

Lifetime data analysis

Professor Jane L Hutton, J.L.Hutton@warwick.ac.uk,
Department of Statistics, The University of Warwick, Coventry, UK

March 2010

Course outline

- Introduction
 - Main definitions and formulae
- Non-parametric survival analysis
 - Kaplan-Meier estimate
 - Actuarial estimate
 - Log-rank test
- Parametric survival analysis
 - Survival functions
 - Likelihood methods
 - Diagnostic plots
- Regression models for survival data
 - Accelerated life models
 - Proportional hazards models

1 Introduction

Let T be a random variable, which represents the failure time of an individual. The origin must be precisely defined. The scale on which time is measured is often real time, or clock time, but could be the length of cloth up to the first flaw, or the cumulative load on a bean. A failure is a point event which must be precisely defined. For example, a ‘failure’ might be the death of a person from cancer, the occurrence of a heart attack, or the first day in full time employment. A failure time is censoring if failure has not occurred before observation finished. Censoring is often a point event, but interval censoring is also possible. The time when the individual was last observed must be recorded.

Main definitions and formulae

- Well defined start point,
- Well defined failure event or end point,
- Random variable T is time from start point to end point,
- Random variable Δ indicates failure or censoring
- Observed time: $\delta = 1$; Censored: $\delta = 0$

Censoring

Censoring can arise in a variety of ways. We will concentrate on right censoring, where the observed time is less than the actual time.

Type 1 Experiments which start a set of items on test at time zero, and end at a prespecified time give rise to type 1 censoring. The number of observed failure times is random.

Type II Experiments which start a set of items on test at time zero, and end when a prespecified number of items, say r , have failed give rise to type II censoring. Only the first r failure times are observed. The time at which the experiment ends is random.

Random censoring Consider experiments in which people enter a study at different times. If the failure event has not occurred when the study ends, people's failure time will be censored. If a person moves away and cannot be traced, her time must be censored at the last time she was observed. If the study is a drug trial, a person might drop-out because of side effects of the drug. The number of observed failure times is random.

Interval censored data, 1 If the failure event cannot be observed directly, we might only know that the event occurred before or after a particular time. To observe a tumour in a rat, we might have to kill the rat and dissect it. If there is a tumour, we know the time to tumour is less than the time at which the rat was killed. The event has already occurred, so is left censored. If there is no tumour, we know the time to tumour is greater than the time at which the rat was killed, right censored.

Interval censored data, 2 If the failure event cannot be observed directly, but observing the event does not kill or destroy the individual, then individuals can be checked at time intervals $i, i+1, i+2, \dots$. If at i the person has not had the event, but at $i+1$, they have had the event, then we know that $i < T < i+1$. For example, if a child loses a tooth between visits to the dentist, the dental records give an interval in which the tooth was lost.

Doubly censored data If data can be left censored or right censored, or observed, the data are doubly-censored. Suppose the time of interest is when a child can write their own name. If we observe a class of first year primary children for their first term, some will be able to write their name when they arrive, so their time is left censored. Some will learn during the term, so we know the exact time. Some children will not manage to write their name by the end of the term, so will be right censored.

Truncation

Truncation refers to individuals who cannot be observed. For example, a child with cerebral palsy cannot be included in a study of cerebral palsy before they are diagnosed. A child with mild or moderate cerebral palsy cannot be reliably diagnosed before she is five years old. If such a child dies aged 18 months, we will not know of her existence. If we consider time from entry to a retirement home to death, we have to remember that only people who live until retirement age can die in a retirement home.

Survival functions

The probability functions which are used in survival analysis include the familiar density function; the survival function, which is the complement of the distribution function; and the hazard function, also known as the instantaneous death rate or the force of mortality.

Density function: $f(t, \theta)$

Survival function: $S(t, \theta) = Pr(T > t) = 1 - F(t, \theta)$

Hazard function:

$$h(t) = \lim_{\epsilon \rightarrow 0^+} Pr(T < t + \epsilon | T > t) / \epsilon$$
$$h(t, \theta) = f(t, \theta) / S(t, \theta)$$

2 Non-parametric survival analysis

2.1 Kaplan-Meier estimate

The notation used is given below. It is easier to understand with an example.

1. t_1, t_2, \dots, t_k are k observed failure times in increasing order, i.e. times at which at least one individual fails (not censored). Let $t_0 = 0$.
2. $I_1 = [0, t_1], I_2 = (t_1, t_2], I_3 = (t_2, t_3], \dots, I_k = (t_{k-1}, t_k]$.
3. For $i = 1, 2, \dots, k$ let $d(t_i)$ = number in study who fail at t_i .
4. for $i = 1, 2, \dots, k$ let $n(t_i)$ = number in study at t_i and 'at-risk' (able to fail at $t - i$).
5. $c(t_i)$ = numbered censored in $[t_{i-1}, t_i)$. **NB** Notice the change in the interval! If a unit fails at t_i , this is counted in the interval $I_i = (t_{i-1}, t_i]$. If a unit is censored at t_i , this is counted in the interval $[t_{i-1}, t_i)$.
6. From the above definitions, $n(t_i) = n(t_{i-1}) - d(t_{i-1}) - c(t_i)$, and, of course, $n(0) =$ initial sample size. The Kaplan-Meier estimate of $S(t)$ is

$$S(t) = \prod_{t_i < t} \left(1 - \frac{d(t_i)}{n(t_i)} \right) = \prod_{t_i < t} S(t_i),$$

where $s(0) = 1$ by definition. This uses $\text{Prob}(\text{survive } I_i | \text{Survive } I_{i-1}) = 1 - \frac{d(t_i)}{n(t_i)}$, which is applied sequentially, i.e.

$$S(t) = \left(1 - \frac{d(t_i)}{n(t_i)} \right) S(t_{i-1}).$$

For $t \in (t_{i-1}, t_i)$, $S(t) = S(t_{i-1})$. This is also called a product-limit estimator.

We also need to have some estimate of the variance. Three estimators are given; Greenwood's formula is not ideal for a hand calculator, but is used in software. The notation $\hat{S}(t)$ is used to indicate that these are estimators.

1. Simple approximation, when there is no censoring:

$$\text{Var}\{\hat{S}(t) = \hat{S}(t)\{1 - \hat{S}(t)\}/n(0)\}.$$

For a binary random variable, $Z \sim B(1, \pi)$, $\text{Var}(Z) = \pi(1 - \pi)$.

2. Simple approximation, with censoring:

$$\text{Var}\{\hat{S}(t) = \hat{S}(t)^2\{1 - \hat{S}(t)\}/n(t_j)\},$$

where $t_j = \max(t_i < t; i \in (1, 2, \dots, k))$.

3. Greenwood's formula (1926):

$$\text{Var}\{\hat{S}(t) = \hat{S}(t)^2 \sum_{t_i < t} \frac{d(t_i)}{n_i\{n(t_i) - d(t_i)\}}\}.$$

2.2 Actuarial estimator

It is sometimes more convenient or practical to estimate the survival function using intervals of fixed length, e.g. one or five years. Generally, we use actuarial life tables to summarise large data sets, such as a population of people with cerebral palsy. The definitions are similar, but not identical, to those for the Kaplan-Meier estimator.

1. $0 = a_0, a_1, a_2, \dots, a_n$ are increasing times.
2. $I_i = [a_{i-1}, a_i)$, are fixed length intervals.
3. d_i = number of failures in interval I_i .
4. c_i = number of censored observations in interval I_i .
5. r_{i-1} = number of subjects entering interval I_i .
6. $r'_i = r_{i-1} - c_i/2$ = adjusted number of subjects at risk in interval I_i .

Note that $r_i = r_{i-1} - d_i - c_i$.

The actuarial estimator for $S(t)$ is $S(a_i) = \prod_{j \leq i} (1 - d_j/r'_j)$.

Table heading for an actuarial survival estimate are given below.

Interval	Number at risk	Number failed	Number censored	Adjusted number at risk	Conditional probability of survival	Survival estimate

The variance of $\hat{S}(t)$ can be estimated using a modification of Greenwood's formula for the Kaplan-Meier estimate.

2.3 The log rank test and Wilcoxon test

These are non-parametric tests of the difference in survival between 2 groups.

The log rank test

The null hypothesis is $H_0 : S_1(t) = S_2(t) \forall t$, where $S_1(t), S_2(t)$ are the survival functions for the two groups. At each failure time, $t_i, i = 1, 2, \dots, k$, in either group, we can form a table as shown below.

	Failed	Still at risk	At risk
Group 1	$d_{1,t}$	$n_{1,t} - d_{1,t}$	$n_{1,t}$
Group 2	$d_{2,t}$	$n_{2,t} - d_{2,t}$	$n_{2,t}$
Totals	d_t	$n_t - d_t$	n_t

The notation is: $d_{j,t}$, is the number in group j who failed at time t_i , and $n_{j,t}$ is the number at risk in group j at time $t_i, j = 1, 2$. The totals are $d_t = d_{1,t} + d_{2,t}$ and $n_t = n_{1,t} + n_{2,t}$. The subscript i is suppressed. Under H_0 , the expected number failing in Group 1 at time t_i is $e_{1,t} = d_t n_{1,t} / n_t$. The variance is

$$Var(d_{1,t}) = v_{1,t} = \frac{n_{1,t} n_{2,t} (n_t - d_t)}{n_t^2 (n_t - 1)}.$$

We define: $E_j = \sum_{i=1}^k e_{j,t}$ = Sum over all failure times of $e_{j,t}$, $V_j = \sum_{i=1}^k v_{j,t}$, and $O_j = \sum_{i=1}^k d_{j,t}$. The log rank test statistic is $Z = (O_1 - E_1) / \sqrt{V_1}$, which is referred to the standard Normal distribution.

An alternative, simpler version does not require V_1 , but only the calculation of E_2 and O_2 . The test statistic in this case is

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2},$$

which is referred to a χ_1^2 distribution. This test is more conservative, and easier to extend to more than 2 groups.

Wilcoxon test

The Wilcoxon test can be derived from the usual Wilcoxon test for uncensored data, which compares the ranks of the two groups.

The test statistic is based on $U_W = \sum_{i=1}^k n_t (d_{1,t} - e_{1,t})$, with the same notation as used for the log rank test. In this case, the difference between observed and expected deaths at each time is weighted by the number at risk at that time. More weight is therefore given to differences in the earlier part of the distributions than in the tails.

The variance of U_W is given by $V_W = \sum_{i=1}^k n_t^2 v_{1,t}$, and the Wilcoxon test statistic is $W_W = U_W^2 / V_W$, which has a χ_1^2 distribution under the null hypothesis.

The log rank test is more suitable, i.e. powerful, when the alternative is that the hazard rates are proportional. For other departures from the null hypothesis, the Wilcoxon test is more appropriate.

3 summary for parametric regression models

Main formulae and definitions for parametric survival analysis.

Let T be a random variable, which represents the failure time of an individual.

Density function : for T is $f_T(t) = f(t)$ if the random variable is obvious.

Survival function : for T is $S_T(t) = S(t) = Pr(T \geq t) = 1 - F_T(t)$.

Hazard function : for T is $h_T(t) = h(t)$. ($\lambda_T(t)$ is also common.)

$$h_T(t) = \lim_{\delta t \rightarrow 0^+} \frac{Pr(t \leq T \leq t + \delta t \mid T \geq t)}{\delta t}$$

(age-specific failure rate, force of mortality, instantaneous failure rate)

Integrated hazard function :

$$H_T(t) = \int_0^t h_T(u) du.$$

Distributions for survival data

Exponential $S(t) = \exp(-\lambda t)$, $f(t) = \lambda \exp(-\lambda t)$, scale parameter $\lambda > 0$.

Weibull $S(t) = \exp\{-(\lambda t)^\kappa\}$, $f(t) = \lambda \kappa (\lambda t)^{\kappa-1} \exp\{-(\lambda t)^\kappa\}$,
scale parameter $\lambda > 0$ and shape parameter $\kappa > 0$.

Log-logistic $S(t) = \{1 + (\lambda t)^\kappa\}^{-1}$, $f(t) = \kappa \lambda^\kappa t^{\kappa-1} \{1 + (\lambda t)^\kappa\}^{-2}$,
scale parameter $\lambda > 0$ and shape parameter $\kappa > 0$.

Others log-normal, gamma, Gompertz-Makeham ...

Notation

n individuals, indexed by i

distinct independent failure times $t_1 < \dots < t_n$

z_i be a $q \times 1$ vector of explanatory variables

vector regression parameters:

β proportional hazards model, γ accelerated life model

Regression models

Accelerated life model

- Baseline probability density function: $f_0(t)$
- Baseline survival function: $F_0(t)$
- Model: $F_z(t) = F_0(te^{\gamma^T z})$
- Covariates scale the expected survival time
- Probability density function | covariates: $e^{\gamma^T z} f_0(te^{\gamma^T z})$
- Hazard function | covariates:
 $e^{\gamma^T z} f_0(te^{\gamma^T z}) / F_0(e^{\gamma^T z}) = e^{\gamma^T z} h_0(te^{\gamma^T z})$.
- In general, replace $e^{\gamma^T z}$ by $\psi(\gamma, z)$

Proportional hazards model

- Baseline probability density function: $g_0(t)$
- Baseline survival function: $G_0(t)$
- Baseline hazard function: $\lambda_0(t)$
(unspecified in Cox semi-parametric regression model)
- Model: $\lambda(t; z) = e^{\beta^T z} \lambda_0(t)$
- Covariates scale the hazard rate
- Survival functions, given the covariates: $\{G_0(t)\}^{e^{\beta^T z}}$
- Prob. density fn. | covariates: $g_0(t) e^{\beta^T z} \{G_0(t)\}^{e^{\beta^T z} - 1}$
- In general, replace $e^{\beta^T z}$ by $\psi(\beta, z)$

4 Parametric survival analysis: model fitting

4.1 Exponential distribution

$$f(t) = \lambda e^{-\lambda t}; F(t) = 1 - e^{-\lambda t}; S(t) = e^{-\lambda t}$$

Thus

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \text{ a constant.}$$

$$\mathbb{E}(T) = \int_0^{\infty} e^{-\lambda t} dt = \left[-\frac{e^{-\lambda}}{\lambda} \right]_0^{\infty} = \frac{1}{\lambda}.$$

We have a Kaplan-Meier estimator $\hat{S}(t)$. How might we explore whether sample data can be summarised by an exponential distribution? We can try looking at the Kaplan-Meier estimator and drawing a curve, but the human eye is much better at straight lines. How to we get a straight line from $S(t) = e^{-\lambda t}$? We take logs.

$$\log S(t) = -\lambda t$$

So, plot $\log S(t)$ against t and look for a negative slope through the origin.

We need a better method than by eye.

Statistical models mainly use the *likelihood*. Consider a sample of k times t_1, \dots, t_k , which are assumed to be independent and identically distributed (iid), with Exponential distribution function $F(t, \lambda)$ and pdf $f(t, \lambda)$, with all t_i observed.

The likelihood function is defined as

$$L(\lambda, t) = \prod_{i=1}^k f(t_i, \lambda) = \prod_{i=1}^k \lambda e^{-\lambda t_i},$$

where $t = (t_1, \dots, t_k)$ and $\lambda > 0$. This is a function of λ , the parameter, with the sample t regarded as given. We regard $f(\lambda, t_i) = \lambda e^{-\lambda t_i}$ as the *likelihood* of t_i for a particular value of λ . For example, if we observe $t = 1$, if the underlying distribution had $\lambda = 1$ we would not be surprised. If λ were 20, so that $\mathbb{E}(T) = 1/20$, we might wonder if there had been a mistake. If the mean life time of mice is assumed to be one year, and we were told that the mouse is 20 years old, we would be surprised.

The standard method of defining the best fitting model is to choose the value of the parameter, λ , in the parameter space which maximizes $L(\lambda, t)$. As differentiating a product is less easy than differentiating a sum, we consider not $L(\lambda, t)$ but

$$\begin{aligned} \log \{L(\lambda, t)\} &= \ell(\lambda, t) \\ &= \sum_{i=1}^k \log f(t_i, \lambda) \\ &= \sum_{i=1}^k (\log \lambda - \lambda t_i). \end{aligned}$$

Aside: Why is this possible? How do we know that the value $\hat{\lambda}$ which maximizes $\ell(\lambda, t)$ will also maximize $L(\lambda, t)$? The value $\hat{\lambda}$ is, by definition, the *maximum likelihood estimator*. For the exponential case, we have

$$\ell(\lambda, t) = k \log \lambda - \lambda \sum_{i=1}^k t_i.$$

So

$$\frac{d\ell}{d\lambda}(\lambda, t) = \frac{k}{\lambda} - \sum_{i=1}^k t_i,$$

which is 0 at $\hat{\lambda}$ if and only if

$$\hat{\lambda} = \frac{k}{\sum_{i=1}^k t_i} = \text{frac}1\bar{t}.$$

The mle is the inverse of the sample mean, or $\hat{\lambda}^{-1} = \bar{t}$. Further,

$$\frac{d^2\ell}{d\lambda^2} = -\frac{k}{\lambda^2},$$

so this is a maximum.

I shall not be discussing maximum likelihood estimators in detail – that would be a different course. I shall simply give some useful properties of mles.

Consider a parameter θ , which might be, for example, $\theta = (\mu, \sigma^2)$ for a Normal distribution. The maximum likelihood estimator is:

1. Asymptotically unbiased, i.e.

$$\lim_{k \rightarrow \infty} \mathbb{E}(\hat{\theta}) = \theta$$

2. Consistent, i.e.

$$\lim_{k \rightarrow \infty} \text{var}(\hat{\theta}) = 0$$

3. Invariant: the mle $\hat{\tau}$ of a continuous 1-1 function of θ , say $\tau(\theta)$, is $\tau(\hat{\theta})$.
4. Asymptotically normal (multivariate normal if θ is a vector), i.e.

$$\hat{\theta} \sim N(0, I^{-1}(\theta))$$

where $I(\theta)$ is the Fisher information matrix.

The Fisher information matrix is denoted by $I(\theta)$ and is defined by its (i, j) th element

$$I(\theta)_{j,k} = -\mathbb{E} \left[\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right].$$

In order to calculate an expectation, we have to know the values of the parameters θ – but if we knew the values, we would not be estimating them! So we have to approximate $I(\theta)$ by the observed information matrix $i(\theta) = I(\hat{\theta})$, so that the (i, j) th element of $i(\theta)$ is

$$-\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \Big|_{\theta = \hat{\theta}}.$$

The derivative of the log-likelihood is called the *score function*

$$u(\theta) = \frac{\partial \ell}{\partial \theta}.$$

Continuing the exponential example, we have $\hat{\lambda}^{-1} = \bar{t}$, and $\mathbb{E}(\hat{\lambda}^{-1}) = \mathbb{E}(\bar{T}) = \lambda^{-1}$ from the ordinary definition of the sample mean. The variance of the sample mean is $\text{var}(\bar{T}) = \text{var}(T)/k$, so

$$\text{var}(\bar{T}) = \frac{\lambda^{-2}}{k}.$$

Obviously as $k \rightarrow \infty$ (as sample size increases) $\text{var}(\bar{T}) \rightarrow 0$. Now, we have

$$\frac{\partial \ell}{\partial \lambda^2} = -\frac{k}{\lambda^2},$$

so the Fisher information is

$$I(\lambda) = -\mathbb{E}\left(-\frac{k}{\lambda^2}\right) = \frac{k}{\lambda^2}$$

and

$$\text{var}(\hat{\lambda}) = I(\lambda)^{-1} = \frac{\lambda^2}{k}.$$

$I(\lambda)^{-1}$ is estimated by

$$I(\hat{\lambda})^{-1} = \frac{1}{k} \times \left(-\left(-\frac{k}{\sum t_i} \right)^2 \right) = \frac{k}{(\sum t_i)^2}.$$

This does tend to zero, though with the sample size in the numerator it might not be obvious. For the Gehan control group, consider $\log \hat{S}(t)$ plotted against t . Now, $\sum t_i = 182$, so

$$\hat{\lambda} = \frac{21}{182} = 0.115,$$

and

$$\text{var}(\hat{\lambda}) = \frac{(0.115)^2}{21} = 0.0006,$$

so

$$\text{s. e.}(\hat{\lambda}) = 0.025.$$

Plot of $\log \hat{S}(t)$ against t shows the line $y = 0.115x$ – not too bad, but perhaps not a really good fit.

4.2 Confidence intervals and significant tests for maximum likelihood estimators

There are three methods which can be used to define significance tests and confidence limits for mles. As the sample size increases, the limits of these methods converge. We wish to test if $\theta = \theta_0$, some defined value, where θ has dimension d (there are d parameters).

4.2.1 Wald Test

For scalar θ ,

$$(\hat{\theta} - \theta_0)^2 I(\hat{\theta}) \approx \left[\frac{\hat{\theta} - \theta_0}{\text{s. e.}(\hat{\theta})} \right]^2 \sim \chi_d^2$$

so that

$$\frac{\hat{\theta} - \theta_0}{\text{s. e.}(\hat{\theta})} \sim N(0, 1).$$

From this we also get a $100(1 - \alpha)\%$ confidence interval in a familiar form:

$$[\hat{\theta} - z_{\alpha/2} \text{s. e.}(\hat{\theta}), \hat{\theta} + z_{\alpha/2} \text{s. e.}(\hat{\theta})]$$

where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha$.

This approximate result can give upper or lower limits which are outside the parameter space of θ .

4.2.2 Likelihood ratio test

$$\begin{aligned} -2 \left\{ \log L(\theta_0) - \log L(\hat{\theta}) \right\} &= -2 \log \left\{ \frac{L(\theta_0)}{L(\hat{\theta})} \right\} \\ &= -2 \left\{ \ell(\theta_0) - \ell(\hat{\theta}) \right\} \\ &\sim \chi_d^2. \end{aligned}$$

To define confidence intervals from this, we have to consider the values for which

$$-2 \left\{ \ell(\theta) - \ell(\hat{\theta}) \right\} \leq \chi_{d,\alpha}^2,$$

where $\chi_{d,\alpha}^2$ is the upper α th quantile of the χ_d^2 distribution. We can rewrite the inequality as

$$\begin{aligned} \ell(\theta) - \ell(\hat{\theta}) &\geq -\chi_{d,\alpha}^2/2 \\ \ell(\hat{\theta}) - \ell(\theta) &\leq \chi_{d,\alpha}^2/2, \end{aligned}$$

ie the difference between $\ell(\theta)$ and $\ell(\hat{\theta})$ must be sufficiently small. These intervals always have values for θ lying within the parameter space, and are not necessarily symmetrical about $\hat{\theta}$.

4.2.3 Score test

$$\frac{u(\theta_0)^2}{I(\theta_0)}$$

The two methods we shall use are the Wald and likelihood ratio tests.

Continuing the exponential example

95% Wald confidence interval:

$$(0.115 - 2 \times 0.025, 0.115 + 2 \times 0.025) = (0.065, 0.165)$$

95% confidence interval, likelihood ratio:

$$\begin{aligned}\ell(\lambda) &= k \log \lambda - \lambda \sum t_i = 21 \log \lambda - 182\lambda \\ \ell(\hat{\lambda}) &= 21 \log(21/182) - (21/182) \times 182 = -66.349 \\ \chi_{1,0.95}^2 &= 3.84\end{aligned}$$

so the confidence interval is $(0.073, 0.171)$, a little different from the Wald interval.

In this case, we could also get an exact confidence interval, because for T_i iid $\text{Exp}(\lambda)$,

$$\sum_{i=1}^k T_i \sim \Gamma(k, \lambda),$$

and statistical theory shows that

$$2\lambda \sum_{i=1}^k T_i = 2k \frac{\lambda}{\hat{\lambda}} \sim \chi_{2k}^2.$$

Thus, with probability $1 - \alpha$,

$$\chi_{2k, 1-\alpha/2}^2 \leq 2k \frac{\lambda}{\hat{\lambda}} \leq \chi_{2k, \alpha/2}^2,$$

and rearranging gives

$$\frac{\hat{\lambda} \chi_{2k, 1-\alpha/2}^2}{2k} \leq \lambda \leq \frac{\hat{\lambda} \chi_{2k, \alpha/2}^2}{2k}$$

For the Gehan control group, this gives $(0.071, 0.170)$. So, even with only 21 observations, the asymptotic approximation is quite good. The mean survival time is 8.7 weeks; we invert the confidence interval: $(5.9, 14.0)$.

4.3 Maximum likelihood estimation for censored data

Consider a sample of k times, t_1, \dots, t_k , which are iid $\text{Exp}(\lambda)$, where n times are observed and the remaining times are censored. For the observed times, the contribution to the likelihood of t_i is

$$f(\lambda, t_i) = \lambda e^{-\lambda t_i}$$

For the censored times, we know that $T_j > t_j$, and the contribution to the likelihood is $S(t_j) = \mathbb{P}(T_j > t_j) = e^{-\lambda t_j}$. So the likelihood for the sample is

$$L(\lambda, t) = \prod_o \lambda e^{-\lambda t_i} \prod_c e^{-\lambda t_j}$$

where \prod_o is the product over observed failure times, and \prod_c the product over censored times. If we use the notation (t_i, δ_i) with

$$\delta_i = \begin{cases} 1 & \text{if } t_i \text{ fails or dies} \\ 0 & \text{if } t_i \text{ is censored,} \end{cases}$$

we get a more elegant expression

$$\begin{aligned} L(\lambda, t) &= \prod_{i=1}^k (\lambda e^{-\lambda t_i})^{\delta_i} (e^{-\lambda t_i})^{1-\delta_i} \\ &= \prod_{i=1}^k \lambda^{\delta_i} e^{-\lambda t_i} \\ &= \lambda^m e^{-\lambda \sum_{i=1}^k t_i} \end{aligned}$$

where m is the number of failed events.

Notice that $\{(t_1, \delta_1), \dots, (t_k, \delta_k)\}$ is like the `Surv` object in R (`Surv(time, dead)`).

In general, for a random sample of times,

$$\begin{aligned} L(\theta, t) &= \prod_{i=1}^k f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^k (h(t_i)S(t_i))^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^k h(t_i)^{\delta_i} S(t_i) \end{aligned}$$

Exponential example, Gehan 6MP group:

$$\begin{aligned} \ell(\lambda) &= m \log \lambda - \lambda \sum_{i=1}^k t_i \\ \frac{d\ell(\lambda)}{d\lambda} &= \frac{m}{\lambda} - \sum_{i=1}^k t_i \implies \hat{\lambda} = \frac{m}{\sum_{i=1}^k t_i} \\ \frac{d^2\ell}{d\lambda^2} &= -\frac{m}{\lambda^2} \end{aligned}$$

$m = 9$, $\sum_{i=1}^2 1t_i = 359$, so $\hat{\lambda} = 0.0251$, $\hat{\lambda}^{-1} = 39.9$.

$$I(\hat{\lambda}) = \frac{m}{\hat{\lambda}^2}$$

so

$$\begin{aligned} \text{var}(\hat{\lambda}) &= \frac{\lambda^2}{m} = \frac{m^2}{m(\sum t_i)^2} = \frac{m}{(\sum t_i)^2} \\ &= \frac{9}{359^2} = 0.00006, \end{aligned}$$

and

$$\text{s. e.}(\lambda) = 0.008.$$

Wald confidence interval: $0.0251 \pm 2 \times 0.008 = (0.0091, 0.0411)$, which corresponds to a survival time of (24.3, 109.9) – well beyond the data.

Log-likelihood confidence interval: (0.0121, 0.0452), which corresponds to survival times (22.1, 82.64).

We can also estimate medians using the fitted distributions:

$$0.5 = e^{-\hat{\lambda}t_m} \iff t_m = \frac{\log \frac{1}{2}}{\hat{\lambda}}$$

4.4 Maximum likelihood estimator with two groups

So, we have fitted exponential distribution to two separate groups, but how should we compare them?

Define an indicator function X_i with $X_i = 1$ if individual i is in 6MP group and $X_i = 0$ if not, i.e. individual i is in the control group. What we have so far is an estimate for the 6MP groups, $\hat{\lambda}(= 9/359)$ and an estimate for the control group $\hat{\lambda}_0$. As the two groups are independent, the log-likelihood function for the whole study is the product of the individual group likelihood functions. We now want the data as (T_i, Δ_i, X_i) for $i = 1, \dots, n_0 + n_1$, where n_0 is the number in control group, 0 and n_1 is the number in 6MP group, 1. So

$$\begin{aligned} \ell(\lambda_0, \lambda_1, \underline{t}, \underline{\sigma}, \underline{x}) &= \sum_{i=1}^{n_0+n_1} x_i \delta_i \log(\lambda_{x_i}) - \sum_{i=1}^{n_0+n_1} x_i t_i \lambda_{x_i} \\ &= m_0 \log(\lambda_0) - \lambda_0 t_0 + m_1 \log(\lambda_1) - \lambda_1 t_1, \end{aligned}$$

where m_0 is the number of observed control times, m_1 is the number of observed 6MP times, t_0 is the sum of all times in group 0, and t_1 is the sum of all times in group 1.

If we let λ_0, λ_1 vary freely, we will have same estimates as before, and $\ell(\lambda_0, \lambda_1, \underline{t}, \underline{\sigma}, \underline{x})$ is sum of two separate likelihoods. If we wish to test the hypothesis $\lambda_0 = \lambda_1 = \lambda_n$ we fit this model, i.e.

$$\begin{aligned} \ell(\lambda_n, \underline{t}, \underline{\delta}, \underline{x}) &= \sum_{i=1}^{n_0+n_1} \delta_i \log(\lambda_{\lambda_n}) - \lambda_{\lambda_n} \sum_{i=1}^{n_0+n_1} t_i \\ &= (m_0 + m_1) \log(\lambda_n) - \lambda_n (t_0 + t_1), \end{aligned}$$

Hence,

$$\hat{\lambda}_n = \frac{m_0 + m_1}{t_0 + t_1}.$$

as we would require. We can use the likelihood ratio test $2 \left\{ \ell(\hat{\lambda}_n) - \ell(\hat{\lambda}_0, \hat{\lambda}_1) \right\} \sim \chi_1^2$

$$-2 \left\{ \ell(\hat{\lambda}_n) - \ell(\hat{\lambda}_0, \hat{\lambda}_1) \right\} = -2 \{-116.8 - 108.5\} = 16.6$$

4.5 Alternative parameterisation

Instead of using parameters λ_0, λ_1 for groups 0 and 1, without any relation, consider using ρ_0 for group 0 and $\rho_1 = b\rho_0$ for group 1. This gives the effect of group 1, 6MP in our example, as scaling the mean, or the hazard rate. In terms of λ_0, λ_1 , we have $\rho_0 = \lambda_0, b = \lambda_1/\lambda_0$. Then we have $\log \rho_1 = \log \rho_0 + \log b$, and using the group indicator x_i , we have

$$\begin{aligned} \log(\mathbb{E}(T_i)) &= \log \rho_0 + x_i \log b \\ &= -(\mu + \beta x_i), \end{aligned}$$

where $\mu = -\log \rho_0$ and $\beta = -\log b$. The log-likelihood becomes

$$\begin{aligned} \ell(\lambda_n, \underline{t}, \underline{\delta}, \underline{x}) &= - \sum_{i=1}^{n_0+n_1} \delta_i(\mu + \beta x_i) - e^{-(\mu+\beta)} \sum_{i=1}^{n_0+n_1} t_i \\ &= -(m_0 + m_1)\mu - m_1\beta - t_0e^{-\mu} - t_1e^{-(\mu+\beta)} \\ \frac{\partial \ell}{\partial \mu} &= -(m_0 + m_1) + t_0e^{-\mu} + t_1e^{-(\mu+\beta)} \\ \frac{\partial \ell}{\partial \beta} &= -m_1 + t_1e^{-(\mu+\beta)}. \end{aligned}$$

Setting both derivatives to zeros, we have simultaneous equations

$$m_0 + m_1 = t_0e^{-\hat{\mu}} + t_1e^{-(\hat{\mu}+\hat{\beta})} \quad (1)$$

$$m_1 = t_1e^{-(\hat{\mu}+\hat{\beta})}. \quad (2)$$

Equation (1)-equation (2) gives

$$\begin{aligned} m_0 &= t_0e^{-\hat{\mu}} & e^{\hat{\mu}} &= \frac{m_0}{t_0} \\ \hat{\mu} &= -\log\left(\frac{m_0}{t_0}\right) & m_1 &= t_1e^{-(\hat{\mu}+\hat{\beta})}. \end{aligned}$$

Subtracting $e^{-\hat{\mu}}$ gives

$$e^{-\hat{\beta}} = \frac{m_1}{t_1}e^{\hat{\mu}} = \frac{m_1}{t_1} \frac{t_0}{m_0},$$

as expected. For the Gehan example, (Look at output of summary (ghn.surv))

$$\begin{aligned} \hat{\mu} &= -\log(21/182) = 2.16, \\ \hat{\beta} &= -\log(9/359 \times 182/21) = 1.53. \end{aligned}$$

5 Model choice and Model checking

We have already asked about choice of covariates to include. If we can discuss with practitioner, that is best. If not, there are various methods, e.g the so-called forward, backward and stepwise selection. The general aim is always to include variables which sufficiently explain or predict the outcome of interest, but not to include any variables unnecessarily.

The analysis of the Feigl and Zelen data is a small example of forward selection. I suggested you consider each covariate separately.

Backward Selection starts with all variables in model and drops out the variables which is least significant, i.e with largest p-value.

Stepwise Selection allows for correlation between covariates. It might that we start with forward selection, and include covariates x_1, x_2, x_3 . We then add x_4 , and find that x_2 is no longer significant, so drop it.

For the purpose of this course, I suggest the following guideline: First fit each covariate separately and see how significant they are. Decide if you think any covariate must be included.

For example, if we are analyzing the result of a clinical trial of a drug, we should always include the effect of the drug.

After including the essential covariates, add each of the other covariates in the order of significance. keep the most significant one. Then consider the models with each of the remaining covariates.

If you want to read more, the terms to look for in contents or index or web are ‘statistical model selection’, ‘forward selection’, ‘backward selection’, ‘stepwise selection’, ‘selection of covariates’.

Graphical methods

In the file ‘R-survfit.txt’, I show how to extract time survival estimates, which you can use in plotting estimated survival curves to see whether a distribution is reasonable.

fun= “cloglog” in a plot of survfit allows us to check for Weibull fit. (For loglogistic, the relevant function is in Splus, but in R you have to write your own).

To check whether an accelerated life model is sensible, i.e. to see whether the effect of covariates is to scale the median and the mean, we can use quantile-quantile plots. (Next week we will consider proportional hazards models).

The model $\log(T_x) = \beta'x + \log(T_0)$ implies $T_x = T_0e^{\beta'x}$

$$S_{T_x}(t) = S_{T_0}(te^{-\beta'x})$$

$$[\mathbb{P}_{T_x}(T_x > t) = \mathbb{P}_{T_0}(T_x > t) = \mathbb{P}_{T_0}(T_0e^{\beta'x} > t)]$$

If we define t_{0q} by $S_{T_0}(t_{0q}) = q$, the q th quantile, and t_{xq} by $S_{T_x}(t_{xq}) = q$, for $S(t)$ invertible, we have

$$S_{T_x}(t_{xq}) = S_{T_0}(t_{0q})$$

$$S_{T_0}^{-1}\{S_{T_x}(t_{xq})\} = S_{T_0}^{-1}\{S_{T_0}(t_{xq}e^{-\beta'x})\} = S_{T_0}^{-1}\{S_{T_0}(t_{0q})\}$$

$$t_{xq} = t_{0q}e^{\beta'x}$$

i.e the quantiles lies on a straight line through 0, I have not mentioned a specific form of $S(t)$, the distribution, so this is true for all AFT models.

If there are sufficient points in (all) groups of a binary or categorical variable, we can plot quantiles estimated by KM curves for two groups against one another.

Residuals

For the ordinary least squares models, $Y = \underline{\beta}'\underline{x} + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, you remember that the residual are

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}'x_i,$$

and $r_i \sim N(0, \sigma^2)$. There are various plots to consider- qq plots to see whether normal distribution is reasonable, residuals against fitted values to check for heteroscedasticity, residuals against control variable to check for linearity, outliers and leverage points.

Sometimes the studentized residuals are considered

$$r_i^* = \frac{y_i - \hat{y}_i}{\sqrt{\text{var}(y_i - \hat{y}_i)}} .$$

Residuals in survival analysis

For accelerated lifetime models, the obvious thought is

$$r_i = \log t_i - \hat{\beta}' x_i.$$

If there were no censoring, and T followed a lognormal distribution, we would be in familiar territory. In fact, without censoring, even with the error extreme value or logistic, we would have a known distribution. However, because of censoring, there are complications.

Cox-Snell residuals

The i th Cox-Snell residual is defined as

$$r_i = \hat{H}_0(t_i) \times \exp(x_i' \hat{\beta}),$$

where $\hat{H}_0(t_i)$ is the maximum likelihood estimator of the baseline cumulative hazard. These residuals are positive – both terms must be positive. Remember

$$H(t) = -\log S(t) = -\log(1 - F(t)).$$

Now, $F(T)$ is always uniform on $[0, 1]$, for all distributions (even on the full real line). This means that $H(T)$ is a standard exponential (you can prove this for yourself). So r_i are scaled exponential variables. Under proportional hazards assumption (see next week), we can compare these with the exponential distribution (this will be better with plots).

The Martingale residuals are

$$\hat{M}_i = \delta_i - r_i,$$

where δ_i is the usual indicator

$$\delta_i = \begin{cases} 1 & \text{if } i \text{ dead,} \\ 0 & \text{if } i \text{ censored.} \end{cases}$$

Hence the Martingale residuals are always negative for censored values. Very approximately, a Martingale residual compares alive or dead with the cumulative hazard, i.e. the accumulated risk of death. So these residuals are in the range $(-\infty, 1)$, but are not symmetric about 0.

Deviance residuals

First, we have to define the deviance. It is related to the analysis of variance in an ordinary linear regression. We mentioned the proportion of variance explained by the LWBC in the F2

data. We also have an estimate of σ^2 , which indicates how widely the points are scattered about the fitted line. If the line fitted exactly, $\sigma^2 = 0$. If we use a more flexible model than a straight line, we could get a closer fit. The simplest model would be

$$y_i = a + \epsilon_i,$$

with $\epsilon_i \sim N(0, \sigma^2)$, where there is only an intercept term. The other extreme is to have as many parameters as there are data points, n say, with

$$y_i = \alpha_i x_i + \epsilon_i.$$

Then for each point we can find $\hat{\alpha}_i = x_i/y_i$, so the model fits exactly and ϵ_i becomes null (i.e. $\sigma^2 = 0$, $L(\cdot) = 1$ and $\log L(\cdot) = 0$, the maximum possible (log)likelihood). Why consider such a (silly) extreme model? It provides a useful reference point when the error distribution is not normal.

In general, the model deviance is

$$D = -2 \left\{ \ell(\hat{\theta}) - \ell_s(\tilde{\theta}) \right\}$$

(different books use different notation), i.e. the difference between a model with the number of parameters we have for our maximum likelihood estimator of θ , and a model with n parameters.

The log-likelihood is a sum over terms for each individual i , so we can write

$$-2 \left\{ \ell(\hat{\theta}) - \ell_s(\tilde{\theta}) \right\} = -2 \sum_{i=1}^n \left\{ \ell_i(\hat{\theta}) - \ell_{s_i}(\tilde{\theta}) \right\},$$

with $\ell_i(\hat{\theta}) - \ell_{s_i}(\tilde{\theta})$ the contribution from individual i . This is another measure of how the individual varies from the “centre” or “average”.

The i th deviance residual is

$$D_i = \text{sign}(\hat{M}_i) \sqrt{-2 \left\{ \ell_i(\hat{\theta}) - \ell_{s_i}(\tilde{\theta}) \right\}}$$

There is symmetry about ϕ if model fits. The D_i do not always sum to zero, unlike the ordinary least squares residuals.

If there is not too much censoring, then the D_i will be quite near to iid normal. Of course, CP there is heavy censoring (i.e. a lot of censoring).

There are three plots of residuals which are useful.

1. *Probability plots.* If there is light censoring, one can plot the k th ordered D_i of a sample of size n against its normal score $Z \left\{ (k - 0.375)/(n + 0.25) \right\}$. $Z(a)$ is the a th quantile of a standard normal distribution (`qnorm()`). Individual points which don't fit the model will fall away from the straight line. If the model does not fit well, there will not be a straight line.
2. *Deviance residuals against risk scores.* Plot D_i against $\hat{\beta}'x_i$. This should give a random scatter about zero. We look for points separate from general scatter.
3. *Deviance residuals against observation number.* Plot D_i against i , and look for outliers.