Lecture 32: Survivor and Hazard Functions

(Text Section 10.2)

Let Y denote survival time, and let $f_Y(y)$ be its probability density function. The cdf of Y is then

$$F_Y(y) = \mathrm{P}(Y \leq y) = \int_0^y f_Y(t)dt.$$

Hence, $F_Y(y)$ represents the probability of failure by time y.

The *survivor function* is defined as

$$S_Y(y) = \mathrm{P}(Y > y) = 1 - F_Y(y).$$

In other words, the survivor function is the probability of survival beyond time y.

One use of the survivor function is to predict quantiles of the survival time. For example, the median survival time (say, y_{50}) may be of interest. (The median may be preferable to the mean as a measure of centrality if the data are highly skewed.) We can compute y_{50} as the solution to

$$S_Y(y) = 1 - 0.5 = 0.5.$$

Likewise, the time by which 90% of the population will have failed (say, y_{90}) is given by the solution to

$$S_Y(y) = 1 - 0.9 = 0.1.$$

The *hazard function* is defined as

$$h_Y(y) = \frac{f_Y(y)}{S_Y(y)}. \tag{1}$$

The hazard function is **not** a density or a probability. However, we can think of it as the probability of failure in an infinitesimally small time period between y and $y + \partial y$ *given* that the subject has survived up till time y. In this sense, the hazard is a measure of risk: the greater the hazard between times y_1 and y_2, the greater the risk of failure in this time interval.

In particular, since by definition

$$f_Y(y) = \lim_{\partial y \to 0} \frac{F_Y(y + \partial y) - F_Y(y)}{\partial y},$$

we can write the hazard function as

$$\begin{aligned} h_Y(y) &= \lim_{\partial y \to 0} \frac{F_Y(y + \partial y) - F_Y(y)}{\partial y \cdot S_Y(y)} \\ &= \lim_{\partial y \to 0} \frac{\mathrm{P}(y < Y \leq y + \partial y)}{\partial y \cdot S_Y(y)} \\ &= \lim_{\partial y \to 0} \frac{\mathrm{P}(y < Y \leq y + \partial y \mid Y > y)}{\partial y}. \end{aligned}$$

Using (1), we can determine the connection between the hazard and survivor functions. We have

$$\begin{aligned} h_Y(y) &= \frac{f_Y(y)}{S_Y(y)} \\ &= \frac{f_Y(y)}{1 - F_Y(y)} \\ &= -\frac{\partial}{\partial y}\log[1 - F_Y(y)] \\ &= -\frac{\partial}{\partial y}\log[S_Y(y)]. \end{aligned}$$

Therefore,

$$S_Y(y) = \exp[-H_Y(y)]$$

where

$$H_Y(y) = \int_0^y h_Y(t)dt.$$

The function $H_Y(y)$ is called the *cumulative hazard function* or the *integrated hazard function.* Like the hazard function, the cumulative hazard function is **not** a probability. However, it is also a measure of risk: the greater the value of $H_Y(y)$, the greater the risk of failure by time y.

Example: Exponential distribution

The exponential density with mean parameter θ is

$$f_Y(y) = \frac{1}{\theta}\exp\left(-\frac{y}{\theta}\right).$$

The survivor function is

$$S_Y(y) = 1 - F_Y(y) = \exp\left(-\frac{y}{\theta}\right).$$

The hazard function is then

$$\begin{aligned} h_Y(y) &= \frac{f_Y(y)}{S_Y(y)} \\ &= \frac{\frac{1}{\theta}\exp\left(-\frac{y}{\theta}\right)}{\exp\left(-\frac{y}{\theta}\right)} \\ &= \frac{1}{\theta}. \end{aligned}$$

In other words, the hazard function is constant when the survival time is exponentially distributed.

The cumulative hazard is then

$$H_Y(y) = \frac{y}{\theta}.$$

The constant hazard function is a consequence of the *memoryless property* of the exponential distribution: the distribution of the subject's remaining survival time *given* that s/he has survived till time t does not depend on t. In other words, the probability of death in a time interval $[t, t+y]$ does not depend on the starting point, t.

Example: Weibull distribution

The Weibull density with shape parameter λ and scale parameter θ is

$$f_Y(y) = \frac{\lambda y^{\lambda-1}}{\theta^\lambda} \exp\left[-\left(\frac{y}{\theta}\right)^\lambda\right].$$

The survivor function is

$$\begin{aligned} S_Y(y) &= \int_y^\infty \frac{\lambda t^{\lambda-1}}{\theta^\lambda} \exp\left[-\left(\frac{t}{\theta}\right)^\lambda\right] dt \\ &= \exp\left[-\left(\frac{y}{\theta}\right)^\lambda\right]. \end{aligned}$$

The hazard function is

$$\begin{aligned} h_Y(y) &= \frac{f_Y(y)}{S_Y(y)} \\ &= \frac{\frac{\lambda y^{\lambda-1}}{\theta^\lambda} \exp\left[-\left(\frac{y}{\theta}\right)^\lambda\right]}{\exp\left[-\left(\frac{y}{\theta}\right)^\lambda\right]} \\ &= \left(\frac{\lambda}{\theta^\lambda}\right) y^{\lambda-1}. \end{aligned}$$

The cumulative hazard is then

$$H_Y(y) = \left(\frac{1}{\theta^\lambda}\right) y^\lambda.$$

For the Weibull distribution, the hazard function depends on y. We can see that, depending on whether λ is greater than or less than 1, the hazard can increase or decrease with increasing y. This is often more realistic than the assumption of a constant hazard function (as in the exponential case). Since the exponential distribution is a special case of the Weibull with $\lambda = 1$, one way of analyzing the hazard rate is to fit the (more general) Weibull model and then test whether $\lambda = 1$.

Using the Weibull and Exponential Distributions to Model Survival Data

Typically survival times will depend on covariates. Usually (e.g. in S-PLUS), we incorporate these covariates in the following way. First, we assume that λ is constant across subjects.

(This is given in the exponential case, where $\lambda \equiv 1$ for all subjects.) We then allow θ to vary across subjects (i.e. we assume that subject i has scale parameter θ_i). For subject i with covariates $\mathbf{x}_i$, we assume that

$$\log \theta_i = \sum_{j=1}^{p} x_{ij}\beta_j \equiv \eta_i.$$

In this case, the hazard function for the Weibull distribution becomes

$$\begin{aligned} h_{Y_i}(y) &= \left(\frac{\lambda}{\theta_i^\lambda}\right) y^{\lambda-1} \\ &= \left(\lambda e^{-\lambda\eta_i}\right) y^{\lambda-1}. \end{aligned}$$

Say that $x_{i1} \equiv 1$ so that β_1 is the intercept. The hazard function when $x_{i2} = \cdots = x_{ip} = 0$ is called the *baseline hazard function*. We will denote the baseline hazard by h_0. We have that

$$h_0(y) = \left(\lambda e^{-\lambda\beta_1}\right) y^{\lambda-1}.$$

The *hazard ratio* is defined as

$$\begin{aligned} \frac{h_{Y_i}(y)}{h_0(y)} &= \frac{\left(\lambda e^{-\lambda\eta_i}\right) y^{\lambda-1}}{\left(\lambda e^{-\lambda\beta_1}\right) y^{\lambda-1}} \\ &= \frac{e^{-\lambda\eta_i}}{e^{-\lambda\beta_1}} \\ &= \exp\left(-\lambda \sum_{j=2}^{p} x_{ij}\beta_j\right). \end{aligned}$$

The hazard ratio does not depend on y in this case. This is an example of a more general class of models, *proportional hazards models*, where the hazard function can be written as

$$h_{Y_i}(y) = h_0(y)\; g(\mathbf{x}_i).$$

Here, g is a non-negative function that depends on the covariates ($\mathbf{x}_i$) but not on time (y).