Chapter g12 – Survival Analysis

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,

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

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

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.

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

Computes Kaplan–Meier estimates of the survivor function and their standard deviations `g12aac`

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).