Chapter 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   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

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

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.

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

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 the proportional hazard model can be fitted to censored data using the method described by Aitkin and Clayton [1] 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

Cox [2] considered the case when $\lambda_0(t)$ was an unspecified function of time. For fixed covariates the model can be fitted by maximising the approximate marginal likelihood.

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

where n_d it the number of distinct failure times, s_i is the sum of the covariates of the d_i 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 approximation, due to Peto in [2], concerns situations where there are ties in the data, i.e. $d_i \neq 1$ for all i. The approximation is reasonable provided d_i is small compared to the size of $R(t_i)$.

Having fitted the model an estimate of the base-line survivor function, derived from $\lambda_0(t)$ and 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.

G12BAF   fits the Cox proportional hazards model for fixed covariates.

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 [1]).

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 Kabfleisch and Prentice [4]).

4 Routines Withdrawn or Scheduled for Withdrawal

None since Mark 13.

5 References

[**1**] 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

[**2**] Cox D R (1972) Regression models in life tables (with discussion) *J. Roy. Statist. Soc. Ser. B* **34** 187–220

[**3**] Gross A J and Clark V A (1975) *Survival Distributions: Reliability Applications in the Biomedical Sciences* Wiley

[**4**] Kalbfleisch J D and Prentice R L (1980) *The Statistical Analysis of Failure Time Data* Wiley