NAG Library Chapter Introduction

G12 – Survival Analysis

Contents

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 routines which are also used to analyse this type of data. Chapter G02 contains generalized linear models, Chapter G07 contains routines 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

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

where S_j is the survivor function for group j, g is the number of groups being tested and τ 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 $i = 1, 2, \ldots, n_d$, 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 jth 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 jth 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 $x = \big(O_1 - E_1, O_2 - E_2, \ldots, O_g - E_g\big)$

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

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

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

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

Under the null hypothesis, $T \sim \chi^2_\nu$ where the degrees of freedom, ν, 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 $\tilde{S}(t_i) = \prod_{t_j \le t_i} \frac{n_j - d_j + 1}{n_j + 1}$

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,

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

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

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

A plot of $\hat{\Lambda}(t)$ or $\log\Big(\hat{\Lambda}(t)\Big)$ 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

$$\lambda(t, z) = \lambda_0(t) \exp\left(\beta^{\mathrm{T}} z\right)$$

where $\lambda_0(t)$ is a baseline hazard function, z is a vector of covariates and β 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

$$\frac{\exp\left(s_i^{\mathrm{T}} \beta\right)}{\sum \exp\left(z_l^{\mathrm{T}} \beta\right)} \tag{1}$$

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

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

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:

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

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 Routines

The following routines are available.

G12AAF computes Kaplan–Meier estimates of the survivor function and their standard deviations.

G12ABF comparison of survival curves using rank statistics.

G12BAF fits the Cox proportional hazards model for fixed covariates.

G12ZAF 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 G12ABF 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 routines from other chapters may also be useful in the analysis of survival data.

G01MBF the reciprocal of Mills' Ratio, that is the hazard rate for the Normal distribution.

G02GCF fits generalized linear model with Poisson errors (see Aitkin and Clayton (1980)).

G02GDF fits generalized linear model with gamma errors.

G07BBF fits Normal distribution to censored data.

G07BEF fits Weibull distribution to censored data.

G08RBF fits linear model using likelihood based on ranks to censored data (see Kalbfleisch and Prentice (1980)).

G11CAF fits a conditional logistic model. When applied to the risk sets generated by G12ZAF the Cox proportional hazards model is fitted by exact marginal likelihood in the presence of tied observations.

4 Functionality Index

5 Auxiliary Routines Associated with Library Routine Parameters

None.

6 Routines 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