right]} \quad (2)$$

where  $n_d$  is the number of distinct failure times. For cases where there are ties the following approximation, due to Peto [2], can be used:

$$L = \prod_{i=1}^{n_d} \frac{\exp(s_i^T \beta)}{\left[ \sum_{l \in R(t_i)} \exp(z_l^T \beta) \right]^{d_i}} \quad (3)$$

Having fitted the model an estimate of the baseline survivor function (derived from  $\lambda_0(t)$  and the residuals) can be computed to examine the suitability of the model, in particular the proportional hazard assumption.

## 3 Recommendations on Choice and Use of Available Functions

The following functions are available.

`nag_prod_limit_surviv_fn (g12aac)` computes Kaplan–Meier estimates of the survivor function and their standard deviations.

`nag_surviv_logrank (g12abc)` comparison of survival curves using rank statistics.

nag\_surviv\_cox\_model (g12bac) fits the Cox proportional hazards model for fixed covariates.  
 nag\_surviv\_risk\_sets (g12zac) creates the risk sets associated with the Cox proportional hazards model for fixed covariates.

Depending on the rank statistic required, it may be necessary to call nag\_surviv\_logrank (g12abc) twice, once to calculate the number of failures ( $d_i$ ) and the total number of observations ( $n_i$ ) at time  $t_i$ , to facilitate in the computation of the required weights, and once to calculate the required rank statistics.

The following functions from other chapters may also be useful in the analysis of survival data.

nag\_mills\_ratio (g01mbc) the reciprocal of Mills' Ratio, that is the hazard rate for the Normal distribution.  
 nag\_glm\_poisson (g02gcc) fits generalized linear model with Poisson errors (see Aitkin and Clayton (1980)).  
 nag\_glm\_gamma (g02gdc) fits generalized linear model with gamma errors.  
 nag\_censored\_normal (g07bbc) fits Normal distribution to censored data.  
 nag\_estim\_weibull (g07bec) fits Weibull distribution to censored data.  
 nag\_rank\_regsn\_censored (g08rbc) fits linear model using likelihood based on ranks to censored data (see Kalbfleisch and Prentice (1980)).  
 nag\_condl\_logistic (g11cac) fits a conditional logistic model. When applied to the risk sets generated by nag\_surviv\_risk\_sets (g12zac) the Cox proportional hazards model is fitted by exact marginal likelihood in the presence of tied observations.

#### 4 Functionality Index

Cox's proportional hazard model,  
 create the risk sets ..... nag\_surviv\_risk\_sets (g12zac)  
 parameter estimates and other statistics ..... nag\_surviv\_cox\_model (g12bac)  
 Survival,  
 Rank statistics ..... nag\_surviv\_logrank (g12abc)  
 Survivor function ..... nag\_prod\_limit\_surviv\_fn (g12aac)

#### 5 Auxiliary Functions Associated with Library Function Arguments

None.

#### 6 Functions Withdrawn or Scheduled for Withdrawal

None.

#### 7 References

Aitkin M and Clayton D (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM *Appl. Statist.* **29** 156–163

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