NAG C 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.

2 Background

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 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,

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

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: $\log(\Lambda(t)) = \lambda(t - \gamma)$.

2.3 Proportional Hazard Models

Often in the analysis of survival data the relationship between the hazard function and the a 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

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

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 non-parametric 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 lognormal distribution.

2.4 Cox's Proportional Hazard Model

Rather than using a specified form for the hazard function, Cox (1972b) 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 know 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

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

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

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

where n_d is the number of distinct failure times. For cases where there are ties the following approximation, due to Peto in Cox (1972b), can be used:

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

Having fitted the model an estimate of the base-line 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 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 (1972b) 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

4 Available Functions

g12aac Kaplan–Meier (product-limit) estimates of survival probabilities

g12bac Fits Cox's proportional hazard model

The function nag_glm_poisson (g02gcc) (which fits a generalized linear model with Poisson errors) may be used to fit models to censored data from the exponential, Weibull and extreme value distributions, see Aitkin and Clayton (1980).