Semester Project
Winter 2005-2006

Survival and censored data

Lefteris Samartzis

Professor : A.C. Davison
Assistant : Vahid Partovi Nia

Contents

Chapter 1

Introduction

Survival time is a main topic in medical statistics, and many reasons makes it difficult to get complete data in studies of survival time. A study is often finished before the death of all patients, and we may keep only the information that some patients were still alive at the end of the study, disregarding when they really died. That is a motivation of studing theory of censored data.

We will see that a first possibility to deal with non-complete (we will say censored) data is to be unaware of them, and compute the statistic only on the rest of the data. However we may lose some information ignoring censored data, and actually our estimator will also be biased because ignoring right-censored data for example is ignoring data which has the property to be greater than a given value. In this case, the expectation of our estimator is smaller than the real value of survival time.

We introduce in this report some model for the survival time and see how we can handle with censored data. We may first deal with right- and left-censored data, and then we will show how we could use an algorithm given by Turnbull [4] to get a nonparametric estimate for interval-censored data. Some applications on data will also be given in order to illustrate the theory.

Chapter 2

Survival models

2.1 A trivial exponential model

In order to analyse survival time, we will study the following model:

$$T_i = \exp(\beta_0 + \beta_1 x_i) Z_i,\ Z_i \sim \exp(1),\ \beta_0, \beta_1 \in \mathbb{R},\ i = 1, ..., n. \qquad (1)$$

In this model, x_i is an explanatory covariate, which may be interpreted in our further example as the age of the patient i. A particularity of this model is that the covariate acts multiplicatively on the time scale, it accelerates or deccelerates the survival time depending on the sign of β_1. This kind of models is called accelerated failure time models [1]. This model is reasonable because it takes only positive values for any given parameters $\beta_0, \beta_1 \in \mathbb{R}$. Another advantage is that the expected survival time is :

$$E[T_i|x_i] = E[\exp(\beta_0 + \beta_1 x_i) Z_i | x_i] = \exp(\beta_0 + \beta_1 x_i) E[Z_i] = \exp(\beta_0 + \beta_1 x_i).$$

Distribution of T_i

The cumulative distribution function of T_i is:

$$F_T(t) = P(T_i \leq t) = P(\exp(\beta_0 + \beta_1 x_i) Z_i \leq t) = P(Z_i \leq \frac{t}{\exp(\beta_0 + \beta_1 x_i)})$$

$$= F_{\exp(1)}(\frac{t}{\exp(\beta_0 + \beta_1 x_i)}) = 1 - e^{-\frac{1}{\exp(\beta_0 + \beta_1 x_i)} t}, \quad t \geq 0.$$

In other words:

$$T_i \sim \exp\left(\frac{1}{\exp(\beta_0 + \beta_1 x_i)}\right), \quad \beta_0, \beta_1 \in \mathbb{R},\ i = 1, ..., n.$$

Linearisation of the model

Studing $\log(T_i)$ instead of T_i has the advantage of giving a linear model. Indeed we have:

$$\log(T_i) = \beta_0 + \beta_1 x_i + \log(Z_i), \quad \beta_0, \beta_1 \in \mathbb{R},\ i = 1, ..., n.$$

It is possible to estimate β_0, β_1 using the well-known least squares method. Let $\varepsilon_i = \log(Z_i)$, we find the distribution of ε_i:

$$F_{\varepsilon_i}(t) = P(\log(Z_i) \leq t) = P(Z_i \leq e^t) = F_Z(e^t) = 1 - e^{-e^t}, \quad t \in \mathbb{R}.$$

We remember that the smallest extreme value distribution [11] is given by

$$\text{sEV}(\mu, \sigma): \quad F(t) = 1 - e^{-e^{\frac{t-\mu}{\sigma}}}, \quad t \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma > 0,$$

so, ε_i follows a smallest extreme value distribution with parameters $\mu = 0$ and $\sigma = 1$. That is $\varepsilon_i \sim \text{sEV}(0, 1)$. It is now easy to find the density function of ε_i:

$$f_{\varepsilon_i}(t) = e^t \exp(-e^t), \quad t \in \mathbb{R}.$$

A generalisation of the exponential model

We find a simple generalisation of the exponential model adding a scale parameter $\sigma > 0$ to ε_i. The model becomes $\log(T_i) = \beta_0 + \beta_1 x_i + \sigma\varepsilon_i$. We show that $\sigma\varepsilon_i \sim \text{sEV}(0, \sigma)$:

$$F_{\sigma\varepsilon}(t) = P(\sigma\varepsilon \leq t) = P(\varepsilon \leq \frac{t}{\sigma}) = 1 - \exp(-e^{\frac{t}{\sigma}}), \quad t \in \mathbb{R},\ \sigma > 0.$$

Coming back to the exponential form of the model, we have

$$F_{\exp(\sigma\varepsilon)}(t) = P(\exp(\sigma\varepsilon) \leq t) = P(\sigma\varepsilon \leq \log t) = 1 - \exp(-t^{\frac{1}{\sigma}}), \quad t > 0,\ \sigma > 0$$

that is:

$$T_i = \exp(\beta_0 + \beta_1 x_i) Z_i,\ Z_i \sim \text{Wei}(1, \frac{1}{\sigma}),\ \beta_0, \beta_1 \in \mathbb{R},\ \sigma > 0,\ i = 1, ..., n.$$

and we remember that the general form of the Weibull [11] cumulative distribution function is:

$$F_W(t) = 1 - \exp\Big[- \Big(\frac{t}{b}\Big)^c \Big], \quad t > 0,\ b, c > 0.$$

We notice too that $T_i \sim \text{Wei}(\exp(\beta_0 + \beta_1 x_i), \frac{1}{\sigma})$:

$$P(T_i \leq t) = P\Big\{ Z_i \leq \frac{t}{\exp(\beta_0 + \beta_1 x_i)} \Big\} = 1 - \exp\Big\{ - \Big(\frac{t}{\exp(\beta_0 + \beta_1 x_i)} \Big)^{\frac{1}{\sigma}} \Big\}.$$

Another generalisation is possible adding new covariates into the model:

$$T_i = \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) Z_i,\ Z_i \sim \text{Wei}(1, \frac{1}{\sigma}), \quad i = 1, ..., n. \qquad (2)$$

2.2 Survival and hazard functions

We usually use two different ways to define a distribution: we may give the cumulative distribution function or the density function. In survival time, we prefer to use the survival function which is related to the cdf:

$$S(t) = P(T > t) = 1 - F(t).$$

For survival time we have $t > 0$ in general. This function represents for a given t the probability to live until at least t. That is the reason which we prefer to talk about survival function rather than distribution function.

Another function useful in survival analysis is the hazard function [1]:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\big(t < T \leq t + \Delta t | T > t\big)}{\Delta t} = \frac{f(t)}{S(t)},$$

where T has to be a non negative continous random variable. This function represents the probability of dying really soon, knowing we have live until t. We see that for an exponential distribution, the hazard function is:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda, \quad t > 0.$$

We expected this result as we know [3] that the exponential distribution is memoryless. That means the probability of dying soon doesn't depend on the time. This hazard function is useful, because it gives another interpretation, which may be more interpretable for non-statisticians. Moreover, we could imagine that a model has good reasons that we assume some properties on its hazard function, and then find the survival function corresponding to the particular hazard function with the relation:

$$S(t) = \exp\Big(- \int_0^t h(u)du\Big).$$

Hazard functions is also interesting to fit a model. Cox proposed a distribution-free model where the regression is made on the hazard function [16]. We could compute the hazard function of our model using the distribution we had given. The exponential model (1) has a hazard function given by:

$$h_i(t) = \lambda_i = \frac{1}{\exp(\beta_0 + \beta_1 x_i)}, \quad i = 1, ..., n,$$

and the Weibull model (2):

$$\begin{aligned} h_i(t) &= \frac{\exp -\big(\frac{t}{b}\big)^c c\big(\frac{t}{b}\big)^{c-1}\frac{1}{b}}{\exp -\big(\frac{t}{b}\big)^c}, \\ &= ct^{c-1}\Big(\frac{1}{b}\Big)^c, \\ &= \Big(\frac{1}{\exp(\beta_0 + \beta_1 x_i)}\Big)^{\frac{1}{\sigma}} \frac{1}{\sigma} t^{\frac{1}{\sigma}}. \end{aligned}$$

2.3 Maximum likelihood estimator

Here we are interested in fitting a model to the data. We will use the MLE [9] to find the most likely parameters of our model. In this Section, we show how to fit the exponential model (1) to the data. The parameter that we want to estimate is $\boldsymbol{\beta} = (\beta_0, \beta_1) \in \mathbb{R}^2$. We first need to find the log-likelihood function:

$$f_T(t) = \frac{1}{\exp(\beta_0 + \beta_1 x_i)} \exp\left(\frac{-t}{\exp(\beta_0 + \beta_1 x_i)}\right), \quad t > 0,\ i = 1, ..., n,$$

so

$$l(\beta_0, \beta_1) = \log\Big(\prod_{i=1}^{n} f_T(t)\Big) = \sum_{i=1}^{n}\left[-(\beta_0 + \beta_1 x_i) - \frac{t_i}{\exp(\beta_0 + \beta_1 x_i)}\right].$$

We will see that there is a special case in solving the partial derivatives equations which gives a closed form, and there is no closed form in the general case.

A special case : $\beta_1 = 0$

If we assume that β_1 is fixed to 0, there is a closed form for $\hat{\beta}_0$, given by:

$$\frac{\partial l}{\partial \beta_0}(\beta_0) = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{t_i}{\exp(\beta_0)} = n \Leftrightarrow \sum_{i=1}^{n} \frac{t_i}{\exp(\beta_0)} = n \Leftrightarrow \hat{\beta}_0 = \log(\bar{t}).$$

General case : $\beta_1 \neq 0$

In order to find the maximum of the log-likelihood function, we try to set the partial derivatives to 0 :

$$\begin{cases} \frac{\partial l}{\partial \beta_0}(\hat{\beta}) = 0 \\ \frac{\partial l}{\partial \beta_1}(\hat{\beta}) = 0 \end{cases} \Leftrightarrow \begin{cases} -n + \sum_{i=1}^{n} \frac{t_i}{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = 0 \\ \sum_{i=1}^{n} \frac{x_i t_i}{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)} - \sum_{i=1}^{n} x_i = 0 \end{cases}.$$

We note that we have no closed form for the solution of the system of equations. Hence, we intoduce an iterative method to find the solution. We propose Newton's method.

2.4 Newton's method

Newton's method [8] works on the principle that it is easy to find the maximum of a quadratic approximation of our function $l(\boldsymbol{\beta}) = l(\beta_0, \beta_1)$. Let $l_Q(\boldsymbol{\beta})$ be the quadratic approximation of $l(\boldsymbol{\beta})$ at $\boldsymbol{\beta^0}$, where $\boldsymbol{\beta^0}$ is the vector of

initial values of the parameters. We define a $1\times p$ vector: $\nabla l(\boldsymbol{\beta^0})_j = \frac{\partial l}{\partial \beta_j}(\boldsymbol{\beta^0})$, and a $p \times p$ matrix: $H(\boldsymbol{\beta^0})_{ij} = \frac{\partial^2 l}{\partial \beta_i \beta_j}(\boldsymbol{\beta^0})$. We have:

$$l_Q(\boldsymbol{\beta}) = l(\boldsymbol{\beta^0}) + \nabla l(\boldsymbol{\beta^0})(\boldsymbol{\beta} - \boldsymbol{\beta^0}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta^0})H(\boldsymbol{\beta^0})(\boldsymbol{\beta} - \boldsymbol{\beta^0}).$$

We look for the maximum of the quadratic approximation:

$$\max l_Q(\boldsymbol{\beta}) \Leftrightarrow \nabla l_Q(\boldsymbol{\beta}) = 0 \Leftrightarrow \nabla l(\boldsymbol{\beta^0}) + H(\boldsymbol{\beta^0})(\boldsymbol{\beta} - \boldsymbol{\beta^0}) = 0$$

$$\Leftrightarrow \boldsymbol{\beta} = \boldsymbol{\beta^0} - H^{-1}(\boldsymbol{\beta^0})\nabla l(\boldsymbol{\beta^0}).$$

Starting from the initial parameter $\boldsymbol{\beta^0}$, we use the iterative formula

$$\boldsymbol{\beta^{k+1}} = \boldsymbol{\beta^k} - H^{-1}(\boldsymbol{\beta^k})\nabla l(\boldsymbol{\beta^k}),$$

which would converge to the global maximum of $l(\boldsymbol{\beta})$ assuming that $l(\boldsymbol{\beta})$ is concave and has at least a local maximum. For practical reasons, we will set a stopping condition to our iterative algorithm, which could be for example the absolute convergence criterion: $|x^{k+1} - x^k| < \epsilon$, for some $\epsilon > 0$.

Application of Newton's method

In order to apply Newton's method to the exponential model (1), we have to calculate:

$$\nabla l(\beta) = \left(-n + \sum_{i=1}^{n} \frac{t_i}{\exp(\beta_0 + \beta_1 X_i)}, \sum_{i=1}^{n} \frac{x_i t_i}{\exp(\beta_0 + \beta_1 X_i)} - \sum_{i=1}^{n} x_i \right),$$

$$H(\beta) = \begin{pmatrix} -\sum_{i=1}^{n} \frac{t_i}{\exp(\beta_0+\beta_1 X_i)} & -\sum_{i=1}^{n} \frac{x_i t_i}{\exp(\beta_0+\beta_1 X_i)} \\ -\sum_{i=1}^{n} \frac{x_i t_i}{\exp(\beta_0+\beta_1 X_i)} & -\sum_{i=1}^{n} \frac{x_i^2 t_i}{\exp(\beta_0+\beta_1 X_i)} \end{pmatrix}.$$

Of course, this method works also for the Weibull model proposed in Section 2.1, in which a scale parameter $\sigma > 0$ should be contained in the parameter vector $\boldsymbol{\beta}$ as well.

2.5 Variance and covariance of MLE's

We know that MLE's are asymptotically unbiased [9]. However we are interested in the variance of MLE's in order to compute confidence interval and test our model. We may find the asymptotic variance of MLE's [16] using the Fisher information matrix:

$$I_{ij} = \mathbf{E}\Big[-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \Big].$$

The inverse matrix I^{-1} is the asymptotic variance-covariance matrix of the estimated parameters $\boldsymbol{\beta}$.

2.6 Likelihood ratio tests

For the moment, we have presented two models: the exponential (1) and the Weibull (2). In order to choose which one is better, we introduce the likelihood ratio test. Assume that we have two vectors of parameters corresponding to two nested models: $\boldsymbol{\beta_f} = (\beta_1, ..., \beta_{r_1})$ for the full model and $\boldsymbol{\beta_r} = (\beta_1, ..., \beta_{r_2})$ for the restricted model. Assume that $r_1 > r_2$ and let $d = r_2 - r_1$. Then we have :

$$-2 \log \left[\frac{L(\hat{\beta}_r)}{L(\hat{\beta}_f)} \right] \sim \chi^2_d.$$

In our case, we have that the exponential model is a special case of the Weibull model. Thus, these are nested models.

Chapter 3

Censored data

3.1 Definitions

Right-censored data

A datum T_i is said to be right-censored if the event occurs at a time after a right bound, but we don't know when. The only information we have is this right bound. This is very important in study of survival time, because data are often right-censored.

Left-censored data

A datum T_i is said to be left-censored if we know that the event occurs at a time before a left bound, but we don't know when. It happens, for example, when we know the date of a medical exam that revealed a disease, but we don't know when the patient has been infected.

Interval-censored data

A datum T_i is said to be interval-censored if we know that the event occurs in a time interval $\big(L_i, R_i\big]$, but we don't know exactly when in this interval. It could occurs, for example, when a patient is regularly checked, and one time we discover a medical deterioration. The only information we have is that the deterioration appears between two checks.

Generalisation of interval-censored data

We could easily generalize the definition of interval-censored data in censoring the data in a union of intervals or even in any given set C, and call a datum T_i censored if we only know that the event occurs in the set C_i of possible survival time of patient i.

3.2 Likelihood function of censored data

We are interested in the likelihood function of censored data, because it give us the possibility to compute the MLE in order to fit a model to censored data. In order to analyse such data we write c_i the constant (it may be a random variable too) defined:

$$c_i = \begin{cases} 1, & i\text{th data is not censored,} \\ 0, & i\text{th data is censored.} \end{cases} .$$

The likelihood function of an estimator based on censored data looks like the usual likelihood function, but we have to add the information given by censored data [2]. For right-censored data, we have:

$$L(\beta) = \prod_{i=1}^{n} f_\beta^{c_i}(t_i) S^{(1-c_i)}(t_i),$$

where $S(t) = P(T > t) = 1 - F(t)$ is the survival function. We could do the same for left-censored data:

$$L(\beta) = \prod_{i=1}^{n} f_\beta^{c_i}(t_i) F^{(1-c_i)}(t_i).$$

Then for interval censored data, it becomes:

$$L(\beta) = \prod_{i=1}^{n} f_\beta^{c_i}(t_i) \big[S(L_i) - S(R_i)\big]^{(1-c_i)}.$$

And for general censored data, we have generalized this last expression:

$$L(\beta) = \prod_{i=1}^{n} f_\beta^{c_i}(t_i) P(C_i)^{(1-c_i)},$$

where $P(C_i)$ is the probability of C_i according to the distribution f_β:

$$P(C_i) = \int_{C_i} f_\beta(t) dt.$$

Now, we could theorically fit a model to any censored data. However, it might not be so easy to find the MLE of general-censored data.

3.3 Parametric estimator for censored data

Fitting a model

We have given definitions of censored data, and we will see in this section how we can fit a parametric model to them. In this Section, we assume the Weibull model (2) we have introduced in Section 2.1:

$$T_i = \exp(\beta_0 + \beta_1 x_i) Z_i, \quad Z_i \sim \text{Wei}(1, \frac{1}{\sigma}), \quad \beta_0, \beta_1 \in \mathbb{R},\ \sigma > 0,\ i = 1, ..., n.$$

We want to estimate β_0, β_1 and σ using right-censored data by the MLE:

$$L(\beta_0, \beta_1, \sigma) =$$

$$\prod_{i=1}^{n} \Big[\frac{(1/\sigma)t_i^{(\frac{1}{\sigma}-1)}}{\exp(\beta_0+\beta_1 x_i)^{\frac{1}{\sigma}}} \exp -(\frac{t_i}{\exp(\beta_0+\beta_1 x_i)})^{\frac{1}{\sigma}}\Big]^{c_i} \Big[\exp -(\frac{t_i}{\exp(\beta_0+\beta_1 x_i)})^{\frac{1}{\sigma}}\Big]^{(1-c_i)}$$

$$= \prod_{i=1}^{n} \left\{ \Big[\frac{(1/\sigma)t_i^{(\frac{1}{\sigma}-1)}}{\exp(\beta_0+\beta_1 x_i)^{\frac{1}{\sigma}}}\Big]^{c_i} \exp -\Big[(\frac{t_i}{\exp(\beta_0+\beta_1 x_i)})^{\frac{1}{\sigma}}\Big] \right\}.$$

The log-likelihood $l(\beta_0, \beta_1, \sigma)$ is

$$\sum_{i=1}^{n} \left\{ c_i \Big[-\log(\sigma) + (\frac{1}{\sigma} - 1)\log(t_i) - \frac{1}{\sigma}(\beta_0 + \beta_1 x_i)\Big] - \frac{t_i}{\sigma \exp(\beta_0 + \beta_1 x_i)} \right\}.$$

We notice that setting $c_i = 1$ for $i = 1, ..., n$ and $\sigma = 1$, we find the particular case that we have already met in Section 2.3. Thus we won't find any closed form by solving the equations of the partial derivatives set to zero. In order to find the maximum of the log-likelihood, we will use the Newton's method, as outlined in section 2.4. We could follow the same principle to fit a given model to left-censored, interval-censored or even general-censored data. However, it might be more difficult to maximize the log-likelihood in a general-censored case, and it requires some other optimisation techniques [10].

Special case: $\beta_1 = 0$ and $\sigma = 1$

In the very simple model

$$T_i = \exp(\beta_0) Z_i, \quad Z_i \sim \exp(1),\ \beta_0 \in \mathbb{R},\ i = 1, ..., n,$$

we could find a closed form for β_0 in the case of right-censored data solving the equation:

$$\frac{\partial l}{\partial \beta_0}(\beta_0) = 0.$$

We introduce the following notation :

- n_n : number of non-cenored data
- n_c : number of censored data
- $n = n_n + n_c$: total number of data

Here we have :

$$l(\beta_0) = \sum_{i=1}^{n} -c_i\beta_0 - \frac{t_i}{\exp(\beta_0)},$$

$$\begin{aligned}
\frac{\partial l}{\partial \beta_0}(\beta_0) = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{n} \Big[- c_i + \frac{t_i}{\exp(\beta_0)} \Big] = 0 \\
&\Leftrightarrow \quad \exp(\beta_0) = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} c_i} = \frac{\sum_{i=1}^{n} t_i}{n_n} \\
&\Leftrightarrow \quad \hat{\beta}_0 = \log\Big(\frac{\sum_{i=1}^{n} t_i}{n_n}\Big).
\end{aligned}$$

where $\sum_{i=1}^{n} t_i$ involve the value of the t_i for non-censored data and the censor-bound for censored data. We remember that considering only non-censored data, we found 2.3 the estimator

$$\hat{\beta}_0 = log\bar{t}.$$

Chapter 4

Nonparametric estimator for the survival function

4.1 Trivial estimator

We want to estimate the survival function $S(t)$ of a data set. The first idea is to find the empirical cumulative distribution function of the data and write $\hat{S}(t) = 1 - \hat{F}(t)$. The empirical distribution is:

$$\hat{F}(t) = \frac{|\{t_i \leq t\}|}{n}, \qquad t \in \mathbb{R}$$

We ask ourselves what happens if we compute this estimator on right-censored data. The answer is that it is impossible to compute it for t greater than the smallest right-censored bound R_i, because if we have a data right-censored by R_i and $t > R_i$ we don't know where we have to count it. One solution to this problem is to ignore the censored data and to compute the empirical distribution only for the non-censored data, but the estimate we get is biased as we miss some information. We call this estimator the *trivial* estimator.

4.2 Kaplan-Meier estimator

The great advantage of the Kaplan-Meier (K-M) estimator [18] is that it is computable for right-censored data. The idea of the K-M estimator is given by the conditional probability. Let $t_i \leq t_{i+1}$:

$$\begin{aligned} S(t_i) &= P(T > t_i) \\ &= P(T > t_i, T > t_{i-1}) \\ &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1}) \\ &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-2}) ... P(T > t_0 = 0). \end{aligned}$$

We assume that at the start of the study all subjects were alive, so $P(T > T_0 = 0) = 1$. The conditional probability is

$$P(T > t_i | T > t_{i-1}) = \frac{n_i - d_i}{n_i},$$

where n_i is the number of subjects at risk in the study at the time t_i, and d_i is the number of subject dying at time t_i. The Kaplan-Meier estimator is :

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i \leq t} \Big(1 - \frac{d_i}{n_i}\Big).$$

Example

In order to illustrate the computation of the K-M estimator, we give a very simple example. We have some data on survival time of dogs in years:

$$3, 5^*, 9, 9^*, 10^*, 12,$$

where "$*$" means that the datum is right-censored. To compute K-M estimate, we need to fill the table 4.1. Then we give the corresponding graph of the K-M estimate in Figure 4.1.

t	n	d	s(t)
0	6	0	1
3	6	1	$\frac{5}{6}$
9	4	1	$\frac{5}{6}\frac{3}{4} = \frac{5}{8}$
12	1	1	0

Table 4.1: Survival time of dogs

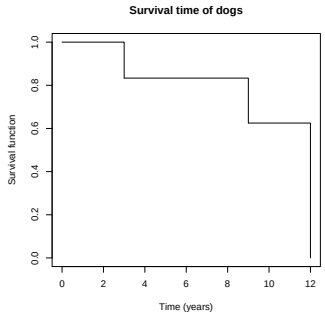

Figure 4.1: K-M estimate for survival time of dogs

Delta-method

In order to estimate the variance of the Kaplan-Meier estimator, we need to introduce the delta method. The delta method uses the first order Taylor expansion of a function f of a random variable X around $\mu = E(X)$ to approximate the variance of $f(X)$:

$$\begin{aligned} f(X) &\simeq f(\mu) + f'(\mu)(X-\mu), \\ \mathrm{VAR}\Big(f(X)\Big) &\simeq \mathrm{VAR}\big(f(\mu) + f'(\mu)(X-\mu)\big) \\ &= f'^2(\mu)\mathrm{VAR}(X-\mu) \\ &= f'^2(\mu)\mathrm{VAR}(X) \\ &= f'^2(\mu)\sigma^2, \end{aligned}$$

where $\sigma^2 = \mathrm{VAR}(X)$. The delta method estimator is:

$$\widehat{VAR}\big(f(X)\big) = f'^2(\hat{\mu})\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is an estimator of $\mathrm{VAR}(X)$ and $\hat{\mu}$ is an estimator of $E[X]$.

Variance of the Kaplan-Meier estimator

The estimate of the variance is given by Greenwood's formula:

$$\widehat{VAR}(\hat{S}(t)) = \hat{S}^2(t)\sum_{t_i<t}\frac{d_i}{n_i(n_i-d_i)}.$$

Here we show how to find the Greenwood's formulae using the delta method. We need to use the delta method two times:

$$\log(X) \simeq \log(\mu) + (X-\mu)\frac{1}{\mu} \Rightarrow \widehat{VAR}(\log(X)) \simeq \hat{\sigma}^2\frac{1}{\hat{\mu}^2},$$

and

$$\exp(X) \simeq \exp(\mu) + (X-\mu)\exp(X) \Rightarrow \widehat{VAR}(\exp(X)) \simeq \exp^2(\hat{\mu})\hat{\sigma}^2.$$

First we look at $\log\hat{S}(t)$:

$$\log\hat{S}_{KM}(t) = \log\prod_{i:t_i\leq t}\big[1-\frac{d_i}{n_i}\big] = \sum_{i:t_i\leq t}\log\big[1-\frac{d_i}{n_i}\big].$$

Let $p_i = P(T > t_i | T > t_{i-1})$ then $\hat{p}_i = \big[1-\frac{d_i}{n_i}\big]$ is an estimate of this conditional probability. That means we assume that $d_i \sim B(n, 1-p_i)$. Hence, the variance of $\hat{p}_i$ is estimated by $\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$. Moreover, the Binomial variables are independent for all subjects in the study. We have then:

$$\widehat{VAR}\Big\{\sum_{i:t_i\leq t}\log(\hat{p}_i)\Big\} = \sum_{i:t_i\leq t}\widehat{VAR}(\log(\hat{p}_i)).$$

A first use of the delta method gives:

$$\widehat{VAR}(log(\hat{p}_i)) \simeq \frac{p_i(1-p_i)}{n}\frac{1}{\hat{p}_i^2} = \frac{1-(1-\frac{d_i}{n_i})}{n_i(1-\frac{d_i}{n_i})} = \frac{\frac{d_i}{n_i}}{n_i-d_i} = \frac{d_i}{n_i(n_i-d_i)}$$

$$\Rightarrow \log\Big[\widehat{VAR}(\hat{S}(t))\Big] \simeq \sum_{i:t_i\leq t}\frac{d_i}{n_i(n_i-d_i)}.$$

We use the delta method for the second time and finally find:

$$\begin{aligned}\widehat{VAR}(\hat{S}(t)) &= \widehat{VAR}\Big\{\exp\Big[\log(\hat{S}(t))\Big]\Big\} \\ &= \exp^2\Big[\log(\hat{S}(t))\Big]\sum_{i:t_i\leq t}\frac{d_i}{n_i(n_i-d_i)} \\ &= \hat{S}^2(t)\sum_{i:t_i\leq t}\frac{d_i}{n_i(n_i-d_i)}.\end{aligned}$$

This last formula had been given in 1926 by Greenwood [17] before Kaplan and Meier [18] published their estimator in 1958.

The Kaplan-Meier estimator for left-censored data

If we have left-censored data, we have to estimate the cumulative distribution function instead of the survival function. We could use an estimator derived from the idea of the Kaplan-Meier estimator. However, here we are interested in the infection time instead of the dead time. We have the following statement (assuming $t_i \leq t_{i+1}$):

$$\begin{aligned}F(t_i) &= P(T_i \leq t_i) \\ &= P(T_i \leq t_i | T_i \leq t_{i+1})P(T_i \leq t_{i+1}) \\ &= P(T \leq t_i | T \leq t_{i+1})P(T \leq t_{i+1}|T \leq t_{i+2})...P(T \leq t_n).\end{aligned}$$

We assume that we have only non-censored or left-censored data. Then we have $P(T \leq t_n) = 1$, as t_n is the greatest time of realisation of all random variables. This suggests the following estimator:

$$\hat{F}_L(t) = \prod_{i:t_i>t}\frac{n_i-d_i}{n_i} = \prod_{i:t_i>t}[1-\frac{d_i}{n_i}]$$

where d_i is the number of subjects getting infected at time t_i and n_i is the number of data in the study at time t_i. If a datum is left-censored, it would enter in the study at the left-censor bound time, and may be count among n_i. If a datum t_i is not left-censored, it would enter in the study at time t_i and count among d_i.
For the variance of this estimator, we may adapt the Greenwood's formula changing $t_i \leq t$ by $t_i > t$.

4.3 Turnbull's algorithm

For the moment, we know how to compute a non-parametric estimator for non-censored, right-censored or left-censored data, but we have nothing for interval-censored data. Turnbull gives an algorithm [4] that we will use to find a nonparametric estimator for interval-censored data. This algorithm works on the principle of Expectation-Maximisation algorithm [5]. Assume that we have some incomplete data, and we want to estimate a parameter. The EM-algorithm starts from an initial parameter, and expects the missing values knowing the initial parameter. Then it finds the parameters that maximize the likelihood considering expected data as the given data. Then the algorithm alternates these two phases until it reaches some stopping conditions. From [5], we know that this algorithm improves the likelihood at every step. We haven't see any package in R using the Turnbull algorithm, so we will implement this.

How does the Turnbull algorithm work?

For this algorithm, we make a partition of $\mathbb{R}_+$ such that each censored set during which an event could happens is a union of intervals of the partition. Actually, we have some knots $k_1, ..., k_{m-1}$ and the partition is formed by the intervals : $J_1 = (0, k_1], J_2 = (k_1, k_2], ..., J_{m-1} = (k_{m-2}, k_{m-1}], J_m = (k_{m-1}, \infty)$.

Let $\mathbf{s} = (s_1, ..., s_m)$ be a m-vector with $s_i \geq 0$, $i = 1, ..., m$ and $\sum_{i=1}^m s_i = 1$. Here $\mathbf{s}$ represents the step of the distribution function $\hat{F}$ between each knot of the partition. We may find $\mathbf{s}$ in order to get an nonparametric estimate of the distribution of the survival time using the relation:

$$\hat{S}(t) = 1 - \sum_{\{i \in K\} \cap \{i \leq t\}} \mathbf{s}_i, \quad t > 0.$$

Now, using the n given data, we define a $n \times m$ matrix α such that

$$\alpha_{ij} = \begin{cases} 1 & \text{if the event } i \text{ could have occured during interval } J_j, \\ 0 & \text{if the event } i \text{ could not have occured during interval } J_j, \end{cases}$$

$$i = 1, ..., n, \; j = 1, ..., m.$$

Then let

$$I_{ij} = \begin{cases} 1 & \text{if the event } i \text{ occurs during interval } J_j, \\ 0 & \text{if the event } i \text{ doesn't occurs at time } J_j, \end{cases}$$

$$i = 1, ..., n, \; j = 1, ..., m.$$

Because of censored data, it is impossible to know I_{ij} for all values of i and j, so we work with its expectation assuming we know $\mathbf{s}$:

$$\mu_{ij} = E_s[I_{ij}] = \alpha_{ij}s_j \Big/ \sum_{k=1}^{m} \alpha_{ik}s_k,\ i = 1, ..., n,\ j = 1, ..., m.$$

Here μ_{ij} represents the probability the event i occurs during interval J_j, assuming survival time has the distribution induced by $\mathbf{s}$.

If we assume now that μ_{ij}'s are observed data, the proportion of data in interval J_j is

$$\pi_j(\mathbf{s}) = (1/n) \sum_{i=1}^{n} \mu_{ij}, j = 1, ..., m.$$

We say that $\mathbf{s}$ is self-consistent if $s_j = \pi_j(\mathbf{s})$, and we show that a self consistent estimate $\mathbf{s}$ is actually an MLE.
Proof: The likelihood function of $\mathbf{s}$ is given by

$$L(\mathbf{s}) = \prod_{i=1}^{n} \Big(\sum_{j=1}^{m} \alpha_{ij}s_j \Big),$$

so the log-likelihood is

$$l(\mathbf{s}) = \sum_{i=1}^{n} \log \Big(\sum_{j=1}^{m} \alpha_{ij}s_j \Big).$$

Now assume that we increase a chosen s_j with $\epsilon > 0$, then we have to divide $s_k\,(k = 1, ..., m)$ by $1 + \epsilon$ in order to keep the sum of $s_k\,(k = 1, .., m)$ equal to one. Then let $d_j(\mathbf{s})$ be the derivative of $l(s)$ with respect to ϵ evaluated at $\epsilon = 0$. Using the chain rule [15] we get:

$$\begin{aligned}
d_j(\mathbf{s}) &= \frac{\partial}{\partial \epsilon} l\Big(\frac{s_1}{1+\epsilon}, ..., \frac{s_{j-1}}{1+\epsilon}, \frac{s_j + \epsilon}{1+\epsilon}, \frac{s_{j+1}}{1+\epsilon}, ..., \frac{s_m}{1+\epsilon} \Big) \Big|_{\epsilon=0} \\
&= \sum_{k=1}^{m} \frac{\partial l}{\partial s_k} \frac{\partial s_k}{\partial \epsilon} \Big|_{\epsilon=0} \\
&= \sum_{k=1}^{m} \Big[\frac{\partial l}{\partial s_k} - \frac{s_k}{(1+\epsilon)^2} \Big] + \frac{\partial l}{\partial s_j} \frac{(1+\epsilon) - \epsilon}{(1+\epsilon)^2} \Big|_{\epsilon=0} \\
&= -\sum_{k=1}^{m} \frac{\partial l}{\partial s_k} s_k + \frac{\partial l}{\partial s_j} \\
&= -\sum_{k=1}^{m} \frac{\alpha_{ik}s_k}{\sum_{k=1}^{m} \alpha_{ik}s_k} + \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{k=1}^{m} \alpha_{ik}s_k} \\
&= -1 + \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{k=1}^{m} \alpha_{ik}s_k}.
\end{aligned}$$

Thus we obtain

$$\pi_j(\mathbf{s}) = (1/n)\sum_{i=1}^{n}\mu_{ij} = (1/n)\sum_{i=1}^{n}\frac{\alpha_{ij}s_j}{\sum_{k=1}^{m}\alpha_{ik}s_k} = s_j + \frac{1}{n}d_j(\mathbf{s})s_j = \Big(1+\frac{d_j(\mathbf{s})}{n}\Big)s_j.$$

From this last relation, we conclude

$$\begin{aligned}\hat{\mathbf{s}} \text{ is a MLE} &\Leftrightarrow d_j(\mathbf{s}) = 0 \text{ OR } (d_j(\mathbf{s}) \leq 0 \text{ AND } s_j = 0), j = 1, ..., m \\ &\Leftrightarrow \pi_j(\mathbf{s}) = s_j, j = 1, ..., m.\end{aligned}$$

Indeed $d_j(\mathbf{s}) \leq 0$ means we could increase the likelihood decreasing by s_j, what is impossible if $s_j = 0$, and using a continuity argument Turnbull claims that $d_j(\mathbf{s}) \leq 0$ if $s_j = 0$.
Now we have shown that a self-consistent estimate $\mathbf{s}$ is an MLE. We will then implement the Turnbull algorithm to find a self-consistent estimate of $\mathbf{s}$.

Input:

From a data set, we construct the partition of $\mathbb{R}_+$, and a $n \times m$ matrix α with coefficients α_{ij} as described.

Output:

We get a vector $\hat{\mathbf{s}}$ of dimension m which represents the step of the survival function made in interval J_j, for $j = 1, ..., m$.

Initialisation:

$\mathbf{s^0} = (\frac{1}{m}, ..., \frac{1}{m})$, $k = 0$.

Loop:

Until stopping conditions are reached do:

- Compute

$$s_j^{k+1} = \pi_j(\mathbf{s^k}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_{ij}s_j^k}{\sum_{l=1}^{m}\alpha_i s_l^k}, \quad \text{for } j = 1, ..., m.$$

- $k := k + 1$.

A stopping condition could be

$$\sum_{j=1}^{m}\left(s_j^{k+1} - s_j^k\right)^2 < \epsilon, \text{or } \max_{j=1,...,n}\left|s_j^{k+1} - s_j^k\right| < \epsilon$$

for a given $\epsilon > 0$.

Chapter 5

Applications

5.1 HIV data

We will now apply the theory on HIV infection data provided by David W. Hosmer and Stanley Lemeshow [1]. We are interested in the survival time of patients who have been infected by HIV, and we will see how age of the patient influence the survival time.

Non-parametric K-M estimator

It is interesting to start with nonparametric estimator in order to have an idea of the shape of the survival function that we may compare then with some fitted models. We first estimate the survival function using the trivial estimator, that ignores the censored data. Then we estimate it with the Kaplan-Meier estimator. Actually, ignoring censored data the trivial estimator is equal to the K-M.

Comparing graphs, we see in Figure 5.1 that the trivial and K-M estima-

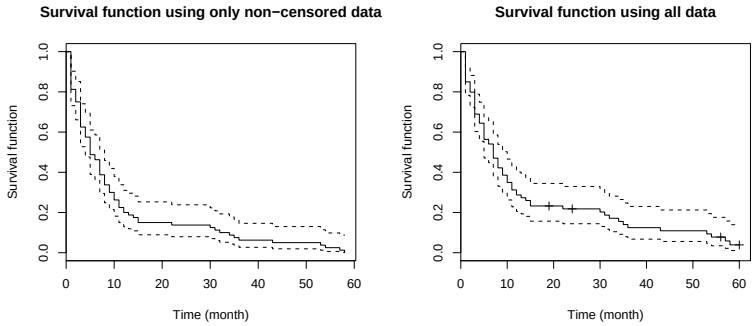

Figure 5.1: Comparison between survival function given by nonparametric estimator, with 95% confidence interval.

tor are quite close to each other. As we expected, the K-M estimator gives a higher estimate of the survival function. That show the bias of the trivial estimator ignoring the censored-data.

Non-parametric Turnbull estimator

We may now compare the K-M estimator with the Turnbull estimator, see Section 4.3. This estimate is used for interval censored data, but we may use it on right-censored data, telling that every censored data is an interval censored data with right interval bound equal to infinity. We compute it using 60 knots corresponding to the first 60 months, and we obtain the estmate given in Figure 5.2.

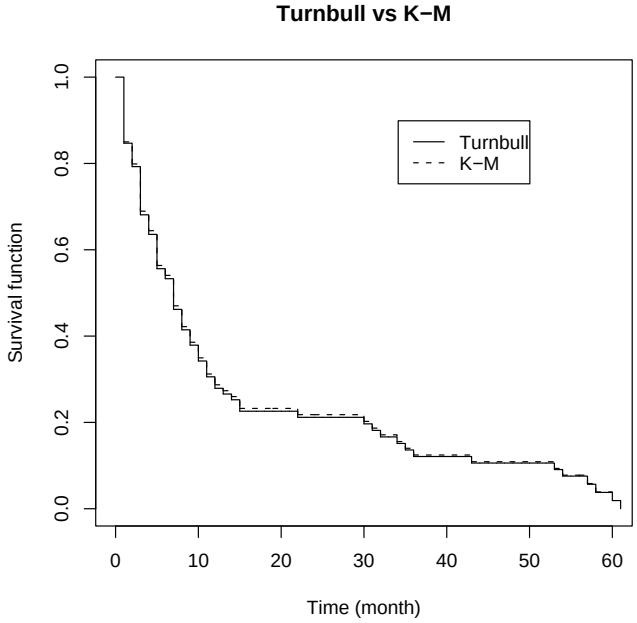

Figure 5.2: Comparison of the Turnbull and K-M estimates.

The two estimators of Figure 5.2 are quite close to each other. Actually they are both MLE, so they should be exactly the same. The little difference that we see here comes from a computation approximation. Actually, the Turnbull algorithm consider non-censored data as interval-censored data, where the interval are $(T_i - 1, T_i]$.

Parametric exponential model

The first model we fit is the next simple model :

$$T_i = \exp(\beta_0) Z_i, \quad Z_i \sim \exp(1), \quad \beta_0 \in \mathbb{R}, \quad i = 1, ..., n.$$

This is a very simple model taking only positive value what is reasonable to explain survival time. In order to fit the model, we find the MLE by maximising the log-likelihood as explained in Section 2.4. However, using the R software, there are some ready-to-use package *Survival* which provides some tools we need to fit our model [6]. In order to test this package, we have also use the function *optim* directly on the log-likelihood function we found and found the same results. More details about *optim* are provided in [7].

```
Considering only non-censored data:
            Value   Std. Error    z       p
(Intercept)  2.36      0.112   21.1   5.66e-99
Scale fixed at 1


Considering all data:
            Value Std. Error     z        p
(Intercept)  2.65      0.112   23.7   1.71e-124
Scale fixed at 1
```

We draw the graph of the survival time using the formula:

$$\hat{S}(t) = 1 - F(t) = \exp\left(\frac{t}{\exp(\beta_0)}\right), \quad t \geq 0, \quad \beta_0 = 2.36, 2.65.$$

The Figure 5.3 shows the estimated survival function according to the simple model fitted one time using all the data and the other time using only non-censored data. We note that they are quite close to each other, and that the estimate survival function using all data is greater than the other. That shows the bias of non-censored data estimate. Then we compare the simple model with the K-M nonparametric estimate, and see in Figure 5.4 that this simple model is a first approximation.

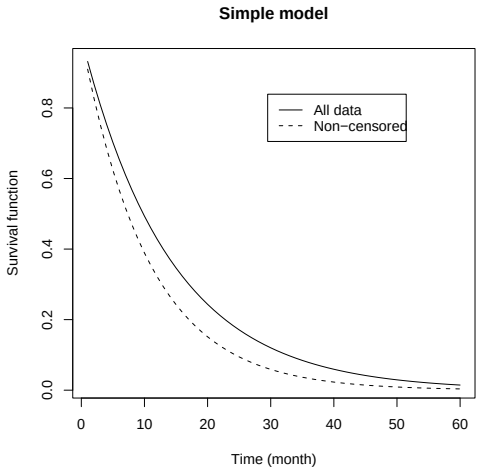

Figure 5.3: Survival function given by the simple exponential model.

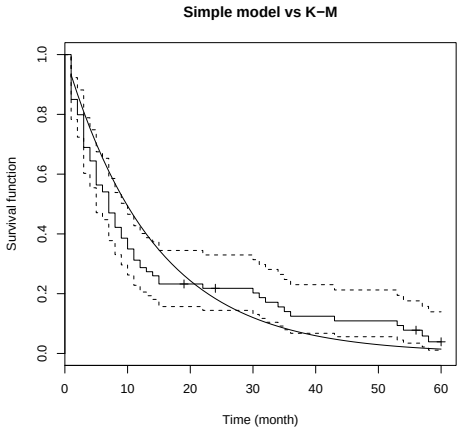

Figure 5.4: Comparison between the simple exponential model and K-M estimate.

Parametric Weibull model

The first model we fitted was really simple, maybe we could explain more using the Weibull model (2) considering the age of patients:

$$T_i = \exp(\beta_0 + \beta_1 \text{AGE}_i) Z_i,\ Z_i \sim \text{Wei}(1, \frac{1}{\sigma}), \quad \beta_0, \beta_1 \in \mathbb{R},\ \sigma > 0,\ i = 1, ..., n.$$

This model is reasonable for HIV survival time. We think that there should be a correlation between age and survival time. We expect that younger patients survive longer than older. We also relax the hypothesis that the scale parameter is fixed at one, which correspond to the exponential model. We fit the Weibull model and get these next results:

```
Considering only non-censored data:
              Value   Std. Error    z      p
(Intercept)  4.9387     0.6266   7.881  3.25e-15,
AGEN        -0.0746     0.0169  -4.418  9.95e-06,
Log(scale)   0.0289     0.0818   0.353  7.24e-01.

Considering all data:
              Value   Std. Error    z      p
(Intercept)  5.8607     0.5918   9.904  4.02e-23,
AGE         -0.0941     0.0160  -5.889  3.88e-09,
Log(scale)   0.0110     0.0817   0.135  8.93e-01.
```

Analysing these results, we keep the hypothesis that age is a significant covariate in the model. We see that the sign of AGE is negative and that is what we expected. Another result we note is that the Log(scale) parameter is not significantly different from 0. That means the scale paramter is not significantly different from 1. Thus, we should prefer an exponential model including the explainatory covariate AGE.

Parametric exponential model considering AGE

We fit the model :

$$T_i = \exp(\beta_0 + \beta_1 AGE_i)Z_i, \quad Z_i \sim \exp(1), \quad \beta_0 \in \mathbb{R}, \quad i = 1, ..., n.$$

and find the results:

```
             Value Std. Error     z        p
(Intercept)  5.859     0.5853 10.01 1.37e-23,
AGE         -0.094     0.0158 -5.96 2.59e-09.

Scale fixed at 1.
```

We could perform a likelihood ratio test to show that the Weibull model is not significantly better. Here the exponential model is a special case of the Weibull model with one fixed parameters : $\sigma = 1$. The log-likelihood of the exponential model is -275 and -275 too for the Weibull model. It's obvious to say that the exponential model is better, as the likelihood of both models are the same.
We could also perform a likelihood ratio test between the exponential model considering age and the other without covariates. Here, the score of the test is

$$-2 * (-292.3 + 275) = 34.6 > \chi^2_{1,\,0.95} = 3.84,$$

which confirm that the covariates AGE should be added. Some difficulties appears when we want to draw the graph of the survival function of this last model, because the survival time doesn't only depend on the time, but it depends on the age too. In order to draw the graph, we compute the expected survival time $\hat{T}_i = \exp(5.86 - 0.09 AGE_i)$ for each AGE_i in the data set. Then we have a set of survival times $\hat{T}_i$ and we draw the empirical survival function:

$$\hat{S}(t) = \frac{|\hat{T}_i > t|}{n}.$$

In the Figure 5.5 we see that, like the simple exponential model, the model with the AGE covariate seems to overestimate the survival function for little time and then under-evaluates it. If we are just looking at the graph, we may think that the simple model gives a survival function which is closer to the K-M estimate. This model is still not very adapted. Then note in Figure 5.6 how the model explains the difference of expected survival time between young and older patients.

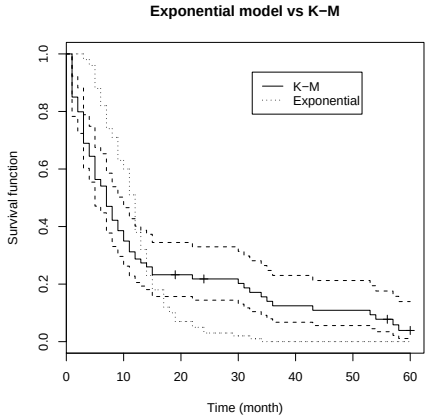

Figure 5.5: Comparison between survival functions given by the exponential model considering age and K-M estimate.

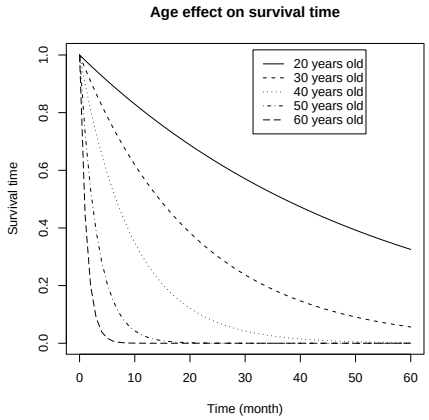

Figure 5.6: Comparison between survival functions at differents ages (according to the exponential model).

In Figure 5.7, we show the non-censored data (bullet) and the relation between age and survival time deducted by the exponential model (line). For a given age, we note that there is a great difference of survival time. For example at 26 years old, there is a patient dying after one month and another after 43. To improve the prediction, it seems necessary to add new covariates into the model.

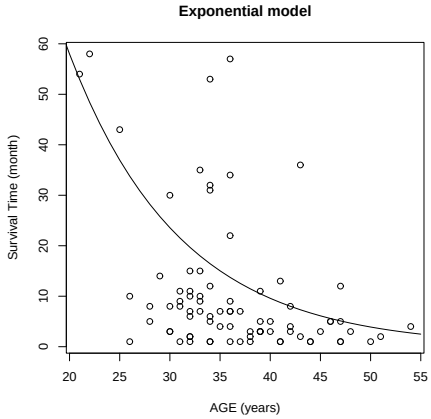

Figure 5.7: Relation between age and survival time.

We remember that in Figure 5.5, the model seems to give a survival function which does not fit very well to the K-M estimate. It might come from a bad assumption on the distribution of the error term. In order to test if the exponential distribution is a good hypothesis, we will look at the non-censored residuals and performed a Kolmogorov-Smirnov test [9] to see if they follow an exponential distribution.

```
        One-sample Kolmogorov-Smirnov test

data:  resexp
D = 0.4611, p-value = 3.442e-15
alternative hypothesis: two.sided
```

We conclude that the exponential distribution is not a good choice for our data. To get a better fit, we should maybe take a look to other models as the log logistic [2] distibution which provides another accelarate failure time model or the log normal distibution or even a generalized gamma distribution. However, we may remember that the most important is not to find a model that perfectly fit the data, but to be able to say why a model is adequate for some data.

5.2 Applications on cosmetic deterioration of breast cancer patients data

Here we give an application of the Turnbull algorithm on interval-censored data. John P. Klein and Melvin L. Moeschberger give data [2] about the cosmetic deterioration of the breast after the beginning of a cancer treatment. These data were obtained in order to compare the effect of two treatments on cosmetic deterioration. The study checks patients every 4-6 month at the beginning and then less frequently. When a deterioration occurs, it happens in an interval of time between two medical checks. If deterioration has never happend, we consider the last medical check as a right-censored bound. We compute a nonparametric estimator of both treatments (with or without chemotherapy) using the Turnbull algorithm.

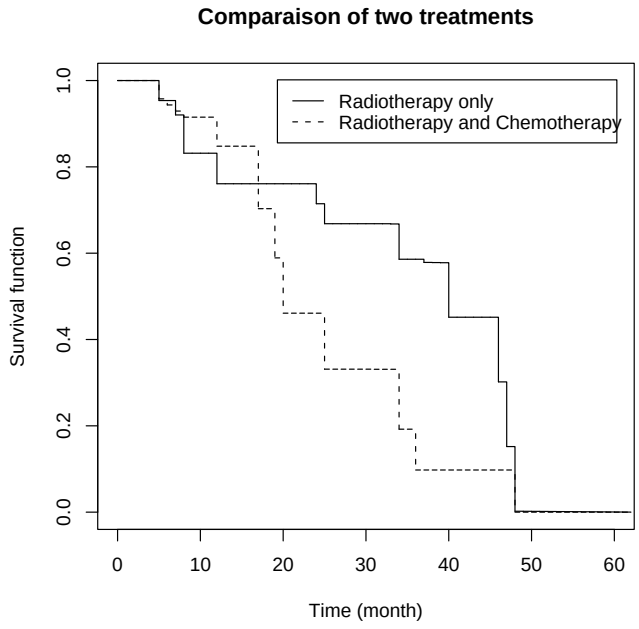

Figure 5.8: Nonparametric comparison between two treatments.

We see in Figure 5.8 that the treatment with chemotherapy seems to accelerate the deterioration, and that's what we expected. In order to test if the treatment with chemotherapy really accelarate the time of deterioration, we could look at the Kolmogorov-Smirnov test [9] for two samples. The problem is to deal with censored data in the sample, so an idea is to simulate data samples according to the estimated survival function, and then compute a 2-sample Kolmogorov-Smirnov test on these simulated sample.

We have simulated a sample of 95 survival time according to each estimated distributions, computed a Kolmogorov-Smirnov test, and found:

```
        Two-sample Kolmogorov-Smirnov test

data:  s1 and s2
D = 0.5474, p-value = 8.702e-13
alternative hypothesis: two.sided
```

We could also perform the test with the hypothesis that treatment without chemotherapy is better:

```
data:  s1 and s2
D^- = 0.5474, p-value = 4.351e-13
alternative hypothesis: less
```

We conclude that the treatment without chemoterapy improves highly significantly the time of deterioration. This test confirms our first idea given by Figure 5.8.

Chapter 6

Conclusion

We have seen in this report some fundamental theories on study of survival time using censored data. We are able now to construct non-parametric estimate for right-, left- or interval-censored data. Some given examples shows how to use non-parametric estimators as a good start to guess the shape of the survival function. These estimators are useful for comparisons with fitted parametric models.
We have introduced some reasonable accelerated life time models for survival time, and of course there is a lot of more complicated models, but the basic ideas are given here.

Appendix A

HIV data

Here we give the survival time in month of each patient, his age and a censoring information. CENSOR = 1 if the datum is not censored and CENSOR = 0 if the datum is right-censored. This data set is provided by [1].

TIME	AGE	CENSOR	TIME	AGE	CENSOR	TIME	AGE	CENSOR	TIME	AGE	CENSOR
5	46	1	11	32	1	11	31	1	3	37	0
6	35	0	2	42	0	56	20	0	43	25	1
8	30	1	5	47	1	2	44	0	1	38	0
3	30	1	4	30	0	3	39	0	6	32	0
22	36	1	1	47	0	15	33	1	53	34	1
1	32	0	13	41	1	1	31	0	14	29	1
7	36	1	3	40	0	10	33	1	4	36	0
9	31	1	2	43	0	1	50	0	54	21	1
3	48	1	1	41	0	7	36	0	1	26	0
12	47	1	30	30	1	3	30	0	1	32	0
2	28	0	7	37	0	3	42	0	8	42	0
12	34	1	4	42	0	2	32	0	5	40	0
1	44	1	8	31	0	32	34	1	1	37	0
15	32	1	5	39	0	3	38	0	1	47	0
34	36	1	10	32	1	10	33	0	2	32	0
1	36	1	2	51	0	11	39	1	7	41	0
4	54	1	9	36	0	3	39	0	1	46	0
19	35	0	36	43	1	7	33	0	10	26	1
3	44	0	3	39	0	5	34	0	24	30	0
2	38	1	9	33	0	31	34	1	7	32	0
2	40	0	3	45	0	5	46	0	12	31	0
6	34	1	35	33	1	58	22	1	4	35	0
60	25	0	8	28	0	1	44	0	57	36	1
7	35	0	1	34	0	2	35	0	1	41	0
60	29	0	5	28	0	1	34	1	12	36	0

Appendix B

Cosmetic Deterioration of Breast Cancer data

On the left side, you find data in which the deterioration has appeared in the time interval (low,up], and on the right side data where no deterioration has appeared until the last check. Treatment 1 is without chemotherapy and treatment 2 is with. Time are given in month, and the data set come from [2].

low	up	Treatment	last	Treatment
0	5	1	15	1
0	7	1	17	1
0	8	1	18	1
4	11	1	22	1
5	11	1	24	1
5	12	1	24	1
6	10	1	32	1
7	14	1	33	1
7	16	1	34	1
11	15	1	36	1
11	18	1	36	1
17	25	1	37	1
17	25	1	37	1
18	26	1	37	1
19	35	1	38	1
25	37	1	40	1
26	40	1	45	1
27	34	1	46	1
36	44	1	46	1
36	48	1	46	1
37	44	1	46	1

low	up	Treatment
0	5	2
0	22	2
4	8	2
4	9	2
5	8	2
8	12	2
8	21	2
10	17	2
10	35	2
11	13	2
11	17	2
11	20	2
12	20	2
13	39	2
14	17	2
14	19	2
15	22	2
16	20	2
16	24	2
16	24	2
16	60	2
17	23	2
17	26	2
17	27	2
18	24	2
18	25	2
19	32	2
22	32	2
24	30	2
24	31	2
30	34	2
30	36	2
33	40	2
34	34	2
35	39	2
44	48	2
48	48	2

last	Treatment
46	1
46	1
46	1
46	1
11	2
11	2
13	2
13	2
13	2
21	2
23	2
31	2
32	2
34	2
34	2
35	2

Appendix C

R Code

```
*************** MASS, Survival ***********************
library(MASS)
library(survival)

*************** NON-CENSORED DATA ******************
k=1
TIMEN=0
AGEN=0
for(i in 1:100) if(CENSOR[i] == 0) i else
{TIMEN[k]<-TIME[i];k<-k+1}
k=1
for(i in 1:100) if(CENSOR[i] == 0) i else
{AGEN[k]<-AGE[i];k<-k+1}

**************FIT TRIVIAL KAPLAN-MEIER ****************
s_empirical_all<-survfit(Surv(TIME,CENSOR))
plot(s_empirical_all,xlab="Time (month)"
,ylab="Survival function")
title(main="Survival function using all data")
s_empirical_n<-survfit(Surv(TIMEN))
plot(s_empirical_n,xlab="Time (month)"
,ylab="Survival function")
title(main="Survival function using only non-censored data")

************** COMPARAISON SELF-made K-M *************
plot(s_empirical_all)
points(s_empirical_n,lty=1,type='S',pch=21)
legend(locator(1),legend=c("All data","Non-censored")
,lty=c(1,1))
title(main="Comparaison between trivial and K-M estimator")
```

```
plot(1:60,km,type="S",xlab="Time (month)"
,ylab="Survival function")
points(s_empirical_all)

**************** SIMPLE MODEL ***********************
s_sim_all<-survreg(Surv(TIME,CENSOR)~1
,dist="exponential")
s_sim_n<-survreg(Surv(TIMEN)~1 ,dist="exponential")
summary(s_sim_all)
summary(s_sim_n)
plot(survfit(Surv(ntimes)),log=T)


******************* WEIBULL MODEL *******************
s_wei_all<-survreg(Surv(TIME,CENSOR)~AGE,dist="weibull")
s_wei_n<-survreg(Surv(TIMEN)~AGEN,dist="weibull")
summary(s_wei_all)
summary(s_wei_n)



************* TURNBULL ALGORITHM **********************
****Init*****
a=1
noeud=1
a<-read.table("alpha.dat")
nc<-length(a)
nr=length(t(a))/nc
noeud<-read.table("knot.dat")
alpha<-matrix(c(t(a)),nrow=nr,ncol=nc,byrow=TRUE)
mu<-matrix(0,nrow=nr,ncol=nc,byrow=TRUE)
s0=1
for (i in 1:nc) s0[i]=1/nc

****Loop******
expe <- function(mu){
num<-matrix(t(t(alpha)*s0),nrow=nr,ncol=nc,byrow=FALSE)
den<-c(alpha %*% s0)
mu<-num/den
}
maxi <- function(s0){
for(i in 1:nc) s0[i]=sum(mu[,i])/nr
s0
```

```
}
***** ALGORITHM *****
for(i in 1:50)
{
mu<-expe(mu)
s0<-maxi(s0)
}
survival.fct=1
for(i in 2:(nc+1)) survival.fct[i-1]-s0[i-1]
->survival.fct[i]
survival.fct
***** GRAPH **********
plot(c(0,t(noeud)),survival.fct,type="s",xlab="Time"
,ylab="Survival function")
title(main="Turnbull estimation")

*********** COMPARAISON K-M, TURNBULL **************
**** ALPHA ****
nr=100
nc=61
alpha<-matrix(1:nc*nr,nrow=nr,ncol=nc,byrow=TRUE)
mu<-matrix(0,nrow=nr,ncol=nc,byrow=TRUE)
one=0
for (i in 1:nr)
{
 one=0;
 for(j in 1:nc)
 {
  if(TIME[i]==j) one=1;
  alpha[i,j]<-one;
  if(CENSOR[i]==1) one=0;
 }
}
noeud=1:nc

s0=1
for (i in 1:nc) s0[i]=1/nc


for(i in 1:150)
{
mu<-expe(mu)
s0<-maxi(s0)
}
```

```
survival.fct=1
for(i in 2:(nc+1)) survival.fct[i-1]-s0[i-1]
->survival.fct[i]
plot(c(0,t(noeud)),survival.fct,type="s"
,xlim=c(0,60),xlab="Time (month)"
,ylab="Survival function")
title(main="Turnbull vs K-M")
points(s_empirical_all,lty=2,type='s')
legend(locator(1),legend=c("Turnbull","K-M")
,lty=c(1,2))
plot(c(0,t(noeud)),survival.fct,type="s"
,xlab="Time (month)",ylab="Survival function")
title(main="Turnbull estimation")


********** KAPLAN-MEIER Self-made ******************
c(j):
for(i in 1:60) kmc[i]<- {t<-0;for(j in 1:100)
if(TIME[j]==i &&
CENSOR[j]==0) t+1->t else t->t}

d(j):
for(i in 1:60) kmd[i]<- {t<-0;for(j in 1:100)
if(TIME[j]==i) t+1->t
else t->t}
kmd-kmc->kmd

n(j):
for(i in 2:60) kmn[i]<-kmn[i-1]-kmd[i-1]-kmc[i-1]

km=1
for(i in 2:60) km[i]<-( km[i-1]*
( 1- (kmd[i-1]/kmn[i-1]) ) )

************* EMPIRICAL Self made *****************
for(i in 1:60) nn[i]<-0
for(i in 1:60) for(j in 1:80) if(timenc[j]>i)
nn[i]<-nn[i]+1
else nn[i]->nn[i]
for(i in 1:60) S[i]<-nn[i]/80

************** MLE USING OPTIM ******************
 fn<-function(timefull,beta) { sum(  CENSOR
 * ( - log( beta[3] ) +
```

```
( 1/beta[3] - 1 ) * log(TIME) - 1/beta[3]
* (beta[1] + beta[2] * AGE) ) -
(TIME / ( beta[3] * exp( beta[1]+beta[2]*AGE ) ) ) ) }
ffn<-function(beta){-fn(beta)}
optim(c(1,1,1),ffn)

fnc<-function(beta) {- sum(  CENSOR *
( - log( beta[3] ) + ( 1/beta[3] - 1 )
 * log(TIME) - 1/beta[3] * (beta[1] + beta[2] * AGE)
 - (TIME / ( beta[3] *
exp( beta[1]+beta[2]*AGE ) ) ) ) )  }

Tc=exp(5.85-0.09*AGE)
Tnc=exp(4.94-0.07*AGE)

************* GRAPHIC *************************
***** NON PARAMETRIC ******
plot(1:60,km,type="s",xlab="Time"
,ylab="Probabilities")
points(1:60,S,type="s",lty=2)
legend(locator(1),legend=c('Trivial','Kaplan-Meier')
,lty=c(2,1))

***** SIMPLE MODEL *********
Ssimple<-function(t){exp(-(1/exp(2.65))*t)}
Ssimplew<-function(t){exp(-(1/exp(2.36))*t)}

plot(1:60,Ssimple(1:60),type="l",xlab="Time (month)"
,ylab="Survival function")
points(1:60,Ssimplew(1:60),type="l",lty=2)
legend(locator(1),legend=c("All data","Non-censored")
,lty=c(1,2))
title(main="Simple model")


****** EXP MODEL - AGE ************
twei=0
for(i in 1:100) twei[i]=exp(5.8607-0.0941*AGE[i])
swei<-survfit(Surv(twei))
plot(s_empirical_all,lty=1,xlab="Time (month)"
,ylab="Survival function",)
points(swei,lty=3,type='s')
title(main="K-M vs EXP model")
legend(locator(1),legend=
```

```
c("K-M with confidence interval","Exp")
,lty=c(1,3))

***** DIFFERENT AGES ************
plot(0:60,c(1,exp(-(1/exp(5.85-0.09*40))*1:60)^(1/1.75))
,type="l",lty=3,xlab="Time (month)",ylab="Survival time")
points(0:60,c(1,exp(-(1/exp(5.8607-0.0941*60))*1:60)
^(1/1.75)),type="l",lty=5)
points(0:60,c(1,exp(-(1/exp(5.8607-0.0941*20))*1:60)
^(1/1.75)),type="l",lty=1)
points(0:60,c(1,exp(-(1/exp(5.8607-0.0941*30))*1:60)
^(1/1.75)),type="l",lty=2)
points(0:60,c(1,exp(-(1/exp(5.8607-0.0941*50))*1:60)
^(1/1.75)),type="l",lty=4)
legend(locator(1),legend=c("20 years old","30 years old"
,"40 years old",
"50 years old","60 years old"),lty=c(1,2,3,4,5))
title(main="Age effect on survival time")


***** SIMPLE MODEL vs K-M *****
plot(s_empirical_all,xlab="Time (month)"
,ylab="Survival function")
points(1:60,Ssimple(1:60),type="l")
title(main="Simple model vs K-M")


****** Exp-AGE vs K-M ********
expti<-exp(5.859-0.094*AGE)
empexp<-survfit(Surv(expti))
plot(s_empirical_all,xlab="Time (month)"
,ylab="Survival function")
points(empexp,type="s",lty=3)
legend(locator(1),legend=c("K-M","Exponential")
,lty=c(1,3))
title(main="Exponential model vs K-M")



***************    BREAST CANCER    ****************
cancer<-read.table("cancer1.dat",header=TRUE)
canc<-read.table("cancer2.dat",header=TRUE)
attach(cancer)
attach(canc)
```

```
*** TREATMENT 1 ****
nr=46
ncc=61
nc=ncc+1
alpha<-matrix(0,nrow=nr,ncol=nc,byrow=TRUE)
mu<-matrix(0,nrow=nr,ncol=nc,byrow=TRUE)

one=0
for (i in 1:21)
{
 one=0;
 for(j in 1:ncc)
 {
  if(up[i]>=j && low[i]<j) one=1;
  alpha[i,j]<-one;
  one=0;
 }
}
for (i in 22:nr)
{
 one=0;
 for(j in 1:nc)
 {
  if(bound[i-21]==j) one=1;
  alpha[i,j]<-one;
 }
}
noeud=1:nc
s0=1
for (i in 1:nc) s0[i]=1/nc
for(i in 1:150)
{
mu<-expe(mu)
s0<-maxi(s0)
}
surv.t1=1
for(i in 2:(nc+1)) surv.t1[i-1]-s0[i-1] ->surv.t1[i]
xt1<-noeud

***** TREATMENT 2 *****
nr=49
nc=62
alpha<-matrix(0,nrow=nr,ncol=nc,byrow=TRUE)
```

```
mu<-matrix(0,nrow=nr,ncol=nc,byrow=TRUE)
one=0
for (i in 22:58)
{
 one=0;
 for(j in 1:nc)
 {
  if(up[i]>=j && low[i]<j) one=1;
  alpha[i-21,j]<-one;
  one=0;
 }
}
for (i in 26:37)
{
 one=0;
 for(j in 1:nc)
 {
  if(bound[i]==j) one=1;
  alpha[i+12,j]<-one;
 }
}
alpha[34,34]=1;
alpha[37,48]=1;
noeud=1:nc
s0=1
for (i in 1:nc) s0[i]=1/nc
for(i in 1:150)
{
mu<-expe(mu)
s0<-maxi(s0)
}
surv.t2=1
for(i in 2:(nc+1)) surv.t2[i-1]-s0[i-1] ->surv.t2[i]
xt2<-noeud

***** PLOT GRAPH *******
plot(c(0,xt1),surv.t1,type="s",xlim=c(0,60)
,xlab="Time (month)",ylab="Survival function")
title(main="Comparaison of two treatments")
points(c(0,xt2),surv.t2,lty=2,type='s')
legend(locator(1),legend=c("Radiotherapy only"
,"Radiotherapy and Chemotherapy"),lty=c(1,2))

plot(c(0,t(noeud)),survival.fct,type="s"
```

```
,xlab="Time (month)",ylab="Survival function")
title(main="Turnbull estimation")

******    COMPARE TREATMENT      *********
t1=surv.t1[1:61]
t2=surv.t2[1:61]
pt1=0
pt2=0
for (i in 1:60) pt1[i]=t1[i]-t1[i+1]
for (i in 1:60) pt2[i]=t2[i]-t2[i+1]
s1<-sample(1:60,95,replace=TRUE,prob=pt1)
s2<-sample(1:60,95,replace=TRUE,prob=pt2)
ks.test(s1,s2,alternative="greater")
ks.test(s1,s2,alternative="less")
ks.test(s1,s2,alternative="two.sided")
```

Bibliography

[1] David W. Hosmer & Stanley Lemeshow, *Applied Survival Analysis*, Wiley Interscience, 1999

[2] John P. Klein & Melvin L. Moeschberger, *Survival Analysis*, Springer, 1997

[3] Sheldon M. Ross, *Initiation aux probabilités*, PPUR, 1994

[4] Bruce W. Turnbull, *The empirical distribution function with arbitrarily grouped, censored and truncated data*,Journal of the Royal Statistical Society 38, 1976

[5] A.P. Dempster, N.M. Laird, & D.B. Rubin *Maximum likelihood from incomplete data via the EM algorithm*,Journal of the Royal Statistical Society 39, 1977

[6] W.N.Venables & B.D. Ripley, *Modern Applied Statistics with S*, Springer, fourth edition 2002

[7] The R Development Core Team, *R: A Language and Environment for Statistical Computing*, Version 2.2.0 (2005-10-06)

[8] A.Quarteroni, *Méthode numériques pour le calcul scientifique*, Springer, 2000

[9] S.Morgenthaler, *Introduction à la statistique*, PPUR, 1997

[10] Elijah Polak, *Optimization*, Springer, 1997

[11] Wayne B. Nelson, *Applied Life Data Analysis*, Wiley, 2004

[12] Elisa T. Lee, *Statistical Methods for Survival Data Analysis*, second edition, Wiley, 1992

[13] Regina C. Elandt-Johnson & Norman L. Johnson, *Survival Models and Data Analysis*, Wiley, 1980

[14] E.J. Gumbel, *Statistics of Extremes*, Columbia University press, 1958

[15] Srichti D. Chatterji, *Cours d'Analyse 1*, PPUR, 1997

[16] D.R. Cox, *Regression models and life tables*, Journal of the Royal Statistical Society B 34, 1972

[17] M. Greenwood, *The natural duration of cancer*, Reports on Public Health and Medical Subjects 33, 1926

[18] E.L. Kaplan & P. Meier, *Nonparametric estimation from incomplete observations*, Journal of The American Statistical Association 53, 1958