Chapter 6

Survival Analysis

6.1 An introduction to survival analysis

6.1.1 What is survival data?

Data where a set of 'individuals' are observed and the failure time or lifetime of that individual is recordered is usually called survival data. We note that individual does not necessarily need to be a person but can be an electrical component etc. Examples include:

- Lifetime of a machine component.
- Time until a patient's cure, remission, passing.
- Time for a subject to perform a task.
- Duration of an economic cycle.
- Also it may not be 'time' we are interested in but:
  - Length of run of a particle.
  - Amount of usage of a machine, eg. amount of petrol used etc.

In the case that we do not observe any regressors (explanatory variables) which influence the survival time (such as gender/age of a patient etc), we can model the survival times as iid random variables. If the survival times are believed to have the density $f(x;\theta_0)$, where $f(x;\theta)$ is known but θ_0 is unknown, then the maximum likelihood can be used to

estimate θ. The standard results discussed in Section 2.2 can be easily applied to this type of data.

6.1.2 Definition: The survival, hazard and cumulative hazard functions

Let T denote the survival time of an individual, which has density f. The density f and the distribution function $F(x) = \int_0^x f(u)du$ are not particularly informative about the chance of survival at a given time point. Instead, the survival, hazard and cumlative hazard functions, which are functions of the density and distribution function, are used instead.

- **The survival function**.

  This is $\mathcal{F}(x) = 1 - F(x)$. It is straightforward to see that $\mathcal{F}(x) = P(T > x)$ (observe that the strictly greater than sign is necessary). Therefore, $\mathcal{F}(x)$ is the probability of survival over x.

- **The hazard function**

  The hazard function is defined as

$$\begin{aligned} h(x) &= \lim_{\delta x \to 0} \frac{P(x < T \leq x + \delta x | T > x)}{\delta x} = \lim_{\delta x \to 0} \frac{P(x < T \leq x + \delta x)}{\delta x P(T > x)} \\ &= \frac{1}{\mathcal{F}(x)} \lim_{\delta x \to 0} \frac{F(x + \delta x) - F(x)}{\delta x} = \frac{f(x)}{\mathcal{F}(x)} = -\frac{d \log \mathcal{F}(x)}{dx}. \end{aligned}$$

  We can see from the definition the hazard function is the 'chance' of failure (though it is a normalised probability, not a probability) at time x, given that the individual has survived until time x.

  We see that the hazard function is similar to the density in the sense that it is a positive function. However it does not integrate to one. Indeed, it is not integrable.

- **The cumulative Hazard function**

  This is defined as

$$H(x) = \int_{-\infty}^{x} h(u)du.$$

It is straightforward to see that

$$H(x) = \int_{-\infty}^{x} -\frac{d \log \mathcal{F}(x)}{dx} \rfloor_{x=u} du = -\log \mathcal{F}(x).$$

This is just the analogue of the distribution function, however we observe that unlike the distribution function, $H(x)$ is unbounded.

It is straightforward to show that $f(x) = h(x)\exp(-H(x))$ and $\mathcal{F}(x) = \exp(-H(x))$.

It is useful to know that given any one of $f(x)$, $F(x)$, $H(x)$ and $h(x)$, uniquely defines the other functions. Hence there is a one-to-one correspondence between all these functions.

Example 6.1.1 • **The Exponential distribution**

Suppose that $f(x) = \frac{1}{\theta}\exp(-x/\theta)$.

Then the distribution function is $F(x) = 1 - \exp(-x/\theta)$. $\mathcal{F}(x) = \exp(-x/\theta)$, $h(x) = \frac{1}{\theta}$ *and* $H(x) = x/\theta$.

The exponential distribution is widely used. However, it is not very flexible. We observe that the hazard function is constant over time. This is the well known memoryless property of the exponential distribution. In terms of modelling it means that the chance of failure in the next instant does not depend on on how old the individual is. The exponential distribution cannot model 'aging'.

- **The Weibull distribution**

We recall that this is a generalisation of the exponential distribution, where

$$f(x) = \left(\frac{\alpha}{\theta}\right)\left(\frac{x}{\theta}\right)^{\alpha-1} \exp(-(x/\theta)^{\alpha}); \alpha, \theta > 0, \quad x > 0.$$

For the Weibull distribution

$$\begin{aligned} F(x) &= 1 - \exp\big(-(x/\theta)^{\alpha}\big) \qquad \mathcal{F}(x) = \exp\big(-(x/\theta)^{\alpha}\big) \\ h(x) &= (\alpha/\theta)(x/\theta)^{\alpha-1} \qquad H(x) = (x/\theta)^{\alpha}. \end{aligned}$$

Compared to the exponential distribution the Weibull has a lot more flexibility. Depending on the value of α, *the hazard function* $h(x)$ *can either increase over time or decay over time.*

- **The shortest lifetime model**

  Suppose that $Y_1, \ldots, Y_k$ are independent life times and we are interested in the shortest survival time (for example this could be the shortest survival time of k sibling mice in a lab when given some disease). Let g_i, $\mathcal{G}_i$ H_i and h_i denote the density, survival function, cumalitive hazard and hazard function respectively of Y_i (we do not assume they have the same distribution) and $T = \min(Y_i)$. Then the survival function is

$$\mathcal{F}_T(x) = P(T > x) = \prod_{i=1}^{k} P(Y_i > x) = \prod_{i=1}^{k} \mathcal{G}_i(x).$$

  Since the cumulative hazard function satisfies $H_i(x) = -\log \mathcal{G}_i(x)$, the cumulative hazard function of T is

$$H_T(x) = -\sum_{i=1}^{k} \log \mathcal{G}_i(x) = \sum_{i=1}^{k} H_i(x)$$

  and the hazard function is

$$h_T(x) = \sum_{i=1}^{k} \frac{d(-\log \mathcal{G}_i(x))}{dx} = \sum_{i=1}^{k} h_i(x)$$

- **Survival function with regressors** *See Section 3.2.2.*

Remark 6.1.1 (Discrete Data) *Let us suppose that the survival time are not continuous random variables, but discrete random variables. In other words, T can take any of the values $\{t_i\}_{i=1}^{\infty}$ where $0 \leq t_1 < t_2 < \ldots$. Examples include the first time an individual visits a hospital post operation, in this case it is unlikely that the exact time of visit is known, but the date of visit may be recorded.*

Let $P(T = t_i) = p_i$, using this we can define the survival function, hazard and cumulative hazard function.

(i) **Survival function** *The survival function is*

$$\mathcal{F}_i = P(T > t_i) = \sum_{j=i+1}^{\infty} P(T = t_j) = \sum_{j=i+1}^{\infty} p_j.$$

(ii) **Hazard function** *The hazard function is*

$$\begin{aligned} h_i &= P(t_{i-1} < T \leq t_i | T > t_{i-1}) = \frac{P(T = t_i)}{P(T > T_{i-1})} \\ &= \frac{p_i}{\mathcal{F}_{i-1}} = \frac{\mathcal{F}_{i-1} - \mathcal{F}_i}{\mathcal{F}_{i-1}} = 1 - \frac{\mathcal{F}_i}{\mathcal{F}_{i-1}}. \end{aligned} \tag{6.1}$$

Now by using the above we have the following useful representation of the survival function in terms of hazard function

$$\mathcal{F}_i = \prod_{j=2}^{i} \frac{\mathcal{F}_i}{\mathcal{F}_{i-1}} = \prod_{j=2}^{i} \big(1 - h_j\big) = \prod_{j=1}^{i} \big(1 - h_j\big), \tag{6.2}$$

since $h_1 = 0$ *and* $\mathcal{F}_1 = 1$.

(iii) **Cumulative hazard function** *The cumulative hazard function is* $H_i = \sum_{j=1}^{i} h_j$.

These expression will be very useful when we consider nonparametric estimators of the survival function $\mathcal{F}$.

6.1.3 Censoring and the maximum likelihood

One main feature about survival data which distinguishes, is that often it is "incomplete". This means that there are situations where the random variable (survival time) is not completely observed (this is often called incomplete data). Usually, the incompleteness will take the form as *censoring* (this will be the type of incompleteness we will consider here).

There are many type of censoring, the type of censoring we will consider in this chapter is right censoring. This is where the time of "failure", may not be observed if it "survives" beyond a certain time point. For example, is an individual (independent of its survival time) chooses to leave the study. In this case, we would only know that the individual survived beyond a certain time point. This is called right censoring. Left censoring arises when the start (or birth) of an individual is unknown (hence it is known when an individual passes away, but the individuals year of birth is unknown), we will not consider this problem here.

Let us suppose that T_i is the survival time, which may not be observed and we observe instead $Y_i = \min(T_i, c_i)$, where c_i is the potential censoring time. We *do know* if the data

has been censored, and together with Y_i we observe the indicator variable

$$\delta_i = \begin{cases} 1 & T_i \leq c_i \quad \text{(uncensored)} \\ 0 & T_i > c_i \quad \text{(censored)} \end{cases}.$$

Hence, in survival analysis we typically observe $\{(Y_i, \delta_i)\}_{i=1}^n$. We use the observations $\{(Y_i, \delta_i)\}_{i=1}^n$ to make inference about unknown parameters in the model.

Let us suppose that T_i has the distribution $f(x; \theta_0)$, where f is known but θ_0 is unknown.

Naive approaches to likelihood construction

There are two naive approaches for estimating θ_0. One method is to ignore the fact that the observations are censored and use time of censoring as if the were failure times. Hence define the likelihood

$$\mathcal{L}_{1,n}(\theta) = \sum_{i=1}^n \log f(Y_i; \theta),$$

and use as the parameter estimator $\hat{\theta}_{1,n} = \arg\max_{\theta \in \Theta} \mathcal{L}_{1,n}(\theta)$. The fundemental problem with this approach is that it will be biased. To see this consider the expectation of $n^{-1}\mathcal{L}_{1,n}(\theta)$ (for convenience let $c_i = c$). Since

$$Y_i = T_i I(T_i \leq c) + cI(T_i > c) \Rightarrow \log f(Y_i; \theta) = [\log f(T_i; \theta)]I(T_i \leq c) + [\log f(c; \theta)]I(T_i > c)$$

this gives the likelihood

$$\mathrm{E}(\log f(Y_i; \theta)) = \int_0^c \log f(x; \theta) f(x; \theta_0) dx + \underbrace{\mathcal{F}(c; \theta_0)}_{\text{probability of censoring}} \log f(c; \theta).$$

There is no reason to believe that θ_0 maximises the above. For example, suppose f is the exponential distribution, using the $\mathcal{L}_{1,n}(\theta)$ leads to the estimator $\widehat{\theta}_{1,n} = n^{-1} \sum_{i=1}^n Y_i$, which is clearly a biased estimator of θ_0. Hence this approach should be avoided since the resulting estimator is biased.

Another method is to construct the likelihood function by ignoring the censored data. In other words use the log-likelihood function

$$\mathcal{L}_{2,n}(\theta) = \sum_{i=1}^n \delta_i \log f(Y_i; \theta),$$

and let $\hat{\theta}_{2,n} = \arg\max_{\theta\in\Theta} \mathcal{L}_{2,n}(\theta)$ be an estimator of θ. It can be shown that if a fixed censor value is used, i.e. $Y_i = \min(T_i, c)$, then this estimator is not a consistent estimator of θ, it is also biased. As above, consider the expectation of $n^{-1}\mathcal{L}_{2,n}(\theta)$, which is

$$\mathrm{E}(\delta_i \log f(Y_i;\theta)) = \int_0^c \log f(x;\theta) f(x;\theta_0) dx.$$

It can be shown that θ_0 does not maximise the above. Of course, the problem with the above "likelihood" is that it is not the correct likelihood (if it were then Theorem 2.6.1 tells us that the parameter will maximise the expected likelihood). The correct likelihood conditions on the non-censored data being less than c to give

$$\mathcal{L}_{2,n}(\theta) = \sum_{i=1}^{n} \delta_i \left(\log f(Y_i;\theta) - \log(1 - \mathcal{F}(c;\theta))\right).$$

This likelihood gives a consistent estimator of the θ; this see why consider its expectation

$$\mathrm{E}\left(n^{-1}\mathcal{L}_{2,n}(\theta)\right) = \mathrm{E}\left(\log \frac{f(Y_i;\theta)}{\mathcal{F}(c;\theta)} \big| T_i < c\right) = \int_0^c \log \frac{f(x;\theta)}{\mathcal{F}(c;\theta)} \log \frac{f(c;\theta_0)}{\mathcal{F}(c;\theta_0)} dx.$$

Define the "new" density $g(x;\theta) = \frac{f(x;\theta)}{\mathcal{F}(c;\theta)}$ for $0 \leq x < c$. Now by using Theorem 2.6.1 we immediately see that the $\mathrm{E}\left(n^{-1}\mathcal{L}_{2,n}(\theta)\right)$ is maximised at $\theta = \theta_0$. However, since we have not used all the data we have lost "information" and the variance will be larger than a likelihood that includes the censored data.

The likelihood under censoring (review of Section 1.2)

The likelihood under censoring can be constructed using both the density and distribution functions or the hazard and cumulative hazard functions. Both are equivalent. The log-likelihood will be a mixture of probabilities and densities, depending on whether the observation was censored or not. We observe (Y_i, δ_i) where $Y_i = \min(T_i, c_i)$ and δ_i is the indicator variable. In this section we treat c_i as if they were deterministic, we consider the case that they are random later.

We first observe that if $\delta_i = 1$, then the log-likelihood of the individual observation Y_i is $\log f(Y_i;\theta)$, since

$$P(Y_i = x|\delta_i = 1) = P(T_i = x|T_i \leq c_i) = \frac{f(x;\theta)}{1 - \mathcal{F}(c_i;\theta)} dx = \frac{h(y;\theta)\mathcal{F}(x;\theta)}{1 - \mathcal{F}(c_i;\theta)} dx. \tag{6.3}$$

On the other hand, if $\delta_i = 0$, the log likelihood of the individual observation $Y_i = c|\delta_i = 0$ is simply one, since if $\delta_i = 0$, then $Y_i = c_i$ (it is given). Of course it is clear that

$P(\delta_i = 1) = 1 - \mathcal{F}(c_i;\theta)$ and $P(\delta_i = 0) = \mathcal{F}(c_i;\theta)$. Thus altogether the joint density of $\{Y_i, \delta_i\}$ is

$$\left(\frac{f(x;\theta)}{1-\mathcal{F}(c_i;\theta)} \times (1-\mathcal{F}(c_i;\theta))\right)^{\delta_i}\left(1 \times \mathcal{F}(c_i;\theta)\right)^{1-\delta_i} = f(x;\theta)^{\delta_i}\mathcal{F}(c_i;\theta)^{1-\delta_i}.$$

Therefore by using $f(Y_i;\theta) = h(Y_i;\theta)\mathcal{F}(Y_i;\theta)$, and $H(Y_i;\theta) = -\log \mathcal{F}(Y_i;\theta)$, the joint log-likelihood of $\{(Y_i, \delta_i)\}_{i=1}^n$ is

$$\begin{aligned}\mathcal{L}_n(\theta) &= \sum_{i=1}^n \Big(\delta_i \log f(Y_i;\theta) + (1-\delta_i)\log\big(1 - F(Y_i;\theta)\big)\Big) \\ &= \sum_{i=1}^n \delta_i\big(\log h(T_i;\theta) - H(T_i;\theta)\big) - \sum_{i=1}^n (1-\delta_i)H(c_i;\theta) \\ &= \sum_{i=1}^n \delta_i \log h(Y_i;\theta) - \sum_{i=1}^n H(Y_i;\theta). \end{aligned} \tag{6.4}$$

You may see the last representation in papers on survival data. Hence we use as the maximum likelihood estimator $\hat{\theta}_n = \arg\max \mathcal{L}_n(\theta)$.

Example 6.1.2 The exponential distribution *Suppose that the density of T_i is $f(x;\theta) = \theta^{-1}\exp(x/\theta)$, then by using (6.4) the likelihood is*

$$\mathcal{L}_n(\theta) = \sum_{i=1}^n \Big(\delta_i\big(-\log\theta - \theta^{-1}Y_i\big) - (1-\delta_i)\theta^{-1}Y_i\Big).$$

By differentiating the above it is straightforward to show that the maximum likelihood estimator is

$$\hat{\theta}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)c_i}{\sum_{i=1}^n \delta_i}.$$

6.1.4 Types of censoring and consistency of the mle

It can be shown that under certain censoring regimes the estimator converges to the true parameter and is asymptotically normal. More precisely the aim is to show that

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)^{-1}), \tag{6.5}$$

where

$$I(\theta) = -\mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^n \delta_i \frac{\partial^2 \log f(Y_i;\theta)}{\partial\theta^2} + \frac{1}{n}\sum_{i=1}^n (1-\delta_i)\frac{\partial^2 \log \mathcal{F}(c_i;\theta)}{\partial\theta^2}\Big).$$

Note that typically we replace the Fisher information with the observed Fisher information

$$\widetilde{I}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial^2\log f(Y_i;\theta)}{\partial\theta^2} + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\partial^2\log \mathcal{F}(c_i;\theta)}{\partial\theta^2}.$$

We discuss the behaviour of the likelihood estimator for different censoring regimes.

Non-random censoring

Let us suppose that $Y_i = \min(T_i, c)$, where c is some deterministic censoring point (for example the number of years cancer patients are observed). We first show that the expectation of the likelihood is maximum at the true parameter (this under certain conditions means that the mle defined in (6.4) will converge to the true parameter). Taking expectation of $\mathcal{L}_n(\theta)$ gives

$$\begin{aligned}\mathrm{E}\Big(n^{-1}\mathcal{L}_n(\theta)\Big) &= \mathrm{E}\Big(\delta_i\log f(T_i;\theta) + (1-\delta_i)\log\mathcal{F}(T_i;\theta)\Big)\\ &= \int_0^c \log f(x;\theta)f(x;\theta_0)dx + \mathcal{F}(c;\theta_0)\log\mathcal{F}(c;\theta).\end{aligned}$$

To show that the above is maximum at θ (assuming no restrictions on the parameter space) we differentiate $\mathrm{E}\big(\mathcal{L}_n(\theta)\big)$ with respect to θ and show that it is zero at θ_0. The derivative at θ_0 is

$$\begin{aligned}\frac{\partial\mathrm{E}\big(n^{-1}\mathcal{L}_n(\theta)\big)}{\partial\theta}\rfloor_{\theta=\theta_0} &= \frac{\partial}{\partial\theta}\int_0^c f(x;\theta)dx\rfloor_{\theta=\theta_0} + \frac{\partial\mathcal{F}(c;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}\\ &= \frac{\partial(1-\mathcal{F}(c;\theta))}{\partial\theta}\rfloor_{\theta=\theta_0} + \frac{\partial\mathcal{F}(c;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0} = 0.\end{aligned}$$

This proves that the expectation of the likelihood is maximum at zero (which we would expect, since this all fall under the classical likelihood framework). Now assuming that the standard regularity conditions are satisfied then (6.5) holds where the Fisher information matrix is

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2}\Big(\int_0^c f(x;\theta_0)\log f(x;\theta)dx + \mathcal{F}(c;\theta_0)\log\mathcal{F}(c;\theta)\Big).$$

We observe that when $c = 0$ (thus all the times are censored), the Fisher information is zero, thus the asymptotic variance of the mle estimator, $\hat{\theta}_n$ is not finite (which is consistent with out understanding of the Fisher information matrix). It is worth noting that under this censoring regime the estimator is consistent, but the variance of the estimator will

be larger than when there is no censoring (just compare the Fisher informations for both cases).

In the above, we assume the censoring time c was common for all individuals, such data arises in several studies. For example, a study where life expectancy was followed for up to 5 years after a procedure. However, there also arises data where the censoring time varies over individuals, for example an individual, i, may pull out of a study at time c_i. In this case, the Fisher information matrix is

$$\begin{aligned} I_n(\theta) &= -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \Big(\int_0^{c_i} f(x;\theta_0) \log f(x;\theta) dx + \mathcal{F}(c_i;\theta_0) \log \mathcal{F}(c_i;\theta) \Big) \\ &= \sum_{i=1}^{n} I(\theta; c_i). \end{aligned} \tag{6.6}$$

However, if there is a lot of variability between the censoring times, one can "model these as if they were random". I.e. that c_i are independent realisations from the random variable C. Within this model (6.6) can be viewed as the Fisher information matrix conditioned on the censoring time $C_i = c_i$. However, it is clear that as $n \to \infty$ a limit can be achieved (which cannot be when the censoring in treated as deterministic) and

$$\frac{1}{n} \sum_{i=1}^{n} I(\theta; c_i) \stackrel{\text{a.s.}}{\to} \int_{\mathbb{R}} I(\theta; c) k(c) dc, \tag{6.7}$$

where $k(c)$ denotes the censoring density. The advantage of treating the censoring as random, is that it allows one to understand how the different censoring times influences the limiting variance of the estimator. In the section below we formally incorporate random censoring in the model and consider the conditions required such that the above is the Fisher information matrix.

Random censoring

In the above we have treated the censoring times as fixed. However, they can also be treated as if they were random i.e. the censoring times $\{c_i = C_i\}$ are random. Usually it is assumed that $\{C_i\}$ are iid random variables which are *independent* of the survival times. Furthermore, it is assumed that the distribution of C does not depend on the unknown parameter θ.

Let k and K denote the density and distribution function of $\{C_i\}$. By using the arguments given in (6.3) the likelihood of the joint distribution of $\{(Y_i, \delta_i)\}_{i=1}^n$ can be

obtained. We recall that the probability of $(Y_i \in [y - \frac{h}{2}, y + \frac{h}{2}], \delta_i = 1)$ is

$$\begin{aligned} & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 1\right) \\ = \; & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right] \big| \delta_i = 1\right) P\left(\delta_i = 1\right) \\ = \; & P\left(T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right] \big| \delta_i = 1\right) P\left(\delta_i = 1\right). \end{aligned}$$

Thus

$$\begin{aligned} & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 1\right) = P\left(T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 1\right) \\ = \; & P\left(\delta_i = 1 | T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) P\left(T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) \\ \approx \; & P\left(\delta_i = 1 | T_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) f_{T_i}(y) h \\ = \; & P(T_i \leq C_i | T_i = y) f_{T_i}(y) h \\ = \; & P(y \leq C_i) f_{T_i}(y) h = f_{T_i}(y) \left(1 - K(y)\right) h. \end{aligned}$$

It is very important to note that the last line $P(T_i \leq C_i | T_i = y) = P(y \leq C_i)$ is due to *independence* between T_i and C_i, if this does not hold the expression would involve the joint distribution of Y_i and C_i.

Thus the likelihood of $(Y_i, \delta_i = 1)$ is $f_{T_i}(y)\left(1 - K(y)\right)$. Using a similar argument the probability of $(Y_i \in [y - \frac{h}{2}, y + \frac{h}{2}], \delta_i = 0)$ is

$$\begin{aligned} & P\left(Y_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 0\right) = P\left(C_i \in \left[y - \frac{h}{2}, y + \frac{h}{2}\right], \delta_i = 0\right) \\ = \; & P\left(\delta_i = 0 | C_i = \left[y - \frac{h}{2}, y + \frac{y}{2}\right]\right) f_{C_i}(y) h \\ = \; & P\left(C_i < T_i | C_i = \left[y - \frac{h}{2}, y + \frac{h}{2}\right]\right) f_{C_i}(y) h = k(C_i) \mathcal{F}(C_i; \theta) h. \end{aligned}$$

Thus the likelihood of (Y_i, δ_i) is

$$\left[f_{T_i}(Y_i)\left(1 - K(Y_i)\right)\right]^{\delta_i} \left[k(Y_i)\mathcal{F}(Y_i; \theta)\right]^{1 - \delta_i}.$$

This gives the log-likelihood

$$\begin{aligned}
\mathcal{L}_{n,R}(\theta) &= \sum_{i=1}^{n}\Big(\delta_i\big[\log f(Y_i;\theta)+\log(1-K(Y_i))\big]+(1-\delta_i)\big[\log\big(1-F(C_i;\theta)\big)+\log k(C_i)\big]\Big)\\
&= \underbrace{\sum_{i=1}^{n}\Big(\delta_i\log f(Y_i;\theta)+(1-\delta_i)\log\big(1-F(C_i;\theta)\big)\Big)}_{=\mathcal{L}_n(\theta)}\\
&\quad+\sum_{i=1}^{n}\Big(\delta_i\log(1-K(Y_i))+(1-\delta_i)\log k(C_i)\Big)\\
&= \mathcal{L}_n(\theta)+\sum_{i=1}^{n}\Big(\delta_i\log(1-K(Y_i))+(1-\delta_i)\log k(C_i)\Big). \qquad (6.8)
\end{aligned}$$

The interesting aspect of the above likelihood is that if the censoring density $k(y)$ *does not depend on* θ, then the maximum likelihood estimator of θ_0 is identical to the maximum likelihood estimator using the non-random likelihood (or, equivalently, the likelihood conditioned on C_i) (see (6.3)). In other words

$$\widehat{\theta}_n = \arg\max \mathcal{L}_n(\theta) = \arg\max \mathcal{L}_{n,R}(\theta).$$

Hence the estimators using the two likelihoods are the same. The only difference is the limiting distribution of $\widehat{\theta}_n$.

We now examine what $\widehat{\theta}_n$ is actually estimating in the case of random censoring. To ease notation let us suppose that the censoring times follow an exponential distribution $k(x) = \beta\exp(-\beta x)$ and $K(x) = 1-\exp(-\beta x)$. To see whether $\hat{\theta}_n$ is biased we evaluate the derivative of the likelihood. As both the full likelihood and the conditional yield the same estimators, we consider the expectation of the conditional log-likelihood. This is

$$\begin{aligned}
\mathrm{E}\big(\mathcal{L}_n(\theta)\big) &= n\mathrm{E}\Big(\delta_i\log f(T_i;\theta)\Big)+n\mathrm{E}\Big((1-\delta_i)\log\mathcal{F}(C_i;\theta)\Big)\\
&= n\mathrm{E}\Big(\log f(T_i;\theta)\,\underbrace{\mathrm{E}(\delta_i|T_i)}_{=\exp(-\beta T_i)}\Big)+n\mathrm{E}\Big(\log\mathcal{F}(C_i;\theta)\,\underbrace{\mathrm{E}\big(1-\delta_i|C_i\big)}_{=\mathcal{F}(C_i;\theta_0)}\Big),
\end{aligned}$$

where the above is due to $\mathrm{E}(\delta_i|T_i) = P(C_i > T_i|T_i) = \exp(-\beta T_i)$ and $\mathrm{E}(1-\delta_i|C_i) = P(T_i > C_i|C_i) = \mathcal{F}(C_i;\theta_0)$. Therefore

$$\mathrm{E}\big(\mathcal{L}_n(\theta)\big) = n\Big(\int_0^\infty \exp(-\beta x)\log f(x;\theta)f(x;\theta_0)dx+\int_0^\infty \mathcal{F}(c;\theta_0)\beta\exp(-\beta c)\log\mathcal{F}(c;\theta)dc\Big).$$

It is not immediately obvious that the true parameter θ_0 maximises $\mathrm{E}\big(\mathcal{L}_n(\theta)\big)$, however by using (6.8) the expectation of the true likelihood is

$$\mathrm{E}\left(\mathcal{L}_{R,n}(\theta)\right) = \mathrm{E}\left(\mathcal{L}_n(\theta)\right) + nK.$$

Thus the parameter which maximises the true likelihood also maximises $\mathrm{E}\big(\mathcal{L}_n(\theta)\big)$. Thus by using Theorem 2.6.1, we can show that $\widehat{\theta}_n = \arg\max \mathcal{L}_{R,n}(\theta)$ is a consistent estimator of θ_0. Note that

$$\sqrt{n}\left(\hat{\theta}_{n,R} - \theta_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)^{-1})$$

where

$$\begin{aligned} I(\theta) &= n\int_0^\infty \exp(-\beta x)\left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta_0)^{-1}dx - \\ & n\int_0^\infty \exp(-\beta x)\frac{\partial^2 f(x;\theta)}{\partial\theta^2}dx \\ & +n\int_0^\infty \beta\exp(-\beta c)\left(\frac{\partial \mathcal{F}(c;\theta)}{\partial\theta}\right)^2 \mathcal{F}(c;\theta)^{-1}dc \\ & -n\int_0^\infty \beta\exp(-\beta c)\frac{\partial^2 \mathcal{F}(c;\theta)}{\partial\theta^2}dc. \end{aligned}$$

Thus we see that the random censoring does have an influence on the limiting variance of $\widehat{\theta}_{n,R}$.

Remark 6.1.2 *In the case that the censoring time C depends on the survival time T it is tempting to still use (6.8) as the "likelihood", and use the parameter estimator the parameter which maximises this likelihood. However, care needs to be taken. The likelihood in (6.8) is constructed under the assumption T and C, thus it is not the true likelihood and we cannot use Theorem 2.6.1 to show consistency of the estimator, in fact it is likely to be biased.*

In general, given a data set, it is very difficult to check for dependency between survival and censoring times.

Example 6.1.3 *In the case that T_i is an exponential, see Example 6.1.2, the MLE is*

$$\hat{\theta}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)C_i}{\sum_{i=1}^n \delta_i}.$$

Now suppose that C_i is random, then it is possible to calculate the limit of the above. Since the numerator and denominator are random it is not easy to calculate the expectation. However under certain conditions (the denominator does not converge to zero) we have by Slutsky's theorem that

$$\widehat{\theta}_n \xrightarrow{\mathcal{P}} \frac{\sum_{i=1}^n \mathrm{E}(\delta_i T_i + (1-\delta_i)C_i)}{\sum_{i=1}^n \mathrm{E}(\delta_i)} = \frac{\mathrm{E}(\min(T_i, C_i))}{P(T_i < C_i)}.$$

Definition: Type I and Type II censoring

- *Type I sampling* In this case, there is an upper bound on the observation time. In other words, if $T_i \leq c$ we observe the survival time but if $T_i > c$ we do not observe the survival time. This situation can arise, for example, when a study (audit) ends and there are still individuals who are alive. This is a special case of non-random sampling with $c_i = c$.

- *Type II sampling* We observe the first r failure times, $T_{(1)}, \ldots, T_{(r)}$, but do not observe the $(n-r)$ failure times, whose survival time is greater than $T_{(r)}$ (we have used the ordering notation $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)}$).

6.1.5 The likelihood for censored discrete data

Recall the discrete survival data considered in Remark 6.1.1, where the failures can occur at $\{t_s\}$ where $0 \leq t_1 < t_2 < \ldots$. We will suppose that the censoring of an individual can occur only at the times $\{t_s\}$. We will suppose that the survival time probabilities satisfy $P(T = t_s) = p_s(\theta)$, where the parameter θ is unknown but the function p_s is known, and we want to estimate θ.

Example 6.1.4

(i) The geometric distribution $P(X = k) = p(1-p)^{k-1}$ for $k \geq 1$ (p is the unknown parameter).

(ii) The Poisson distribution $P(X = k) = \lambda^k \exp(-\lambda)/k!$ for $k \geq 0$ (λ is the unknown parameter).

As in the continuous case let Y_i denote the failure time or the time of censoring of the *ith* individual and let δ_i denote whether the *ith* individual is censored or not. Hence, we

observe $\{(Y_i, \delta_i)\}$. To simplify the exposition let us define

$$\begin{aligned} d_s &= \text{number of failures at time } t_s \qquad q_s = \text{number censored at time } t_s \\ N_s &= \sum_{i=s}^{\infty} (d_i + q_i)\,. \end{aligned}$$

So there data would look like this:

Time	No. Failures at time t_i	No. censored at time t_i	Total Number
t_1	d_1	q_1	$N_1 = \sum_{i=1}^{\infty}(d_i + q_i)$
t_2	d_2	q_2	$N_2 = \sum_{i=2}^{\infty}(d_i + q_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Thus we observe from the table above that

$$N_s - d_s = \text{Number survived just before time } t_{s+1}.$$

Hence at any given time t_s, there are d_s "failures" and $N_s - d_s$ "survivals".

Hence, since the data is discrete observing $\{(Y_i, \delta_i)\}$ is equivalent to observing $\{(d_s, q_s)\}$ (i.e. the number of failures and censors at time t_s), in terms of likelihood construction (this leads to equivalent likelihoods). Using $\{(d_s, q_s)\}$ and Remark 6.1.1 we now construct the likelihood. We shall start with the usual (not log) likelihood. Let $P(T = t_s|\theta) = p_s(\theta)$ and $P(T \geq t_s|\theta) = \mathcal{F}_s(\theta)$. Using this notation observe that the probability of (d_s, q_s) is $p_s(\theta)^{d_s} P(T \geq t_s)^{q_s} = p_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s}$, hence the likelihood is

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} p_{Y_i}(\theta)^{\delta_i} \mathcal{F}_{Y_i}(\theta)^{1-\delta_i} = \prod_{s=1}^{\infty} p_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s} \\ &= \prod_{s=1}^{\infty} p_s(\theta)^{d_s} [\sum_{j=s}^{\infty} p_j(\theta)]^{q_s}. \end{aligned}$$

For most parametric inference the above likelihood is relatively straightforward to maximise. However, in the case that our objective is to do nonparametric estimation (where we do not assume a parametric model and directly estimate the probabilities without restricting them to a parametric family), then rewriting the likelihood in terms of the hazard function greatly simplies matters. By using some algebraic manipulations and Remark 6.1.1 we now rewrite the likelihood in terms of the hazard functions. Using that $p_s(\theta) = h_s(\theta)\mathcal{F}_{s-1}(\theta)$ (see equation (6.1)) we have

$$L_n(\theta) = \prod_{s=1} h_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s} \mathcal{F}_{s-1}(\theta)^{d_s} = \prod_{s=1} \underbrace{h_s(\theta)^{d_s} \mathcal{F}_s(\theta)^{q_s + d_{s+1}}}_{\text{realigning the } s} \text{ (since } \mathcal{F}_0(\theta) = 1).$$

Now, substituting $F_s(\theta) = \prod_{j=1}^{s}(1-h_j(\theta))$ (see equation (6.2)) into the above gives

$$\begin{aligned} L_n(\theta) &= \prod_{s=1}^{n} h_s(\theta)^{d_s}\left[\prod_{j=1}^{s}(1-h_j(\theta))\right]^{q_s+d_{s+1}} \\ &= \prod_{s=1}^{n} h_s(\theta)^{d_s}\prod_{j=1}^{s}(1-h_j(\theta))^{q_s+d_{s+1}} \end{aligned}$$

Rearranging the multiplication we see that $h_1(\theta)$ is multiplied by $(1-h_1(\theta))^{\sum_{i=1}(q_i+d_{i+1})}$, $h_2(\theta)$ is multiplied by $(1-h_1(\theta))^{\sum_{i=2}(q_i+d_{i+1})}$ and so forth. Thus

$$L_n(\theta) = \prod_{s=1}^{n} h_s(\theta)^{d_s}\,(1-h_s(\theta))^{\sum_{m=s}^{n}(q_m+d_{m+1})}\,.$$

Recall $N_s = \sum_{m=s}^{n}(q_m+d_m)$. Thus $\sum_{m=s}^{\infty}(q_m+d_{m+1}) = N_s - d_s$ (number survived just before time t_{s+1}) the likelihood can be rewritten as

$$\begin{aligned} L_n(\theta) &= \prod_{s=1}^{\infty} p_s(\theta)^{d_s}[\sum_{j=s}^{\infty} p_j(\theta)]^{q_s} \\ &= \prod_{s=1} h_s(\theta)^{d_s}(1-h_s(\theta))^{N_s-d_s}. \end{aligned}$$

The corresponding log-likelihood is

$$\begin{aligned} \mathcal{L}_n(\theta) &= \sum_{s=1}^{\infty}\left\{d_s\log p_s(\theta)^{d_s} + \log\left[\sum_{j=s}^{\infty} p_j(\theta)]^{q_s}\right]\right\} \\ &= \sum_{s=1}^{\infty}\Big(d_s\log h_s(\theta) + (N_s-d_s)\log(1-h_s(\theta))\Big). \end{aligned} \tag{6.9}$$

Remark 6.1.3 *At time t_s the number of "failures" is d_s and the number of survivors is $N_s - d_s$. The probability of "failure" and "success" is*

$$h_s(\theta) = P(T=s|T\geq s) = \frac{p_s(\theta)}{\sum_{i=s}^{\infty} p_s(\theta)} \qquad 1-h_s(\theta) = P(T>s|T\geq s) = \frac{\sum_{i=s+1}^{\infty} p_i(\theta)}{\sum_{i=s}^{\infty} p_s(\theta)}.$$

Thus $h_s(\theta)^{d_s}(1-h_s(\theta))^{N_s-d_s}$ can be viewed as the probability of d_s failures and $N_s - d_s$ successes at time t_s.

Thus for the discrete time case the mle of θ is the parameter which maximises the above likelihood.

6.2 Nonparametric estimators of the hazard function - the Kaplan-Meier estimator

Let us suppose that $\{T_i\}$ are iid random variables with distribution function F and survival function $\mathcal{F}$. However, we do not know the class of functions from which F or $\mathcal{F}$ may come from. Instead, we want to estimate $\mathcal{F}$ nonparametrically, in order to obtain a good idea of the 'shape' of the survival function. Once we have some idea of its shape, we can conjecture the parametric family which may best fit its shape. See `https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator` for some plots.

If the survival times have *not* been censored the 'best' nonparametric estimator of the cumulative distribution function F is the empirical likelihood

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x \leq T_i).$$

Using the above the empirical survival function $\mathcal{F}(x)$ is

$$\hat{\mathcal{F}}_n(x) = 1 - \hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > x),$$

observe this is a left continuous function (meaning that the limit $\lim_{0<\delta\to 0} F_n(x-\delta)$ exists). We use the notation

$$\begin{aligned} d_s &= \text{number of failures at time } t_s \\ N_s &= n - \sum_{i=1}^{s-1} d_i = \sum_{i=s}^{\infty} d_i = N_{s-1} - d_{s-1} \text{ (corresponds to number of survivals just before } t_s). \end{aligned}$$

If $t_s < x \leq t_{s+1}$, then the empirical survival function can be rewritten as

$$\hat{\mathcal{F}}_n(x) = \frac{N_s - d_s}{n} = \prod_{i=1}^{s}\left(\frac{N_i - d_i}{N_i}\right) = \prod_{i=1}^{s}\left(1 - \frac{d_i}{N_i}\right) \qquad x \in (t_i, t_{i+1}]$$

where $N_1 = n$ are the total in the group. Since the survival times usually come from continuous random variable, $d_i = \{0, 1\}$, the above reduces to

$$\hat{\mathcal{F}}_n(x) = \prod_{i=1}^{s}\left(1 - \frac{1}{N_i}\right)^{d_i} \qquad x \in (t_i, t_{i+1}].$$

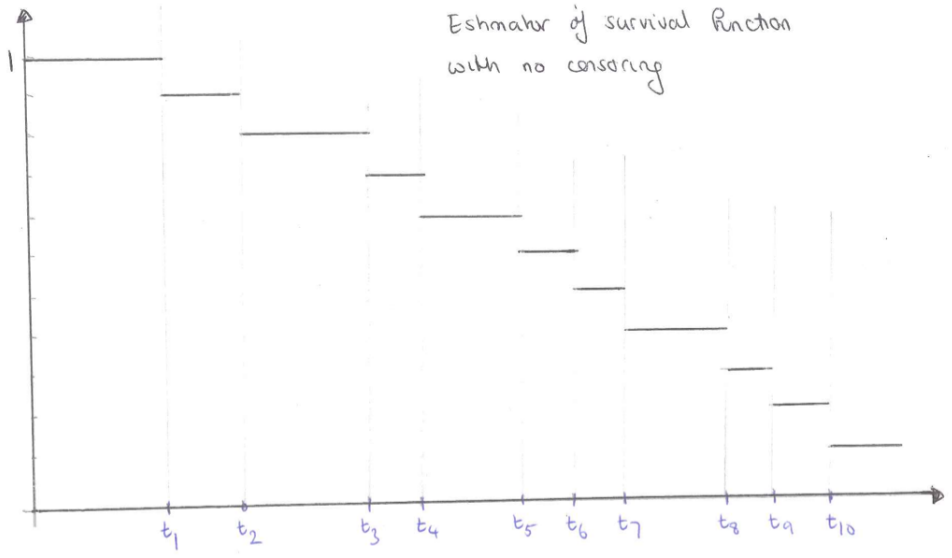

Figure 6.1: The noparametric estimator of the survival function based on the empirical distribution function (with no censoring).

However, in the case, that the survival data is censored and we observe $\{Y_i, \delta_i\}$, then some adjustments have to made to $\hat{\mathcal{F}}_n(x)$ to ensure it is a consistent estimator of the survival function. This leads to the Kaplan-Meier estimator, which is a nonparametric estimator of the survival function $\mathcal{F}$ that takes into account censoring. We will now derive the Kaplan-Meier estimator for discrete data. A typical data set looks like this:

Time	No. Failures at time t_i	No. censored at time t_i	Total Number
t_1	0	0	$N_1 = \sum_{i=1}^{\infty} (d_i + q_i)$
t_2	1	0	$N_2 = \sum_{i=2}^{\infty} (d_i + q_i)$
t_3	0	1	$N_3 = \sum_{i=3}^{\infty} (d_i + q_i)$
t_4	0	1	$N_4 = \sum_{i=4}^{\infty} (d_i + q_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

It is important to note that because these are usually observations from a continuous random variable and observation which as been censored at time t_{s-1} may not have survived up to time t_s. This means that we cannot say that the total number of survivors at time $t_s - \varepsilon$ is $N_s + q_{s-1}$, all we know for sure is that the number of survivors at $t_s - \varepsilon$ is N_s.

The Kaplan-Meier estimator of the hazard function $h_s = P(T = t_s)/P(T > t_s - \varepsilon)$ is

$$\widehat{h}_s = \frac{d_s}{N_s},$$

where d_s are the number of failures at time t_s and N_s are the number of survivors just before time t_s (think $t_s - \varepsilon$). The corresponding estimator of the survival function $P(T > t_s) = \mathcal{F}(t_s)$ is

$$\hat{\mathcal{F}}(t_s) = \prod_{j=1}^{s} \left(1 - \frac{d_j}{N_j}\right).$$

We show below that this estimator maximises the likelihood and in many respects, this is a rather intuitive estimator of the hazard function. For example, if there is *no censoring* then it can be shown that maximum likelihood estimator of the hazard function is

$$\widehat{h}_s = \frac{d_s}{\sum_{i=s}^{\infty} d_s} = \frac{\text{number of failures at time } s}{\text{number who survive just before time } s},$$

which is a very natural estimator (and is equivalent to the nonparametric MLE estimator discussed in Section **??**).

For continuous random variables, $d_j \in \{0, 1\}$ (as it is unlikely two or more survival times are identical), the Kaplan-Meier estimator can be extended to give

$$\hat{\mathcal{F}}(t) = \prod_{j;t>Y_j} \left(1 - \frac{1}{N_j}\right)^{d_j},$$

where Y_j is the time of an event (either failure or censor) and d_j is an indicator on whether it is a failure. One way of interpreting the above is that only the failures are recorded in the product, the censored times simply appear in the number N_j. Most statistical software packages will plot of the survival function estimator. A plot of the estimator is given in Figure 6.2.

We observe that in the case that the survival data is not censored then $N_j = \sum_{s=j}^{m} d_s$, and the Kaplan-Meier estimator reduces to

$$\hat{\mathcal{F}}(t) = \prod_{j;t>Y_j} \left(1 - \frac{1}{N_j}\right).$$

Comparing the estimator of the survival function with and without censoring (compare Figures 6.1 and 6.2) we see that one major difference is the difference between step sizes. In the case there is no censoring the difference between steps in the step function is always n^{-1} whereas when censoring arises the step differences change according to the censoring.

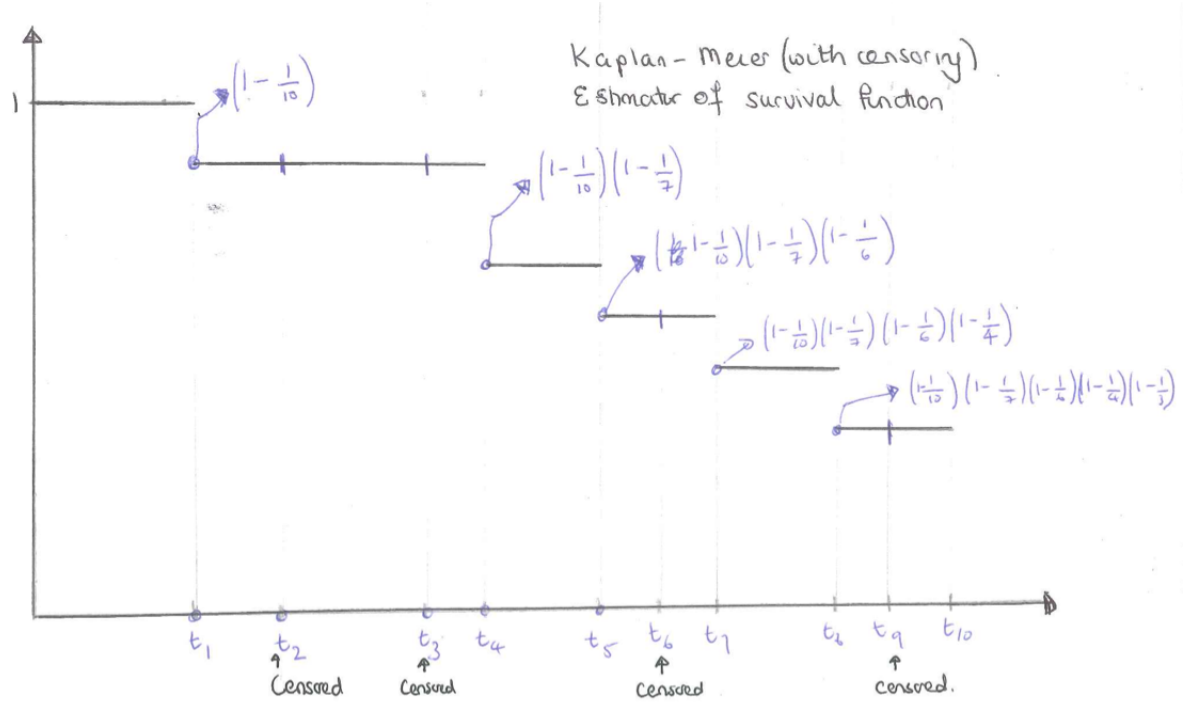

Figure 6.2: An example of the Kaplan-Meier estimator with censoring. The small vertical lines in the plot correspond to censored times.

Derivation of the Kaplan-Meier estimator

We now show that the Kaplan-Meier estimator is the maximum likelihood estimator in the case of censoring. In Section **??** we showed that the empirical distribution is the maximimum of the likelihood for non-censored data. We now show that the Kaplan-Meier estimator is the maximum likelihood estimator when the data is censored. We recall in Section 6.1.5 that the discrete log-likelihood for censored data is

$$\begin{aligned}\mathcal{L}_n(\theta) &= \sum_{s=1}\left(d_s \log p_s(\theta)^{d_s} + q_s \log[\sum_{j=s}^{\infty} p_j(\theta)]\right) \\ &= \sum_{s=1}^{\infty}\left(d_s \log h_s(\theta) + (N_s - d_s)\log(1-h_s(\theta)\right),\end{aligned}$$

where $P(T = t_s) = p_s(\theta)$, d_s are the number of failures at time t_s, q_s are the number of individuals censored at time t_s and $N_s = \sum_{m=s}^{\infty}(q_m + d_m)$. Now the above likelihood is constructed under the assumption that the distribution has a parametric form and the

only unknown is θ. Let us suppose that the probabilities p_s do not have a parametric form. In this case the likelihood is

$$\mathcal{L}_n(p_1, p_2, \ldots) = \sum_{s=1} \left(d_s \log p_s + q_s \log[\sum_{j=s}^{\infty} p_j] \right)$$

subject to the condition that $\sum p_j = 1$. However, it is quite difficult to directly maximise the above. Instead we use the likelihood rewritten in terms of the hazard function (recall equation (6.9))

$$\mathcal{L}_n(h_1, h_2, \ldots) = \sum_{s=1} \Big(d_s \log h_s + (N_s - d_s) \log(1 - h_s) \Big),$$

and maximise this. The derivative of the above with respect to h_s is

$$\frac{\partial \mathcal{L}_n}{\partial h_s} = \frac{d_s}{h_s} - \frac{(N_s - d_s)}{1 - h_s}.$$

Hence by setting the above to zero and solving for h_s gives

$$\hat{h}_s = \frac{d_s}{N_s}.$$

If we recall that d_s = number of failures at time t_s and N_s = number of alive just before time t_s. Hence the non-parametric estimator of the hazard function is rather logical (since the hazard function is the chance of failure at time t, given that no failure has yet occured, ie. $h(t_i) = P(t_i \leq T < t_{i+1} | T \geq t_i)$). Now recalling (6.2) and substituting $\hat{h}_s$ into (6.2) gives the survival function estimator

$$\hat{\mathcal{F}}_s = \prod_{j=1}^{s} \big(1 - \hat{h}_j\big).$$

Rewriting the above, we have the Kaplan-Meier estimator

$$\hat{\mathcal{F}}(t_s) = \prod_{j=1}^{s} \big(1 - \frac{d_j}{N_j}\big).$$

For continuous random variables, $d_j \in \{0, 1\}$ (as it is unlikely two or more survival times are identical), the Kaplan-Meier estimator cab be extended to give

$$\hat{\mathcal{F}}(t) = \prod_{j; Y_j \leq t} \big(1 - \frac{1}{N_j}\big)^{d_j}.$$

Of course given an estimator it is useful to approximate its variance. Some useful approximations are given in Davison (2002), page 197.

6.3 Problems

6.3.1 Some worked problems

Problem: Survival times and random censoring

Example 6.3.1 *Question*

Let us suppose that T and C are exponentially distributed random variables, where the density of T is $\frac{1}{\lambda}\exp(-t/\lambda)$ and the density of C is $\frac{1}{\mu}\exp(-c/\mu)$.

(i) Evaluate the probability $P(T-C<x)$, where x is some finite constant.

(ii) Let us suppose that $\{T_i\}_i$ and $\{C_i\}_i$ are iid survival and censoring times respectively (T_i and C_i are independent of each other), where the densities of T_i and C_i are $f_T(t;\lambda)=\frac{1}{\lambda}\exp(-t/\lambda)$ and $f_C(c;\mu)=\frac{1}{\mu}\exp(-c/\mu)$ respectively. Let $Y_i=\min(T_i,C_i)$ and $\delta_i=1$ if $Y_i=T_i$ and zero otherwise. Suppose λ and μ are unknown. We use the following "likelihood" to estimate λ

$$\mathcal{L}_n(\lambda)=\sum_{i=1}^{n}\delta_i\log f_T(Y_i;\lambda)+\sum_{i=1}^{n}(1-\delta_i)\log\mathcal{F}_T(Y_i;\lambda),$$

where $\mathcal{F}_T$ denotes is the survival function.

Let $\hat{\lambda}_n=\arg\max\mathcal{L}_n(\lambda)$. Show that $\hat{\lambda}_n$ is an asymptotically, unbiased estimator of λ (you can assume that $\hat{\lambda}_n$ converges to some constant).

(iii) Obtain the Fisher information matrix of λ.

(iv) Suppose that $\mu=\lambda$, what can we say about the estimator derived in (ii).

Solutions

(i) $P(T>x)=\exp(-x/\lambda)$ and $P(C>c)=\exp(-c/\mu)$, thus

$$\begin{aligned}P(T<C+x) &= \int \underbrace{P(T<C+x|C=c)}_{\text{use independence}} f_C(c)dc\\ &= \int P(T<c+x)f_C(c)dc\\ &= \int_0^\infty\left[1-\exp(-\frac{c+x}{\lambda})\right]\frac{1}{\mu}\exp(-\frac{c}{\mu})dc = 1-\exp(-x/\lambda)\frac{\lambda}{\lambda+\mu}.\end{aligned}$$

(ii) Differentiating the likelihood

$$\frac{\partial \mathcal{L}_n(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \delta_i \frac{\partial \log f_T(T_i;\lambda)}{\partial \lambda} + \sum_{i=1}^{n} (1-\delta_i) \frac{\partial \log \mathcal{F}_T(C_i;\lambda)}{\partial \lambda},$$

substituting $f(x;\lambda) = \lambda^{-1}\exp(-x/\lambda)$ *and* $\mathcal{F}(x;\lambda) = \exp(-x/\lambda)$ *into the above and equating to zero gives the solution*

$$\hat{\lambda}_n = \frac{\sum_{i=1} \delta_i T_i + \sum_i (1-\delta_i) C_i}{\sum_i \delta_i}.$$

Now we evaluate the expectaton of the numerator and the denominator.

$$\begin{aligned}
\mathrm{E}(\delta_i T_i) &= \mathrm{E}(T_i I(T_i < C_i)) = \mathrm{E}\left(T_i \mathrm{E}(I(C_i > T_i|T_i))\right) \\
&= \mathrm{E}\left(T_i P(C_i > T_i|T_i)\right) = \mathrm{E}\left(T_i P(C_i - T_i > 0|T_i)\right) \\
&= \mathrm{E}(T_i \exp(-T_i/\mu)) = \int t \exp(-t/\mu) \frac{1}{\lambda} \exp(-t/\lambda) dt \\
&= \frac{1}{\lambda} \times \big(\frac{\mu\lambda}{\mu+\lambda}\big)^2 = \frac{\mu^2\lambda}{(\mu+\lambda)^2}
\end{aligned}$$

Similarly we can show that

$$\mathrm{E}((1-\delta_i)C_i) = P\left(C_i P(T_i > C_i|C_i)\right) = \frac{\mu\lambda^2}{(\mu+\lambda)^2}.$$

Finally, we evaluate the denominator $\mathrm{E}(\delta_i) = P(T < C) = 1 - \frac{\lambda}{\mu+\lambda} = \frac{\mu}{\mu+\lambda}$. *Therefore by Slutsky's theorem we have*

$$\hat{\lambda}_n \xrightarrow{\mathcal{P}} \frac{\frac{\mu\lambda^2}{(\mu+\lambda)^2} + \frac{\mu^2\lambda}{(\mu+\lambda)^2}}{\frac{\mu}{\mu+\lambda}} = \lambda.$$

Thus $\hat{\lambda}_n$ *converges in probability to* λ.

(iii) Since the censoring time does not depend on λ *the Fisher information of* λ *is*

$$\begin{aligned}
I(\lambda) &= n\mathrm{E}\left(-\frac{\partial^2 \mathcal{L}_{n,R}(\lambda)}{\partial \lambda^2}\right) = n\mathrm{E}\left(-\frac{\partial^2 \mathcal{L}_n(\lambda)}{\partial \lambda^2}\right) = \frac{n}{\lambda^2}\mathrm{E}[\delta_i] = \frac{n}{\lambda^2} P(T < C) \\
&= \frac{n}{\lambda^2}\frac{\mu}{\lambda+\mu},
\end{aligned}$$

where $\mathcal{L}_{N,R}$ *is defined in (6.8). Thus we observe, the larger the average censoring time* μ *the more information the data contains about* λ.

(iv) It is surprising, but the calculations in (ii) show that even when $\mu = \lambda$ (but we require that T and C are independent), the estimator defined in (ii) is still a consistent estimator of λ. However, because we did not use $\mathcal{L}_{n,R}$ to construct the maximum likelihood estimator and the maximum of $\mathcal{L}_n(\lambda)$ and $\mathcal{L}_{n,R}(\lambda)$ are not necessarily the same, the estimator will not have optimal (smallest) variance.

Problem: survival times and fixed censoring

Example 6.3.2 *Question*

Let us suppose that $\{T_i\}_{i=1}^n$ are survival times which are assumed to be iid (independent, identically distributed) random variables which follow an exponential distribution with density $f(x;\lambda) = \frac{1}{\lambda}\exp(-x/\lambda)$, where the parameter λ is unknown. The survival times may be censored, and we observe $Y_i = \min(T_i, c)$ and the dummy variable $\delta_i = 1$, if $Y_i = T_i$ (no censoring) and $\delta_i = 0$, if $Y_i = c$ (if the survival time is censored, thus c is known).

(a) State the censored log-likelihood for this data set, and show that the estimator of λ is

$$\hat{\lambda}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)c}{\sum_{i=1}^n \delta_i}.$$

(b) By using the above show that when $c > 0$, $\hat{\lambda}_n$ is a consistent of the the parameter λ.

(c) Derive the (expected) information matrix for this estimator and comment on how the information matrix behaves for various values of c.

Solution

(1a) Since $P(Y_i \geq c) = \exp(-c\lambda)$, the log likelihood is

$$\mathcal{L}_n(\lambda) = \sum_{i=1}^n \Big(\delta_i \log\lambda - \delta_i \lambda Y_i - (1-\delta_i)c\lambda \Big).$$

Thus differentiating the above wrt λ and equating to zero gives the mle

$$\hat{\lambda}_n = \frac{\sum_{i=1}^n \delta_i T_i + \sum_{i=1}^n (1-\delta_i)c}{\sum_{i=1}^n \delta_i}.$$

(b) To show that the above estimator is consistent, we use Slutsky's lemma to obtain

$$\hat{\lambda}_n \xrightarrow{\mathcal{P}} \frac{\mathrm{E}\big[\delta T + (1-\delta)c\big]}{\mathrm{E}(\delta)}$$

To show tha $\lambda = \frac{\mathrm{E}[\delta T + (1-\delta)c]}{\mathrm{E}(\delta)}$ *we calculate each of the expectations:*

$$\begin{aligned}
\mathrm{E}(\delta T) &= \int_0^c y \frac{1}{\lambda} \exp(-\lambda y) dy = c\exp(-c/\lambda) - \frac{1}{\lambda}\exp(-c/\lambda) + \lambda \\
\mathrm{E}((1-\delta)c) &= cP(Y > c) = c\exp(-c/\lambda) \\
\mathrm{E}(\delta) &= P(Y \leq c) = 1 - \exp(-c/\lambda).
\end{aligned}$$

Substituting the above into gives $\hat{\lambda}_n \xrightarrow{\mathcal{P}} \lambda$ *as* $n \to \infty$.

(iii) To obtain the expected information matrix we differentiate the likelihood twice and take expections to obtain

$$I(\lambda) = -n\mathrm{E}\left(\delta_i \lambda^{-2}\right) = -\frac{1-\exp(-c/\lambda)}{\lambda^2}.$$

Note that it can be shown that for the censored likelihood $\mathrm{E}(\frac{\partial \mathcal{L}_n(\lambda)}{\partial})^2 = -\mathrm{E}(\frac{\partial^2 \mathcal{L}_n(\lambda)}{\partial \lambda^2})$. *We observe that the larger* c*, the larger the information matrix, thus the smaller the limiting variance.*

6.3.2 Exercises

Exercise 6.1 *If* $\{\mathcal{F}_i\}_{i=1}^n$ *are the survival functions of independent random variables and* $\beta_1 > 0, \ldots, \beta_n > 0$ *show that* $\prod_{i=1}^n \mathcal{F}_i(x)^{\beta_i}$ *is also a survival function and find the corresponding hazard and cumulative hazard functions.*

Exercise 6.2 *Let* $\{Y_i\}_{i=1}^n$ *be iid random variables with hazard function* $h(x) = \lambda$ *subject to type I censoring at time* c*.*

Show that the observed information for λ *is* m/λ^2 *where* m *is the number of* Y_i *that are non-censored and show that the expected information is* $I(\lambda|c) = n[1 - e^{-\lambda c}]/\lambda^2$*.*

Suppose that the censoring time c *is a reailsation from a random variable* C *whose density is*

$$f(c) = \frac{(\lambda\alpha)^\nu c^\nu}{\Gamma(\nu)} \exp(-c\lambda\alpha) \qquad c > 0, \alpha, \nu > 0.$$

Show that the expected information for λ after averaging over c is

$$I(\lambda) = n\left[1-(1+1/\alpha)^{-\nu}\right]/\lambda^2.$$

Consider what happens when

(i) $\alpha \to 0$

(ii) $\alpha \to \infty$

(iii) $\alpha = 1$ *and* $\nu = 1$

(iv) $\nu \to \infty$ *but such that* $\mu = \nu/\alpha$ *is kept fixed.*

In each case explain quantitively the behaviour of $I(\lambda)$.

Exercise 6.3 *Let us suppose that* $\{T_i\}_i$ *are the survival times of lightbulbs. We will assume that* $\{T_i\}$ *are iid random variables with the density* $f(\cdot;\theta_0)$ *and survival function* $\mathcal{F}(\cdot;\theta_0)$*, where* θ_0 *is unknown. The survival times are censored, and* $Y_i = \min(T_i, c)$ *and* δ_i *are observed* ($c > 0$)*, where* $\delta_i = 1$ *if* $Y_i = T_i$ *and is zero otherwise.*

(a) *(i) State the log-likelihood of* $\{(Y_i, \delta_i)\}_i$.

(ii) We denote the above log-likelihood as $\mathcal{L}_n(\theta)$*. Show that*

$$-\mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial\theta^2}\rfloor_{\theta=\theta_0}\right) = \mathrm{E}\left(\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}\right)^2,$$

stating any important assumptions that you may use.

(b) Let us suppose that the above survival times satisfy a Weibull distribution $f(x;\phi,\alpha) = (\frac{\alpha}{\phi})(\frac{x}{\phi})^{\alpha-1}\exp(-(x/\phi)^\alpha)$ *and as in part (a) we observe and* $Y_i = \min(T_i, c)$ *and* δ_i*, where* $c > 0$.

(i) Using your answer in part 2a(i), give the log-likelihood of $\{(Y_i, \delta_i)\}_i$ *for this particular distribution (we denote this as* $\mathcal{L}_n(\alpha,\phi)$*) and derive the profile likelihood of* α *(profile out the nusiance parameter* ϕ*).*

Suppose you wish to test $H_0 : \alpha = 1$ *against* $H_A : \alpha \neq 1$ *using the log-likelihood ratio test, what is the limiting distribution of the test statistic under the null?*

(ii) Let $\hat{\phi}_n, \hat{\alpha}_n = \arg\max \mathcal{L}_n(\alpha, \phi)$ *(maximum likelihood estimators involving the censored likelihood). Do the estimators* $\hat{\phi}_n$ *and* $\hat{\alpha}_n$ *converge to the true parameters* ϕ *and* α *(you can assume that* $\hat{\phi}_n$ *and* $\hat{\alpha}_n$ *converge to some parameters, and your objective is to find whether these parameters are* ϕ *and* α*).*

(iii) Obtain the (expected) Fisher information matrix of maximum likelihood estimators.

(iv) Using your answer in part 2b(iii) derive the limiting variance of the maximum likelihood estimator of $\hat{\alpha}_n$*.*

Exercise 6.4 *Let* T_i *denote the survival time of an electrical component. It is known that the regressors* x_i *influence the survival time* T_i*. To model the influence the regressors have on the survival time the Cox-proportional hazard model is used with the exponential distribution as the baseline distribution and* $\psi(x_i; \beta) = \exp(\beta x_i)$ *as the link function. More precisely the survival function of* T_i *is*

$$\mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i;\beta)},$$

where $\mathcal{F}_0(t) = \exp(-t/\theta)$*. Not all the survival times of the electrical components are observed, and there can arise censoring. Hence we observe* $Y_i = \min(T_i, c_i)$*, where* c_i *is the censoring time and* δ_i*, where* δ_i *is the indicator variable, where* $\delta_i = 0$ *denotes censoring of the* i*th component and* $\delta_i = 1$ *denotes that it is not censored. The parameters* β *and* θ *are unknown.*

(i) Derive the log-likelihood of $\{(Y_i, \delta_i)\}$*.*

(ii) Compute the profile likelihood of the regression parameters β*, profiling out the baseline parameter* θ*.*