1 Preliminaries

Definition 1.1. The survival function S is defined as ;

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

where $T; \Omega \to [0, \infty)$ denotes the random survival time.

Properties of $S(t)$

1. S is a nonincreasing function.
2. S is right continuous.
3. $S(t) \to 0$ as $t \to \infty$.
4. $S(0) \leq 1$ or, $S(0) = 1$ if $P(T = 0) = 0$.

Definition 1.2. The hazard function λ is defined as ;

$$\begin{aligned}\lambda(t) &= \lim_{h \downarrow 0} \frac{P(t \leq T < t + h | T \geq t)}{h} \\ &= \frac{1}{1 - F(t-)} \lim_{h \downarrow 0} \frac{P(T < t + h) - P(T < t)}{h}.\end{aligned}$$

If T is a continuous random variable with a density function f, i.e. $F' = f$,

$$\begin{aligned}\lambda(t) &= \frac{1}{1 - F(t)} \lim_{h \downarrow 0} \frac{F(t + h) - F(t)}{h} = \frac{f(t)}{1 - F(t)} \\ &= -\frac{d}{dt} \log S(t).\end{aligned}$$

The hazard function can alternatively be represented in term of the cumulative hazard function ;

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log S(t) = -\log[1 - F(t)].$$

Notes

1. If for each of the n subjects in a study one observes their survival times, denoted by $T_1, \cdots, T_n$, then this will be referred to as the complete data case.

2 Types of Censoring

Definition 2.1. Type I censoring occurs if an experiment is **started at a given time** for a set of subjects or items, and the experiment is **stopped at a predetermined time**. That is, censoring is a nonrandom value.

Definition 2.2. Type II censoring occurs when an experiment is **continued until a predetermined** number of the subjects under the study have failed.

For example, If the ordered observations are denoted by $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(r)}$ for $1 \leq r \leq n$, then for the remaining $n-r$ subjects one only knows that their failure time is after $T_{(r)}$. We say that the remaining $n-r$ lifetimes are **censored** by the largest observed lifetime $T_(r)$.

3 Random Censorship

3.1 Right Random Censorship

Denote by

$$\begin{aligned}
E_i &= \text{the calender time at which individual } i \text{ enters the study} \\
D_i &= \text{the calender time of the failure of individual } i \\
T_i &= D_i - E_i = \text{individual survival time} \\
C_i &= \text{date of the end of the study} - E_i.
\end{aligned}$$

And the observations consist of

$$(Z_1, \delta_1), \cdots, (Z_n, \delta_n)$$

where

$$\begin{aligned}
&Z_i := \min(T_i, C_i) \quad \& \quad \delta_i := I(T_i \leq C_i) \\
(\therefore) \quad &Z_i = (T_i - C_i)\delta_i + C_i
\end{aligned}$$

Assumptions T_i is independent of C_i for any i.

3.2 Left Random Censorship

The observations consist of

$$\begin{aligned}
&Z_i := \max(T_i, C_i) \quad \& \quad \delta_i := I(C_i \leq T_i) \\
(\therefore) \quad &Z_i = (T_i - C_i)\delta_i - C_i
\end{aligned}$$

and also, T_i is independent of C_i for any i.

3.3 Doubly Random Censored Data

The actual survival time is only observed when it exceeds the left censoring time $C_{l,i}$ and when it does not exceeds the right censoring time $C_{r,i}$.

The observations are now

$$(Z_1, \delta_1), (Y_2, \delta_2), \cdots, (Y_n, \delta_n)$$

where

$$Z_i = \min[\max(T_i, C_{l,i}), C_{r,i}]$$

$$\delta_i = \begin{cases} 1 & if \quad C_{l,i} \le T_i \le C_{r,i} \\ 2 & if \quad C_{r,i} < T_i \\ 3 & if \quad T_i < C_{l,i} \end{cases}$$

3.4 Interval-Censored Data

Interval-censored data typically result from studies in which the objects (subjects) of interest are not constantly followed (or monitored).

4 Estimation Based on Right Censored Data

An observation (Z_i, δ_i) is said to be uncensored if $\delta_i = 1$ and censored otherwise. Denote the distribution function of C as G.

4.1 ML Estimation of a Parametric Model

Suppose that T and C are continuous random variables with density functions f and g, respectively. Denote the parameters of the distribution of T as $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^\top$.

The contribution of an observation $(Z_i, \delta_i = 1)$ to the likelihood is

$$\lim_{\varepsilon \to 0} \frac{P(Z_i - \varepsilon < Z < Z_i + \varepsilon, C \ge T)}{2\varepsilon} = \lim_{\varepsilon \to 0} \frac{P(Z_i - \varepsilon < T < Z_i + \varepsilon, C \ge Z_i)}{2\varepsilon}$$

which can be reduced as

$$f(Z_i)\{1 - G(Z_i)\}$$

because of independence between T and C. Also, for observation $(Z_i, \delta_i = 0)$,

$$\begin{aligned} \lim_{\varepsilon \to 0} \frac{P(Z_i - \varepsilon < Z < Z_i + \varepsilon, C < T)}{2\varepsilon} &= \lim_{\varepsilon \to 0} \frac{P(Z_i - \varepsilon < C < Z_i + \varepsilon, T \ge Z_i)}{2\varepsilon} \\ &= g(Z_i)\{1 - F(Z_i)\}. \end{aligned}$$

In summary, the contribution of an observation (Z_i, δ_i) to the likelihood function is given by

$$[f(Z_i)\{1 - G(Z_i)\}]^{\delta_i}[g(Z_i)\{1 - F(Z_i)\}]^{1-\delta_i}$$

and the entire likelihood is ;

$$L_n[(Z_1, \delta_1), \cdots, (Z_n, \delta_n)] = \prod_{i \in \mathbf{C}^c} [f(Z_i)\{1 - G(Z_i)\}] \prod_{i \in \mathbf{C}} [g(Z_i)\{1 - F(Z_i)\}]$$

where $\mathbf{C}$ is a set of indexes of censored observations. If $\mathbf{C}$ does not involve the parameter $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) = \prod_{i \in \mathbf{C}^c} f(Z_i) \prod_{i \in \mathbf{C}} \{1 - F(Z_i)\}$$

4.2 Nonparametric Estimation of a Survival Function

Suppose that the sample consists of all distinct observations and that it contains n_u uncensored observations and n_c censored observations. Denote the ordered **uncensored** observations by

$$T_{(0)} := 0 < T_{(1)} \leq \cdots \leq T_{(n_u)}.$$

For an arbitrary uncensored observation $T_{(k)}$, we write

$$\begin{aligned} S(T_{(k)}) = P(T > T_{(k)}) &= P(T > T_{(k)}|T > T_{(k-1)})P(T > T_{(k-1)}) \\ &= P(T > T_{(k)}|T > T_{(k-1)})P(T > T_{(k-1)}|T > T_{(k-2)})P(T > T_{(k-2)}) = \cdots \\ &= \prod_{j=1}^{k} \{P(T > T_{(j)}|T > T_{(j-1)})\}P(T > 0) = \prod_{j=1}^{k} \{P(T > T_{(j)}|T > T_{(j-1)})\}. \end{aligned}$$

Now, $P_j := P(T \leq T_{(j)}|T > T_{(j-1)})$ can be estimated empirically by

$$\hat{P}_j = \frac{\sum_{i=1}^{n} I(T_{(j-1)} < T_i \leq T_{(j)})}{\sum_{i=1}^{n} I(T_i \geq T_{(j)})} = \frac{1}{\sum_{i=1}^{n} I(Z_i \geq T_{(j)})} = \frac{1}{\#\{i; Z_i \geq T_{(j)}\}}$$

since $Z_i = T_i$ for $\delta_i = 1$. So,

$$\hat{S}(T_{(k)}) = \prod_{j=1}^{k}(1 - \hat{P}_j) = \prod_{j=1}^{k} \left(1 - \frac{1}{\#\{i; Z_i \geq T_{(j)}\}}\right) = \prod_{j; Z_j \leq T_{(k)}} \left(1 - \frac{1}{\#\{i; Z_i \geq Z_j\}}\right)^{\delta_j}$$

and note that $S(t) = S(T_{(k)})$ for any $t \in [T_{(k)}, T_{(k+1)}]$.

The Kaplan-Meier estimator of $S = 1 - F$ is given by

$$\begin{aligned}\hat{S}_{KM}(t) = 1 - \hat{F}_{KM}(t) &= \prod_{j;Z_j \le t} \left(1 - \frac{1}{\#\{i; Z_i \ge Z_j\}}\right)^{\delta_j} \\ &= \prod_{j;Z_{(j)} \le t} \left(1 - \frac{1}{\#\{i; Z_i \ge Z_{(j)}\}}\right)^{\delta_{(j)}} = \prod_{j;Z_{(j)} \le t} \left(1 - \frac{1}{n-j+1}\right)^{\delta_{(j)}} \\ &= \prod_{j;Z_{(j)} \le t} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}\end{aligned}$$

where $Z_{(1)} \le \cdots \le Z_{(n)}$ and $\delta_{(i)}$ are ordered statistics and corresponding indicator.

Properties of KM estimator

i. Suppose that all observations are uncensored. It is easily seen that in this case the estimator for F reduces to the usual empirical distribution function. Indeed for any $t \in [T_{(k)}, T_{(k+1)}]$, we have

$$\hat{F}_{KM}(t) = 1 - \prod_{j;Z_{(j)} \le t} \left(\frac{n-j}{n-j+1}\right) = 1 - \prod_{j=1}^{k} \left(\frac{n-j}{n-j+1}\right) = 1 - \frac{n-k}{n} = \frac{k}{n}.$$

ii. KM estimator is in fact a generalized likelihood estimator, where maximization of the likelihood is done over a space of functions(p-measures).

KM estimator for ties

When ties between a censored and an uncensored observation occur, the convention is that the uncensored observation happened just before the censored observation. When there are further tied observations, say there are d_j times that Z_j has been observed, then denoting the different observed times by $Z'_{(1)} \le \cdots \le Z'_{(r)}$, and $\delta'_{(i)}$ the associated indicator function, the KaplanMeier estimator can be written as:

$$\hat{F}_{KM}(t) = 1 - \prod_{j;Z'_{(j)} \le t} \left(1 - \frac{d_j}{n_j}\right)^{\delta'_{(j)}}$$

where n_j denotes the number of subjects in the sample at risk at time point $Z'_{(j)}$.

In the case when largest observation is censored

Some special attention is needed when the largest observation $Z_{(n)}$ is censored. In that case, the KaplanMeier estimator is no proper distribution function because

$$\hat{F}_{KM}(Z_{(n)}) = 1 - \prod_{j=1}^{n} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} = 1 - \prod_{j=1}^{n-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} < 1.$$

The usual convention is therefore to treat the largest observation always as uncensored, in other words to define the estimator to be equal to one from the largest observation $Z_{(n)}$ onwards.

4.3 Nonparametric Estimation of a Cumulative Hazard Function

$$\hat{\Lambda}(t) = -\log \hat{S}_{KM}(t) = -\log \prod_{j;Z_{(j)}\leq t} \left(1 - \frac{1}{n-j+1}\right)^{\delta_{(j)}} = \sum_{j;Z_{(j)}\leq t} -\delta_{(j)} \log\left(1 - \frac{1}{n-j+1}\right)$$
$$\approx \sum_{j;Z_{(j)}\leq t} \frac{\delta_{(j)}}{n-j+1} := \hat{\Lambda}_{Nelson}(t)$$

since $-\log(1-x) \approx x$.

5 Regression Models for Right Censored Data

We now look into the regression case when data are subject to right random censorship, and observations on a covariate X are available. The interest is in studying the influence of the covariate X on the survival time T. In this regression context, the observations are

$$(Z_1, \delta_1, X_1), \cdots, (Z_n, \delta_n, X_n).$$

Consider the following general regression model

$$Y \equiv g(T) = m(X) + \sigma(X)\varepsilon$$

where g is a certain given transformation function (e.g., a logarithmic transformation). The (unknown) regression function m describes the influence of X on T. The term $\sigma(X)\varepsilon$ is the error term.

In this regression setup in the context of survival analysis, it is convenient to introduce the concept of conditional hazard function. For a given value x of X, the conditional hazard function at the point t is defined as:

$$\lambda(t|x) = \lim_{h\downarrow 0} \frac{P(t \leq T < t+h | T \geq t, X = x)}{h}.$$

This quantity describes the instantaneous risk of failure giving survival up to the moment t for an individual with value of the covariate $X = x$, knowing that this individual has survived until time t.

5.1 Accelerated Failure Time Model & Proportional Hazards Model

Two very popular regression models for censored data have been around since many decades : the proportional hazards model and the accelerated failure time model. In the **accelerated failure time model**,

$$\lambda(t|x) = \lambda_0\{t\psi(x)\}\psi(x)$$

where $\psi(x)$ is a function describing the influence of X on T. The function λ_0 is a hazard function of reference, referred to as the baseline hazard function. When $\psi(x)$ is such that $\psi(0) = 1$, the function λ_0 represents the hazard function associated with an individual for whom $x = 0$.

In the **proportional hazards model**, the conditional hazard function can be written as :

$$\lambda(t|x) = \lambda_0(t)\exp\{\psi(x)\}.$$

Note that under this model the hazard function associated with $X = x_1$ and the hazard function associated with $X = x_2$ behave as :

$$\frac{\lambda(t|x_2)}{\lambda(t|x_1)} = \exp\{\psi(x_2) - \psi(x_1)\} \quad \forall t$$

In other words, under the proportional hazards model, the conditional hazard functions are proportional to each other.

5.2 Conditional Survival Function

Additional interpretations for the two models can be given. The conditional hazard function can be written in terms of the conditional density function $f(t|x)$ and the conditional survival function $S(t|x) = 1 - F(t|x)$, where $F(t|x)$ is the cumulative distribution function describing the conditional distribution of T given $X = x$. We have

$$S(t|x) = \exp\left[-\int_0^t \lambda(u|x)du\right].$$

Under the **proportional hazards model**, we further obtain

$$S(t|x) = \exp\left[-\exp\{\psi(x)\}\int_0^t \lambda_0(u)du\right] = S_0(t)^{\exp\{\psi(x)\}}$$

where $S_0(\cdot)$ is the survival function associated with the covariate value $X = 0$.

Consider $Y = g(T)$. Then $F_Y(y|x) = F_T\{g^{-1}(y)|x\}$ and $F_Y(y|x) = F_T\{g^{-1}(y)|x\}$ and $f_Y(y|x) = \frac{f_T\{g^{-1}(y)|x\}}{g'\{g^{-1}(y)\}}$. The conditional hazard function of $Y = g(T)$, given $X = x$, is then given by

$$\lambda_Y(y|x) = \frac{\lambda_T\{g^{-1}(y)|x\}}{g'\{g^{-1}(y)\}}.$$

This general statement allows to give an additional interpretation to the accelerated failure time model. Let T_0 be a survival time for which the hazard function is $\lambda_0(\cdot)$, independent of x. From the above expression, it is then easily seen that the hazard function associated with the random variable $T = \frac{T_0}{\psi(x)}$ is given by

$$\lambda(t|x) = \lambda_0\{t\psi(x)\}\psi(x)$$

In other words, in an accelerated failure time model, the covariate X reduces the survival time of an individual with a factor $\psi(x)$. The random variable T admits a regression model

$$\log T = -\log\{\psi(X)\} + \varepsilon$$

where $\varepsilon = \log T_0$. On the other hand, when T satisfies a regression model with $\sigma(X)$ independent of X and with $g(T) = \log T$, then T satisfies an accelerated failure time model.