Chapter 9

Parametric Estimation for Lifetime Models with Covariates

Parameter estimation for the accelerated life and proportional hazards models, which were introduced in Chapter 5, is considered in this chapter. Since there is now a vector of covariates in addition to a failure or censoring time for each item on test, special notation must be established to accommodate the covariates. The accelerated life and proportional hazards models are considered in separate sections because they require different approaches for parameter estimation. The proportional hazards model has the unique feature that the baseline distribution need not be defined in order to estimate the regression coefficients associated with the covariates.

9.1 Model Formulation

The purpose of a lifetime model that incorporates a vector of covariates $\boldsymbol{z} = (z_1, z_2, \ldots, z_q)'$ is to determine the impact of the covariates on survival. The reason for including this vector may be to determine which covariates significantly affect the survival of an item, to determine the probability distribution of the lifetime of an item for a particular setting of the covariates, or to fit a more complicated distribution from a small data set, as opposed to fitting separate distributions for each level of the covariates. As indicated in Section 5.3, one way to define the accelerated life model is through the survivor function

$$S(t, \boldsymbol{z}) = S_0\big(t\psi(\boldsymbol{z})\big),$$

for $t \geq 0$, where $S_0(\cdot)$ is a baseline survivor function and $\psi(\boldsymbol{z})$ is a link function satisfying $\psi(\boldsymbol{0}) = 1$ and $\psi(\boldsymbol{z}) > 0$ for all $\boldsymbol{z}$. The covariate vector $\boldsymbol{z}$ has been added as an argument to the survivor function because the probability of survival to time t is a function of both time and the covariate values. When $\psi(\boldsymbol{z}) > 1$, the covariates increase the rate at which the item moves through time. When $\psi(\boldsymbol{z}) < 1$, the covariates decrease the rate at which the item moves through time. For simplicity and mathematical tractability, the link function is assumed to have the log linear form $\psi(\boldsymbol{z}) = e^{\boldsymbol{\beta}'\boldsymbol{z}}$, throughout this chapter, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ is a vector of regression coefficients. This assumption is not necessary for some of the derivations, so many of the results apply to a wider range of link functions.

Recall that the proportional hazards model was defined in Section 5.4 by

$$h(t, \boldsymbol{z}) = \psi(\boldsymbol{z})h_0(t),$$

for $t \geq 0$, where $h_0(t)$ is a baseline hazard function. The covariates increase the hazard function when $\psi(\boldsymbol{z}) > 1$ or decrease the hazard function when $\psi(\boldsymbol{z}) < 1$. For both the accelerated life and proportional hazards models, the other lifetime distribution representations are given in Table 5.2. The purpose of this chapter is to estimate the $q \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ from a data set consisting of n items on test and r observed failure times.

The notation used to describe a data set in a lifetime model involving covariates will borrow some notation from the previous two chapters but also establish some new notation. The failure time of the ith item on test, t_i, is either observed or right censored at time c_i. As before, let $x_i = \min\{t_i, c_i\}$ and δ_i be a censoring indicator variable (1 for an observed failure and 0 for a right-censored value), for $i = 1, 2, \ldots, n$. In addition, a $q \times 1$ vector of covariates $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iq})'$ is collected for each item on test, for $i = 1, 2, \ldots, n$. Thus, z_{ij} is the value of covariate j for item i, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, q$. This formulation of the problem can be stated in matrix form as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1q} \\ z_{21} & z_{22} & \cdots & z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nq} \end{bmatrix}.$$

Each row in the $\boldsymbol{Z}$ matrix consists of the values of the q covariates collected on a particular item. The matrix approach is useful because complicated systems of equations can be expressed compactly and operations on data sets can be performed efficiently by a computer. For parameter estimation, the survivor, density, hazard, and cumulative hazard functions now have the extra arguments $\boldsymbol{z}$ and $\boldsymbol{\beta}$ associated with them:

$$S(t, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) \qquad f(t, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) \qquad h(t, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) \qquad H(t, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}),$$

for $t \geq 0$, where the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ consists of the p unknown parameters associated with the baseline distribution, which must be estimated along with the regression coefficients $\boldsymbol{\beta}$. In the case of random right censoring, the likelihood function can now be written in the usual form:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i \in U} f(x_i, \boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) \prod_{i \in C} S(x_i, \boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}),$$

where U is the set of indexes of uncensored observations and C is the set of indexes of right censored lifetimes. The log likelihood function is

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i \in U} \ln f(x_i, \boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) + \sum_{i \in C} \ln S(x_i, \boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}),$$

or, equivalently,

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i \in U} \ln h(x_i, \boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) - \sum_{i=1}^{n} H(x_i, \boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}).$$

Two observations with respect to this model formulation are important. First, the maximum likelihood estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ typically cannot be expressed in closed form (as was the case for the exponential distribution in Section 8.2), so numerical methods typically need to be used to find the values of the estimates. Second, the choice of whether to use a model that explicitly includes covariates or to examine each population separately is dependent on the number of unique covariate vectors $\boldsymbol{z}$ and the number of items on test, n. If n is large and there is only a single binary covariate (that is, only two unique covariate vectors, $z_1 = 0$ and $z_1 = 1$), for example, it is probably wiser to analyze each of the two populations separately by the techniques described in Chapter 8.