Chapter 1

Introduction

Bivariate failure time data arise when study subject units are paired. Examples of paired units include eyes or ears from the same person, twins, and father and son. Since there exist natural relationships between the two subjects in one pair, the two failure times within the same pair might be correlated. Moreover, in such studies, either or both failure times might not be observed because of censoring. A well-know study involving paired failure times is the Diabetic Retinoparthy Study, Diabetic Retinopathy Study Research Group (1981). The dependence, along with the censoring, greatly complicates the analysis of the data. This data set will be discussed in Chapter 4.

In univariate survival analysis, classical methods such as the Kaplan-Meier estimator and the Cox proportional regression model are based on the hazard function $h(t)$,

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

where S is the survivor function and f is the density function, see Section 2.1. If T is a discrete random variable and T can take on values $0 \leq t_1 < t_2 < \cdots$, then the hazard function is

$$h(t_j) = P(T = t_j | T \geq t_j) = \frac{P(T = t_j)}{S(t_j)} = 1 - \frac{S(t_{j+1})}{S(t_j)}, j = 1, 2, \cdots.$$

And thus the survival function can be written as

$$S(t) = \prod_{j: t_j \leq t} (1 - h(t_j)),$$

see Lawless (1982). An analogous formula relating S and h exists for T continuous, see Gill (1992). But in the bivariate case, as Gill discussed, we do not have a nice formula relating the hazard and survivor functions, since there is no canonical way to define past, present, and future at "time" t.

Traditionally, there are two main approaches to analyze censored paired data. One approach is to try non-parametric estimation of the bivariate survivor function. The typical works on this approach are Dabrowska (1988) and Prentice and Cai (1992). Dabrowska extended the univariate Kaplan-Meier approach by defining a bivariate hazard and using it to estimate the joint survival function. The marginals of Dabrowska's estimate are given by the univariate Kaplan-Meier estimates. She proved the consistency of the estimates but did not give estimates for the covariance of the estimates. Prentice and Cai (1992) did not define a joint hazard. Instead, they gave a representation of the bivariate survivor function in terms of the marginal survivor functions and the covariance

$$\mathrm{cov}\Big(\int_0^{T_1} h_1(t)dt, \int_0^{T_2} h_2(t)dt\Big),$$

where T_1 and T_2 are paired failure times and h_k is the hazard of T_k, $k = 1, 2$. They proposed an estimate of this bivariate survivor function and proved the consistency of their estimate but they did not give estimates of standard errors. Lin and Ying (1993) presented a simple estimator of the bivariate survival function and also estimated the covariance function of their estimates. They assumed univariate censoring, that is, that there is one censoring time and it affects both failure times.

The other approach to bivariate survival analysis is to extend the Cox proportional hazards model, Cox (1972), to paired data. Recall the Cox proportional model has hazard function

$$h(t|z) = h_0(t) \exp(\boldsymbol{\beta} z)$$

where h_0 is the baseline hazard, z is a covariate vector, and $\boldsymbol{\beta}$ is unknown. The Cox proportional model assumes that all observations are independent. Holt and Prentice (1974) assume that within a pair, hazards are proportional according to a covariate, but the baseline hazard can depend on the pair. The effect of the covariate, that is, the value of $\boldsymbol{\beta}$, is the same across pairs.

Clayton (1978) and Oakes (1982) presented a fully parametric model for paired survival data. This model assumes that each marginal follows the Cox proportional model with respect to some covariates. They used an additional parameter to describe the association within a pair. Huster, Brookmeyer, and Self (1989) extended the Clayton-Oakes model to allow censoring. They also discussed another approach for inducing the correlation within a pair. They obtained the parameter estimates from an independence working model, that is, univariate estimates are used in the model to get the parameter estimates, and then they estimated variance robustly.

Wei, Lin, and Weissfeld (1989) proposed a model for multivariate failure times. They assumed each marginal distribution of the failure times follows a Cox proportional hazards model with respect to some covariates. They showed that the resulting estimators of the parameters for covariate effects are asymptotically jointly normal and gave a consistent estimate of the covariance matrix.

Lee, Wei, and Amato (1991) used a different way to extend the Cox proportional model to paired data. For paired data (T_{1i}, T_{2i}), with treatment indicator z_i, they showed the consistency and asymptotic normality of the estimated regression coefficient in the Cox proportional model. In this case the z_i's might be dependent. But the corresponding variance-covariance estimate may no longer be valid due to the dependence between the members in the pairs. They proposed a correct variance-covariance estimate taking account of the correlation within a pair.

All of the proposed models do not estimate the baseline hazard functions, treating it as a nuisance parameter.

With the development of smoothing theory and methods, many people have used splines in survival analysis, that is, to approximate density functions, survival functions, hazard functions, or baseline hazard functions, in the presence of censored data. For example, Abrahamowicz, Ciampi, and Ramsay (1992) used B-splines to estimate the density and O'Sullivan (1988) used smoothing splines to estimate log hazard functions.

Kooperberg, Stone, and Troung (1995) used linear B-splines and their tensor products to estimate the conditional log-hazard function as a function of t and a covariate vector z. Their model contains the Cox proportional hazards model as a submodel. Since cubic splines can provide a better approximation than linear splines, Kooperberg et. al. also proposed a model in which the log hazard function is estimated with cubic B-splines. Unfortunately, because of the complication in estimation and model selection, they did not consider covariates in this model. Nor did they provide standard errors for the estimated log hazard function.

In this thesis, we use cubic B-splines to estimate log hazards in univariate survival data and log hazard ratios in bivariate data. We use Kooperberg's approach to get estimates, but we provide estimates of standard errors.

In Chapter 2, we define the log hazard regression model, the estimates of the log hazard functions and ratios, and prove the asymptotic properties of the estimates. In Chapter 3, we discuss the properties of cubic B-splines functions and how to choose knots which define these functions. In Chapter 4, we use our model to analyze the data set from the Diabetic Retinoparthy Study. We present our simulation study in Chapter 5.