5

The Cox proportional hazard regression model and advances

As indicated in Chapter 4, a common criticism of using a parametric regression model to analyze survival data is the difficulty in ascertaining the validity of an underlying parametric distribution in survival times. With the selection effect operating throughout the life course, a parametric distribution may not necessarily reflect the true individual trajectories of survival processes; therefore, parameter estimates, derived from a regression model associated with a misspecified distribution of lifetimes, can be biased. Additionally, many researchers are more interested in the effects of covariates on the risk of an event occurrence than in the shape of a specific failure time distribution. While the proportional hazard model is highly welcomed in many applied disciplines, the application of parametric regression models is sometimes inconvenient because only the exponential and the Weibull functions can be formulated as a proportional hazard model. Given these concerns, it is highly useful to create a regression model that yields valid estimates of covariate effects on the hazard function while avoiding the specification of an underlying distributional function. Technically, developing such a regression model is not unrealistic. If all individuals are subject to a common baseline distribution of survival and differences in the hazard rate can be primarily reflected through the multiplicative effect of covariates, the underlying distribution of survival times becomes a constant and thus can potentially cancel out from a specific likelihood function. In the year 1972, David Cox masterfully developed a proportional hazard model, which derives robust, consistent, and efficient estimates of covariate effects using the proportional hazards assumption while leaving the baseline hazard rate unspecified (Cox, 1972). Since then, the *Cox proportional hazard model*, often simply referred to as the *Cox model*, has become the most widely applied regression perspective in survival analysis. Technically, the Cox model uses the maximum likelihood algorithm for a partial likelihood function, with the estimating approach referred to as a *partial likelihood*.

In this chapter, I first describe basic specifications of the Cox model and the detailed procedures of a partial likelihood. Next, the statistical techniques handling tied observations

Survival Analysis: Models and Applications, First Edition. Xian Liu.

in estimating the Cox model are delineated, followed by a section portraying how to estimate a survival function from the hazard rate estimates of the Cox model without specifying an underlying hazard function. Some other advances of the proportional hazard model are also introduced in the chapter, such as the hazard model with time-dependent covariates, the stratified proportional hazard model, and the management of left truncated survival data. Additionally, in Section 5.7, I describe several coding schemes for specifying qualitative factors and the statistical inference of local tests in the Cox model. In each section except 5.1, I illustrate an empirical example with one or two SAS programs. Lastly, I summarize the chapter with comments on the Cox model and its refinements.

5.1 The Cox semi-parametric hazard model

In this section, I describe the basic specifications of the Cox proportional hazard model, the partial likelihood, and the estimating procedures in the presence of right censoring.

5.1.1 Basic specifications of the Cox proportional hazard model

I start the description of the Cox model with the general specification of the proportional hazard rate model given in Subsection 4.1.1. Each observation under investigation is assumed to be subject to an instantaneous hazard rate, $h(t)$, of experiencing a particular event, where $t = 0, 1, \ldots, \infty$. Analogous to the parametric proportional hazard model, the effect of covariates in the Cox model is specified by a multiplicative term $\exp(\boldsymbol{x}'\boldsymbol{\beta})$, given the nonnegative nature of the hazard function. The basic equation of the Cox model is exactly the same as Equation (4.1), given by

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}). \tag{5.1}$$

Though identical in form, the detailed specification of Equation (5.1) differs from Equation (4.1). In the Cox model, the term $h_0(t)$ represents an arbitrary and unspecified baseline hazard function for continuous time T, whereas in the parametric perspective it represents a specific distributional function. The coefficient vector $\boldsymbol{\beta}$ provides a set of covariate effects on the hazard rate, with the same length as $\boldsymbol{x}$. The effects of $\boldsymbol{x}$ on the hazard rate are assumed to be multiplicative, so that the predicted value of $h(t)$ given $\boldsymbol{x}$, denoted *by* $\hat{h}(t|\boldsymbol{x})$, varies in the range $(0, \infty)$. Exponentiation of a specific regression coefficient generates the hazard ratio (HR) of covariate x_m.

Suppose covariate x_m is a dichotomous variable with $x_{m1} = 1$ and $x_{m0} = 0$. Let other covariates all take the value 0; the hazard ratio of covariate x_m is

$$\begin{aligned} HR_m &= \frac{h_0(t)\exp\left(x_{m1}\hat{\beta}_m\right)}{h_0(t)\exp\left(x_{m0}\hat{\beta}_m\right)} \\ &= \exp\left[\left(x_{m1} - x_{m0}\right)\hat{\beta}_m\right] \\ &= \exp\left(\hat{\beta}_m\right), \end{aligned} \tag{5.2}$$

where $\hat{\beta}_m$ is the estimate of the regression coefficient for covariate x_m. This definition of relative risk, independent of $h_0(t)$, holds when other covariates are not zero because additional

terms appearing in both the numerator and the denominator would cancel out in Equation (5.2). This specification can be readily extended to the case of a continuous covariate.

Independent of the unspecified baseline hazard function, the hazard ratio provides an uncomplicated indicator to measure the effect of a given covariate on the hazard rate. If a specific covariate is a dichotomous variable, the hazard ratio displays an intuitive meaning of the relative risk, as also discussed in Chapter 4. For example, if the regression coefficient of cancer (1 = yes, 0 = no) on mortality is 1.60, the hazard ratio is exp(1.60) = 4.95, suggesting that those with cancer are about 5 times as likely to die as those without cancer, other covariates being equal. For a continuous covariate, the hazard ratio displays the extent to which the risk increases ($HR > 1$) or decreases ($HR < 1$) with a 1-unit increase in the value of that covariate.

Broadly, the hazard ratio can also be calculated to reflect the proportional change in the hazard rate with a $\dot{w}$-unit increase in x_m, given by

$$\begin{aligned} HR_{\dot{w}} &= \frac{\exp\left[(x_{m0}+\dot{w})\hat{\beta}_m\right]}{\exp\left(x_{m0}\hat{\beta}_m\right)} \\ &= \exp\left[(x_{m0}+\dot{w}-x_{m0})\hat{\beta}_m\right] \\ &= \exp\left(\dot{w}\hat{\beta}_m\right) \\ &= \exp\left(\hat{\beta}_m\right)^{\dot{w}}. \end{aligned} \tag{5.3}$$

In the above equation, the measure $\exp\left(\dot{w}\hat{\beta}_m\right)$ is referred to as the $\dot{w}$-unit hazard ratio, reflecting the multiplicative change in the hazard rate with a $\dot{w}$-unit increase in covariate x_m. This multiunit hazard ratio is useful for displaying the effect of a continuous or an interval covariate with a small metric unit. For example, a 1-unit hazard ratio of age may not be meaningful enough to reflect the true influence because age has a very small unit given a range of $(0, \omega)$. Under such circumstances, the measure $\exp\left(10\times\hat{\beta}_{\text{age}}\right)$ is a more appropriate choice for displaying the impact an individual's age has on the relative risk to the occurrence of a particular event.

In empirical studies, some researchers prefer to use the equation $\{100 \times [\exp(b_m) - 1]\}$, which gives rise to the percentage change in the hazard rate with a 1-unit change in covariate x_m. Consider the cancer example: if the hazard ratio is 5 for cancer patients, the value of $[100 \times (5 - 1)]$ is 400, in turn indicating a 400 % increase in the mortality rate for those diagnosed with cancer as compared to those not diagnosed.

In the Cox model, all individuals are assumed to follow an unspecified nuisance function $h_0(t)$, and therefore population heterogeneity is primarily reflected in proportional scale changes as a function of $\exp(\boldsymbol{x}'\boldsymbol{\beta})$. I want to emphasize again, however, that the concept of *proportionality* here addresses a statistical perception, and the specification of an interaction term or the addition of an arbitrarily assigned random effect can easily spoil the proportionality hypothesis.

Given the specification of the hazard function, the survival probability, given $\boldsymbol{x}$, can be developed in the same fashion as Equation (4.5):

$$S(t;\boldsymbol{x}) = [S_0(t)]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}, \tag{5.4}$$

where $S_0(t)$ is the baseline survival function, defined as

$$S_0(t) = \exp\left[-\int_0^t h_0(u)\,\mathrm{d}u\right] = \exp[-H_0(t)].$$

In the Cox model, the baseline survival function is unspecified because $h_0(t)$ is an unspecified nuisance function. Practically, the baseline survival function can be understood as the survival function when all covariates are scaled 0, as can be easily deduced from Equation (5.4). Therefore, the substantive meaning of $S_0(t)$ is related to how covariates are created. If covariates are rescaled to be centered at sample means, for example, the baseline survival function indicates a mean survival function for the population that an underlying random sample represents, or the survival probability for an 'average' individual in that population.

Equation (5.4) also shows that the survival function at t, given covariate vector $\boldsymbol{x}$, is simply the baseline survival function $S_0(t)$ raised to the power of the multiplicative effects of covariates on the hazard function. Therefore, with an estimate of the underlying survival function $S_0(t)$, the survival probability for each individual can be estimated given the values of covariates and the estimated regression coefficients $\hat{\boldsymbol{\beta}}$. While not explicitly specified in Equation (5.4), the $S_0(t)$ estimate can be obtained from some other lifetime indicators, thus generating estimates of the complete survival function. The implications of Equation (5.4) will be discussed further in Section 5.3.

Given Equations (5.1) and (5.4), the density function of T given $\boldsymbol{x}$ is

$$f(t;\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})\exp\left[-\exp(\boldsymbol{x}'\boldsymbol{\beta})\int_0^t h_0(u)\,\mathrm{d}u\right]. \tag{5.5}$$

Equation (5.5) indicates that, in the Cox model, the density function is not explicitly specified either, because $h_0(t)$ is unspecified.

The Cox model can be conveniently expressed in terms of a log-linear model, with covariates assumed to be linearly associated with the log $h(t)$ function. Taking log values on both sides of Equation (5.1) yields

$$\log[h(t|\boldsymbol{x})] = \log[h_0(t)] + \boldsymbol{x}'\boldsymbol{\beta} = i^* + \boldsymbol{x}'\boldsymbol{\beta}, \tag{5.6}$$

where i^* is the log of the baseline hazard function in the proportional hazard model. While explicitly specified as the intercept factor in the parametric hazard regression model, this log transformed baseline hazard function is unspecified in the Cox model because $h_0(t)$ is unspecified. Hence, the Cox model is a log-linear regression model without specifying an intercept. As the above specifications differ from both the nonparametric and the parametric perspectives, the Cox model is also called the *semi-parametric regression model*.

As summarized, the Cox model is a simplified perspective of parametric regression modeling described in Chapter 4. The development of the partial likelihood function, presented below, facilitates the validity and reliability of the simplification.

5.1.2 Partial likelihood

The estimation of the original Cox proportional hazard model starts with ordering a random sample of n individuals according to the rank of survival times, assuming no observation ties, given by

$$t_1 < t_2 < t_3 < t_4 \cdots < t_n.$$

Rather than specifying a full likelihood function for Equation (5.1), Cox (1972) considers the conditional probability that an individual experiences a particular event at time t_i ($i = 1, 2, \ldots, n$) given that he or she is one of r individuals at risk to the event at t_i (those with $T \geq t_i$). Let $\mathcal{R}(t_i)$ be the risk set, consisting of all individuals not experiencing the event and uncensored just prior to t_i. In a discrete interval $(t_i, t_i + \Delta)$, the conditional probability that individual i experiences the event given the risk set $\mathcal{R}(t_i)$ can be written by

$$\frac{\mathrm{P}[\text{individual } i \text{ with covariates } \boldsymbol{x}_i \text{ experiences event in time interval} (t_i, t_i + \Delta)]}{\sum_{l \in \mathcal{R}(t_i)} \mathrm{P}[\text{invidual } l \text{ experiences event in } (t_i, t_i + \Delta)]}. \tag{5.7}$$

Because the baseline hazard has an arbitrary form in the Cox model, intervals between successive survival times do not provide information on the multiplicative effects of covariates on the hazard function. Let t be a continuous function. Then the conditional probability transforms to a continuous hazard function, as specified in Chapter 1. Hence, given $\Delta \to 0$, Equation (5.7) becomes

$$\frac{\text{hazard rate at } t_i \text{ for individual } i \text{ with covariates } \boldsymbol{x}_i}{\sum_{l \in \mathcal{R}(t_i)} \text{hazard rate at } t_i \text{ for individual } l}. \tag{5.8}$$

As both the numerator and the denominator are represented by the hazard function given $\Delta \to 0$, Equation (5.8) maintains the characterization of the conditional probability that an individual experiences an event at time t_i given r individuals who are exposed to the risk at t_i. This specification agrees with the Nelson–Aalen estimator described in Chapter 2.

Equation (5.8) can be written in the form of a regression model with a parameter vector $\boldsymbol{\beta}$, written by

$$\begin{aligned} \frac{h(t_i; \boldsymbol{x}_i)}{\sum_{l \in \mathcal{R}(t_i)} h(t_i; \boldsymbol{x}_l)} &= \frac{h_0(t_i) \exp(\boldsymbol{x}_i' \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} h_0(t_i) \exp(\boldsymbol{x}_l' \boldsymbol{\beta})} \\ &= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})}. \end{aligned} \tag{5.9}$$

In the above formulation, the baseline hazard function appears in both the numerator and the denominator of the equation, so that it can cancel out in the final equality. Essentially, the second equation in (5.9) still denotes the conditional probability that an individual experiences a particular event at time t_i given all the individuals exposed to that risk at t_i. With the common terms eliminated, it provides an incomplete hazard function without specifying the distribution of h_0.

The joint likelihood for $\boldsymbol{\beta}$, the only parameters to be estimated in the Cox model, is simply specified as the product of Equation (5.9) over all t_i values, given by

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})} \right]^{\delta_i}, \tag{5.10}$$

where $L_p(\cdot)$ represents an incomplete likelihood function and δ_i is the censoring indicator such that $\delta_i = 1$ if t_i is an event time and $\delta_i = 0$ if t_i is a censored time. As defined, the partial likelihood function $L_p(\boldsymbol{\beta})$ is the probability of a set of regression coefficient values given n observed lifetime outcomes, actual or censored survival times. When $\delta_i = 1$, $L_{pi}(\boldsymbol{\beta})$ is the conditional probability of an event occurrence given the risk set $\mathcal{R}(t_i)$. As no information about h_0 contributes to it, the above incomplete likelihood function is generally referred to as the *partial likelihood function* in survival analysis.

As $L_{pi}(\boldsymbol{\beta})$ is simply 1 when $\delta_i = 0$, the product of the conditional probabilities for all right-censored observations are also 1, thus suggesting that the conditional probabilities over all right-censored cases do not make any contribution to the partial likelihood. Consequently, these censored cases can be unaccounted for in the partial likelihood equation without influencing the total partial likelihood. Therefore, it is statistically tenable to simplify Equation (5.10) by only multiplying the conditional probabilities over all events:

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{d} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})}, \tag{5.11}$$

where d is the total number of events, ordered by rank. Equation (5.11) reflects the fact that each actual event, or failure, contributes a factor to the above simplified likelihood function while the conditional probabilities of censored observations do not play a role in partial likelihood. This equation is the standard partial likelihood function originally created by Cox. Notice that this simplification does not mean that all censored observations are excluded because, at each t_i, the risk set $\mathcal{R}(t_i)$ includes all observations censored at times later than t_i.

Taking log values on both sides of Equation (5.11), a log partial likelihood function is

$$\log L_p(\boldsymbol{\beta}) = \sum_{i=1}^{d} \left\{ \boldsymbol{x}_i'\boldsymbol{\beta} - \sum_{i=1}^{d} \log \left[\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta}) \right] \right\}. \tag{5.12}$$

Like maximization of a complete likelihood function in parametric regression modeling, the above log partial likelihood function can be mathematically maximized with respect to the unknown parameter vector $\boldsymbol{\beta}$ that contains the regression coefficients of covariates without an intercept factor. As the partial likelihood depends on the ordering of actual survival times, hazard rates are unrelated to survival times and intervals between events. This characteristic indicates that the baseline hazard distribution is inherent in the ordering of survival times without a known shape. As the partial likelihoods are based on the number

of events, rather than on the total number of individuals in a random sample, the application of the Cox model requires a considerably larger sample for estimating model parameters than does parametric regression modeling, especially in analyzing events that do not frequently occur.

5.1.3 Procedures of maximization and hypothesis testing on partial likelihood

Maximization of the partial likelihood function is performed with respect to the parameter vector $\boldsymbol{\beta}$ by using the standard procedures described in Chapter 4. As generally applied, the process starts with the construction of a score statistic vector, denoted by $\tilde{\boldsymbol{U}}_i(\boldsymbol{\beta})$ and mathematically defined as the first partial derivatives of the log partial likelihood function with respect to $\boldsymbol{\beta}$. The total score statistic for the mth covariate ($m = 1, \ldots, M$) is

$$\tilde{U}(\beta_m) = \frac{\partial}{\partial \beta_m} \log L_p(\boldsymbol{\beta}) = \sum_{i=1}^{d} \left[x_{im} - \frac{\sum_{l \in \mathcal{R}(t_i)} x_{lm} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})} \right]$$

$$= \sum_{i=1}^{d} \{ x_{im} - \mathrm{E}[x_m | \mathcal{R}(t_i)] \}, \tag{5.13}$$

where

$$\mathrm{E}[x_m | \mathcal{R}(t_i)] = \frac{\sum_{l \in \mathcal{R}(t_i)} x_{lm} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})}.$$

It is worth noting that the score statistic can be conveniently simplified as a permutation test based on residuals computed for the regression on covariates, rather than on the regression coefficients themselves, as also indicated in Chapter 4. In modern survival analysis, the score statistic plays an extremely important role in developing advanced techniques on the hazard model residuals and the asymptotic variance estimators for analyzing clustered survival data, as will be described extensively in some of the later chapters.

Following the standard maximum likelihood estimating procedures, the coefficient vector $\boldsymbol{\beta}$ in the partial likelihood function can be estimated by solving the equation

$$\tilde{\boldsymbol{U}}(\boldsymbol{\beta}) = \frac{\partial}{\partial} \log L_p(\boldsymbol{\beta}) = \mathbf{0}, \tag{5.14}$$

where

$$\tilde{\boldsymbol{U}}(\boldsymbol{\beta}) = \left(\partial \log L_p / \partial \beta_1, \ldots, \partial \log L_p / \partial \beta_M \right)'.$$

The above generalization is a typical maximum likelihood estimator (MLE). Hence, statistically the partial likelihood estimator is essentially no different from the standard maximum likelihood perspective, the only difference being that maximization is performed on a partial rather than a complete function.

As indicated in Chapter 4, for a large sample, $\hat{\boldsymbol{\beta}}$ is the unique solution of $\tilde{\boldsymbol{U}}(\boldsymbol{\beta}) = \mathbf{0}$, so that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and distributed as multivariate normal, which, conventionally, can be expressed as

$$\hat{\boldsymbol{\beta}} \sim N\left[\boldsymbol{\beta}, \tilde{\boldsymbol{V}}(\boldsymbol{\beta})^{-1}\right]. \tag{5.15}$$

Operationally, the estimation of $\boldsymbol{\beta}$ is generally performed through an iterative scheme, with $\hat{\boldsymbol{\beta}}^0 = \mathbf{0}$ as the initial step, using the Newton–Raphson method. The series of $\hat{\boldsymbol{\beta}}$ in the iterative scheme is generally denoted by $\hat{\boldsymbol{\beta}}^{\tilde{j}}$ ($^{\tilde{j}}$= 1, 2, . . .). The iterative scheme terminates when $\hat{\boldsymbol{\beta}}^{\tilde{j}+1}$ is sufficiently close to $\hat{\boldsymbol{\beta}}^{\tilde{j}}$. Consequently, the maximum likelihood estimate of $\boldsymbol{\beta}$ can be operationally defined as $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\tilde{j}+1}$.

Like estimating a parametric regression model, the asymptotic normal distribution of $\hat{\boldsymbol{\beta}}$ facilitates hypothesis testing on $\boldsymbol{\beta}$. In particular, the variance estimator of $\boldsymbol{\beta}$ is based on an $M \times M$ observed information matrix (M is the number of covariates), denoted by $\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)$ and mathematically defined as

$$\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right) = -\left(\frac{\partial^2 \log L\left(\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}^2}\right)_{M\times M}, \tag{5.16}$$

with the (m, m')th element specified as

$$\begin{aligned} I_{mm'}(\boldsymbol{\beta}) = &\sum_{i=1}^{d} \{\mathrm{E}[x_{im} x_{im'} | \mathcal{R}(t_i)]\} \\ &- \sum_{i=1}^{d} \{\mathrm{E}[x_{im} | \mathcal{R}(t_i)] E[x_{im'} | \mathcal{R}(t_i)]\}, \end{aligned} \tag{5.17}$$

where

$$\mathrm{E}[x_{im} x_{im'} | \mathcal{R}(t_i)] = \frac{\sum_{l \in \mathcal{R}(t_i)} x_{lm} x_{lm'} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})}.$$

If Equation (5.16) is correctly assumed, then $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ converges to an M-dimensional normal vector with mean 0 and a covariance matrix $\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}\right)$ that can be consistently estimated by

$$\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)^{-1}. \tag{5.18}$$

Therefore, the variance estimator of $\boldsymbol{\beta}$ for the Cox model is simply the inverse of the Fisher observed information matrix.

In form, Equation (5.18) is exactly the same as Equation (4.33), with both estimators following the standard procedures of maximization. The standard errors of $\boldsymbol{\beta}$ can be obtained by taking the square roots of the diagonal elements in $\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}\right)$. Similarly, the squared z-score, the ratio of each component in $\hat{\boldsymbol{\beta}}$ over its standard error, follows a chi-square distribution with one degree of freedom under the null hypothesis that $\hat{\boldsymbol{\beta}} = \mathbf{0}$. The information matrix evaluated at this null hypothesis is defined as $\boldsymbol{I}(\mathbf{0})$. The corresponding vector of scores, denoted by $\tilde{\boldsymbol{U}}(\mathbf{0})$, tends to be distributed as multivariate normal with mean 0 and a covariance matrix $\boldsymbol{I}(\mathbf{0})^{-1}$ for large samples. As a result, the score test statistic on $\hat{\boldsymbol{\beta}} = \mathbf{0}$ is

$$\boldsymbol{U}(\mathbf{0})' \boldsymbol{I}(\mathbf{0})^{-1} \boldsymbol{U}(\mathbf{0}), \tag{5.19}$$

which is distributed asymptotically as chi-square with M degrees of freedom under the null hypothesis. If there are no tied survival times, this score test in the Cox model is identical to the logrank test (Klein and Moeschberger, 2003; Lawless, 2003).

Given the same null hypothesis, $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean 0 and the covariance matrix $\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)^{-1}$ given the large sample approximation theory and the central-limit theorem. Accordingly, the Wald test statistic is defined by

$$\hat{\boldsymbol{\beta}}' \boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)^{-1} \hat{\boldsymbol{\beta}}, \tag{5.20}$$

which is also distributed asymptotically as chi-square with M degrees of freedom under the null hypothesis. Based on the maximum likelihood estimation procedures, specifications of both the score and the Wald test statistics have a tremendous resemblance to those for parametric hazard models. The Wald test statistic can be used to test a subtest of the estimated regression coefficients in $\hat{\boldsymbol{\beta}}$, which will be described in Section 5.7.

Also analogous to parametric regression modeling, hypothesis testing of $\hat{\boldsymbol{\beta}}$ in the Cox model can be performed by using the partial likelihood ratio test. For example, the partial likelihood ratio with respect to all elements in $\hat{\boldsymbol{\beta}}$ is defined as

$$LR(\boldsymbol{\beta}) = \frac{L_p(\mathbf{0})}{L_p\left(\hat{\boldsymbol{\beta}}\right)}, \tag{5.21}$$

where $L_p(\mathbf{0})$ is the partial likelihood function for the model without covariates (termed the empty model) and $L_p\left(\hat{\boldsymbol{\beta}}\right)$ is the partial likelihood function containing all covariates.

The null hypothesis about $\boldsymbol{\beta}$ can be written as H_0: $\hat{\boldsymbol{\beta}} = \mathbf{0}$ for the whole set of coefficients or H_0: $\hat{\beta}_m = 0$ for an individual component in $\hat{\boldsymbol{\beta}}$. To test the null hypothesis for all elements in $\hat{\boldsymbol{\beta}}$, the partial likelihood ratio statistic is

$$\begin{aligned} \Lambda_p &= 2 \log L_p\left(\boldsymbol{\beta}\right) - 2 \log L_p(\mathbf{0}) \\ &= -2\left[\log L_p(\mathbf{0}) - \log L_p\left(\boldsymbol{\beta}\right)\right], \end{aligned} \tag{5.22}$$

where

$$\log L_p(\mathbf{0}) = -\sum_{i=1}^{d} \log(n_i).$$

In survival analysis, Equation (5.22) is generally referred to as the *minus two log partial likelihood ratio statistic*, often used for testing model fitness of the Cox model.

The statistic Λ_p is asymptotically distributed as $\chi^2_{(M)}$ as well. Consequently, the corresponding p-value can be obtained from the distribution which, in turn, provides information for testing the H_0 hypothesis. Specifically, if Λ_p is associated with a p-value smaller than α, the null hypothesis about $\hat{\boldsymbol{\beta}}$ should be rejected; otherwise, the null hypothesis is to be accepted.

The $100(1 - \alpha)$ % partial likelihood based confidence interval for $\hat{\boldsymbol{\beta}}$ can be approximated by applying Equation (5.22), defined as

$$\Lambda_p = \left[2 \log L_p\left(\boldsymbol{\beta}\right) - 2 \log L_p(0)\right] \leq \chi^2_{(1-\alpha\alpha)} \tag{5.23}$$

The above three main tests on the null hypothesis of the estimated regression coefficients generally yield fairly close statistics and therefore usually lead to the same testing conclusion. In most cases, the Wald and the likelihood ratio tests are more conservative than the score test. The Wald test is frequently used to test the statistical significance of an individual or a subset of coefficients, whereas the log partial likelihood ratio test provides a highly efficient statistic for the model fit of the Cox model. There are some other model fit statistics, such as Akaike's information criterion (AIC) and Schwartz's Bayesian criterion (SBC), used in the Cox model. These modified statistics generally yield fairly close scores, as does the partial likelihood ratio test, so are not particularly described in this text.

The variance of the hazard ratio can be estimated using the delta method. If covariances in $\hat{\boldsymbol{\beta}}$ are weak, the univariate approximation $V[g(X)] \approx [g'(\mu)]^2\sigma^2$ can be applied to transform the variance of the regression coefficient for a given covariate to the variance of its hazard ratio. As the derivative of $\exp(X)$ is still $\exp(X)$, the variance of the hazard ratio for covariate x_m can be approximated by

$$V(HR_m) \approx \left[\operatorname{dexp}\left(\hat{\beta}_m\right)\right]^2 V\left(\hat{\beta}_m\right) = \left[\exp\left(\hat{\beta}_m\right)\right]^2 V\left(\hat{\beta}_m\right). \tag{5.24}$$

Therefore, the variance of the hazard ratio is simply the square of the hazard ratio multiplied by the variance of $\hat{\beta}_m$. The standard error of the hazard ratio is approximated by the square root of $V(HR_m)$, based on which the confidence interval of the hazard ratio can be easily calculated. Given the nonlinearity after transformation, however, this confidence interval is not symmetric. Accordingly, some statisticians suggest that the confidence interval of the hazard ratio should be computed by exponentiation of the two ending points of the confidence interval for β_m (Klein and Moeschberger, 2003).

As the reader might be aware, the inference and the estimation of the partial likelihood parameters are much like the general maximum likelihood procedure described in Chapter 4. Here, Cox develops an innovative likelihood function that eliminates the baseline functional parameters and derives the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ that are completely analogous to analytic results from the complete likelihood function (Kalbfleisch and Prentice, 2002). With the baseline hazard function canceling out of the likelihood function, the partial likelihood perspective provides substantial convenience and flexibility to

the researcher who is doubtful of the validity of an observed parametric distribution in survival times.

5.2 Estimation of the Cox hazard model with tied survival times

The original Cox model is based on the assumption that there are no tied observations in survival data. This assumption, however, is not realistic in most survival datasets because ties in failure times frequently occur, particularly when a discrete time scale is used and the size of a random sample is large. In many observational surveys, events are reported during intervals (such as weeks, months, or years), making it extremely demanding to obtain data with exact event times. When two or more individuals experience a particular event at the same time, ordering their shared survival time is difficult. As a result, the application of the original Cox model is restricted by the difficulty in finding survival data without tied observations. Some statisticians, including Cox himself, have developed a number of approaches to handle ties in estimating the Cox model (Breslow, 1974; Cox, 1972; Efron, 1977; Kalbfleisch and Prentice, 1973).

This section describes four approximation methods for handling ties, in the order of their respective publication years. Then I compare advantages and disadvantages of these approximation approaches. Lastly, I provide an empirical illustration to demonstrate how to apply these methods when using the Cox model.

5.2.1 The discrete-time logistic regression model

In his original article on the proportional hazard model, Cox (1972) proposes to handle tied survival times by adapting the partial likelihood function to a discrete-time regression model, given by

$$\frac{h(t;\boldsymbol{x})\mathrm{d}t}{1-h(t;\boldsymbol{x})\mathrm{d}t}=\exp(\boldsymbol{x}'\boldsymbol{\beta})\frac{h_0(t)\mathrm{d}t}{1-h_0(t)\mathrm{d}t}. \tag{5.25}$$

If T is a continuous distributional function, Equation (5.25) reduces to Equation (5.1). If event times are considered a step function, however, the above equation becomes a logistic regression function, as can be identified from its formation.

Let $t_1 < t_2 < \cdots < t_{n'}$ be the distinct failure times, d_i be the number of events at t_i, and $\mathcal{D}(t_i)$ be the set of individuals sharing tied event time t_i. Given discrete event times, the contribution to the partial likelihood function becomes

$$\frac{\exp(\text{`}\boldsymbol{s}_i'\boldsymbol{\beta})}{\sum_{l\in\mathcal{R}_{d_l}(t_l)}\exp[\text{`}\boldsymbol{s}_l'\boldsymbol{\beta}]}, \tag{5.26}$$

where

$$\text{`}\boldsymbol{s}_i=\sum_{\mathrm{j}\in\mathcal{R}_i}\boldsymbol{x}_{\mathrm{j}},$$

and the nonitalicized j represents jth person in $\mathcal{D}(t_i)$. Note that when there are tied observations, the notation i no longer can be used to indicate an individual and an event time simultaneously; therefore, in the following text the nonitalic letter j is used to indicate an individual at a tied event time.

The vector $`\boldsymbol{s}_i$ is the sum of the covariate values for individuals who fail at the discrete time t_i, represented by $\mathcal{D}(t_i)$. As the reader familiar with generalized linear regression might recognize, Equation (5.26) displays a typical logistic regression model, somewhat inconsistent with the proportional hazard function proposed by Cox himself. Additionally, in this logistic regression model, $\boldsymbol{\beta}$ includes an intercept factor, whereas the intercept cancels out in the partial likelihood function. Hence, this discrete-time function essentially represents a replacement of the Cox model (Kalbfleisch and Prentice, 1973, 2002).

The likelihood function for the above logistic model is given by

$$L_p^{\text{discrete}}(\boldsymbol{\beta}) = \prod_{i=1}^{n'} \frac{\exp(`\boldsymbol{s}_i'\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_{d_i}(t_i)} \exp(`\boldsymbol{s}_l'\boldsymbol{\beta})}. \tag{5.27}$$

Accordingly, the partial log-likelihood function is

$$\log L_p^{\text{discrete}}(\boldsymbol{\beta}) = -\sum_{i=1}^{n'} `\boldsymbol{s}_i'\boldsymbol{\beta} - \sum_{i=1}^{n'} \log\left[\sum_{l \in \mathcal{R}_{d_i}(t_i)} \exp(`\boldsymbol{s}_l'\boldsymbol{\beta})\right]. \tag{5.28}$$

Maximization procedures for the above log-likelihood function are the same as the maximum partial likelihood estimator described in Subsection 5.1.3. As commented by Kalbfleisch and Prentice (2002), the discrete-time logistic partial likelihood is difficult computationally if the number of ties is large. Additionally, the parameters specified in this logistic model do not have exactly the same interpretation as those in the Cox model.

5.2.2 Approximate methods handling ties in the proportional hazard model

The complete solution of handling tied survival times is to consider all possible orders of tied observations in likelihoods, as proposed by Kalbfleisch and Prentice (1973, 2002). As it accounts for all ordering arrangements and then takes the average contribution, this approximation method is generally referred to as the *exact partial likelihood* or the *average partial likelihood method*.

Mathematically, the full set of all possible orders for d_i is simply a matter of $[d_i \times (d_i - 1) \times (d_i - 2) \times \ldots]$ commutations, mathematically denoted by $d_i!$. For example, if there are three individuals with tied survival time at t_i, we have $3! = 6$ possible orders; if there are five tied observations, there are $5! = 120$ commutations; and so forth. Here, I write the set of $d_i!$ as $`\mathcal{Q}_i$ with elements $(`p_1, `p_2, \ldots, `p_{d_i})$, following the notation of Kalbfleisch and Prentice (1973). The average partial likelihood contribution at t_i is then defined as

$$\frac{1}{d_i!}\exp\left[\left(\sum_{l \in \mathcal{D}_i} \boldsymbol{x}_l\right)'\boldsymbol{\beta}\right]\left\{\sum_{`p \in `\mathcal{Q}_{di}} \prod_{r=1}^{d_i}\left[\sum_{l \in \mathcal{R}(t_i, `p, r)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})\right]\right\}^{-1}. \tag{5.29}$$

The average partial likelihood function for all survival times is proportional to the product of likelihoods across all t_i values ($i = 1, 2, \ldots, n'$), given by

$$L_p^{\text{exact}}(\boldsymbol{\beta}) \propto \prod_{i=1}^{n'} \left(\exp\left[\left(\sum_{l \in \mathcal{D}_i} \boldsymbol{x}_l \right)' \boldsymbol{\beta} \right] \left\{ \sum_{p \in Q_{ii}} \prod_{r=1}^{d_i} \left[\sum_{l \in \mathcal{R}(t_i, p, r)} \exp(\boldsymbol{x}_l' \boldsymbol{\beta}) \right] \right\}^{-1} \right). \tag{5.30}$$

Clearly, in the above exact partial likelihood function, the denominator involves all ordering arrangements at each risk set, whereas the numerator adds up the covariates of all individuals who experience the event at t_i.

Essentially, Equation (5.30) takes tremendous similarities with the discrete-time logistic regression method because both approaches use all ties without explicit ordering. The exact method, however, may slightly overestimate $\boldsymbol{\beta}$ because it assumes ties without genuine ordering in tied survival times. If a superficial tie covers up some unobservable ordering, the value of the denominator in Equation (5.30) may be underestimated, thus leading to overestimation of the overall partial likelihood. Additionally, this partial likelihood method is computationally cumbersome and tedious if there are a substantial number of ties at given observed survival times.

Breslow (1974) provides a much simpler method to handle ties. Specifically, this method is a tremendous simplification of Equation (5.26), given by

$$L_p^{\text{Breslow}}(\boldsymbol{\beta}) = \prod_{i=1}^{n'} \left[\frac{\exp\left[\left(\sum_{l \in \mathcal{D}_i} \boldsymbol{x}_l \right)' \boldsymbol{\beta} \right]}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})^{d_i}} \right]. \tag{5.31}$$

In Equation (5.31), the numerator considers the sum of covariates over the d_i individuals who fail at t_i, whereas the denominator is simplified as, compared to the exact approximation, the sum of the partial hazard function in the risk set $\mathcal{R}(t_i)$ raised to the power d_i. Clearly, the Breslow approximation method accounts for the contributions of d_i events simply by multiplying the conditional probabilities over all events at t_i.

The Efron (1977) approximation method advances the Breslow method by proposing the following approximation:

$$L_p^{\text{Efron}}(\boldsymbol{\beta}) = \prod_{i=1}^{n'} \frac{\exp\left[\left(\sum_{l \in \mathcal{D}_i} \boldsymbol{x}_l \right)' \boldsymbol{\beta} \right]}{\prod_{r=0}^{d_i - 1} \left[\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l' \boldsymbol{\beta}) - \frac{r-1}{d_i} \sum_{l \in \mathcal{D}_i} \exp(\boldsymbol{x}_l' \boldsymbol{\beta}) \right]}. \tag{5.32}$$

In Equation (5.32), Efron reduces weight of the denominator by introducing ordering into the partial likelihood function. Some statisticians consider the Efron method to be a more

accurate approximation than both the exact and the Breslow methods after conducting some simulations (Hertz-Picciotto and Rockhill, 1997). This method is considered to be particularly preferable when sample size is small either from the outset or due to heavy censoring.

Notice that the above three partial likelihood functions have the same numerator, so that their differences reside completely in the way of specifying the denominator of the partial likelihood. In each of those estimators, $\boldsymbol{\beta}$ and the variances can be efficiently and consistently estimated by expanding the maximum partial likelihood estimator. When there are no tied survival times, the three partial likelihood functions converge to Equation (5.11).

Of the above three approximations, the exact or average likelihood method provides a complete partial likelihood function, arguably yielding the least amount of bias in $\hat{\boldsymbol{\beta}}$. Nevertheless, this method is computationally difficult when there are heavy ties in survival data. In the dataset of older Americans used for illustrative purposes in this book, the number of months since the baseline survey is used as the time scale, so that there are very heavy ties at many observed event times. In such circumstances, the application of the exact method is cumbersome and unnecessary. Consider 20 tied observations at a particular month: you need 20! steps to compute the likelihood for just one event time!

By contrast, the Breslow (1974) approximation method provides a much simpler approach and therefore becomes a popular first choice in analyzing tied survival data. Specifically, the Breslow approximation is used as default in many statistical software packages including SAS. This method, however, is considered to have some distinct limitations and weaknesses. Compared to the exact and the Efron approximation methods, the denominator in the Breslow likelihood function is slightly overweighed by failed individuals, so that $\hat{\boldsymbol{\beta}}$ may be biased toward zero. Furthermore, when $\boldsymbol{\beta}$ deviates considerably from zero, the Breslow method is thought to be a poor approximation. While some statisticians consider the Efron method to perform better than the other two approximations (Hertz-Picciotto and Rockhill, 1997), Kalbfleisch and Prentice (2002) contend that both the Breslow and the Efron estimators are subject to asymptotic bias under a grouped continuous Cox model. So far, there is no consensus about which method is the most appropriate approach in analyzing tied survival data.

Below, through an empirical example, I further discuss the similarities and differences in the results generated from these three approximation methods.

5.2.3 Illustration on tied survival data: Smoking cigarettes and the mortality of older Americans

In the present illustration, I conduct a study on the association between smoking cigarettes and the mortality of older Americans, using survival data of older Americans from the baseline survey to the end of 2004. The variable 'smoking cigarettes' is defined as a dichotomous variable with 1 = current or past smoker and 0 = else, given the variable name 'Smoking.' In addition to age and educational attainment, I add gender as a control variable. Because older men and women usually have different mortality rates and smoking behaviors, an individual's gender may confound the association between smoking cigarettes and survival. As conventionally applied, gender is measured dichotomously with 1 = female and 0 = male. The three control variables are rescaled to be centered at sample means, named, respectively, 'Age_mean,' 'Female_mean,' and 'Educ_mean.' The centered variable of gender represents

the propensity score of being a female or the proportion of women for a population group. My working null hypothesis is that smokers have the same mortality rate as do nonsmokers after adjusting for these potential confounders. As planned, I want to estimate and compare three sets of regression coefficients of smoking cigarettes on the hazard function, using the exact, the Breslow, and the Efron approximation methods, respectively.

The PROC PHREG procedure is applied to fit the three Cox models on the survival data of older Americans. As mentioned in Section 5.1, the Cox model does not specify an intercept factor, so only the estimates of regression coefficients are obtained. As the Weibull hazard rate model is a proportional hazard function, I also apply the PROC LIFEREG procedure to derive the corresponding Weibull estimates for comparative purposes. Below is the SAS program for estimating the regression coefficients of three Cox models in an 11–12 year period (the code for the Weibull model is not included given a detailed description in Chapter 4).

SAS Program 5.1:

```
……

data new;
  set pclib.chapter5_data;

Status = 0;
if death=1 then Status = 1;

proc SQL;
  create table new as
  select *, age-mean(age) as age_mean,
    female-mean(female) as female_mean,
    educ-mean(educ) as educ_mean
  from new;
quit;

proc phreg data = new;
  model duration*Status(0) = smoking age_mean female_mean educ_mean /  ties = EXACT ;
run;

proc phreg data = new;
  model duration*Status(0) = smoking age_mean female_mean educ_mean /  ties = EFRON ;
run;

proc phreg data = new;
  model duration*Status(0) = smoking age_mean female_mean educ_mean /  ties = BRESLOW ;
run;
```

SAS Program 5.1 specifies three Cox models with each using a specific approximation method to handle survival data with heavy ties. The SAS PROC SQL procedure creates three means-centered covariates as control variables, and then they are saved into the temporary SAS data file for further analysis. The TIES = <method> option specifies which the approximate likelihood function is used to handle tied survival times.

The resulting output file is fairly sizable because the information of all three separate Cox models is reported. Therefore, I select only to display results of the first model, using the exact method, as follows.

SAS Program Output 5.1:

The PHREG Procedure

Model Information

Data Set	WORK.NEW
Dependent Variable	duration
Censoring Variable	status
Censoring Value(s)	0
Ties Handling	EXACT

Number of Observations Read	2000
Number of Observations Used	2000

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
2000	950	1050	52.50

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	11348.893	11015.399
AIC	11348.893	11023.399
SBC	11348.893	11042.825

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	333.4943	4	<.0001
Score	328.0644	4	<.0001
Wald	334.5104	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
smoking	1	0.42041	0.10136	17.2023	<.0001	1.523
age_mean	1	0.08483	0.00519	266.9936	<.0001	1.089
female_mean	1	-0.42618	0.06670	40.8300	<.0001	0.653
educ_mean	1	-0.02238	0.00875	6.5512	0.0105	0.978

In SAS Program Output 5.1, the information on the dataset, the dependent variable, the censoring status variable, and the censoring value, is reported first. The 'Tied Handling: EXACT' statement indicates the use of the exact method for estimating regression coefficients of the four covariates. Given a long period of observation (11–12 years), 950 out of 2000 individuals are deceased (actual events) and the remaining 1050 older persons are right censored. The Model Fit Statistics section displays three indicators of model fitness, with

their values being very close, obviously generating the same conclusion about model fitting. All three tests in the 'Testing Global Null Hypothesis: BETA = 0' section demonstrate that the null hypothesis $\hat{\boldsymbol{\beta}} = \mathbf{0}$ should be rejected. For example, the chi-square of the likelihood ratio test is 333.49 with 4 degrees of freedom, which is very strongly statistically significant ($p < 0.0001$).

In the table of 'Analysis of Maximum Likelihood Estimates,' the regression coefficient of 'Smoking' is 0.4204 ($SE = 0.1014$), statistically significant at $\alpha = 0.05$ ($\chi^2 = 17.2023$; $p < 0.0001$). Exponentiation of this regression coefficient generates the hazard ratio between smokers and nonsmokers, as presented in the last column. The estimate of this hazard ratio is 1.523, suggesting that the mortality rate for current or past smokers is about 52 % higher than among those who never smoke, other covariates being equal. The regression coefficients of the three control variables are all statistically significant. Age is positively associated with the hazard rate, with a hazard ratio of 1.089 ($\beta_2 = 0.0848$; $\chi^2 = 266.99$; $p < 0.0001$), as expected. Likewise, older women's mortality is about 35 % lower than among their male counterparts, given a hazard ratio of 0.65 ($\chi^2 = 40.83$; $p < 0.0001$). Educational attainment is negatively and significantly linked to the mortality of older Americans with a hazard ratio of 0.978 ($\beta_4 = -0.0224$; $\chi^2 = 6.5512$; $p = 0.0105$). In other words, one additional year in education would lower the mortality of older Americans by 2–3 %, other variables being equal.

The results for the other three hazard models have tremendous similarities. Table 5.1 presents the parameter estimates and the hazard ratios derived from all four models, from which the analytic results of these methods can be compared and assessed.

Table 5.1 Maximum likelihood estimates from three approximation methods and Weibull model: older Americans (n = 2000).

Explanatory variable	Parameter estimate	Standard error	Chi-square	p-value	Hazard ratio
Cox regression with exact method (likelihood ratio $\chi^2 = 333.49$; $p < 0.0001$)					
Smoking	0.4204	0.1014	17.2023	<0.0001	1.523
Age_mean	0.0848	0.0052	266.9936	<0.0001	1.089
Female_mean	–0.4262	0.0667	40.8300	<0.0001	0.653
Educ_mean	–0.0224	0.0088	6.5512	0.0105	0.978
Cox regression with Efron method (likelihood ratio $\chi^2 = 333.49$; $p < 0.0001$)					
Smoking	0.4204	0.1014	17.2022	<0.0001	1.523
Age_mean	0.0848	0.0052	266.9943	<0.0001	1.089
Female_mean	–0.4262	0.0667	40.8297	<0.0001	0.653
Educ_mean	–0.0224	0.0088	6.5509	0.0105	0.978
Cox regression with Breslow method (likelihood ratio $\chi^2 = 331.42$; $p < 0.0001$)					
Smoking	0.4189	0.1014	17.0786	<0.0001	1.520
Age_mean	0.0845	0.0052	265.2549	<0.0001	1.088
Female_mean	–0.4248	0.0667	40.5680	<0.0001	0.654
Educ_mean	–0.0230	0.0088	6.4900	0.0108	0.978
Weibull hazard rate model (–2 log likelihood = 4207.88; $p < 0.0001$)					
Smoking	0.4213	0.1018	17.1500	<0.0000	1.524
Age_mean	0.0849	0.0054	241.3300	<0.0000	1.089
Female_mean	–0.4239	0.0672	39.7200	<0.0001	0.654
Educ_mean	–0.0225	0.0087	6.5800	<0.0001	0.978

As Table 5.1 presents, even with heavy ties, the three approximation methods generate almost identical regression coefficient estimates and model fit statistics, especially between the exact and the Efron methods. As the simplest approximation, the Breslow method is shown to be a highly efficient and consistent perspective to handle tied survival times on this dataset. Not surprisingly, the Weibull proportional hazard model yields extremely close estimates of regression coefficients to those derived from the three partial likelihood functions. Because the estimation is based on the full likelihood, the Weibull hazard model has a much higher model fit statistic than the three partial likelihood models. Therefore, when the information on the intercept factor is needed, the Weibull regression should be the first choice to model the proportional hazard function.

Although the Brelow method proves an efficient and consistent approximation in this illustration, I have reason to believe that when the sample size is small, the Breslow method does not perform as well as the other two approximation approaches (Hertz-Picciotto and Rockhill, 1997; Kalbfleisch and Prentice, 2002). Given this limitation, caution must apply in selecting an approximation method when the sample size is exceptionally small (e.g., the number of events is smaller than 25). My personal recommendation is that when analyzing large-scale survival data, the application of the simple Breslow method is good enough to yield efficient, robust, and consistent estimates of regression coefficients. When the sample size is too small, as is often the case in clinical trials, the Efron method should be the choice. If a full likelihood function is needed, the Weibull model is appropriate as this parametric perspective provides similar estimates of regression coefficients with an intercept estimate.

5.3 Estimation of survival functions from the Cox proportional hazard model

Although it provides a simple, efficient way to estimate regression coefficients, the Cox model is a partial likelihood perspective in which the baseline hazard rate is an unspecified nuisance function and thereby is not parameterized. Therefore, some useful lifetime indicators, such as the survival rate or the cumulative hazard function, are not directly obtainable from analytic results of the partial likelihood. In the meantime, researchers are often interested in comparing some complete lifetime functions for assessing the effect of a particular covariate on the risk of experiencing an event. Here, a statistically significant relative risk on the hazard rate may not necessarily translate into strong differences in the hazard rate itself because such differentials also depend on the magnitude of the baseline hazard rate (Liu, 2000; Teachman and Hayward, 1993). Consider, for example, a hazard ratio of 1.5: this relative risk can come either from the ratio of 0.3 over 0.2 or from the ratio of 0.003 over 0.002. While the first case suggests a very strong difference in the hazard rate, the second ratio implies an ignorable disparity. Therefore, it is inappropriate for the researcher to reach a final conclusion about the effect of a covariate just by looking at the hazard ratio (Liu, 2000).

One useful approach to demonstrate a covariate's effect is to compare survival functions between two or more population subgroups or on some selected covariate values. Comparing group-wise cumulative hazard functions is equally informative. Some statisticians have developed approximation methods for estimating the survival function and other related lifetime indicators from the Cox model estimates. In this section, I describe two popular

methods in this area: the Kalbfleisch–Prentice method (1973) and the Breslow approach (1972), both based on the discrete step function of survival. Lastly, I provide an empirical example to demonstrate how to estimate group-wise survival curves from results of a Cox model using SAS programming.

5.3.1 The Kalbfleisch–Prentice method

The Kalbfleisch–Prentice method (1973) starts with defining the conditional probability of survival in the interval (t_{i-1}, t_i) for the baseline population, given by

$$s_i = \exp\left[-\int_{i-1}^{i} h_0(u)\,\mathrm{d}u\right], \tag{5.33}$$

where s_i is the baseline conditional probability in (t_{i-1}, t_i). By definition, the baseline probability of survival at t_i is simply the product of a series of conditional probabilities prior to t_i:

$$S_0(t_i) = \prod_{j=0}^{i-1} s_j. \tag{5.34}$$

According to Equation (5.4), the survival function with covariate vector $\boldsymbol{x}$ can be written as

$$S(t_i;\boldsymbol{x}) = [S_0(t_i)]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})} = \left(\prod_{j=0}^{i-1} s_j\right)^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}. \tag{5.35}$$

The conditional probability that an individual with covariate vector $\boldsymbol{x}_i$ survives from a particular event beyond t_i, given survival just prior to t_i, can be written as

$$\frac{[S_0(t_{i+1})]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}}{[S_0(t_i)]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}} = s_i^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}. \tag{5.36}$$

Likewise, the conditional probability that the individual with covariate vector $\boldsymbol{x}_i$ experiences an event of interest in interval (t_i, t_{i+1}), given survival just prior to t_i, is

$$\frac{[S_0(t_i)]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})} - [S_0(t_{i+1})]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}}{[S_0(t_i)]^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}} = 1 - s_i^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}. \tag{5.37}$$

As the contribution to the likelihood of an individual with covariates $\boldsymbol{x}$ who fails at t_i is $S(t_{i-1}; \boldsymbol{x}) - S(t_i; \boldsymbol{x})$ and the contribution of a censored observation at t_i is simply $S(t_i; \boldsymbol{x})$, an approximate joint likelihood function can be written as

$$\prod_{i=0}^{n'} \left\{ \prod_{l \in \mathcal{D}_i} \left[1 - s_l^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}\right] \right\} \prod_{l' \in \mathcal{R}_i - \mathcal{D}_i} s_{l'}^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}, \tag{5.38}$$

which is to be maximized with respect to $s_1, s_2, \ldots, s_{n'}$ and $\boldsymbol{\beta}$, or simply to $s_1, s_2, \ldots, s_{n'}$, with estimates $\hat{\boldsymbol{\beta}}$ from the partial likelihood function.

Differentiating the log of Equation (5.38) with respect to s_i gives the maximum likelihood estimate of s_i by solving the following equation:

$$\sum_{l\in\mathcal{D}_i}\frac{\exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}{1-\hat{s}_l^{\exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}}=\sum_{l\in\mathcal{R}(t_i)}\exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right) \tag{5.39}$$

If there are no ties ($d_i = 1$ for all $i = 1, 2, \ldots, \tilde{d}$), the conditional probability of survival is

$$\hat{s}_i=\left[1-\frac{\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)}{\sum_{l\in\mathcal{R}(t_i)}\exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}\right]^{\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)}. \tag{5.40}$$

If there are tied observations, an iterative solution for (5.40) is required. Kalbfleisch and Prentice (1973) suggest that a suitable starting value for the iteration should be

$$\hat{s}_{i0}=\exp\left[\frac{-d_i}{\sum_{l\in\mathcal{R}(t_i)}\exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}\right], \tag{5.41}$$

where $\hat{s}_{i0}$ is expected to approximate $\hat{s}_i$ very closely if the number of distinct event times is large.

Consequently, the baseline survival probability can be easily estimated. If there are no ties, then

$$\hat{S}_0(t)=\prod_{t_i<t}\hat{s}_i=\prod_{t_i<t}\left[1-\frac{\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)}{\sum_{l\in\mathcal{R}(t_i)}\exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}\right]^{\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)}. \tag{5.42}$$

Clearly, the above equation is essentially a discrete step function with $\hat{S}_0(t)$ for $t < t_1$ and $\hat{S}_0(t) = 0$ if $t > t_n$ unless there are censored times after t_n.

When $\hat{\boldsymbol{\beta}}$ in the above equation, $\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)=1$ and

$$\sum_{l\in\mathcal{R}(t_i)}\exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)=n_i,$$

then

$$\hat{s}_i=\frac{n_i-1}{n_i}. \tag{5.43}$$

Consequently, Equation (5.42) reduces to

$$\hat{S}_0(t)=\prod_{t_i<t}\frac{n_i-1}{n_i}. \tag{5.44}$$

Therefore, the above Kalbfleisch and Prentice (1973) method is essentially an extension of the Kaplan–Meier estimator without ties, described in Chapter 2.

With the baseline survival function $S_0(t)$ estimated, the baseline cumulative hazard function is given by

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = -\sum_{t_i<t} \log \hat{s}_i. \tag{5.45}$$

Given the estimates of the baseline hazard and the baseline cumulative hazard functions, the step survival and the cumulative hazard functions, given covariate vector $\boldsymbol{x}_i$, can be readily derived:

$$\hat{S}(t;\boldsymbol{x}_i) = \left[\hat{S}_0(t)\right]^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}, \tag{5.46}$$

$$\hat{H}(t;\boldsymbol{x}_i) = \hat{H}_0(t)\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right). \tag{5.47}$$

The above procedures generate the survival function and the cumulative hazard function, conditional on $\boldsymbol{x}$, as two discrete step functions. If there are ties, Equations (5.46) and (5.47) require an iterative solution, and the estimation then becomes fairly tedious.

Kalbfleisch and Prentice (1973) also propose an approach to estimate a continuous estimate of the survival function by connecting a sequence of straight lines across the log transformed step function at all t_i values; this approach, however, is not well integrated in the formation of the Cox regression.

5.3.2 The Breslow method

Given the complexity of the Kalbfleisch–Prentice method, Breslow (1972) provides a much simpler and more user-friendly approach to estimate the survival function in the Cox model. It starts with the following approximate equation on the baseline hazard rate at t_i, denoted by h_i:

$$\hat{h}_i = \frac{d_i}{t_i - t_{i-1}}\left[\sum_{l\in\mathcal{R}(t_i)} \exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)\right]^{-1}. \tag{5.48}$$

According to the definition given in Chapter 2, this approximate of the baseline hazard rate is actually the average hazard rate in interval (t_{i-1}, t_i), namely, the interval-specific discrete rate divided by the unit of the interval. Accordingly, the baseline conditional probability of experiencing a particular event can be written as

$$\hat{q}_i = \frac{d_{\mathrm{i}}}{\sum_{l\in\mathcal{R}(t_i)} \exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}, \tag{5.49}$$

where the denominator can be understood as the model-based estimate of n_i because it counts the exposure to the risk at t_i.

As mentioned in Chapter 1, in a step function it is valid to view the conditional probability of event q_i as an approximate to h_i because in a step function all hazard rates between t_{i-1} and t_i are 0. Given the equation $S(t) = \exp[-H(t)]$, the baseline survival function at t_i can be estimated from

$$\hat{S}_0(t) = \prod_{t_i<t} \exp\left[\frac{-d_i}{\sum_{l\in\mathcal{R}(t_i)} \exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}\right]. \tag{5.50}$$

Likewise, the baseline cumulative hazard function is given by

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = \sum_{t_i<t} \left[\frac{d_i}{\sum_{l\in\mathcal{R}(t_i)} \exp\left(\boldsymbol{x}_l'\hat{\boldsymbol{\beta}}\right)}\right]. \tag{5.51}$$

Given the estimates of the baseline hazard function and the resulting baseline cumulative hazard function, the full survival function and the cumulative hazard function can be estimated by applying Equations (5.46) and (5.47).

Just like the Kalbfleisch–Prentice method equivalent to the Kaplan–Meier estimator, the Breslow method on the survival function is essentially based on the Nelson–Aalen estimator. More specifically, Equation (5.50) is equivalent to Equation (2.6) when $\boldsymbol{\beta} = 0$ because the denominator becomes n_i.

As it is much simpler and equally efficient compared to the Kalbfleisch–Prentice method, the Breslow estimator on the survival function becomes a first-choice method to estimate survival curves from results of the Cox model.

5.3.3 Illustration: Comparing survival curves for smokers and nonsmokers among older Americans

In this illustration, I follow up the example presented in Subsection 5.2.3. Although the hazard ratio of smoking cigarettes is as high as 1.523, there is not enough evidence to conclude that this relative risk reflects a substantively meaningful difference in the risk of death because it is just a ratio of two absolute rates. Given this concern, in the present analysis I want to compare the survival curves between smokers and nonsmokers throughout the whole observation period, using results of the Cox model. A distinct separation between the two curves would suggest the 'meaningful' significance of the hazard ratio. As the Brelow approximation method for estimating the regression coefficients is the simplest approach that derives almost identical estimates, I use the third regression in SAS Program 5.1 as the final model. In estimating the two survival curves, values of the three control variables – 'Age_mean, 'Female_mean,' and 'Educ_mean' – are fixed as 0 (sample means), so that survival curves for smokers and nonsmokers can be compared effectively (Fox, 1987; Liu, 2000). As a result, the model-based baseline survival function actually describes the expected survival rate for nonsmokers. For smokers, the survival function can be estimated by

$$\hat{S}(t;\text{smokers}) = \left[\hat{S}_0(t)\right]^{\exp\left(\hat{b}_1\right)}.$$

For a further comparison, I also generate the same set of survival curves from analytic results of the Weibull model to examine whether the discrete and the continuous survival perspectives yield different survival curves.

Below is the SAS program for estimating the two discrete survival curves.

SAS Program 5.2:

```
……

data group;
  length Id $30;
  input smoking age_mean female_mean educ_mean Id $12-61;
  datalines;
  0.00 0.00 0.00 0.00 smoking = 0
  1.00 0.00 0.00 0.00 smoking = 1
  ;

ods graphics on;
proc phreg data = new plot(overlay) = survival;
  model duration*Status(0) = smoking age_mean female_mean educ_mean
             / ties = BRESLOW ;
  baseline covariates = group out = pred1 survival=_all_ / rowid = Id;
run;
ods graphics off;
```

In SAS Program 5.2, I use the results of the Cox model to create two predicted survival curves, one for smokers and one for nonsmokers. First, a temporary dataset 'GROUP' is created to specify values of covariates for predicting the two survival functions. As specified, the first group is nonsmokers and the second smokers, with values of the three control variables fixed at zero. Using ODS Graphics, the PROC PHREG procedure plots two survival curves for group = 0 and group = 1, respectively, according to the specification saved in the dataset 'GROUP.' The option PLOTS(OVERPLAY)=SURVIVAL tells SAS to overlay both survival curves in the same plot. In the MODEL statement, the TIES = BRESLOW option specifies the use of the Breslow approximation method to handle tied survival data (this option can be omitted because RESLOW serves as default in SAS). The COVARIATES=GROUP option specifies that the dataset GROUP contains the set of covariates in the Cox model. The ROWID=ID option identifies the curves in the plot. The option SURVIVAL=_ALL_ requests that the estimated survival function, standard error, and lower and upper confidence limits be output into the SAS dataset specified in the OUT=PRED1 option.

The resulting plot is displayed in Figure 5.1, which plots the evolution of the survival curves for, respectively, smokers and nonsmokers, derived from the Breslow approximation method on estimating the survival function. The plot displays distinct separation between the two survival curves, reflecting strong negative effects of smoking cigarettes on the survival probability. As I specify a long observation period with more than 130 time points, the step function approximates a continuous survival curve; therefore the area between the two survival curves approximates the extra person-months lived by nonsmokers as compared to the expected life among smokers. Thus, it can safely be concluded now that the hazard ratio of smoking cigarettes is substantively meaningful after adjusting for confounding effects.

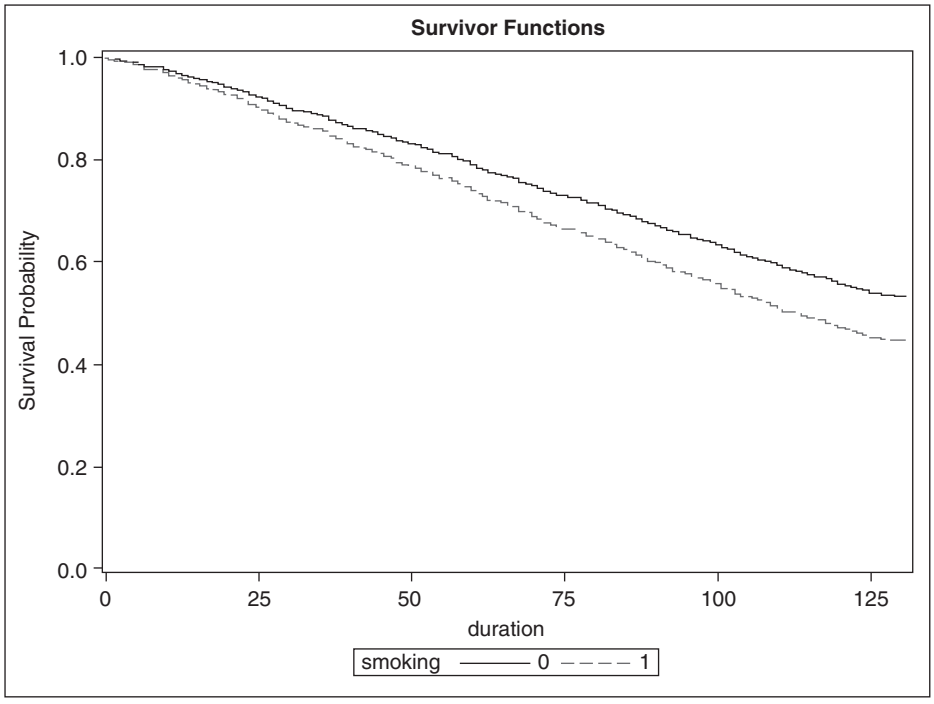


Figure 5.1 Estimated survival functions for smokers and nonsmokers.

Next, I generate two corresponding continuous survival curves using the regression coefficient of 'Smoking' from the Weibull hazard model (see SAS Program Output 5.1).

SAS Program 5.3:

```
options ls=80 ps=58 nodate;
ods select all;
ods trace off;
ods listing;
title1;
run;

data Weibull_smoking;
  do group = 1 to 2;
  do t = 0 to 130;
  if group = 1 then sr = exp(-exp(-6.9969) * (t**1.3237));
  if group = 2 then sr = exp(-exp(-6.9969 + 0.4213)*(t**1.3237));
  output;
  end;
  end;
run;

proc format;
  value group_fmt 1 = 'Nonsmokers'
            2 = 'Smokers';
run;
```

```
filename outgraph '<a subdirectory to save the plot>\Chapter5_332.gif';
goptions gsfname = outgraph;
goptions gsfname = outgraph;

Title "Figure 5.2. Survival functions of smokers and nonsmokers: the Weibull model";

symbol1 c=black i=splines l=1 v=plus;
symbol2 c=red i=splines l=2 v=diamond;

proc gplot data=Weibull_smoking;
  format group group_fmt.;
  plot sr*t = group ;
run;;
run;
quit;
```

SAS Program 5.3 generates the following plot of two continuous survival curves, one for nonsmokers and one for smokers. Figure 5.2 looks similar to Figure 5.1. As the estimated regression coefficients in the Weibull hazard model are almost the same as those estimated for the Cox model, it is useful and informative to use this parametric regression model for assessing the relative risk. One might argue that trajectories generated from a Weibull survival function may be just an artifact resulting from heterogeneous selection. I contend, however, that the Weibull hazard model is applied here to display separation between two continuous survival functions as a supplementary means for examining the effect of a covariate on the hazard rate. Such a relative risk does not depend on the validity of the baseline survival function, and using the true baseline function, assuming it is known, would lead to a similar separation of survival curves, thereby generating the same conclusion. If

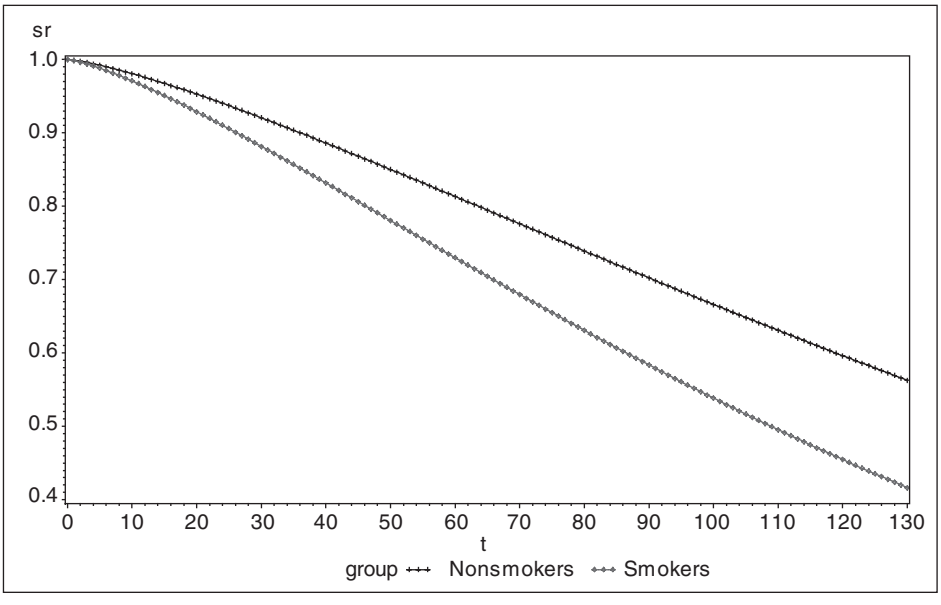


Figure 5.2 Survival functions of smokers and nonsmokers: Weibull model.

the presumed bias in the baseline hazard function truly exists, the step function generated from the Cox model should be subject to the same source of bias. At least, the area between the two Weibull survival curves displays more accurate differentials in survival times than does the Cox model.

In most situations, application of the Cox model is efficient enough to derive group-specific survival curves for large samples, as proved by Andersen *et al*. (1993) and Fleming and Harrington (1991) using the martingale central limit theorem, which will be described in Chapter 6. If the analyst needs to calculate the exact probability difference between two or more survival curves and perform a significance test, the Weibull hazard model is suggested for use.

5.4 The hazard rate model with time-dependent covariates

In previous sections, I described basic specifications of the Cox model and several refined Cox-type methods. All those regression models assume covariate values to be fixed throughout all study time, usually fixed at baseline. Variables thus specified are generally referred to as the *time-independent covariates*. This fixed-value assumption is valid for such demographic factors as sex, race/ethnicity, age at first marriage for ever-married persons, and, arguably, educational attainment among older persons. For some other variables, however, their values may change along the course of a particular life event, in turn posing potential threats to the validity of the time-independent assumption on covariates. Marital status among older persons, for example, varies over time due to high mortality, so that the mortality rate is not only a function of marital status at baseline but it may also be affected by its variations over time. The social environment, individual attitude and behavior, and many health indicators (e.g., blood pressure, cholesterol level, serum glucose), have the same time-varying characteristic. When used as explanatory factors in constructing a survival model, such time-varying variables are called the *time-dependent covariates*.

In this section, I first describe two categories of time-dependent variables, followed by specification of the hazard model using time-dependent variables as covariates and the estimating procedures. An empirical illustration is provided on the association between time-dependent marital status and the mortality of older Americans.

5.4.1 Categorization of time-dependent covariates

In the presence of time-dependent covariates, the covariate vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}(t)$ containing a set of variables $[x_1(t), \ldots, x_M(t)]$. Generally, the vector $\boldsymbol{x}(t)$ can be used to represent both time-independent and time-dependent covariates. If x_m is a time-independent covariate, $x_m(t) \equiv x_m(0)$, as a special case of $x_m(t)$; if it is a time-dependent covariate, $x_m(t)$ denotes the value of covariate x_m just prior to t. Theoretically, $x_m(t)$ as a time-dependent covariate is supposed to be known at every possible time point of an observation period. For some variables, however, measuring their values at every possible time point is unrealistic. A common practice is to measure values of such a time-dependent covariate at time intervals and then choose the measure closest to a specific survival time into analysis.

Here, I use $\tilde{\boldsymbol{x}}(t)$ to indicate values of time-dependent covariates closest to t, mathematically defined as

$$\tilde{\boldsymbol{x}}(t) = [\boldsymbol{x}(\tilde{t}) : 0 \le \tilde{t} \le t].$$

It must be borne in mind that a time-dependent covariate may be closely associated with other time-dependent variables because they tend to vary interactively along the course of survival processes, in turn implying the existence of complicated mechanisms among those variables. For example, currently smoking cigarettes generally predicts the death rate, but in the course of a lifetime event, the presence of a serious disease may influence cigarette smoking behavior, thus confounding the association between smoking cigarettes and mortality (Fisher and Lin, 1999). If currently smoking cigarettes and the severity of disease are used as time-dependent covariates simultaneously, it is very likely that both covariates will not display any significant effects on mortality due to the existence of colinearity. Considerable caution, therefore, must be exercised in using several time-dependent covariates simultaneously unless interactions among covariates are well understood.

Kalbfleisch and Prentice (2002) divide time-dependent covariates into two categories: external and internal. *External time-dependent covariates* are those whose values apply to all individuals thus not directly related to an individual's survival process. Although such time-dependent covariates can influence the hazard rate over time, its future variations are not affected by the occurrence of a particular event. One distinct example in social science is collective efficacy in the community, defined as the linkage of mutual trust and solidarity among neighbors (Sampson, Raudenbush, and Earls, 1997). This factor acts as a time-dependent variable given its frequent variations over time and the potential impact on an individual's attitude, behavior, and practice. Such an environmental variable, however, is external to an individual's survival mechanisms and thereby are not under direct observation of a particular study. In biomedical research, a prominent instance of external time-dependent covariates is the influence of the climate change on the rate of contracting bronchitis. Seasonal weather variation affects the rate of bronchitis, but its future progress is completely independent of an individual's survival from this disease.

Mathematically, an external time-dependent covariate satisfies the condition (Kalbfleisch and Prentice, 2002)

$$\mathrm{P}\{T \in [u, u+\mathrm{d}u) \,|\, \boldsymbol{x}(u), T \ge u\} = \mathrm{P}\{T \in [u, u+\mathrm{d}u) \,|\, \boldsymbol{x}(t), T \ge u\}. \tag{5.52}$$

The above equation indicates that the future path of $\boldsymbol{x}$ to any time greater than u is external to survival processes.

In contrast, the *internal time-dependent covariate* refers to a variable whose value is generated from time to time by an individual and thereby carries information on the hazard rate. Given such internal effects, time-dependent variables require periodic observation. There are many examples for an internal time-dependent covariate, such as marital status, cigarette smoking behavior, alcohol consumption, blood pressure, the presence and severity of a serious disease, and the like. In the field of military medicine, active-duty soldiers change deployments periodically, in turn affecting the prevalence and incidence of various psychiatric disorders among family members. As changes in a covariate's value are closely associated with an individual's survival process, ignoring such subject-specific variations can lead to the loss of important information.

It does not mean, however, that all subject-specific time-dependent covariates are internally time dependent. Age, for example, is a typical time-dependent variable that changes continuously over time. Nevertheless, it progresses simultaneously with time, thus not requiring direct observation. Technically, age even should not be considered an external time-dependent covariate because its time-dependent effect would simply cancel out in the partial likelihood function. As a result, fixing the value of age at baseline generates exactly the same regression coefficient as using it as a time-dependent covariate. Other age-related covariates, such as duration since first marriage, have the same property.

5.4.2 The hazard rate model with time-dependent covariates

The proportional hazard model with time-dependent covariates, as an extension of the Cox model, is defined by

$$h_i(t) = h_0(t)\exp\left[\boldsymbol{x}_i(t)'\boldsymbol{\beta}\right], \tag{5.53}$$

where the vector $\boldsymbol{x}(t)$, indicated earlier, may consist of both time-independent and time-dependent covariates. For analytic convenience, Equation (5.53) is sometimes written as

$$h_i(t) = h_0(t)\exp\left[\sum_{m=1}^{M} x_{im}(t)\beta_m\right], \tag{5.54}$$

where M is the number of covariates considered in the Cox model, among which some are time dependent and the others time independent.

Assuming censoring to be noninformative, the partial likelihood function with time-dependent covariates is

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{d} \frac{\exp\left[\sum_{m=1}^{M} x_{im}(t_i)\beta_m\right]}{\sum_{l\in\mathcal{R}(t_i)} \exp\left[\sum_{m=1}^{M} x_{lm}(t_l)\beta_m\right]}. \tag{5.55}$$

Accordingly, the log-likelihood function is

$$\log\ [L_p(\boldsymbol{\beta})] = \sum_{i=1}^{d}\left(\sum_{m=1}^{M} x_{im}(t_i)\beta_m - \log\left\{\sum_{l\in\mathcal{R}(t_i)} \exp\left[\sum_{m=1}^{M} x_{lm}(t_l)\beta_m\right]\right\}\right). \tag{5.56}$$

As the standard procedure, the estimation of $\boldsymbol{\beta}$ is based on the partial likelihood score function, given by

$$\tilde{\boldsymbol{U}}(\boldsymbol{\beta}) = \sum_{i=1}^{d}\left(\boldsymbol{x}_i(t_i) - \frac{\sum_{l\in\mathcal{R}(t_i)_i} \exp\left[\boldsymbol{x}_l(t_i)'\boldsymbol{\beta}\right]\boldsymbol{x}_l(t_i)}{\sum_{l\in\mathcal{R}(t_i)} \exp\left[\boldsymbol{x}_l(t_i)'\boldsymbol{\beta}\right]}\right). \tag{5.57}$$

The vector $\hat{\boldsymbol{\beta}}$ can be estimated by solving $\tilde{\boldsymbol{U}}(\boldsymbol{\beta}) = \mathbf{0}$. The variance–covariance matrix of $\hat{\boldsymbol{\beta}}$ can be obtained by the inverse of the observed information matrix, denoted by $\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right)$, where

$$\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{d}\left(\frac{\sum_{l\in\mathcal{R}(t_i)_i}\exp\left[\boldsymbol{x}_l(t_i)'\hat{\boldsymbol{\beta}}\right]\boldsymbol{x}_l(t_i)\boldsymbol{x}_l(t_i)'}{\sum_{l\in\mathcal{R}(t_i)}\exp\left[\boldsymbol{x}_l(t_i)'\hat{\boldsymbol{\beta}}\right]}\right) - \frac{\left\{\sum_{l\in\mathcal{R}(t_i)_i}\exp\left[\boldsymbol{x}_l(t_i)'\hat{\boldsymbol{\beta}}\right]\boldsymbol{x}_l(t_i)\right\}\left\{\sum_{l\in\mathcal{R}(t_i)_i}\exp\left[\boldsymbol{x}_l(t_i)'\hat{\boldsymbol{\beta}}\right]\boldsymbol{x}_l(t_i)\right\}'}{\sum_{l\in\mathcal{R}(t_i)}\exp\left[\boldsymbol{x}_l(t_i)'\hat{\boldsymbol{\beta}}\right]}. \tag{5.58}$$

With respect to a time-dependent covariate, time is taken into account in measuring its value, so that its effect on the hazard rate is no longer multiplicative or proportional. Consider, for example, the hazard ratio for two individuals, whose values of covariate x_m differ at time t while values of all other covariates are the same and time independent. Let j^* and l^* denote two individuals; the hazard ratio for those two individuals can be written as

$$\frac{h_{j^*}(t)}{h_{l^*}(t)} = \exp\left\{b_m\left[x_{j^*m}(t) - x_{l^*m}(t)\right]\right\}, \quad j^* \neq l^*. \tag{5.59}$$

As both $x_{j^*m}(t)$ and $x_{l^*m}(t)$ are random over time, $[x_{j^*m}(t) - x_{l^*m}(t)]$ is a function of t, suggesting that this hazard ratio is not constant along the life course. For this reason, the hazard model with time-dependent covariates can be used as a tool to check proportionality of a covariate's effects, which will be further discussed in the next chapter.

With time-dependent covariates included in survival analysis, the hazard model becomes much more complicated both technically and substantively (Fisher and Lin, 1999). Without an extensive understanding of the interrelationship between the survival outcomes and a given time-dependent covariate, the estimated regression coefficient of a time-dependent covariate can lead to misleading conclusions. Consider the example of the association between smoking cigarettes and the mortality of older persons. If current cigarette smoking behavior is specified as a time-dependent covariate, then it is most likely that all deaths would be identified as nonsmokers in the time period just prior to death. Consequently, a misleading conclusion would be derived from the regression coefficient of smoking cigarettes: nonsmokers are more likely to die than smokers, other variables being equal. Obviously, specifying cigarettes smoking as a time-dependent covariate confounds the true relationship between smoking behavior and survival among older persons. Here, both current cigarette smoking status just prior to dying and the occurrence of death are the consequences of worsening health, thereby highlighting a misspecified time-dependent covariate in smoking cigarettes.

In terms of the above example, we may combine information on current and past smoking behaviors to define a smoker, as I did in some of the previous examples. Nevertheless, cigarette smoking behavior thus specified does not distinctively depend on time, particularly among older persons, because those who stop smoking cigarettes before or during a study are still considered smokers. As a result, defining cigarette smoking behavior as a time-independent covariate serves as a more appropriate choice. The above example indicates that whether a hazard rate model should involve any time-dependent covariates needs to be guided by extensive knowledge of the mechanisms involved in the time-varying processes of covariates. If an unclear and vague time-dependent function is misspecified, an underlying hazard model can yield disastrous results (Fisher and Lin, 1999).

5.4.3 Illustration: A hazard model on time-dependent marital status and the mortality of older Americans

In this illustration, I reanalyze the effect of marital status on the mortality of older Americans, replacing the time-independent variable 'Married' with a corresponding time-dependent covariate. In Chapter 4, I examined this association for two observation intervals, 12 months and 48 months, adjusting for the confounding effects of age and educational attainment. The results did not display a significant effect of current marriage on the mortality of older Americans. This lack of association might be due to a relatively short observation interval, during which the impact of marital status on the hazard rate cannot be captured. In the present illustration, therefore, I extend the observation period from baseline to the end of the AHEAD Wave VI survey (the end of 2004), yielding an 11–12-year observation period. Given a much extended interval, I specify marital status as a time-dependent covariate, evaluated at six time points (1993, 1995, 1998, 2000, 2002, and 2004). At each time point, marital status is indexed dichotomously: 1 = 'currently married' and 0 = else, with its variable name given as 'Married $\tilde{t}$,' where $\tilde{t} = 1, 2, \ldots, 6$. Age, gender, and educational attainment are used as the control variables, with centered measures, 'Age_mean,' 'Female_mean,' and 'Educ_mean,' considered in the regression analysis.

As marital status is measured repeatedly over a series of unequally spaced time points, I use some conditional statements to capture the appropriate value for an individual in each risk set. Specifically, I first define six time variables, time1 to time6, operationally defined as the number of months elapsed from the baseline survey to each follow-up investigation. Given this definition, I let time1 = 0, time2 = 24, time3 = 60, time4 = 84, time5 = 108, and time6 = 132. Marital status in the hazard rate model is used as a time-dependent variable, named 'Married_time,' with its value determined via the conditional statements displayed in the following SAS program.

SAS Program 5.4:

```
……
proc phreg data = new ;
  model duration*Status(0) = married_time age_mean female_mean educ_mean /
     ties = BRESLOW;
```

```
  array married{*} married1-married6;
  array time{*} time1-time6;
  time1 = 0; time2 = 24; time3 = 60; time4 = 84; time5 = 108; time6 = 132;
  if duration < time[2] then married_time = married[1];
  else if duration >= time[6] then married_time = married[6];
  else do i = 1 to 5;
    if time[i] <= duration < time[i+1] then married_time = married[i];
  end;
run;

proc phreg data = new ;
  model duration*Status(0) = married age_mean female_mean educ_mean /  ties = BRESLOW ;
run;
```

SAS Program 5.4 specifies that an individual's marital status value, 0 or 1, is assigned according to his or her survival time, actual or censored. If an individual's survival time is located between time1 and time2 (24 months), the value of marital status at baseline is assigned. If it is equal to or greater than time6 (132 months), then the individual's marital status is measured at time6. For other situations, I ask SAS to assign the value of marital status at the beginning of a time interval during which the individual's death or censoring takes place. For comparison, I also estimate a Cox model with time-independent covariate 'Married' in SAS Program 5.4. From differences in the estimated regression coefficients and the model fit statistics, the importance of creating a time-dependent covariate for marital status can be evaluated. The results of the first model are presented below.

SAS Program Output 5.2:

Summary of the Number of Event and Censored Values

Total	Event	Event with Missing	Censored	Percent Censored
2000	864	86	1050	52.50

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	326.5593	4	<.0001
Score	317.0191	4	<.0001
Wald	323.5435	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
married_time	1	-0.22918	0.08009	8.1881	0.0042	0.795
age_mean	1	0.08074	0.00570	200.5965	<.0001	1.084
female_mean	1	-0.56492	0.07518	56.4661	<.0001	0.568
educ_mean	1	-0.02117	0.00924	5.2424	0.0220	0.979

In SAS Program Output 5.2, the information on the dataset, the dependent variable, the censoring status variable, and the censoring value statements, is not presented because it is exactly the same as previously reported. By involving a time-dependent covariate, there are 86 missing cases, 864 events (deaths), and 1050 censored observations. All three tests in the 'Testing Global Null Hypothesis: BETA = 0' section suggest that the null hypothesis $\hat{\boldsymbol{\beta}} = \mathbf{0}$ should be rejected, given a very strong statistical significance for each statistic ($p < 0.0001$).

In the table of 'Analysis of Maximum Likelihood Estimates,' the regression coefficient of 'Married_time' is –0.2292 ($SE = 0.0810$), statistically significant at $\alpha = 0.05$ ($\chi^2 = 8.1881$; $p = 0.0042$). The hazard ratio of this time-dependent variable, 0.795, suggests that the mortality rate for currently married persons is about 20 % lower than among those who are not currently married, other covariates being equal. The regression coefficients of the three control variables are all statistically significant, with values close to those previously reported. Therefore, for a long observation period, currently married older persons are shown to have significantly lower mortality than their counterparts not currently married, after adjusting for confounders.

Unfortunately, when time-dependent covariates are included, the survival curves for the currently married and currently not married cannot be estimated in SAS. Such survival functions, however, can be approximated by using the Cox model with all covariates fixed at baseline. Below are the results from the second model in SAS Program 5.4.

SAS Program Output 5.3:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	327.6730	4	<.0001
Score	315.3265	4	<.0001
Wald	321.5530	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
married	1	-0.25749	0.07548	11.6369	0.0006	0.773
age_mean	1	0.07627	0.00549	193.2284	<.0001	1.079
female_mean	1	-0.53572	0.07187	55.5603	<.0001	0.585
educ_mean	1	-0.01856	0.00887	4.3840	0.0363	0.982

In SAS Program Output 5.3, the model fit statistics are very close to those generated from the hazard model with marital status used as a time-dependent covariate. For example, the likelihood ratio test of the time-independent model is 327.67 compared to 326.56 estimated for the time-dependent hazard model. Obviously, the difference in this statistic is not statistically significant, indicating the use of the time-dependent marital status variable to be unnecessary. There are some variations in the four parameter estimates, but such differences are not sizable and therefore can be ignored.

From the results reported in SAS Program Output 5.3, two survival curves, one for married and one for not married, are created from the same SAS procedure displayed in

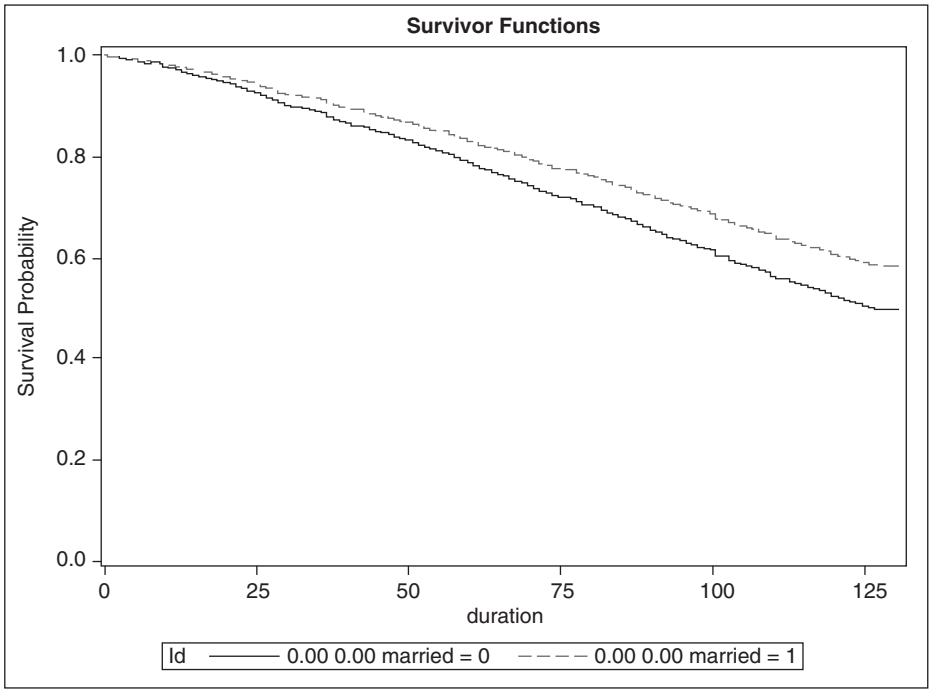


Figure 5.3 Estimated survival functions for currently married and currently not married.

SAS Program 5.2. Figure 5.3 displays distinct separation between the two survival curves, highlighting strong positive effects of current married life on the survival probability. As indicated earlier, the step function approximates a continuous survival curve given a long observation interval, so that the area between the two curves can be used for approximating the probability difference in survival. In view of the sizable separation displayed in Figure 5.3, the hazard ratio of 'Married' is associated with substantively meaningful differences in the probability of survival along the life course, exhibiting a strong impact of married life on the mortality of older Americans, other variables being equal.

As the hazard model involving one or more time-dependent covariates cannot be used to estimate the survival function, it is informative to use the Cox model without time-dependent covariates for plotting group-specific survival functions, even if the hazard model with time-dependent covariates fits survival data statistically better.

5.5 Stratified proportional hazard rate model

So far, the Cox model and its refinements have been described on the assumption that all individuals are subject to a common baseline hazard function, with population heterogeneity primarily reflected in the multiplicative predictor $\exp\left(\hat{\boldsymbol{\beta}}\boldsymbol{x}\right)$; that is, all subjects with the same x's are assumed to have an equivalent predicted hazard rate, given implicit random variation

inherent in the unspecified baseline hazard function. This assumption is not always justified because some subpopulations can be subject to a very different distribution of the baseline hazard, thus making the proportionality assumption sometimes questionable. The cohort effect, for example, is a well-documented factor for differences in the distribution of survival times among older persons (Vaupel, Manton, and Stallard, 1979).

If the assumption of proportionality is violated for a given covariate, one popular approach is to stratify on this covariate, fitting a proportional hazard model for each stratified group. Consequently, this covariate no longer serves as a covariate but is used as a stratification factor. This stratifying technique, referred to as the *stratified Cox model*, permits the underlying hazard function to vary across two or more well-defined strata and thus can yield more efficient regression coefficients of other covariates. Additionally, the stratified hazard model can be used to make graphic checks for the proportional hazards hypothesis; that is, if the proportionality assumption does not hold, a set of stratified log–log survival functions (the log transformation of the cumulative hazard function) will not be parallel.

This section describes the stratified Cox proportional hazard model. As the estimating steps follow the standard procedures described in Section 5.1, I only introduce the basic equations of this technique. An empirical illustration is provided on the association between smoking cigarettes and the mortality of older Americans using age group as a stratification factor.

5.5.1 Specifications of the stratified hazard rate model

The general approach for the stratified proportional hazard rate model is to stratify on a specific covariate that is believed to be nonproportional and then to apply the proportional hazard model for each stratified group. Suppose individuals are assigned to K well-defined and mutually exclusive strata. In stratum k, where $k = 1, \ldots, K$, the proportional hazard rate model with covariate vector $\boldsymbol{x}$ is

$$h_k(t; \boldsymbol{x}_k) = h_{0k}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{5.60}$$

where h_{0k} is the unspecified baseline hazard function for stratum k. As the above specification is actually a standard Cox model for individuals in the same stratum, Equation (5.60) is referred to as the *stratified Cox proportional hazard model*.

According to Equation (5.60), all individuals in stratum k are subject to a common baseline hazard function, whereas individuals in other strata are not. In the formulation of the Cox model, however, differences in the baseline hazard function across strata are unspecified because the primary purpose for applying the stratified Cox model is to derive more efficient regression coefficients. Because $\boldsymbol{\beta}$ in Equation (5.60) is not specific to stratum k, the stratified Cox model assumes consistent proportional hazards for other covariates across all strata. In other words, while the baseline hazard function varies from stratum to stratum, the regression coefficients and the hazard ratios of other covariates are assumed to be identical. Statistical models in which covariate effects vary over strata will be described in the next chapter.

For each stratum, the partial likelihood function is specified in the same fashion as described in Section 5.1. Specifically, the partial likelihood function for stratum k is

$$L_{kp}(\boldsymbol{\beta}) = \prod_{i=1}^{d_k} \frac{\exp(\boldsymbol{x}_{ki}'\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_{ki})} \exp(\boldsymbol{x}_{kl}'\boldsymbol{\beta})}, \tag{5.61}$$

where d_k is the number of events that occur in stratum k and t_{ki} is the ith observed event time in stratum k. Similarly, $\mathcal{R}(t_{ki})$ represents the corresponding risk set at event time t_{ki} and $\boldsymbol{x}_{ki}$ is the covariate vector with respect to stratum k. Accordingly, the log partial likelihood function is given by

$$\log L_{kp}(\boldsymbol{\beta}) = \sum_{i=1}^{d_k} \left\{ \boldsymbol{x}_{ki}'\boldsymbol{\beta} - \sum_{i=1}^{d_j} \log \left[\sum_{l \in \mathcal{R}(t_{ki})} \exp(\boldsymbol{x}_{kl}'\boldsymbol{\beta}) \right] \right\}. \tag{5.62}$$

Given the specification of the partial likelihood function for each stratum, the complete stratified partial likelihood is simply the combination of all contributions from K strata to the likelihood:

$$L_p^{\text{stratified}}(\boldsymbol{\beta}) = \prod_{k=1}^{K} L_{kp}(\boldsymbol{\beta}). \tag{5.63}$$

Accordingly, the complete stratified partial log-likelihood function is given by

$$\log L_p^{\text{stratified}}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \log L_{kp}(\boldsymbol{\beta}). \tag{5.64}$$

The partial derivative of the complete log-likelihood function is obtained by summing the partial derivatives across all strata, and then the complete log partial likelihood function is maximized with respect to $\boldsymbol{\beta}$. The score test, the Wald, and the likelihood ratio test statistics can be readily obtained from using the procedures described in previous sections.

Although the heterogeneous baseline hazard functions are unspecified in estimating the regression coefficients, graphic checks can be made to evaluate whether or not those baseline hazard functions are proportional to each other. As the log–log survival function is the log transformation of the cumulative hazard rate, it is plausible to compare stratum-specific log–log survival curves with all covariates set at zero. If they are roughly parallel, it can be inferred that the baseline hazard rates across strata tend to be proportional and therefore using the stratified Cox model is unnecessary. In contrast, if they are not parallel at all, the baseline hazard functions of different strata differ nonmultiplicatively, in turn suggesting that the inclusion of an underlying stratification factor in the Cox model, as a covariate, would violate the proportional hazards assumption.

5.5.2 Illustration: Smoking cigarettes and the mortality of older Americans with stratification on three age groups

In this illustration, I follow up the example presented in Subsection 5.3.3. Although the distinct separation of two survival curves, one for older smokers and one for nonsmokers, displays the substantive importance of the hazard ratio for the variable 'Smoking' (1.523), this result might be somewhat biased because the effect of smoking cigarettes could be inappropriately adjusted by misspecifying the effect of age as multiplicative. Older persons

at various age ranges are of different birth cohorts surviving from a rigorous 'selection of the fittest' process, so they may not necessarily be subject to a common baseline hazard function.

In this illustration, I want to estimate the effect of smoking cigarettes on the mortality of older Americans by stratifying on different age groups. In particular, I identify three age groups: younger than 75 years of age (young old), 74–84 years (old old), and 85 years or over (oldest old), all measured at baseline. The observation period is still from the baseline survey to the end of the year 2004. Among 2000 older Americans in the dataset, 923 persons are young old (46.15 %; termed 'Group I'), 884 are the old old (44.20 %; 'Group II'), and the remaining 193 older persons are oldest old (9.65 %; 'Group III'). The numbers of deaths in the three age groups are, respectively, 309, 485, and 156 persons. The estimating procedures are exactly the same as described in Subsection 5.3.3, except for the stratification procedure. In estimating the survival curves for older smokers and nonsmokers in each age group, values of the remaining two control variables – 'Female_mean' and 'Educ_mean' – are fixed as 0 (sample means). The SAS program for estimating this stratified proportional hazard model is displayed below.

SAS Program 5.5:

```
.....

if age < 75 then age_group = 1;
  else if age < 85 then age_group = 2;
  else age_group = 3;

.....

data group;
  length Id $30;
  input smoking female_mean educ_mean Id $12-61;
  datalines;
  0.00 0.00 0.00 0.00 smoking = 0
  1.00 0.00 0.00 0.00 smoking = 1
  ;

ods graphics on;
proc phreg data = new plot(overlay)=survival;
  model duration*Status(0) = smoking female_mean educ_mean /  ties = BRESLOW ;
  strata age_group;
  baseline covariates = group out = pred1 survival=_all_ / rowid = Id;
run;
ods graphics off;
```

In SAS Program 5.5, I first create a stratification factor 'Age_group,' according to the aforementioned classification. The DATA statement and the SAS PROC SQL procedure are mostly the same as those of previous examples and are therefore not presented. I continue to use the TIES = BRESLOW option as the approximation method to handle tied survival times. In the PROC PHREG statement, I add a STRATA AGE_GROUP statement to tell SAS that a Cox model is fitted on each age group. The results of the stratified proportional hazard model are used to create two predicted survival curves for each stratum, one for older smokers and one for nonsmokers. A temporary dataset 'GROUP' is created to specify values of covariates for predicting the two survival functions for each stratum. Using ODS Graphics, the PROC PHREG procedure plots two survival curves for group = 0 and

group = 1, respectively. The option PLOTS(OVERPLAY) = SURVIVAL tells SAS to overlay both survival curves in the same plot. Because of stratification on age, SAS generates three sets of survival curves, with each attaching to an age group. In the BASELINE statement, the options are exactly the same as previously presented.

The analytic results of the above stratified Cox model are shown in the following SAS output file.

SAS Program Output 5.4:

Summary of the Number of Event and Censored Values

Stratum	age_group	Total	Event	Censored	Percent Censored
1	1	923	309	614	66.52
2	2	884	485	399	45.14
3	3	193	156	37	19.17
Total		2000	950	1050	52.50

Testing Global Null Hypothesis: BETA=0

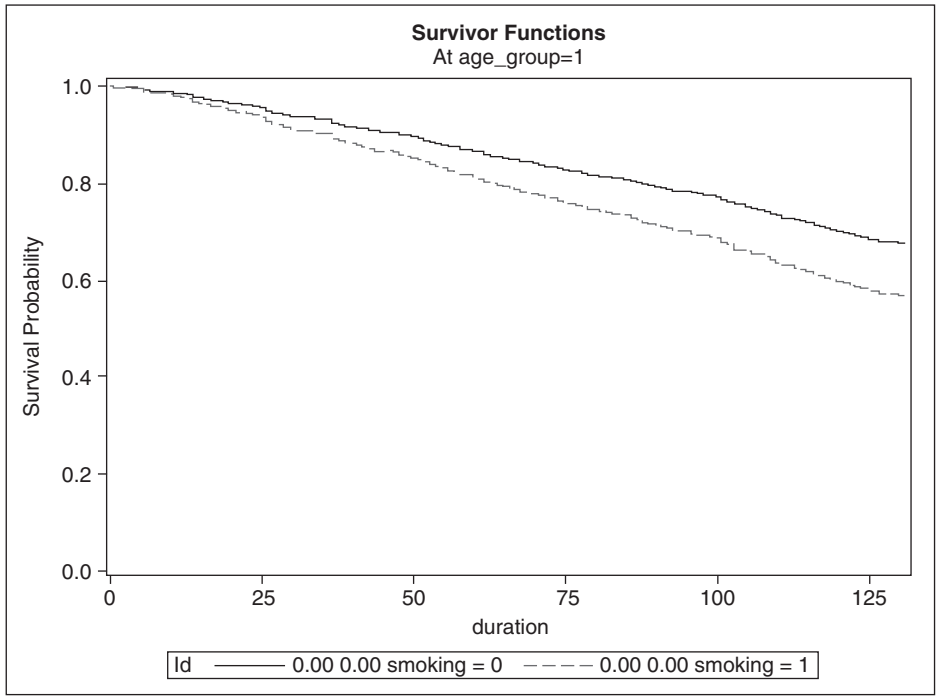
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	65.2319	3	<.0001
Score	70.3170	3	<.0001
Wald	69.3868	3	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
smoking	1	0.37484	0.10133	13.6846	0.0002	1.455
female_mean	1	-0.43271	0.06672	42.0592	<.0001	0.649
educ_mean	1	-0.02732	0.00878	9.6828	0.0019	0.973

In SAS Program Output 5.4, the 'Summary of the Number of Event and Censored Values' statement matches the data reported previously. All three tests in the 'Testing Global Null Hypothesis: BETA = 0' section demonstrate that the null hypothesis $\hat{\boldsymbol{\beta}} = \mathbf{0}$ should be rejected. Nevertheless, the chi-square values from the three tests, each with three degrees of freedom, are substantially lower than those generated from the Cox model using age as a covariate. For example, the chi-square value of the likelihood ratio is 65.23, much lower than the corresponding test statistic of 331.42 for the overall Cox model. As Kalbfleisch and Prentice (2002) caution, when stratification is used unnecessarily, some loss of efficiency is encountered.

In the table of 'Analysis of Maximum Likelihood Estimates,' the regression coefficient of 'Smoking' is 0.3748 ($SE = 0.1013$), statistically significant at $\alpha = 0.05$ ($\chi^2 = 13.68; p = 0.0002$) and close to the estimate obtained from the Cox model using age as a covariate ($\beta_1 = 0.4189$; $SE = 0.1014$; $\chi^2 = 17.08$; $p < 0.0001$). The hazard ratio of smoking cigarettes is 1.455, also



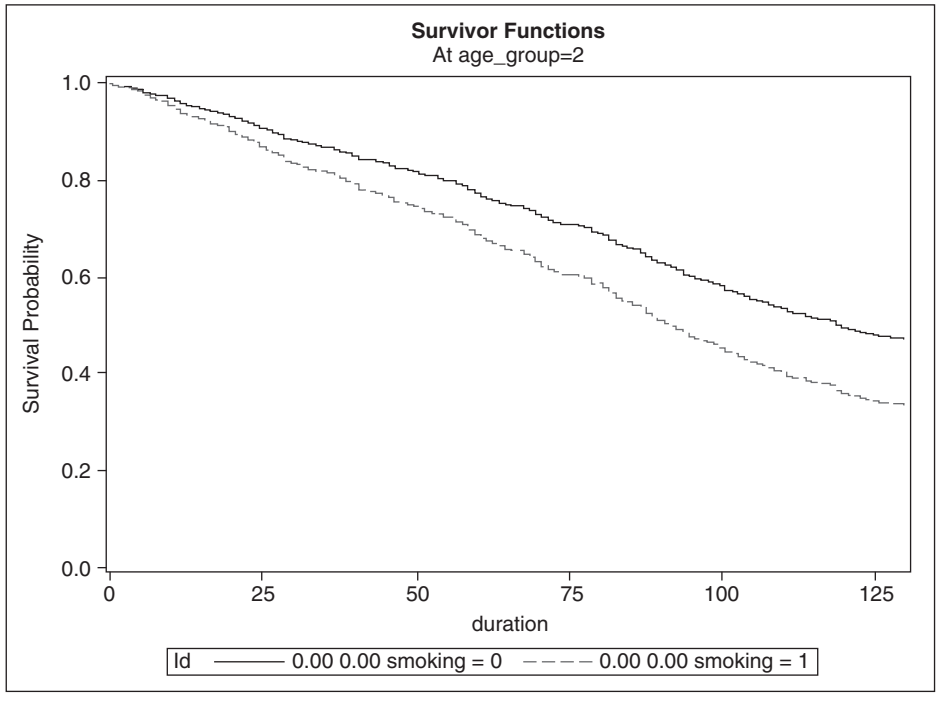


Figure 5.4 Estimated survival curves of smokers and nonsmokers for three age groups. (*Continued*)

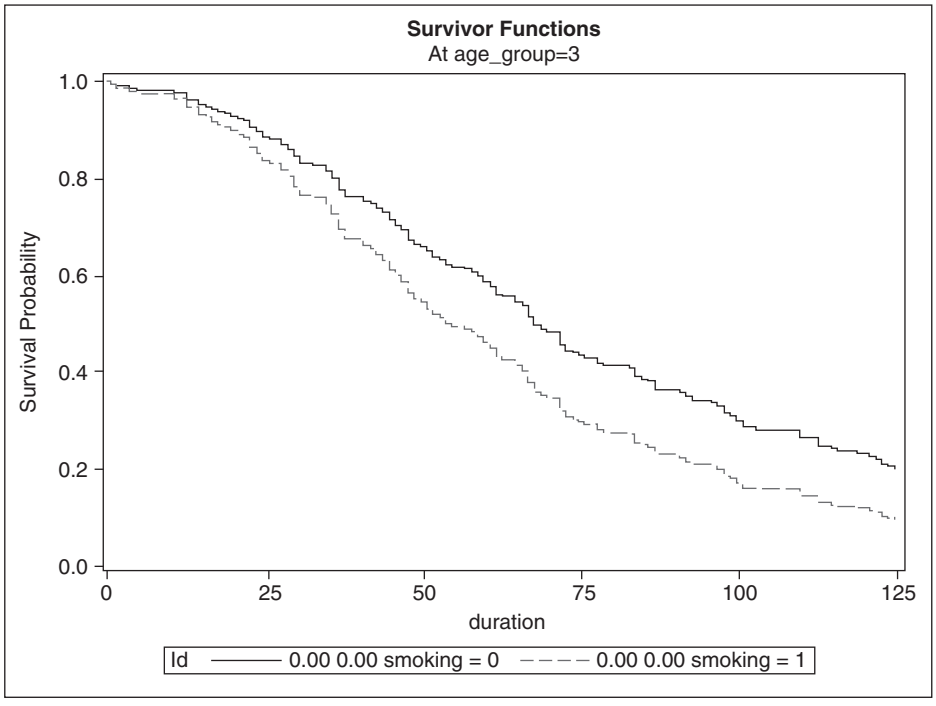


Figure 5.4 (*Continued*)

analogous to the previous score of 1.523. The regression coefficients of gender and educational attainment, both statistically significant, are close to those reported previously.

SAS Program 5.5 produces three graphs. Figure 5.4 plots three sets of survival curves for older smokers and nonsmokers, each for a specific age group. As anticipated, the two survival curves in an older age group decline more sharply than do those of a younger age group, thereby demonstrating higher mortality rates. Nevertheless, all three plots display similar separation of the two survival curves with the dominance of survival among nonsmokers, highlighting strong negative effects of smoking cigarettes on survivorship. Given these analytic results, there is no direct evidence that proportionality assumption on age is misspecified.

As the above results do not directly demonstrate whether or not age can be included in the Cox model as a covariate, I display the $\log\left[-\log\hat{S}_{0j}(t)\right]$ plot against t (the log–log function) for each age group. If the effects of age on the hazard function are proportional, the three curves should appear parallel. The SAS program for this plot is given below.

SAS Program 5.6:

```
……
proc phreg data = new ;
  model duration*Status(0) = smoking_mean female_mean educ_mean /  ties = BRESLOW ;
  strata age_group;
  baseline out = graph loglogs = lls survival = s ;
run;
```

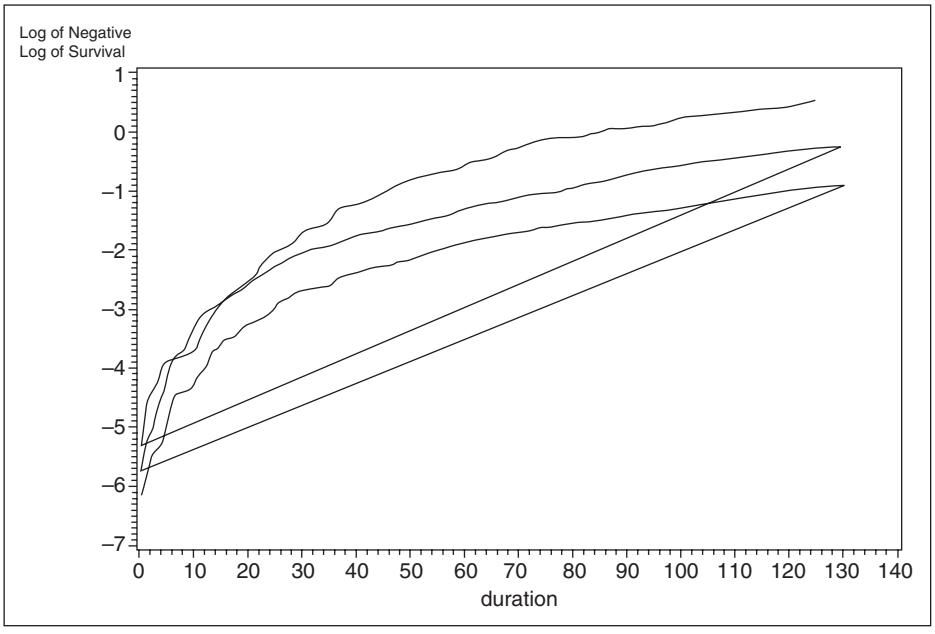


Figure 5.5 Log–log plot of three age groups.

```
proc gplot data = graph;
  plot lls * duration;
  Title "Figure 5.5. Log minus log plot of three age groups";
  symbol1 interol = join color = black line = 1;
  symbol2 interol = join color = black line = 2;
  symbol3 interol = join color = black line = 3;
run;
```

SAS Program 5.6 generates three log–log survival curves in one plot. In the PROC SQL procedure, not presented here, the variable 'Smoking' is rescaled to be centered about its sample mean, named 'Smoking_mean.' Consequently, the three baseline log–log survival curves in fact display the log-transformed cumulative hazard function for three age groups with values of all covariates fixed at sample means. Therefore, the proportionality hypothesis on age can be evaluated effectively. The resulting plot is presented in Figure 5.5, which displays three log–log survival curves with distinct separations that appear approximately parallel over time. Therefore, this graph provides some evidence that age can be considered a covariate in the Cox model for producing more efficient regression coefficients and model fit statistics. Proportionality of other covariates can be examined by plotting log–log survival curves over two or more well-defined strata.

5.6 Left truncation, left censoring, and interval censoring

Previous sections describe the Cox model and its refinements, assuming the presence of noninformative right censoring. As indicated in Chapter 1, right censoring is the most frequently encountered censoring type in survival analysis and, as a result, all textbooks

of survival analysis place tremendous emphasis on the procedures of handling this type of information loss. In empirical research, however, researchers also encounter other types of incomplete data, such as left truncation, left censoring, and interval censoring. This section briefly discusses the mechanisms inherent in those missing value types and then describes the statistical methods used to analyze survival data with their presence. As left truncation is the second most common type of information loss in survival data, the focus of this section is placed upon its description and the statistical techniques handling left truncated data. Lastly, I illustrate an empirical example on how to analyze survival data with left truncation.

5.6.1 The Cox model with left truncation, left censoring, and interval censoring

Left truncation refers to the circumstance in which an individual's survival time is not observed from the starting time t_0 to the occurrence of a particular event, but, rather, from some intermediate time exceeding t_0, denoted by $\tilde{v}$. In biomedical research, the *left truncation time* $\tilde{v}$ is equivalent to the delayed entry time. Because one has to survive to $\tilde{v}$ for being identified as a delayed entry, the survival rate from t_0 to $\tilde{v}$ for left truncated observations is 1. Therefore, left truncation in survival data actually reflects an issue of selection bias, particularly since left truncated observations are selected by an intermediate process from the time origin to $\tilde{v}$. Methodologically, those who are left truncated should have an event time defined as $T \geq t \mid \tilde{v}$, instead of $T \geq t$. If a particular event time T_i is conditional on $\tilde{v}$, this lifetime is left truncated. When left truncated observations are counted into all the risk sets from t_0 to t, the hazard rate is underestimated because the denominator of the partial likelihood function is overrepresented by involving the left truncated observations whose left truncation time exceeds the observation time. For those left truncated with $T > \tilde{v} > t$, it is already known that all of them survive to $\tilde{v} > t$ and thereby are not at risk at t. If left truncation is associated with certain covariates, then regression coefficients of those covariates may be subject to some bias, with extent depending on how left truncated observations are distributed.

Left truncation is frequently observed in demographical research. As Andersen *et al*. (1993) comment, the construction of a classical life table is based on survival data of individuals within some well-specified age intervals; thus they are left truncated at the beginning of the interval. For this reason, a classical life table only reflects survival processes for a 'synthetic' cohort combining a series of left truncated survival data, rather than of a real birth cohort in the life course. In biomedical studies, left truncation arises when a researcher introduces some intermediate event as the eligible condition for further analysis of a lifetime event, thus yielding a unique survival pattern for those who have experienced that intermediate event (Klein and Moeschberger, 2003). In clinical trials, some patients may not have entered a particular study at the time of origin until some weeks or months later; without any adjustments, including them in the analysis for the entire study period can result in erroneous outcomes.

Statisticians believe that it is relatively easy to handle left truncation if the delayed entry time $\tilde{v}$ is conditionally independent of T with its distribution depending on parameters $\boldsymbol{x}$. If left truncation tends to be random, which I believe is often the case, then the partial likelihood given $T > t > \tilde{v}$ is unbiased when those with left truncation times exceeding t are

excluded from the estimation process (Andersen *et al*., 1993). Consequently, estimation of the Cox model with left truncated data only needs to modify the delayed event time in the risk set, given by

$$\mathcal{R}(t) = (r|\tilde{v} < t < T).$$

With this modification, the Cox model can be readily applied to analyze left truncated survival data with a redefined entry time.

Left censoring, briefly discussed in Chapter 1, points to the situation in which a time point is known to be before the start of a particular study but unknown about its exact location. While left truncation is often associated with a selection process, left censoring usually occurs randomly. One popular approach to handle left censoring is to reverse the time scale, using the difference between a large time value (the maximum time point observed) and the original survival time as the event time. By performing this reversion, the left censored survival data become right censored, thereby facilitating the estimation of the Cox model with limited functional adjustments (Ware and DeMets, 1976). Sometimes, survival data consist of both left and right censored survival times, referred to as *double censoring*, and its presence would considerably complicate the estimation of the variance–covariance matrix of the survival function. Given its relative rarity, however, I do not describe further the statistical techniques of handling double censoring in this text. The interested reader is referred to Klein and Moeschberger (2003) and Turnbull (1974).

Interval censoring, described extensively in Chapters 1 and 4, points to the condition in which the exact timing for the occurrence of a particular event is not observed and, instead, the researcher only knows that the actual survival time T_i falls in a time interval (t_{i-1}, t_i). In fact, in survival data with a longitudinal growth design, event times are mostly interval censored because the occurrence of a particular event is usually identified by the status change between two successive time points. When time intervals are narrowly spaced, T_i can be approximated by using the midpoint of an interval, as widely applied by demographers. In Section 4.6, an empirical illustration on the parametric regression model provides strong evidence that when a time interval for an observed event is reasonably limited, the use of the midpoint as the time scale is appropriate and statistically efficient, deriving unbiased parameter estimates. This conclusion also applies to the estimation of the Cox model. When the interval is large, however, regular inference of survival processes might overlook some important information on survival variations within intervals, thus causing some bias in parameter estimates.

5.6.2 Illustration: Analyzing left truncated survival data on smoking cigarettes and the mortality of unmarried older Americans

In this illustration, I analyze the association between smoking cigarettes and the mortality of older Americans who are not currently married prior to death. Specifically, I apply the Cox model to estimate the regression coefficient of 'Smoking' on the hazard function, excluding currently married persons in the risk sets, and then I compare the survival curves between unmarried smokers and unmarried nonsmokers. In estimating this model, I continue to use the centered measures of age, gender, and educational attainment as the control variables. When calculating the two survival curves, I set the values of those

centered variables at 0 (sample means) for an effective comparison. Among those not currently married in the course of the study, some are unmarried at baseline and others are identified as not currently married only in subsequent waves of investigation. As an older person has to survive to a later time to change his or her marital status, those who become unmarried in follow-up surveys are systematically selected and left truncated. As indicated above, the direct application of the Cox model on left truncated data underestimates the hazard rate because individuals who become unmarried at a later time than t are improperly included in the risk set $\mathcal{R}(t)$. Thus, the risk set for estimating the relative risk of death needs to be adjusted by only including those who are unmarried prior to t and are still at risk. In other words, those who become unmarried later than t should be excluded from the risk set at t.

For analytic simplicity and convenience and without loss of generality, I assume currently unmarried persons would not become married in the course of the study, an assumption not unreasonable for older persons. Below is the SAS program to estimating the Cox model adjusting for left truncation.

SAS Program 5.7:

```
……
if married1 = 0 or married2 = 0 or married3 = 0 or married4 = 0 or married5 = 0 or
married6 = 0
  then not_married = 1;
  else not_married = 0;

  time1 = 0; time2 = 24; time3 = 60; time4 = 84; time5 = 108; time6 = 132;
  if married1 = 0 then t1 = time1;
  else if married1 = 1 and married2 = 0 then t1 = time2;
  else if married1 = 1 and married2 = 1 and married3 = 0 then t1 = time3;
  else if married1 = 1 and married2 = 1 and married3 = 1 and married4 = 0 then t1 =
    time4;
  else if married1 = 1 and married2 = 1 and married3 = 1 and married4 = 1 and
    married5 = 0 then t1 = time5;
  else if married1 = 1 and married2 = 1 and married3 = 1 and married4 = 1 and
    married5 = 1 and married6 =0 then t1 = time6;
  else t1 = 0;

……
data group;
  length Id $30;
  input smoking age_mean female_mean educ_mean Id $12-61;
  datalines;
  0.00 0.00 0.00 0.00 smoking = 0
  1.00 0.00 0.00 0.00 smoking = 1
  ;

ods graphics on;
proc phreg data = new plot(overlay)=survival; where not_married = 1;
  model duration*Status(0) = smoking age_mean female_mean educ_mean / entry = t1 ;
  baseline covariates = group out = pred1 survival=_all_ / rowid = Id;
run;
ods graphics off;
```

In SAS Program 5.7, I first create a dichotomous variable 'not_married,' with 1 = not married at any time of the observation period and 0 = married throughout the observation interval.

Next, I construct an entry time variable called 't1' redefining the origin of time for the occurrence of a death. Specifically, I define six time variables, time1 to time6, operationally defined as the number of months elapsed from the baseline survey to each follow-up survey. Given this definition, I let time1 = 0, time2 = 24, time3 = 60, time4 = 84, time5 = 108, and time6 = 132. The value of t1, the entry time, is determined via the conditional statements in SAS Program 5.7. If an individual is identified as currently not married at the baseline survey, t1 is set at 0, indicating the absence of left truncation. If the individual is married at baseline but not currently married at the first follow-up survey, t1 = 24, indicating a left truncation time of 24. Likewise, if the individual is married both at baseline and at the first follow-up but identified as currently not married at the second follow-up survey (Wave III), t1 is 60, and so forth. Obviously, for those who become unmarried later than t_0, t1 is the left truncation time $\tilde{v}$; therefore, by using this new origin of time, an event time can be properly adjusted.

A temporary dataset 'GROUP' is created again to specify values of covariates used for predicting the two survival functions, one for smokers and one for nonsmokers. As previously noted, the three control variables are fixed at zero, representing sample means. Using ODS Graphics, the PROC PHREG procedure plots two survival curves for group = 0 and group = 1 according to the specification saved in the dataset 'GROUP.' In the MODEL statement, the ENTRY = t1 option is now added to specify left truncated survival data. Here, the interval (t1, DURATION) specifies the at-risk time span, thereby taking left truncation time into consideration. Consequently, only those with $\tilde{v} < t < T$ are included in each risk set $\mathcal{R}(t)$. The options included in this procedure are interpreted previously.

The results of this Cox model with left truncated data are shown below.

SAS Program Output 5.5:

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
1214	568	646	53.21

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	132.0691	4	<.0001
Score	136.6246	4	<.0001
Wald	137.3406	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
smoking	1	0.41272	0.13479	9.3750	0.0022	1.511
age_mean	1	0.07271	0.00693	110.1823	<.0001	1.075
female_mean	1	-0.45409	0.09962	20.7774	<.0001	0.635
educ_mean	1	-0.01064	0.01160	0.8409	0.3591	0.989

SAS Program Output 5.5 shows that excluding those who are married throughout the study period and missing cases, 1214 observations are used in estimating this Cox model, among whom 568 persons are dead and 646 are right censored. Again, all three tests in the 'Testing Global Null Hypothesis: BETA = 0' section demonstrate that the null hypothesis $\hat{\boldsymbol{\beta}} = \mathbf{0}$ should be rejected.

In the table of 'Analysis of Maximum Likelihood Estimates,' the regression coefficient of 'Smoking' is 0.4127 ($SE = 0.1348$), statistically significant at $\alpha = 0.05$ ($\chi^2 = 9.38$; $p = 0.0022$), analogous to the estimate obtained from the Cox model including all individuals ($\beta_1 = 0.4189$; $SE = 0.1014$; $\chi^2 = 17.08$; $p < 0.0001$). Such a similarity indicates that marital status is not a modifier on the association between smoking cigarettes and the mortality of older Americans. The hazard ratio of 'Smoking' among unmarried persons is 1.511, very close to the score of 1.523 estimated for the entire sample. The regression coefficients of age and gender, both statistically significant, are also similar to those reported previously. Educational attainment does not have a statistically significant impact on the hazard rate.

SAS Program 5.7 also produces the graph in Figure 5.6, which plots two survival curves for, respectively, unmarried smokers and unmarried nonsmokers, derived from the Breslow method on estimating the survival function. With similar estimates of regression coefficients, the plot below is analogous to Figure 5.1, displaying distinct separation between the two survival curves. Therefore, the hazard ratio of 'Smoking' among unmarried older Americans is associated with significant differences in the survival probability between older smokers and nonsmokers, other covariates being equal. As marital status does not modify the association between smoking cigarettes and mortality, I expect married older Americans have the same pattern of survival differences displayed by Figure 5.6.

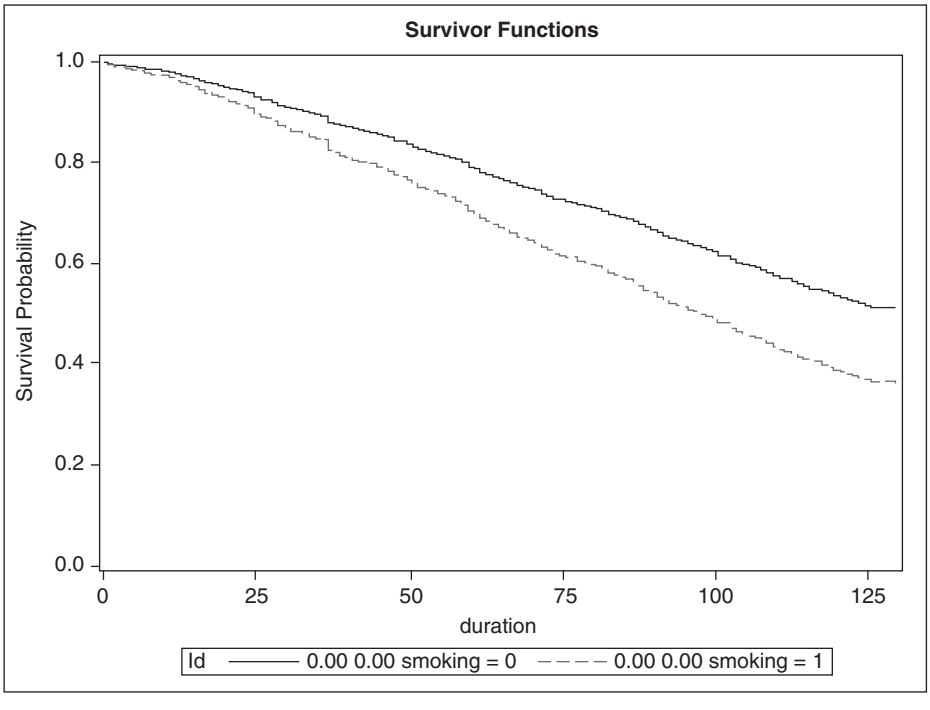


Figure 5.6 Survival functions for smokers and nonsmokers for left truncated survival times.

It may be useful to present the influence of left truncation on estimating the hazard function by using the stratified Cox model. To illustrate this influence, I create a stratification factor 'NOT_MARRIED_NEW,' with 1 = 'becoming unmarried after t_0' and 0 = 'unmarried at baseline.' The first group is left truncated, whereas the second is a regular right censored subsample. I first estimate a Cox model stratifying on this new variable without adjusting for left truncation. If left truncation impacts on the estimation of the Cox model, the survival curve for those becoming married after t_0 will tend to be relatively flat because those with left truncation times exceeding t values are all included in the risk set at t_i. Accordingly, this survival curve would deviate markedly from the survival curve for those unmarried at baseline. The SAS program for this stratified Cox model is presented below.

SAS Program 5.8:

```
……
if married1 ne 0 and not_married = 1 then not_married_new = 1;
  else if married1 = 0 then not_married_new = 0;
  else not_married_new = .;
……
proc phreg data = new ; where not_married = 1 ;
  model duration*Status(0) = smoking_mean age_mean female_mean educ_mean /  ties =
    BRESLOW ;
  strata not_married_new ;
  baseline out = graph loglogs = lls survival = s ;
run;

proc gplot data = graph;
  plot lls * duration;
  Title "Figure 5.7. Log minus log plot of old and new singles";
  symbol1 interpol = join color = black line = 1;
  symbol2 interpol = join color = black line = 2;
run;
```

In SAS Program 5.8, I first create the stratification factor 'NOT_MARRIED_NEW' according to the above classification. When this variable is 1, the survival time is left truncated because the individual has to survive to a given time point, later than t_0, when he or she becomes unmarried. The MODEL statement is conventional, thus indicating the partial likelihood function unadjusted. Given the STRATA NOT_MARRIED_NEW statement, SAS Program 5.8 generates two log–log survival curves in one plot, one for those unmarried at baseline and one for those who become unmarried later than the baseline survey and therefore left truncated. Figure 5.7 displays the result, with two log–log survival curves for those unmarried at baseline and those who become unmarried in later times. The left, upper log–log curve comes from the right censored subsample who are unmarried at baseline and thus unbiased, whereas the right, lower curve derives from the left truncated data. As can be well recognized, for those becoming unmarried after the baseline survey, the cumulative hazard rate prior to month 24 is 0 because all of them have to survive to at least 24 months later to be identified as a 'newly unmarried.' As those with left truncation times exceeding t values are mistakenly included in the risk sets, the underestimation of the log–log survival function for the left truncated stratum is obvious. The two log–log curves are uncommonly separated to an unexpected extent and the curve for the left truncated group looks strangely rectangularized. Without adjusting for left truncation,

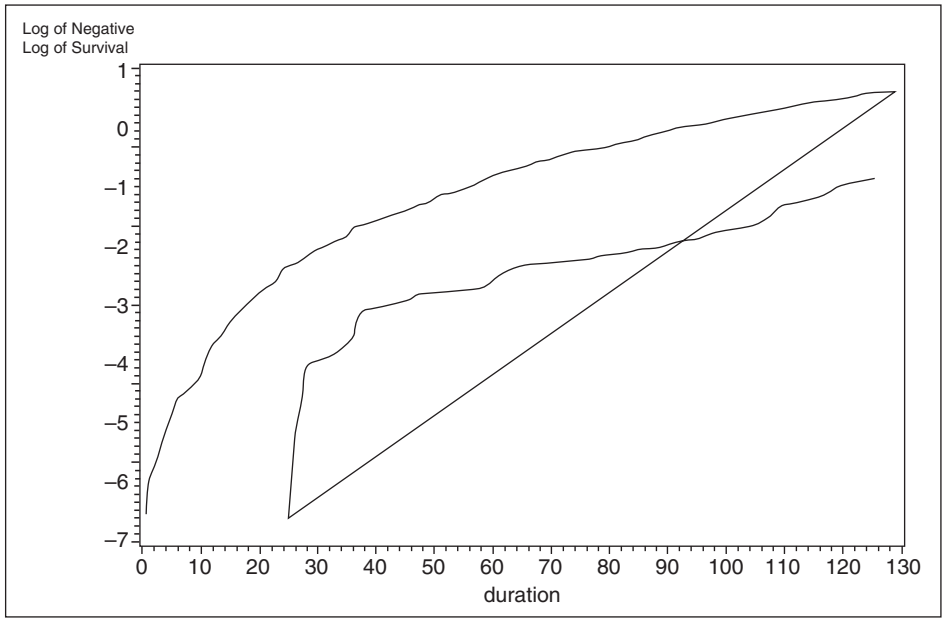


Figure 5.7 Log–log plot of old and new singles.

therefore, the application of the Cox model would inevitably yield erroneous survival differences.

To correct for the bias from left truncation, the following SAS program excludes those whose left truncation times exceed t values from the risk sets.

SAS Program 5.9:

```
……
proc phreg data = new ; where not_married = 1 ;
  model duration*Status(0) = smoking_mean female_mean educ_mean /  entry = t1 ;
  strata not_married_new ;
  baseline out = graph loglogs = lls survival = s ;
run;

proc gplot data = graph;
  plot lls * duration;
  Title "Figure 5.8. Adjusted log minus log plot of unmarried persons";
  symbol1 interpol = join color = black line = 1;
  symbol2 interpol = join color = black line = 2;
run;
```

SAS Program 5.9 specifies that any contribution to the partial likelihood must satisfy the condition that the truncation time has been exceeded. Consequently, the selection bias caused by left truncation is corrected. Figure 5.8 demonstrates the result of such correction and shows that underestimation of the relative risk is corrected after taking the delayed entry into consideration. Now, the event time T and the left truncation time $\tilde{v}$ appear independent of each other, agreeing with the general belief. The two log–log survival curves now appear

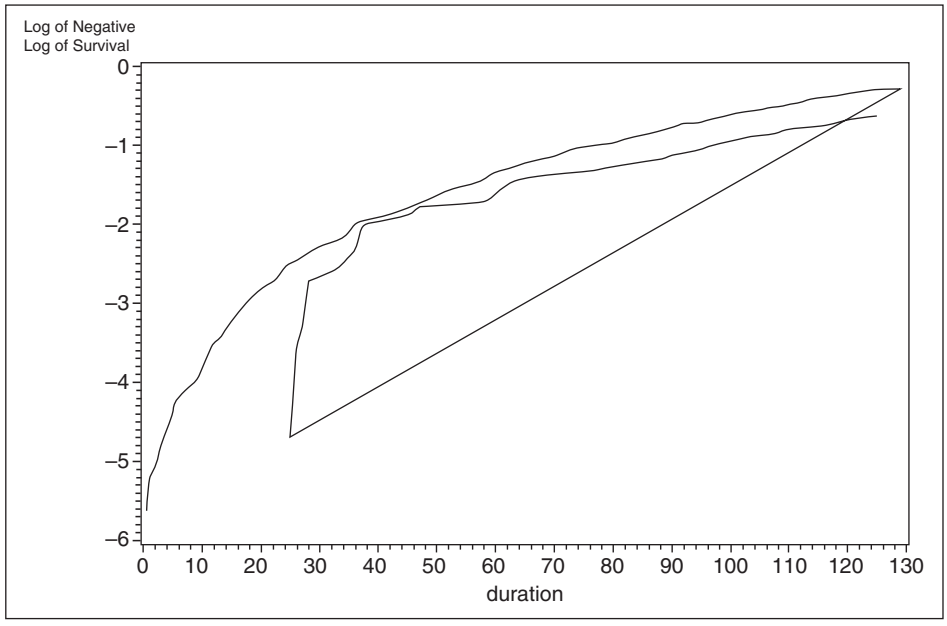


Figure 5.8 Adjusted log–log plot of unmarried persons.

parallel following the same survival pattern, except for the first 24 months, where the risk of dying for those becoming unmarried at later waves is 0.

5.7 Qualitative factors and local tests

In conducting a regression analysis, researchers often need to examine the effects of qualitative factors, represented by two or more nominal categories. For a dichotomous covariate, testing its effect on a lifetime outcome variable can be performed simply by looking at the estimated regression coefficient, the standard error, and the p-value. For a qualitative factor with more than two categories, however, just examining the results of the global tests does not provide an entire answer to the question on its effect because the factor involves a subset of $\boldsymbol{x}$, rather than a single variable. As qualitative factors are frequently used as covariates in survival analysis, it is essential for the reader to comprehend statistical techniques and estimating procedures of local tests in the Cox model.

This section first introduces several basic coding schemes for creating a qualitative factor taking more than two values. Then I describe statistical inference for local tests on the statistical significance of the estimated regression coefficients associated with a single qualitative factor. Lastly, an empirical illustration is provided on the effects of four educational groups on the mortality of older Americans, treating educational attainment as a qualitative factor.

5.7.1 Qualitative factors and scaling approaches

When a qualitative factor with K categories ($K > 2$) is used as a predictor in the Cox model, first this factor needs to be classified into K mutually exclusive levels or groups and then an

appropriate value should be assigned to each level or group. There is a variety of ways for coding those K groups, with each approach depending on the researcher's perspective to compare contrasts. In this text, I describe four basic coding schemes routinely used in regression analysis: *reference coding*, *effect coding*, *ordinal coding*, and *GLM coding*.

For analytic convenience, let X be a qualitative factor with four levels: 1, 2, 3, and 4. Accordingly, I create four design matrices containing elements X_1, X_2, X_3, and X_4, respectively, to designate group membership for each individual. Table 5.2 displays the design matrices for the aforementioned four coding schemes. The first panel displays the scheme for reference coding, also referred to as *binary coding*. This coding scheme is straightforward to apply and easy to understand, therefore serving as the most popular approach for coding a qualitative factor. Briefly, reference coding creates $K - 1$ dichotomous variables, with the remaining level or group as the reference. In Table 5.2, level 4 is used as the reference, so that three dichotomous variables are created for the three nonreference levels – 1, 2, and 3. In particular, the first three columns in the design matrix, X_1, X_2, and X_3, are three dummy variables indicating group membership of the nonreference levels. If an individual belongs to level 1, his or her group membership is defined as $X_1 = 1$, $X_2 = 0$, and $X_3 = 0$. If this person is identified as a member of level 2, the group membership is given by $X_1 = 0$, $X_2 = 1$, and $X_3 = 0$. Likewise, for an individual in level 3, the three columns are coded as $X_1 = 0$, $X_2 = 0$, and $X_3 = 1$. In terms of the reference level, level 4, all three dichotomous variables are coded 0. Therefore, the fourth dummy variable, X_4, does not need to be created for identifying the

Table 5.2 Coding schemes for classification factor $\boldsymbol{X}$.

Group in X	Design matrix			
	X_1	X_2	X_3	X_4
Reference coding				
1	1	0	0	—
2	0	1	0	—
3	0	0	1	—
4	0	0	0	—
Effect coding				
1	1	0	0	—
2	0	1	0	—
3	0	0	1	—
4	−1	−1	−1	—
Ordinal coding				
1	0	0	0	—
2	1	0	0	—
3	1	1	0	—
4	1	1	1	—
GLM coding				
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

membership of the reference group. Within the construct of this coding scheme, the estimated regression coefficient of X_1, X_2, or X_3 estimates the main effect of each nonreference group relative to the effect of the reference level. In the Cox model, for example, a hazard ratio of 1.5 for X_1 indicates that members in group 1 are 1.5 times as likely to experience a particular event as those in level 4, other covariates being equal. Additionally, the difference between two of those estimated regression coefficients displays the main effect of one nonreference membership relative to the other.

The second panel in Table 5.2 presents the scheme for effect coding. This coding scheme is analogous to reference coding, except the code for the reference group. Three dichotomous variables are created for the three nonreference levels 1, 2, and 3, in the same fashion as reference coding. For the reference group, all three dichotomous variables are coded −1, instead of 0. Therefore, the fourth dummy variable, X_4, does not need to be specified for effect coding either. Given this coding scheme, the estimated regression coefficient of X_1, X_2, or X_3 estimates the difference in the main effect of each nonreference group relative to the average effect of all four levels. Consider the Cox model again: a hazard ratio of 1.5 for X_1, given effect coding, indicates an average member in group 1 to be 1.5 times as likely to experience a particular event as is expected for the whole population.

The third coding scheme presented in Table 5.2 is ordinal coding. Under this coding scheme, all three dummy variables are coded 0 for the membership of level 1, used as a control level. With one level up, the value of 1 is added to the membership code successively; as a result, a member in level 4 is assigned to value 1 for each of X_1, X_2, or X_3. Like the above two coding schemes, the fourth dummy variable, X_4, does not need to be created for ordinal coding. The estimated regression coefficient of X_1, X_2, or X_3, using ordinal coding, displays the effect difference between two successive levels. For example, a hazard ratio of 1.5 for X_3 suggests an average member in group 3 to be 1.5 times as likely to experience a particular event as those of level 2, other covariates being equal.

The last coding scheme is used in general linear regression modeling, as its name readily indicates. It is the only coding scheme presented in Table 5.2 that requires a code for the fourth dummy variable. Codes for the first three design matrices are analogous to the reference or effect coding scheme, with the value 1 assigned to a member of the fourth level. Using this coding scheme, the main effect for each of the first three dummy variables presents the effect difference between a nonreference group and the reference level. Operationally, the GLM coding scheme is a reference cell coding, deriving exactly the same estimates of the main effects as the reference coding.

There are some more complicated coding schemes, such as orthogonal contrast coding and polynomial coding. The orthogonal contrast coding is widely used in sociological and economic studies, applied for testing some specific uncorrelated hypotheses by creating uncorrelated variables for a qualitative factor. Statisticians have also developed some mathematically oriented orthogonalization schemes to account for multicollinearity incurred from correlation among multiple covariates for a particular qualitative factor. For more details on those complicated coding schemes, the interest reader is referred to Davis (2010), Hardy and Reynolds (2004), and SAS (2009).

5.7.2 Local tests

In the Cox model, the local tests are based on the null hypothesis on a subset of the coefficient vector $\boldsymbol{\beta}$ with null hypothesis H_0: $\boldsymbol{\beta}_a = \mathbf{0}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_a', \boldsymbol{\beta}_b')'$. Specifically, $\boldsymbol{\beta}_a$ is

defined as a $\bar{q} \times 1$ vector of β values and, accordingly, $\boldsymbol{\beta}_b$ is an $(M - \bar{q}) \times 1$ vector containing coefficients of the remaining covariates.

Analogous to inference for the global test, the local test on the above null hypothesis uses the estimates derived from maximization of the partial likelihood function. Let $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_a', \hat{\boldsymbol{\beta}}_b'\right)'$ be the decomposed maximum likelihood estimator of $\boldsymbol{\beta}$. Accordingly, the information matrix $\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)$ can also be partitioned into

$$\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right) = -\begin{pmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{pmatrix}, \tag{5.65}$$

where $\boldsymbol{I}_{11}$ is the $\bar{q} \times \bar{q}$ submatrix of the second partial derivative of the minus log likelihood with respect to $\boldsymbol{\beta}_a$, $\boldsymbol{I}_{22}$ is the $(M - \bar{q}) \times (M - \bar{q})$ submatrix of the second derivatives with respect to $\boldsymbol{\beta}_b$, and $\boldsymbol{I}_{12}$ and $\boldsymbol{I}_{21}$ are the matrices of mixed second derivatives.

Given Equation (5.65), the Wald test can be used with respect to $\hat{\boldsymbol{\beta}}_a$, given by

$$\chi^2_{\text{Wald}} = \left(\hat{\boldsymbol{\beta}}_a\right)' \left[\boldsymbol{I}_{11}\left(\hat{\boldsymbol{\beta}}_a\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}}_a\right). \tag{5.66}$$

Given the large-sample approximation theory, this statistic is distributed as chi-square with $\bar{q}$ degrees of freedom under the null hypothesis; therefore H_0: $\boldsymbol{\beta}_a = \mathbf{0}$ can be statistically tested, given the value of $\chi^2_{(1-\alpha;q)}$.

Similarly, the other two test statistics can be readily defined for the local test on $\hat{\boldsymbol{\beta}}_a$. Let $\hat{\boldsymbol{\beta}}_b(\mathbf{0})$ be the partial maximum likelihood estimates of $\boldsymbol{\beta}_b$ with values in $\boldsymbol{\beta}_a$ all set at 0. The partial likelihood ratio test on H_0: $\boldsymbol{\beta}_a = \mathbf{0}$ is then given by

$$\chi^2_{\text{LR}} = 2\left\{\left[\log L_p\left(\hat{\boldsymbol{\beta}}\right)\right] - \log L_p\left[\mathbf{0}, \hat{\boldsymbol{\beta}}_b(\boldsymbol{\beta}_a = \mathbf{0})\right]\right\}, \tag{5.67}$$

which also has a chi-square distribution with $\bar{q}$ degrees of freedom. Given the critical value of $\chi^2_{\text{LR},\alpha}$, the null hypothesis that $\boldsymbol{\beta}_a = \mathbf{0}$ can be tested by the partial likelihood ratio statistic described in Section 5.1.

In terms of the score statistic, let $\tilde{\boldsymbol{U}}_a\left[\mathbf{0}, \hat{\boldsymbol{\beta}}_b(\boldsymbol{\beta}_a = \mathbf{0})\right]$ be the $\bar{q} \times 1$ vector of scores for $\boldsymbol{\beta}_a$, evaluated at the hypothesized values of $\boldsymbol{\beta}_a = \mathbf{0}$ and at the restricted partial maximum likelihood estimator for $\boldsymbol{\beta}_b$. Then the score statistic is

$$\chi^2_{\text{SC}} = \tilde{\boldsymbol{U}}_a\left[\mathbf{0}, \hat{\boldsymbol{\beta}}_b(\boldsymbol{\beta}_a = \mathbf{0})\right]' \left\{\boldsymbol{I}_{11}\left[\mathbf{0}, \hat{\boldsymbol{\beta}}_b(\boldsymbol{\beta}_a = \mathbf{0})\right]\right\}^{-1} \tilde{\boldsymbol{U}}_a\left[\mathbf{0}, \hat{\boldsymbol{\beta}}_b(\boldsymbol{\beta}_a = \mathbf{0})\right], \tag{5.68}$$

which, like the other two test statistics, has a large-sample chi-square distribution with $\bar{q}$ degrees of freedom under the null hypothesis. Clearly, a local test on a subset of the coefficient vector $\boldsymbol{\beta}$ is simply the localized realization of the global test (Klein and Moeschberger, 2003).

Attaching to a specific coding scheme, local tests can be performed using a linear combination of parameters. For the Wald test, for example, a matrix of $\bar{q}$ linear combinations can be applied to test the null hypothesis with respect to $\boldsymbol{\beta}_a$. Let $\tilde{\boldsymbol{C}}$ be a vector of contrasts for the ith linear combination of $\boldsymbol{\beta}$ with elements $(c_{i1}, c_{i2}, \ldots, c_{iM})$. The null hypothesis is given by

$$H_0 : \tilde{\boldsymbol{C}}'\boldsymbol{\beta} = \mathbf{0}. \tag{5.69}$$

From the large-sample theory, the Wald test statistic can be written as

$$\left(\tilde{C}'\hat{\beta}\right)'\left[\tilde{C}I^{-1}\left(\hat{\beta}\right)\tilde{C}'\right]^{-1}\left(\tilde{C}'\hat{\beta}\right), \tag{5.70}$$

which is distributed asymptotically as chi-square under the null hypothesis, as mentioned above.

Consider, for example, the qualitative factor with four levels, described above. After applying the reference coding scheme, three dummy variables are created – X_1, X_2, and X_3 (level 4 is used as the reference group, so X_4 does not need to be specified). The application of the Cox model generates the estimates of the main effects of those three dichotomous variables, denoted by b_1, b_2, and b_3, respectively, together with a robust variance–covariance estimator matrix. Suppose that I want to test the null hypothesis on the contrast between b_1 and b_3, given by H_0: $b_1 = b_3$. Then Equation (5.70) can be applied to test the statistical significance of this contrast, given by

$$\chi^2_W = \frac{(b_1 - b_3)^2}{\hat{V}(b_1) + \hat{V}(b_3) - 2\mathrm{cov}(b_1 b_3)},$$

which has a chi-square distribution with one degree of freedom under the null hypothesis. Given the value of α, the significance of the contrast between b_1 and b_3 can be evaluated. Significance tests for other contrasts can be conducted in the same fashion.

5.7.3 Illustration of local tests: Educational attainment and the mortality of older Americans

In previous illustrations, I assume the effects of an older American's educational attainment to be a continuous function on mortality. On certain occasions, it may be more interesting to examine the variation in the death rate by several well-defined educational groups, classifying years of education into distinct categories. In this illustration, I specify educational attainment as a qualitative factor, categorizing older Americans into four educational groups – middle school or lower, high school, college, and higher than college. Here, I want to examine the associations of these four educational levels with the mortality of older Americans, using survival information from the baseline survey to the end of 2004. The null hypothesis of this analysis is that the mortality rate in each of these four educational groups does not differ from any of the others.

Of 2000 older persons in the dataset, 480 are identified as middle school graduates or lower, 944 received education in high school, 457 were in college, and the remaining 119 persons had an education level higher than college. Accordingly, I create a qualitative factor of educational attainment, named 'Educ_group,' with 1 = middle school or lower, 2 = high school, 3 = college, and 4 = more than college. Then this factor is included in the Cox model to replace the continuous variable 'Educ.' Three centered covariates, 'Married_mean,' 'Age_mean,' and 'Female_mean,' are used as control variables in the analysis. Below is the SAS program for estimating this proportional hazard model.

SAS Program 5.10:

```
.....
if educ < 9 then educ_group = 1;
  else if educ < 13 then educ_group = 2;
  else if educ < 17 then educ_group = 3;
  else educ_group = 4;
```

```
proc phreg data = new ;
  class educ_group ;
  model duration*Status(0) = educ_group married_mean age_mean female_mean  ;
run;
```

In SAS Program 5.10, I first create the variable 'Educ_group,' using the classification indicated earlier. In the PROC PHREG procedure, the CLASS EDUC_GROUP statement specifies the variable 'Educ_group' as a classification factor, with its effects on the mortality modeled as the regression coefficients of three dichotomous covariates. As reference coding is applied as default in the SAS PROC PHREG procedure, the highest value of Educ_group, representing 'higher than college,' is used as the reference level. The following output tables display some of the results from SAS Program 5.10.

SAS Program Output 5.6:

Class Level Information

Class	Value	Design Variables		
educ_group	1	1	0	0
	2	0	1	0
	3	0	0	1
	4	0	0	0

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	333.8392	6	<.0001
Score	316.2930	6	<.0001
Wald	324.8221	6	<.0001

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
educ_group	3	9.4744	0.0236
married_mean	1	12.2753	0.0005
age_mean	1	199.5799	<.0001
female_mean	1	60.0868	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
educ_group	1	1	0.43256	0.16908	6.5449	0.0105	1.541	educ_group 1
educ_group	2	1	0.47551	0.16408	8.3987	0.0038	1.609	educ_group 2
educ_group	3	1	0.35007	0.17176	4.1541	0.0415	1.419	educ_group 3
married_mean		1	-0.26381	0.07530	12.2753	0.0005	0.768	
age_mean		1	0.07791	0.00552	199.5799	<.0001	1.081	
female_mean		1	-0.56012	0.07226	60.0868	<.0001	0.571	

In SAS Program Output 5.6, the Class Level Information section indicates the use of reference coding, with group 4 as the reference level. All three tests in the 'Testing Global Null Hypothesis: BETA = **0**' section demonstrate that the null hypothesis $\hat{\boldsymbol{\beta}} = \mathbf{0}$ should be rejected. The 'Type 3 Tests' table shows that the qualitative factor 'Educ_group' is locally statistically significant; that is, at least two of the four educational groups differ significantly in the mortality rate, other covariates being equal.

In the table of 'Analysis of Maximum Likelihood Estimates,' all the regression coefficients of the three dichotomous variables for 'Educ_group' are statistically significant, with each hazard ratio considerably greater than 1 (1.54, 1.61, and 1.42, respectively). These hazard coefficients and the hazard ratios suggest that members in each of the lower three educational groups are expected to have significantly higher mortality rates than those with an education higher than college. The three regression coefficients, however, differ slightly among themselves. The regression coefficients of the three control variables are all statistically significant, with values close to those previously reported.

Before proceeding with the local test on the contrasts of the three estimated regression coefficients, it is necessary to consider whether the effects of educational attainment depend on some other covariates. The presence of a significant interaction would change the test procedure tremendously and complicate the interpretation of the analytic results. In this analysis, I want to check whether education effects are uniform between currently married and currently not married persons. To test this interactive effect, marital status needs to be specified as a classification factor in the Cox model. Therefore, instead of using the centered covariate 'Married_mean,' I use the dichotomous variable 'Married' to construct an interaction term between education and marital status. Below is the SAS program, using Educ_group, Married, Educ_group × Married, and the two remaining control variables as covariates.

SAS Program 5.11:

```
……
proc phreg data = new ;
  class educ_group married ;
  model duration*Status(0) = educ_group married educ_group*married  age_mean
     female_mean ;
run;
```

The CLASS statement now specifies two classification factors – Educ_group and Married. An interaction term between Educ_group and Married is created for assessing whether the effects of an older person's education on mortality depends on marital status. Consequently, three additional regression coefficients are estimated, presented below.

SAS Program Output 5.7:

The PHREG Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	337.3597	9	<.0001
Score	319.9922	9	<.0001
Wald	326.8846	9	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
educ_group 1	1	0.37332	0.20789	3.2247	0.0725
educ_group 2	1	0.30891	0.19595	2.4855	0.1149
educ_group 3	1	0.16471	0.20630	0.6374	0.4246
married 0	1	-0.17407	0.35442	0.2412	0.6233
educ_group*married 1 0	1	0.31359	0.37437	0.7016	0.4022
educ_group*married 2 0	1	0.49307	0.36517	1.8231	0.1769
educ_group*married 3 0	1	0.54644	0.38008	2.0669	0.1505
age_mean	1	0.07802	0.00552	200.0903	<.0001
female_mean	1	-0.56000	0.07252	59.6225	<.0001

In SAS Program Output 5.7, values of all three tests in the 'Testing Global Null Hypothesis: BETA = **0**' section are very close to those in the model without the interaction term. For example, the chi-square of the partial likelihood ratio test here is 337.36 with 9 degrees of freedom, compared to 333.84 with 6 degrees of freedom estimated for the model without the interaction. Obviously, the difference in the partial likelihood ratio statistic between the two Cox models is not statistically significant with 3 degrees of freedom ($337.36 - 333.84 = 3.52$; $p > 0.10$). The other two tests, the score and the Wald, yield the same testing results.

In the table of 'Analysis of Maximum Likelihood Estimates,' none of the three estimated regression coefficients for the interaction term between education and marital status is statistically significant. Furthermore, with three additional covariates, the regression coefficients of the three nonreference educational groups are not statistically significant, an indication of model misspecification compared to the estimates derived from the previous model. Consequently, the Cox model without the interaction is a better choice.

Given the results on statistical significance of the interaction between education and marital status, now I can comfortably perform the local tests on the coefficient contrasts among the three lower educational groups. The following SAS program is provided for this step.

SAS Program 5.12:

```
……
proc phreg data = new ;
  class educ_group ;
  model duration*Status(0) = educ_group married_mean age_mean female_mean  ;
```

```
  contrast 'Middle school vs High School' educ_group 1 -1 ;
  contrast 'Middle School vs College' educ_group 1 0 -1 ;
  contrast 'High school vs College' educ_group 0 1 -1 ;
run;
```

In SAS Program 5.12, I use the CONTRAST statement to perform local tests on three null hypotheses: (1) H_0: $b_1 - b_2 = 0$; (2) H_0: $b_1 - b_3 = 0$; and (3) H_0: $b_2 - b_3 = 0$. I specify three contrasts – 'middle school' versus 'high school,' 'middle school' versus 'college,' and 'high school' versus 'college.' The point estimates of the three relative risks are $\exp(0.4755 - 0.4326) = 1.04$, $\exp(0.3501 - 0.4326) = 0.92$, and $\exp(0.3501 - 0.4755) = 0.88$, respectively. While the partial likelihood ratio test and the Wald statistic can both be applied for testing the three null hypotheses, the asymptotic chi-square distribution of the Wald statistic is used in SAS, with results displayed below.

SAS Program Output 5.8:

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Middle school vs High school	1	0.2923	0.5888
Middle school vs College	1	0.7465	0.3876
High school vs College	1	2.1534	0.1423

SAS Program Output 5.8 demonstrates that none of the three contrasts is statistically significant. It is now safe to conclude that while members in the lower three educational groups each have significantly higher mortality than those with an education higher than college, there are no distinct differences among themselves, other covariates being equal.

5.8 Summary

The Cox model, including the original perspective and later advancements, is the most widely applied regression models in survival analysis. Because of its simplicity, robustness, and efficiency, the Cox model becomes a very popular means to analyze the effects of covariates on the hazard function. Given the proportional hazards assumption, the application of this model has almost covered all applied disciplines, ranging from clinical trials to criminological research, from the analysis of social networks to studies of health services utilization. The development of various refined methods, such as time-dependent covariates and the stratified proportional hazard model, further widens the applicability of the Cox model, particularly in situations where the proportionality assumption in the Cox model is violated.

According to some researchers, one distinct limitation of the Cox model is the restriction of sample size. As the partial likelihood function is essentially based on the number of events, the application of the Cox model requires a much larger sample size than does a parametric regression model, particularly when categorical covariates are used or when an event of interest does not occur frequently. Therefore, the Cox model is thought not to be highly suitable for analyzing survival data with a small sample size or for dynamic events that rarely take place. Andersen, Bentzon, and Klein (1996) contend, however, that the Cox model can produce reliable estimates of survival probabilities with a small sample size. Through a Monte Carlo simulation study, these authors conclude that even with a sample size of 50

and with 25–50 % censoring, the Cox model still performs well and fits the survival data pretty nicely.

In some special situations, the use of parametric survival models is preferred (Allison, 2010; Johansen, 1983). Some advanced techniques in survival analysis, like structural equation modeling, need statistical information on the variation in the baseline hazard function. For example, with the addition or the elimination of one or more covariates, the added or purged effects are not only reflected in the estimated regression coefficients but they may also cause a change in the value of the intercept. Obviously, the Cox model cannot derive such a structural change, thereby making it difficult to derive additional estimators. In these circumstances, a parametric regression model with a complete likelihood function, like the Weibull proportional hazard model, can be used to replace the Cox model, as will be discussed further in Chapter 8.