6

Counting processes and diagnostics of the Cox model

As described in Chapter 5, maximization of the partial likelihood function derives efficient, robust, and consistent estimates of the coefficient vector $\boldsymbol{\beta}$ in the Cox model. In essence, the Cox model describes a stochastic counting process of survival over time without referring to an underlying probability distribution. As a result, it offers a flexible perspective for further advancement of survival models. More recently, some statisticians have solidified the Cox model by developing a unique counting system, referred to as *counting processes*, combining elements in the large sample theory, the martingale theory, and the stochastic integration theory (Aalen, 1975; Andersen *et al*., 1993; Andersen and Gill, 1982; Fleming and Harrington, 1991). This counting process approach, given its tremendous flexibility and the attachment to the traditional probability theory, provides a powerful tool to describe some complex stochastic processes in survival analysis, such as regression residuals of the proportional hazard model and the occurrence of repeated events.

Once the Cox model is fitted with survival data, regression diagnostics are necessary for verifying whether the statistical model fits the data appropriately or meets the proportional hazards assumption. In linear regression models, such assessment of model adequacy is generally focused on checking linearity, normality, homogeneity of variance, independence of errors, and the like. With respect to survival analysis, regression diagnostics become more complicated because an individual's hazard rate is regression derived without an observed value, and heavy censoring regularly exists in survival data. Some general principles in this area, however, are universal for all regression models, and the Cox model is no exception. Standard diagnostic techniques, such as deviation of the expected value from the observed, the overall model fitness, and identification of influential observations, also need to be performed in survival analysis. Through certain functional transformations, advanced methods of regression diagnostics have been developed and used in the Cox model, with some applying formulations of counting processes.

In this chapter, I first introduce basic specifications of counting processes and the martingale theory. Then five types of residuals in the Cox model are described, which is

Survival Analysis: Models and Applications, First Edition. Xian Liu.
© 2012 Higher Education Press. All rights reserved. Published 2012 by John Wiley & Sons, Ltd.

followed by three sections on, respectively, assessment of the proportional hazards assumption, inspection of the functional form for a covariate, and identification of influential observations from results of the Cox model, with each giving an empirical illustration. Lastly, I summarize the chapter with comments on these extended techniques attaching to the Cox model.

6.1 Counting processes and the martingale theory

As it provides a highly flexible and powerful counting system, counting processes and the martingale theory have seen increasing popularity in the past two decades. Inevitably, therefore, the counting process formulation is occasionally used in this book when I describe certain martingale-type techniques, as is seen in much of this chapter and some of the later chapters. This section provides a brief introduction of this unique counting system. In particular, I first describe the basic formulation of counting processes and then present specifications of the martingale theory and the martingale central limit theorems. Lastly, I respecify the Cox model and the partial likelihood inference with the counting process formulation for helping the reader further comprehend this system. To describe counting processes and martingales accurately, I use the original mathematical notations and functions specific to this work as much as possible. For the mathematical notations or symbols applied previously for other mathematical and statistical concepts or functions, I use the regular, nonitalicized fonts to avoid notational confusion.

6.1.1 Counting processes

In the Cox model, survival processes are expressed as a function of three key lifetime indicators – t_i, δ_i, and $\boldsymbol{x}_i$. Analogous to such expressions, a survival event in counting processes is formulated as following a triple $\{\mathrm{N}_i(t), \mathrm{Y}_i(t), \mathbf{Z}_i(t)\}$ of counting paths. Specifically, the nonitalicized N $_i(t)$ is the number of observed events in (0, t), defined as

$$\mathrm{N}_i(t) = \mathrm{I}(T_i \le t, \delta_i = 1), \tag{6.1}$$

where δ_i is the 0/1 (1 = event, 0 = censored) censoring indicator as previously defined; in the counting process formulation, it can be written as

$$\delta_i(t) = \mathrm{I}(T_i \le C_i).$$

As a count variable, $\mathrm{N}_i(t)$ is specified as a right-continuous and piecewise constant step function, with jumps of size +1. Given a single event, it is defined as a stochastic process with N(0) = 0 and N(t) < ∞, with probability summing to one. Consequently,

$$\mathrm{N}(t) = \sum_i \mathrm{N}_i(t) = \sum_{t_i \le t} \delta_i$$

The count Y in the triple denotes the number at risk just before t for failing in the interval (t, t + dt), given by

$$\mathrm{Y}_i(t) = \mathrm{I}\{\tilde{T}_i \ge t\}, \tag{6.2}$$

where Y $_i(t)$ is a left-continuous process and

$$\tilde{T}_i = \min\{T_i, C_i\}.$$

Therefore,

$$\mathrm{Y}(t) = \sum_i \mathrm{Y}_i(t)$$

is the number of individuals at risk at t. The last component in the triple, $\mathbf{Z}$, is the covariate vector, equivalent to $\boldsymbol{x}$ in the Cox model. In counting processes, it is often specified as time dependent, written as $\mathbf{Z}(t)$.

Given the above specifications, counting processes are based on the prior history about survival, censoring, and covariates, referred to as a *filtration*, denoted by $\{\mathcal{F}_t; t \geq 0\}$. Mathematically, $\{\mathcal{F}_t; t \geq 0\}$ is the sub-σ-algebra of the σ-algebra $\mathcal{F}$, continuous if $\mathcal{F}_{t+} = \mathcal{F}_t$:

$$\mathcal{F}_t = \sigma\{\mathrm{N}_i(\mathrm{s}), \mathrm{Y}_i(\mathrm{s}+), \mathbf{Z}_i(\mathrm{s}), i = 1, \ldots, n; 0 \leq \mathrm{s} \leq t\}, \quad t > 0, \tag{6.3}$$

where the nonitalicized σ{·} is the *sigma algebra* generated from the random processes specified in the brace.

The corresponding limit of $\mathcal{F}_t$ from the left is referred to as $\mathcal{F}_{t-}$, defined as the σ-algebra generated by the stochastic process $\{\mathrm{N}(t), \mathrm{Y}(t), \mathbf{Z}(t)\}$ on $[0, t)$. In other words, $\mathcal{F}_{t-}$ indicates the values of $\tilde{T}_i$ and δ_i for all i values such that $\tilde{T}_i < t$, or otherwise just the information that $\tilde{T}_i \geq t$ (Andersen *et al.*, 1993). Note that for any $\mathrm{s} \leq t$, $\mathcal{F}_\mathrm{s} \subset \mathcal{F}_t$, so that $\mathcal{F}$ is increasing.

Let T and C be nonnegative, independent random variables and T is a continuous function; thus its distribution has a density. Given $F(t) = \mathrm{P}\{T \leq t\}$ and $S(t) = 1 - F(t)$ (note that F and $\mathcal{F}$ are different functions), the intensity function at t, denoted by $\lambda(t)$, is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T < t + \Delta t | T \geq t, C \geq T\}}{\Delta t}. \tag{6.4}$$

Clearly, the intensity function in the counting processes formulation, given a single event, is equivalent to the hazard function given that T is independent of C.

Given a limited time interval $[0, \tau)$ where $\tau < \infty$ is a fixed value, the intensity function for individual i at t can be expressed as the probability of an increment in N_i over the infinitesimal time interval $[t, t + \mathrm{d}t)$ conditional on the process history just prior to t:

$$\lambda(t)\mathrm{d}t = \Pr\{\mathrm{dN}(t)\} = \mathrm{E}\{\mathrm{dN}(t) | \mathcal{F}_{t-}\}, \tag{6.5}$$

where $\mathrm{dN}_i(t)$ is the infinitesimal change in the process $\mathrm{N}(t)$ over $[t, t + \mathrm{d}t)$. Assuming no ties, $\mathrm{dN}_i(t) = 1$ for event occurrence and $\mathrm{dN}_i(t) = 0$ otherwise. Therefore, the intensity function acts as a random variable through dependence on the random variables in $\mathcal{F}_{t-}$.

Because the value of $\mathrm{Y}_i(t)$, 0 or 1, is the discrete realization of the σ-algebra $\mathcal{F}_{t-}$ for individual i, Equation (6.5) can be written as

$$\begin{aligned} \mathrm{E}\{\mathrm{dN}_i(t) | \mathcal{F}_{t-}\} &= \mathrm{E}\{\mathrm{dN}_i(t) | \mathrm{Y}_i(t)\} \\ &= \Pr\{t \leq T_i < t + \mathrm{d}t,\ C_i \geq t | \mathrm{Y}_i(t)\} \\ &= \mathrm{Y}_i(t)\lambda(t)\mathrm{d}t. \end{aligned} \tag{6.6}$$

The third equation in (6.6) is an expected value, generally referred as the *compensator* of $N_i(t)$ with respect to the filtration $\mathcal{F}_t$. Individually, if the event has not occurred prior to t, the intensity rate is $\lambda(t)$; if it has already occurred, the intensity rate is 0 at t because individual i is no longer at risk at t. Therefore, $\{Y_i(t), t \geq 0\}$ is also referred to as the at-risk process.

The integrated intensity process for individual i, denoted by the nonitalicized $\Lambda_i(t)$, is given by

$$\Lambda_i(t) = \int_0^t \lambda_i(u)\mathrm{d}u = \int_0^t \mathrm{Y}_i(u)\lambda_i(u)\mathrm{d}u, \quad t \geq 0. \tag{6.7}$$

For a single event, $\Lambda(t)$ is equivalent to the cumulative hazard function defined previously and reflects the information on the prior history of counting processes (the definition of $\Lambda(t)$ for a repeatable event is discussed in Chapter 7). Given the filtration $\mathcal{F}_{t-}$, $\mathrm{Y}(t)$ is fixed; therefore

$$\mathrm{E}[\Lambda(t)|\mathcal{F}_{t-}] = \mathrm{E}[\mathrm{N}(t)|\mathcal{F}_{t-}] = \Lambda(t). \tag{6.8}$$

Therefore, in the counting process system, $\Lambda(t)$, given $\mathcal{F}_{t-}$, is a predictable compensator because it is fixed, not random.

For analytic convenience, the intensity function is often expressed in terms of differential of the integrated intensity process, given by

$$\lambda_i(t)\mathrm{d}t = \mathrm{d}\Lambda_i(t). \tag{6.9}$$

Given the definition of $\Lambda(t)$, the intensity function for $\mathrm{N}_i(t)$ with covariate vector $\mathrm{Z}_i(t)$ can be written as a regression model:

$$\mathrm{Y}_i(t)\mathrm{d}\Lambda\{t,\mathbf{Z}_i(t)\} = \mathrm{Y}_i(t)\exp[\mathbf{Z}_i'(t)\boldsymbol{\beta}]\mathrm{d}\Lambda_0(t). \tag{6.10}$$

Given the flexibility of Equation (6.10), the above specification can be used to model counting processes of repeated events, as will be extensively described and discussed in Chapter 7.

6.1.2 The martingale theory

The processes $\mathrm{Y}_i(t)$ and $\mathbf{Z}_i(t)$ are said to be *adapted* to the filtration $\mathcal{F}_t$ because their values are specified by $\mathcal{F}_{t-}$, thereby $\mathcal{F}_t$ measurable for each $t \in [0, \tau]$. Therefore, given the right continuous counting processes of $\mathrm{N}_i(t)$, the left continuous functions $\mathrm{Y}_i(t)$ and $\mathrm{Z}_i(t)$ are predictable with respect to $\mathcal{F}_t$, which, in turn, highlights the intensity rate to be a deterministic function. According to the *Doob–Meyer decomposition theorem*, the observed event count $\mathrm{N}_i(t)$ can be expressed as the summation of a systematic compensator, represented by Equation (6.10), and a random component, called 'the counting process martingale M':

$$\mathrm{N}_i(t) = \int_0^t \mathrm{Y}_i(u)\exp[\mathbf{Z}_i'(u)\boldsymbol{\beta}]\mathrm{d}\Lambda_0(u) + \mathrm{M}_\mathrm{i}(t), \tag{6.11}$$

where the nonitalicized $\mathrm{M}_i(t)$ is a counting process martingale, defined as a stochastic process in which the expected value for individual i at time t, given its process history $\mathcal{F}_{t-}$, is equal to its value at some earlier time s.

The martingale is mathematically defined as $\{M(t), 0 \le t \le \tau\}$ with respect to the filtration $\{\mathcal{F}_t\}$. From Equation (6.11), the martingale can be easily deduced by

$$\mathrm{M}_i(t) = \mathrm{N}_i(t) - \int_0^t \mathrm{Y}_i(u)\exp[\mathbf{Z}_i'(u)\boldsymbol{\beta}]\mathrm{d}\Lambda_0(u), \tag{6.12}$$

where $\mathrm{M}_i(t)$ is the difference over $(0, t)$ between the observed number of events for individual i and the expected value of the cumulative intensity rate derived from a regression model.

As defined, a martingale must satisfy the following property:

$$\mathrm{E}[\mathrm{M}(t)|\mathcal{F}_s] = \mathrm{M}(\mathrm{s}), \quad \text{for all } \mathrm{s} \le t \le \tau, \tag{6.13}$$

which implies

$$\mathrm{E}[\mathrm{dM}(t)|\mathcal{F}_{t-}] = 0, \quad \text{for all } t \in (0, \tau]. \tag{6.14}$$

If the equality '=' in Equations (6.13) and (6.14) are replaced by the inequality '≥', the process is defined as a *submartingale* or a *local martingale* with respect to $\mathcal{F}_t$. Likewise, when the inequality is reversed as '≤', M is called a *supermartingale*. Given the nature of survival processes, a counting process $\mathrm{N}_i(t)$ is a nondecreasing step function and thus is a submartingale. Additionally, a martingale $\mathrm{M}(t)$ is said to be *square integrable* if $V[\mathrm{M}(t)] < \infty$, where $V[\mathrm{M}(t)]$ is its variance.

From the above two equations, a martingale can be viewed as the discrete random walk in a typical Markovian process, conditional on its prior history $\mathcal{F}_{t-}$. Therefore we have $\Sigma \mathrm{M}_i(t) = 0$ for any t. Additionally, as it is a stochastic process without drift and all increments are uncorrelated, a martingale also satisfies the following properties asymptotically (Andersen *et al.*, 1993; Barlow and Prentice, 1988; Therneau, Grambsch, and Fleming, 1990):

$$\mathrm{cov}[\mathrm{M}(t), \mathrm{M}(t+u) - \mathrm{M}(t)] = 0, \quad \text{for all } t \in (0, \tau], \tag{6.15}$$

$$\mathrm{cov}[\mathrm{M}_i(t), \mathrm{M}_{i'}(t)] = 0, \quad \text{for } i, i' = 1, \ldots, n \text{ and } i \ne i'. \tag{6.16}$$

These two equations indicate that the martingale increment in any t is uncorrelated and that any two right continuous $\mathcal{F}_t$-measurable martingales at any t are conditionally independent of each other. Two martingales are said to be *orthogonal martingales* if they satisfy these two conditions.

A martingale, however, has positive autocorrelation given the $\mathcal{F}_{t-}$-measurable value because, given the independence of martingale increments,

$$\mathrm{cov}[\mathrm{M}(t+u), \mathrm{M}(t)] = V[\mathrm{M}(t)], \quad \text{for all } t \in (0, \tau].$$

Additionally,

$$\mathrm{E}[\mathrm{M}^2(t)|\mathcal{F}_\mathrm{s}] = \mathrm{E}\{[\mathrm{M}(t) - \mathrm{M}(\mathrm{s})]^2 + \mathrm{M}^2(t)|\mathcal{F}_\mathrm{s}\} \ge \mathrm{M}^2(\mathrm{s}), \quad \text{for all } \mathrm{s} \le t.$$

As a result, the variance of a martingale tends to increase over t. Obviously, these two properties do not agree with the random process of ordinary residuals, thereby bringing some difficulty in the decomposition of counting processes.

Differentiation of Equation (6.11) provides a better statistical perspective for the decomposition of counting processes. Let the compensator of the counting process be $\Lambda_i(t)$ because it is the compensator of the $N_i(t)$ process. Then the differentiated counting process is

$$dN_i(t) = d\Lambda_i(t) + dM_i(t). \tag{6.17}$$

Conditional on prior history $\mathcal{F}_{t-}$, the component $dM_i(t)$ is independent of $d\Lambda_i(t)$ with mean 0 and zero autocorrelation, though not satisfying the condition of equal variance. As a result, a differentiated martingale, in many ways, behaves like an ordinary residual of linear regression models with mean 0 and without autocorrelation.

The variance of M(t), denoted V[M(t)] and conditional on $\mathcal{F}_t$, is a submartingale, so that it can also be decomposed into a compensator and a martingale with $M(0) = 0$ and $E[M(t)] = 0$. The compensator component, denoted by $\langle M \rangle(t)$ or $\langle M, M \rangle(t)$ and generally referred to as the *predictable variation* process (Andersen *et al.*, 1993; Kalbfleisch and Prentice, 2002), is

$$\langle M \rangle(t) = \int_0^t \operatorname{var}[dM(u)|\mathcal{F}_{u-}], \tag{6.18}$$

with its differentiated form

$$d\langle M \rangle(t) = \operatorname{var}[dM(t)|\mathcal{F}_{t-}]. \tag{6.19}$$

Behaving as a counting process martingale, $d\langle M \rangle(t)$ can be well decomposed according to the Doob–Meyer decomposition theorem, given by

$$d\langle M \rangle(t) = \operatorname{var}[dN(t) - d\Lambda(t)], \tag{6.20}$$

where dN(t) is a Poisson random variable with variance $\Lambda(t)$. As a consequence, the predictable variation for a counting process martingale is identical to the compensator for the counting process N(t). Using a regression formulation, $\langle M \rangle(t)$ can be written by

$$\langle M \rangle(t) = \Lambda(t) = \int_0^t Y_i(u) \exp[\mathbf{Z}_i'(u)\boldsymbol{\beta}] d\Lambda_0(u). \tag{6.21}$$

Given the above equation, $\langle M \rangle(t)$ can be viewed as an unbiased estimator of V[M(t)]. Correspondingly, a *predictable covariation* process, denoted by $\langle M_1, M_2 \rangle(t)$, where M_1 and M_2 are two martingales, can be defined in the same rationale. Specifically, with $\langle M_1, M_2 \rangle(0) = 0$ and, $E\langle M_1, M_2 \rangle(t) < \infty$ and if $d\langle M_1, M_2 \rangle(t) = \operatorname{cov}\{dM_1(t), dM_2(t)|\mathcal{F}_{t-}\}$, there exists a predictable covariation process for two martingales M_1 and M_2. M_1M_2 is a martingale if and only if $\langle M_1, M_2 \rangle \equiv 0$. This process is important in specifying a covariance matrix for a set of martingales, as will be discussed in Chapter 7.

Another simpler estimator of V[M(t)] is the *quadratic variation process*, denoted by [M] (t) and also referred to as the *optional variation process*, a simple statistical function based on observed data. Let the interval [0, t] be partitioned into J subintervals such that $0 = t_0 < t_1 < \cdots < t_J = t$. Then [M]($t$) is defined by

$$[M](t) = \sum_{s \le t} [\Delta M(s)]^2, \tag{6.22}$$

where $s = 0, t_1, \ldots, t_J$ and $\Delta M(s) = M(s) - M(s-)$.

Because N(t) is a jump function with jump size +1, [M](t) jumps correspondingly with the counting process. Therefore, a counting process martingale with a continuous compensator satisfies [M.](t) = N.(t), indicating that [M](t) is a submartingale with $\langle M\rangle(t)$ as its compensator. When M(t) has continuous sample paths, [M](t) = $\langle M\rangle(t)$.

6.1.3 Stochastic integrated processes as martingale transforms

Let M(t) be a zero mean and square-integrable martingale and H be a predictable stochastic process, both with respect to the filtration $\mathcal{F}_t$. Then the process U(t) is defined as

$$\{U(t), 0 \leq t \leq \tau\} = \int_0^t H(u)\,dM(u). \tag{6.23}$$

Clearly, the process U(t) is also a square-integrable martingale if H(t) is bounded. This martingale transform can be readily verified given Equation (6.14):

$$\begin{aligned} E[dU(t)|\mathcal{F}_{t-}] &= E[H(t)dM(t)|\mathcal{F}_{t-}] \\ &= H(t)E[dM(t)|\mathcal{F}_{t-}] \\ &= 0. \end{aligned}$$

The component H(t) can be taken outside the expectation because it is predictable and fixed by the σ-algebra $\mathcal{F}_{t-}$. The specification of this martingale transform H(t) is important because, in survival analysis, many estimators, such as the Kaplan–Meier, the Nelson–Aalen, and the partial likelihood score functions, can be expressed as stochastic integration processes.

It follows then that the predictable and quadratic processes for U(t) can also be transformed from those of M(t), given by

$$\langle U\rangle(t) = \int_0^t H^2(u)\,d\langle M\rangle(u), \tag{6.24}$$

$$[U](t) = \int_0^t H^2(u)\,d[M](u). \tag{6.25}$$

These martingale transforms are orthogonal martingales if M(t) are orthogonal martingales with respect to a common filtration $\{\mathcal{F}_t, t \geq 0\}$.

The above specifications can be used to model a number of integrated functions. A simple example is the Nelson–Aalen estimator, which can be formulated by using the martingale theory. Specifically, for nonzero Y(t), the conditional probability at t can be expressed as

$$\frac{dN(t)}{Y(t)} = \lambda(t)dt + \frac{dM(t)}{Y(t)}. \tag{6.26}$$

Let J(t) = I{Y(t) > 0} be the number of observations at risk where $Y(t) = \sum_i Y_i(t)$ and let 0/0 be 0. Then, the Nelson–Aalen estimator can be written as

$$\hat{\Lambda}(t) = \int_0^t \frac{J(u)}{Y(u)} dN(u), \quad 0 \leq t \leq \tau. \tag{6.27}$$

As the counting process $\hat{\Lambda}(t)$ has the compensator

$$\Lambda^*(t) = \int_0^t J(u)\lambda(u)du, \quad 0 \leq t \leq \tau, \tag{6.28}$$

the function $\hat{\Lambda}(t) - \Lambda^*(t)$ is a martingale transform with mean 0, with the filtration $\{\mathcal{F}_t, t \geq 0\}$ given by

$$\begin{aligned}\hat{\Lambda}(\mathrm{t}) - \Lambda^*(t) &= \int_0^t \frac{J(u)}{Y(u)}[dN(u) - Y(u)\lambda(u)du] \\ &= \int_0^t H(u)dM(u).\end{aligned}$$

Therefore, $\hat{\Lambda}(t) - \Lambda^*(t)$ is a martingale. Given Equation (6.24), the martingale transform $\hat{\Lambda}(t) - \Lambda^*(t)$ has variance

$$\left\langle \hat{\Lambda} - \Lambda^* \right\rangle(t) = \int_0^t \left[\frac{J(u)}{Y(u)}\right]^2 d\langle M\rangle(u).$$

The above inference indicates that the probability of an event occurrence at *t* is the sum of the predictable intensity rate and a random process. For large samples, the observed conditional probability tends to have continuous sample paths and, correspondingly, a predictable variation process tends to converge to a deterministic function. As a result, the bias in $\hat{\Lambda}(t)$ tends to be asymptotically negligible with an increase in sample size, thereby verifying the robustness of the Nelson–Aalen estimator. Some other stochastic integration processes, such as the Kaplan–Meier and the logrank test estimators, can also be assessed as martingale transforms (Andersen *et al.*, 1993; Fleming and Harrington, 1991), with their asymptotic processes mathematically proved by using the martingale central limit theorems described below.

6.1.4 Martingale central limit theorems

Suppose that normalized sums of orthogonal martingales converge weakly to a time-transformed *Wiener process* W(*t*) as the number of summed martingales increases. As defined, the process W(*t*) satisfies the following three characteristics:

(1) $W(0) = 0$;

(2) W(*t*) has independent increments with distribution $W(t) - W(s) \sim N(0, t - s)$ for $0 \leq s \leq t$;

(3) W(*t*) is an almost surely continuous martingale with $W(0) = 0$ and quadratic variation $[W(t)W(t,)] = t$.

Given the above three conditions, W(t) behaves as a Gaussian process if it has continuous sample paths, following the Brownian motion.

If f is a measurable nonnegative function and $V(t)=\int_0^t \mathrm{f}^2(u)\mathrm{d}u$, then $\int \mathrm{f}\,\mathrm{dW}$ is a process satisfying the three characteristics, with

$$\operatorname{var}\left[\int_0^t \mathrm{f}(u)\mathrm{dW}(u)\right]=V(t). \tag{6.29}$$

If W(t) is a multivariate process, then let {W$_1$, . . . , W$_K$} be a K-variate independent Gaussian process with independent increments and f$_1$, . . . , f$_K$ be K measurable nonnegative functions satisfying $V_\kappa(t)=\int_0^t \mathrm{f}_\kappa^2(u)\mathrm{d}u<\infty$, for all $t>0$ and $\kappa = 1, \ldots, \mathrm{K}$. The above specification is used to establish the weak convergence of a square-integrable martingale $\mathrm{U}^{(n)}$, defined below, to the Wiener process.

Let $\mathrm{U}^{(n)}(t)=\sum_{i=1}^{n}\int_0^t \mathrm{H}_i^{(n)}(u)\ \mathrm{dM}_i^{(n)}(u)$ be the sum of n orthogonal martingale transforms, where the superscript (n) indicates the dependence on sample size n (Fleming and Harrington, 1991). As a square-integrable martingale transform, the process $\mathrm{U}^{(n)}(t)$, given that H(t) is bounded, satisfies the above three conditions for the W(t) process. Consequently, the process $\mathrm{U}^{(n)}(t)$ can be viewed as a large-sample K-variate normal distribution, denoted by $\mathrm{U}_1^{(n)}(t),\ldots,\mathrm{U}_\mathrm{K}^{(n)}(t)$. Then, define

$$\mathrm{U}_{i,\kappa}^{(n)}(\mathrm{t})=\int_0^t \mathrm{H}_{i,\kappa}^{(n)}(u)\mathrm{dM}_{i,\kappa}^{(n)}(u), \tag{6.30}$$

$$\mathrm{U}_{\kappa}^{(n)}(t)=\sum_{i=1}^{n}\int_0^t \mathrm{H}_{i,\kappa}^{(n)}(u)\mathrm{dM}_{i,\kappa}^{(n)}(u), \tag{6.31}$$

where $i = 1, \ldots, n$; $\kappa = 1, \ldots, \mathrm{K}$. Also, for any $\varepsilon > 0$ (size of jumps, say), define

$$\mathrm{U}_{i,\kappa,\varepsilon}^{(n)}(t)=\int_0^t \mathrm{H}_{i,\kappa}^{(n)}(u)\mathrm{I}\left\{\left|\mathrm{H}_{i,\kappa}^{(n)}(u)\right|\geq\varepsilon\right\}\mathrm{dM}_{i,\kappa}^{(n)}(u), \tag{6.32}$$

$$\mathrm{U}_{\kappa,\varepsilon}^{(n)}(t)=\sum_{i=1}^{n}\mathrm{U}_{i,\kappa,\varepsilon}^{(n)}(t). \tag{6.33}$$

According to the specifications of $\mathrm{U}^{(n)}$, the above four processes are local square-integrable martingales with independent increments.

The martingale central limit theorem about the homogeneous process $\mathrm{U}^{(n)}$ considers conditions when $\mathrm{U}^{(n)}$ approaches a normal limit as $n \to \infty$. If the process U$_1$, U$_2$, . . . , U$_K$ are local square-integrable martingales, zero at time 0, with continuous sample paths, then the {U$_1$, U$_2$, . . . , U$_K$} behave like independent Gaussian processes, with independent increments and var{U$_i$(t)} = $V(t)$. Given $V(t) < \infty$, all $t > 0$, and $n \to \infty$, the above conditions lead to the following two characteristics:

$$\left\langle \mathrm{U}^{(n)}\right\rangle(t)\xrightarrow{\mathrm{P}}V(t) \tag{6.34}$$

and

$$\left\langle \mathrm{U}_{\varepsilon}^{(n)}\right\rangle(t)\xrightarrow{\mathrm{P}}0,\quad \text{for any } \varepsilon>0. \tag{6.35}$$

Then, it can be concluded that

$$\mathrm{U}^{(n)} \Rightarrow \mathrm{U}^{\infty} \equiv \int \mathrm{f}\,\mathrm{dW} \text{ on } \mathbb{D}(0,\infty) \text{ as } n \to \infty, \tag{6.36}$$

where W is the Brownian motion defined in the Wiener process and the sign ⇒ denotes weak convergence over the relative interval.

The martingale central limit theorem about the multivariate process $\{\mathrm{U}_1, \mathrm{U}_2, \ldots, \mathrm{U}_\mathrm{K}\}$ is just an extension of the above specifications on the univariate distribution. If the three conditions for the standard W(t) process hold and $n \to \infty$, then

$$\left\langle \mathrm{U}_\kappa^{(n)} \right\rangle(t) \xrightarrow{\mathrm{P}} \int_0^t \mathrm{f}_\kappa^2(u)\,\mathrm{d}u, \tag{6.37}$$

$$\left\langle \mathrm{U}_{\kappa,\epsilon}^{(n)} \right\rangle(t) \xrightarrow{\mathrm{P}} 0, \quad \text{for any } \epsilon > 0, \tag{6.38}$$

$$\left(\mathrm{U}_1^{(n)},\ldots,\mathrm{U}_\mathrm{K}^{(n)}\right) \Rightarrow \mathrm{U}^{\infty} \equiv \left(\int \mathrm{f}_1 \mathrm{dW}_1,\ldots,\mathrm{f}_\mathrm{K} \mathrm{dW}_\mathrm{K}\right) \text{ in } [\mathbb{D}(0,\tau)]^\mathrm{K} \quad \text{as } n \to \infty. \tag{6.39}$$

As $\mathrm{U}_1^{(\infty)}(t),\ldots,\mathrm{U}_\mathrm{K}^{(\infty)}(t)$ are defined as K independent Brownian motion processes, the covariance between $\mathrm{U}_\kappa^{(n)}$ and $\mathrm{U}_{\kappa'}^{(n)}$, where $\kappa \neq \kappa'$, has the property

$$\left\langle \mathrm{U}_\kappa^{(n)}, \mathrm{U}_{\kappa'}^{(n)} \right\rangle(t) \xrightarrow{\mathrm{P}} 0 \text{ as } n \to \infty \text{ for any } \kappa \neq \kappa'. \tag{6.40}$$

The above martingale central limit theorems literally state that when jumps of a martingale converge to a multivariate normal distribution, its sample paths or trajectory tend to an asymptotically transformed Wiener process with mean 0 and variance–covariance vector $\boldsymbol{V}(t)$. Then the variation process becomes deterministic given multivariate normality $[0, \boldsymbol{V}(t)]$.

The aforementioned seven equations compose the martingale central limit theorems when $\mathrm{U}_1^{(\infty)}(t),\ldots,\mathrm{U}_\mathrm{K}^{(\infty)}(t)$ are independent processes. Theorems about dependent time-transformed Brownian motion processes are not included in this book. For mathematical proofs of the theorems, the reader is referred to Fleming and Harrington (1991) and Rebolledo (1980).

Some statisticians have utilized the martingale central limit theorem to develop refined estimators in survival analysis (Andersen and Gill, 1982; Prentice, Williams, and Peterson, 1981). For example, if $\{\mathrm{M}_1(\cdot), \ldots, \mathrm{M}_n(\cdot)\}$ are orthogonal martingales in a partial likelihood regression model, the score function $\tilde{\boldsymbol{U}}(\boldsymbol{\beta}_0,\cdot)$, defined in Chapter 5, is a local martingale; hence it follows directly from the martingale central limit theorems that

$$\sqrt{n}\tilde{\boldsymbol{U}}(\boldsymbol{\beta}_0,\cdot) \to N(0,\boldsymbol{\Sigma}),$$

where Σ is the variance–covariance matrix for $\tilde{\boldsymbol{U}}(\boldsymbol{\beta}_0,\cdot)$. Consequently, $\hat{\boldsymbol{\beta}}$ has asymptotical normality, denoted by

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0,\cdot\right) \to N\left(0,\boldsymbol{\Sigma}^{-1}\right).$$

In some specific counting processes, $\{\mathrm{M}_1(\cdot), \ldots, \mathrm{M}_n(\cdot)\}$ are not martingales given the existence of dependence in survival times. On those occasions, the martingale central limit theorem described above is not directly applicable for estimating a hazard model, so some adjustments are needed for correctly formulating various stochastic processes, especially of $\boldsymbol{V}(t)$, as will be discussed in Chapter 7.

6.1.5 Counting process formulation for the Cox model

As Andersen *et al.* (1993) contend, the Cox model can be viewed as a special case of counting processes as it sequentially counts the number of events according to the rank order of event times. Given the counting process formulation and the powerful martingale theory, counting processes can be readily used to specify the Cox model and its asymptotic stochastic properties. For example, the score function $\tilde{\boldsymbol{U}}(\boldsymbol{\beta}_0,\cdot)$, the partial derivatives of the log partial likelihood, is basically a local martingale, for which the martingale central limit theorem applies.

Let $\mathrm{N}_i \equiv \{\mathrm{N}_i(t),\ t \geq 0\}$ be the number of observed events experienced over time t for individual i, with $\mathrm{N}_i(0) = 0\}$, and the sample paths of the process are step functions. The hazard function for individual i at time t, given an underlying nuisance function λ_0, can be specified as a counting process:

$$\lambda_i(t) = \lambda_0(t)\exp[\mathbf{Z}_i'(t)\boldsymbol{\beta}], \tag{6.41}$$

where $\boldsymbol{\beta}$ is an $M \times 1$ vector of regression coefficients. By definition, $\mathrm{Y}_i(t) = 1$ is specified until the occurrence of a particular event or of censoring and $\mathrm{Y}_i(t) = 0$ otherwise.

The partial likelihood for n independent triples $\{\mathrm{N}_i, \mathrm{Y}_i, \mathrm{Z}_i\}$, where $i = 1, \ldots, n$, is given by

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{t\geq 0}\left\{\frac{\mathrm{Y}_i(t)\exp[\mathbf{Z}_i'(t)\boldsymbol{\beta}]}{\sum_l \mathrm{Y}_l(t)\exp[\mathbf{Z}_l'(t)\boldsymbol{\beta}]}\right\}^{d\mathrm{N}_i(t)}. \tag{6.42}$$

Equation (6.42) is basically analogous to the formation of the Cox model but using a different system of terminology. Likewise, in the counting process formulation the log partial likelihood function is

$$\log L_p(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^{\infty} \mathrm{Y}_i(t)[\mathbf{Z}_i'(t)\boldsymbol{\beta}] - \log\left\{\sum_l \mathrm{Y}_l(t)\exp[\mathbf{Z}_l'(t)\boldsymbol{\beta}]\right\} d\mathrm{N}_i(t). \tag{6.43}$$

As routinely applied, the estimator $\hat{\boldsymbol{\beta}}$ in the partial likelihood function is defined as the solution to the equation

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

where $\tilde{\boldsymbol{U}}(\boldsymbol{\beta}) = (\partial \log L_p / \partial\beta_1, \ldots, \partial \log L_p / \partial\beta_M)'$. Then the total score statistic at time t is given by

$$\tilde{\boldsymbol{U}}(\boldsymbol{\beta},\infty) = \sum_{i=1}^{n}\int_0^{\infty}\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(\boldsymbol{\beta},t)\}d\mathrm{N}_i(t), \tag{6.44}$$

where the second term within the brace represents the expected covariate vector over a given risk set, defined as

$$\bar{\mathbf{Z}}(\boldsymbol{\beta},t)=\frac{\sum_{l=1}^{n}\mathrm{Y}_l(t)\mathbf{Z}_l(t)\exp[\mathbf{Z}_l'(t)\boldsymbol{\beta}]}{\sum_{l=1}^{n}\mathrm{Y}_l(t)\exp[\mathbf{Z}_l'(t)\boldsymbol{\beta}]}.$$

As $\mathrm{dN}(t)$ is a submartingale, the score function with respect to $\boldsymbol{\beta}$ can be expressed as a martingale transform in the form of Equation (6.23):

$$\tilde{\boldsymbol{U}}(\boldsymbol{\beta},\infty)=\sum_{i=1}^{n}\int_0^{\infty}\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}(\boldsymbol{\beta},t)\}\mathrm{dM}_i(t), \tag{6.45}$$

where

$$\mathrm{M}_i(t)=\mathrm{N}_i(t)-\int_0^t \mathrm{Y}_i(u)\exp[\mathbf{Z}_i'(u)\boldsymbol{\beta}]\mathrm{d}\Lambda_0(u).$$

As discussed in Chapter 5, the score function in the Cox model can be conveniently expressed in terms of a permutation test based on residuals computed for the regression on covariates, rather than on the regression coefficients. Here, using the counting process formulation to express the score function provides a flexible instrument to specify some complex refinements in the hazard model, such as the asymptotic variance estimators for handling clustered survival data in the Cox model, as will be described extensively in Chapter 7.

If the N counting process satisfies conditions of the martingale central limit theorem, the $\{\mathrm{M}_1(.), \ldots, \mathrm{M}_n(.)\}$ series are orthogonal martingales; then $\boldsymbol{\beta}$ can be estimated using the conventional estimators (Andersen and Gill, 1982). The case in which the M_i series do not behave as orthogonal martingales will be discussed in Chapter 7.

Given that $\{\mathrm{M}_1(.), \ldots, \mathrm{M}_n(.)\}$ are martingales, the information matrix, which is the minus second partial derivatives of the log partial likelihood function with respect to $\boldsymbol{\beta}$, can be used to approximate $\boldsymbol{V}(t)$, given by

$$\boldsymbol{I}(\boldsymbol{\beta},\infty)=\sum_{i=1}^{n}\int_0^{\infty}\{\tilde{\mathbf{Z}}(\boldsymbol{\beta},t)-\bar{\mathbf{Z}}(\boldsymbol{\beta},t)\}\mathrm{dN}_i(t), \tag{6.46}$$

where

$$\tilde{\mathbf{Z}}(\boldsymbol{\beta},t)=\frac{\sum_{l=1}^{n}\mathrm{Y}_l(t)\mathbf{Z}_l(t)^{\otimes 2}\exp[\mathbf{Z}_l'(t)\boldsymbol{\beta}]}{\sum_{l=1}^{n}\mathrm{Y}_l(t)\exp[\mathbf{Z}_l'(t)\boldsymbol{\beta}]}.$$

Given the above inference, statistical tests on the null hypothesis that H_0: $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ can be based on the estimated values $\hat{\boldsymbol{\beta}}$. According to the martingale central limit theorem, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)$

has approximately a multivariate normal distribution with mean 0 and covariance matrix $n\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0)$, which can be estimated by $n\hat{\boldsymbol{I}}^{-1}\left(\hat{\boldsymbol{\beta}}\right)$. Subsequently, the three test statistics – the score, the Wald, and the partial likelihood ratio tests – can be readily derived using procedures described in Chapter 5.

Expressed in terms of the counting process formulation, the above specifications take a tremendous resemblance to the corresponding formulas in the Cox model. The counting process formulation, however, describes counts at observed times, thereby permitting multiple jumps for an individual as a step function. As a result, this system provides a more flexible perspective to describe repeated events and the proportional rates/means function. The Cox model in the counting process formulation can be estimated in SAS by specifying a semi-closed time interval; given a single event, however, the procedure derives exactly the same results as the conventional Cox approach.

6.2 Residuals of the Cox proportional hazard model

In linear regression models, it is straightforward to compute residuals from the difference between the observed and the expected values for a continuous outcome variable. In the case of the Cox model, however, the hazard rate is unobservable because the outcome data are binary and censored. The same issue also exists for some other generalized linear models such as the logistic and the probit regressions. This lack of adequate information makes computation of regression residuals challenging in the Cox model, in turn prompting the development of a number of residual types for assessing the adequacy of the proportional hazards model. In principle, a considerable deviation from an observed rate transform reflects inadequate specification of the relative risk, thereby alerting the researcher to reconsider the analytic strategy on the model specification.

In this section, I describe five residuals that have been widely used in survival analysis: the Cox and Snell (1968), the Schoenfeld (1982) and the scaled Schoenfeld (Grambsch and Therneau, 1994), the martingale (Andersen *et al.*, 1993; Fleming and Harrington, 1991), the score (Therneau, Grambsch, and Fleming, 1990), and the deviance (Therneau, Grambsch, and Fleming, 1990) residuals. Lastly, I provide an illustration to demonstrate how to derive various types of residuals using SAS programming.

6.2.1 Cox–Snell residuals

The Cox–Snell residual is originally developed to define residuals associated with different model functions, including the log-linear regression. Let the link function of a regression model be g(·) and the outcome variable be *Y*. Then, for individual *i*, a regression model can be expressed by

$$Y_i = \mathrm{g}_i\left(\hat{\boldsymbol{\beta}}, \varepsilon_i\right), \tag{6.47}$$

where $\hat{\boldsymbol{\beta}}$ contains the ML estimates of regression coefficients and ε_i is the residual for individual *i*. Equation (6.47) has a unique solution for ε_i, given by

$$\varepsilon_i = g_i^{-1}\left(Y_i, \hat{\boldsymbol{\beta}}\right), \tag{6.48}$$

where $\mathrm{g}^{-1}(\cdot)$ is the inverse link function of g(·).

In the case of a log-linear regression, Equation (6.47) can be expanded to

$$Y_i = \exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \varepsilon_i\right) = \exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)\exp\left(\varepsilon_i\right). \tag{6.49}$$

Let $\tilde{r}_i = \exp(\varepsilon_i)$ be a transform from the additive residual ε_i. It follows that the transformed residuals of this log-linear regression can be readily specified from Equation (6.49):

$$\tilde{r}_i = \left[\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)\right]^{-1} Y_i, \tag{6.50}$$

where $\tilde{r}_1, \ldots, \tilde{r}_n$ are multiplicative residuals transformed from ε_i, which should have a lognormal distribution with mean $\exp(\sigma^2/2)$ and variance $\exp[2\sigma^2 - \exp(\sigma^2)]$ if ε_i is normally distributed with mean 0 and variance σ^2. We can also say that if the mean of $\tilde{r}_1, \ldots, \tilde{r}_n$ is assumed to be 1, ε_i does not have a normal distribution with zero expectation unless $\sigma^2 = 0$.

While developed prior to the publication of the Cox model, the Cox–Snell residual has been borrowed to specify residuals of the semi-parametric proportional hazard model. If the equation $h_i(t) = \hat{h}_0(t)\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)$ reflects the true hazard rate for individual i at time t, the integrated hazard rate, the cumulative hazard function, is considered to be a random variable evaluated at survival time t_i ($i = 1, \ldots, n'$). The parametric form $H_0(t) = t$ yields a unit exponential distribution. This distribution is referred to as the *unit exponential function*. Given this specification, the estimated cumulative hazard function, represented by $\hat{H}_i(t_i) = \hat{H}_0(t_i)\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)$, should behave approximately as a censored sample with the unit exponential distribution. Therefore, the Cox–Snell residual for a step function is simply given by

$$\tilde{r}_1^{\text{Cox-Snell}} = \hat{H}_0(t_i)\exp\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right), \tag{6.51}$$

where the baseline cumulative hazard rate at t_i, $\hat{H}_0(t_i)$, can be obtained from applying the Breslow estimator defined by Equation (5.51). Given a unit exponential distribution of residuals, the expected cumulative hazard function should display an upward straight line from the origin. Therefore, if the model is fitted correctly, the Cox–Snell residuals should be narrowly scattered around a straight line, behaving approximately as a censored sample with a unit exponential distribution.

The validity of the Cox–Snell residuals is based on the assumption of a unit exponential distribution for survival times, so that the expected straight line of the cumulative hazard function only applies to the special case of the exponential function. For other distributional functions, another transform should be specified (Cox and Snell, 1968). For example, if the true form is $H_0(t) = t^{\tilde{p}}$, a Weibull function, then the Cox–Snell residuals would be misspecified. Even if deviations from the unit exponential distribution demonstrate an unbiased set of random errors, the Cox–Snell residuals are not associated with a numeric statistic that can be used to assess analytically whether the residuals are statistically significant under the null hypothesis.

6.2.2 Schoenfeld residuals

Schoenfeld (1982) proposes an approach to calculate residuals from a different direction. Specifically, such residuals, formally called the *Schoenfeld residuals*, are defined specifically for the proportional hazard model, with computations performed within the context of the

Cox model. Like the specification of the partial likelihood, the derivation of the Schoenfeld residual does not depend on time but, rather, on the rank order of survival times. For fully comprehending this residual type, the reader might want to review general inference on the Cox model, described in Chapter 5, before proceeding with the following description.

Let d be the total number of events, ordered by event time, and $\mathcal{R}(t_i)$ the risk set at t_i. In the Cox model, the coefficient vector $\boldsymbol{\beta}$ can be estimated by maximizing the partial likelihood function:

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{d} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})}. \tag{6.52}$$

From Equation (6.52), $\boldsymbol{x}_i$ can be viewed as a random variable with expected value

$$\mathrm{E}[\boldsymbol{x}_i | \mathcal{R}(t_i)] = \frac{\sum_{l \in \mathcal{R}(t_i)} \boldsymbol{x}_l \exp(\boldsymbol{x}_l'\boldsymbol{\beta})}{\sum_{l \in \mathcal{R}(t_i)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})}. \tag{6.53}$$

Notice that for analytic simplicity and without loss of generality, the above specifications are based on the assumption of no tied observations.

As described in Chapter 5 (and also in Section 6.1 using the counting process formulation), the maximum likelihood estimate of $\boldsymbol{\beta}$ is a solution to

$$\sum_{i=1}^{d} \{\boldsymbol{x}_i - \mathrm{E}[\boldsymbol{x}_i | \mathcal{R}(t_i)]\} = \mathbf{0}.$$

Given the above partial likelihood estimator, Schoenfeld (1982) defines the partial residual at t_i as a vector $\hat{\tilde{\boldsymbol{r}}}_i = \left(\hat{\tilde{r}}_{i1}, \ldots, \hat{\tilde{r}}_{iM}\right)'$, where

$$\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}} = \boldsymbol{x}_i - \mathrm{E}[\boldsymbol{x}_i | \mathcal{R}(t_i)].$$

Therefore, the Schoenfeld residuals are the differences between the covariate vector for the individual at event time t_i and the expectation of the covariate vector over the risk set $\mathcal{R}(t_i)$. Because it comes from the covariate vector $\boldsymbol{x}$, the Schoenfeld residual is an $M \times n$ matrix, computing a series of individual-specific residual values corresponding to each of the M covariates considered in the Cox model. Consequently, each residual needs to be evaluated separately, corresponding to a given covariate. Given the simplicity of the formulation, the variance–covariance matrix of the Schoenfeld residuals at event time t_i, denoted by $\hat{\boldsymbol{V}}\left(\hat{\tilde{\boldsymbol{r}}}_i\right)$, can be easily estimated. Notice that, as in Equation (6.52), $\boldsymbol{x}_i$ represents the covariate vector for an event at t_i; Schoenfeld does not define this type of residual for censored observations.

Let $\tilde{\boldsymbol{U}}_i$ be an $M \times M$ score matrix with the (m, m')th element being $\partial\tilde{r}_{im} / \partial\beta_{m'}$ evaluated at $\hat{\boldsymbol{\beta}}$, and let $\tilde{\boldsymbol{U}} = \sum_{i \in d} \tilde{\boldsymbol{U}}_i$. Using a Taylor expansion series, the Schoenfeld residual can be written by

$$\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}} = \tilde{\boldsymbol{r}}_i^{\text{Schoen}} + \tilde{\boldsymbol{U}}_i\left(\hat{\boldsymbol{\beta}}\right) + 0_{\text{p}}\left(n^{-1}\right), \tag{6.54}$$

where $0_{\text{p}}(n^{-1})$, by definition, indicates the set of $\hat{\tilde{\boldsymbol{r}}}_i$ to converge to 0 in probability as n tends to a large number.

When the score statistic is substituted for $\hat{\boldsymbol{\beta}}$, we have

$$\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}} = \tilde{\boldsymbol{r}}_i^{\text{Schoen}} + \tilde{\boldsymbol{U}}_i \tilde{\boldsymbol{U}}^{-1} \sum_{l \in D} \tilde{\boldsymbol{r}}_l^{\text{Schoen}} + 0_{\text{p}}\left(n^{-1}\right). \tag{6.55}$$

Equation (6.55) displays the dependence of $\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}}$ on $\hat{\boldsymbol{\beta}}$.

Grambsch and Therneau (1994) consider it statistically more efficient to standardize the Schoenfeld residuals by using the coefficients and the variance matrix from a standard time-independent Cox model fit, given by

$$\hat{\tilde{\boldsymbol{r}}}_i^{*,\text{Schoen}} = \frac{\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}}}{\hat{V}\left(\tilde{\boldsymbol{r}}_i^{\text{Schoen}}\right)}. \tag{6.56}$$

It is recognizable that Equation (6.56) is the standard formulation for computing a standardized score. Such standardized scores, contained in the vector $\hat{\tilde{\boldsymbol{r}}}_i^{*,\text{Schoen}}$, are referred to as the *weighted*, or the *scaled*, *Schoenfeld residuals*.

Grambsch and Therneau (1994) found through some empirical analyses that the variance–covariance matrix $\hat{\boldsymbol{V}}\left(\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}}\right)$ varies slowly and remains quite stable over time until the last few event times, so that computation of $\hat{\boldsymbol{V}}\left(\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}}\right)$ at each observed event time is not necessary. As a result, the inverse of $\hat{\boldsymbol{V}}\left(\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}}\right)$ can be approximated by the inverse of the observed information matrix, given by

$$\hat{\tilde{\boldsymbol{r}}}_i^{*,\text{Schoen}} = \boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right)\hat{\tilde{\boldsymbol{r}}}_i^{\text{Schoen}}. \tag{6.57}$$

As a result, the scaled, or the weighted Schoenfeld residuals can be conveniently obtained from the analytic results of the Cox model. Because its plot against the rank order of survival times can display deviations from proportional hazards, the scaled Schoenfeld residuals can be used to assess the time trend of proportionality for a specific covariate, as will be discussed further in next section.

6.2.3 Martingale residuals

The specification of martingale residuals is based on the counting process formulation, so I use the terminology used in Section 6.1. In the system of counting processes, the intensity function is given by

$$\Pr\{\mathrm{dN}_i(t)\} = \mathrm{E}\{\mathrm{dN}_i(t) | \mathcal{F}_{t-}\} = \lambda(t)\mathrm{d}t, \tag{6.58}$$

where $\mathcal{F}_{t-}$ is the process history prior to t about survival, censoring, and covariates. The integrated intensity processes, denoted by $\Lambda(t)$, is given by

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u, \quad t \geq 0. \tag{6.59}$$

$\Lambda(t)$ is equivalent to the cumulative hazard function for a single event per individual.

The intensity function or the hazard rate for the $N_i(t)$ counting process can be more conveniently expressed in terms of $\Lambda(t)$, as also indicated in Section 6.1:

$$Y_i(t)d\Lambda\{t,\mathbf{Z}_i(t)\} = Y_i(t)\exp[\mathbf{Z}_i'(t)\boldsymbol{\beta}]d\Lambda_0(t), \tag{6.60}$$

where Λ_0 is the baseline integrated intensity function or, for a single event per individual, the baseline cumulative hazard function.

The observed event count $N_i(t)$ can be written as the summation of a compensator and a martingale process:

$$N_i(t) = \int_0^t Y_i(u)\exp\left[\mathbf{Z}_i'(u)\hat{\boldsymbol{\beta}}\right]d\Lambda_0(u) + M_i(t), \tag{6.61}$$

where, as defined earlier, $M_i(t)$ is a martingale given the σ-algebra $\mathcal{F}_{t-1}$:

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\exp\left[\mathbf{Z}_i'(u)\hat{\boldsymbol{\beta}}\right]d\hat{\Lambda}_0(u), \tag{6.62}$$

where $\hat{M}_i(t)$ is estimated as the difference over (0, t) between the observed number and the expected number of events from a regression model for individual i. As defined, such martingale residuals have the property that $\sum \hat{M}_i(t) = 0$ for any t because

$$\begin{aligned}\sum_{i=1}^n \hat{M}_i(t) &= \sum_{i=1}^n \left\{N_i(t) - \int_0^t Y_i(u)\exp\left[\mathbf{Z}_i'(u)\hat{\boldsymbol{\beta}}\right]d\hat{\Lambda}_0(u)\right\} \\ &= \sum_{i=1}^n \left(\int_0^t dN(u) - \int_0^t Y_i(u)\exp\left[\mathbf{Z}_i'(u)\hat{\boldsymbol{\beta}}\right]\left\{\frac{\sum_{i=1}^n dN(u)}{\sum_{l=1}^n Y_l(u)\exp\left[\mathbf{Z}_l'(u)\hat{\boldsymbol{\beta}}\right]}\right\}d\hat{\Lambda}_0(u)\right) \\ &= 0.\end{aligned}$$

Also, as two martingales $\langle \hat{M}_i, \hat{M}_{i'} \rangle = 0$, then $E(\hat{M}_i) = \text{cov}(\hat{M}_i, \hat{M}_{i'}) = 0$ asymptotically (Andersen *et al.*, 1993; Barlow and Prentice, 1988; Therneau, Grambsch, and Fleming, 1990). Given these properties, in many ways the martingale residuals resemble the ordinary residuals of linear regression models. Because $N_i(t)$ is a binary count for a single event per individual, however, the distribution of the martingale residuals is not symmetric. Because $N_i(t)$ is 0 or 1 for a single event, for the Cox model without time-dependent covariates the martingale residual reduces to

$$\hat{M}_i(t) = \delta_i - \hat{\Lambda}_0(t_i)\exp\left(\mathbf{Z}_i'\hat{\boldsymbol{\beta}}\right), \tag{6.63}$$

where $\hat{\Lambda}_0(t_i)$ is the estimated baseline cumulative hazard function at the observed survival time t_i.

In logic, Equation (6.63) defines a martingale when the Cox model is correctly specified, so the adequacy of the Cox model can be graphically assessed by the martingale residuals in the presence of censoring. In reality, however, when the survival time is an actual survival time ($\delta_i = 1$), the martingale residual is positive; when it is a censored time ($\delta_i = 0$), the martingale residual is negative. Consequently, individuals with early failure times tend to have positive residuals along the life course because they experience the event too early. Likewise, those with large values of failure times yield more negative scores because the event occurs too late. For survival data with heavy censoring, the majority of martingale residuals are negative, though scattered about zero. Furthermore, in the martingale residuals the Type I right censored observations would be clustered below zero at the ending limit of a given observation interval. Such loss of balance led to the development of more symmetrically distributed martingale transforms.

6.2.4 Score residuals

The martingale residual provides a natural specification of deviations in the counting processes formulation. When Equation (6.63) is applied to the Cox model, the Breslow (1974) method can be used to estimate the baseline cumulative hazard function $\hat{\Lambda}_0(t_i)$. The advantage of using the martingale residuals is that departures from the expected values can be evaluated at every observed survival time for the overall fit of the Cox model.

Therneau, Grambsch, and Fleming (1990) advance the martingale residuals by considering a Kolmogorov-type test based on the cumulative sum of residuals. For the Cox model in which $\hat{\Lambda}_0(t_i)$ is unspecified, the partial derivative of the log partial likelihood L_p with respect to a single covariate β_m can be written as

$$\left[\frac{\partial \log L_p}{\partial \beta_m}\right]_{\boldsymbol{\beta}=\boldsymbol{b}} = \sum_{i=1}^{n} \int_0^\infty \left[\mathrm{Z}_{im}(u) - \bar{\mathrm{Z}}_m(b,u)\right] \mathrm{dN}_i(u), \tag{6.64}$$

where

$$\bar{\mathrm{Z}}_m(b,u) = \frac{\sum_{i=1}^{n} \mathrm{Y}_i(u)\exp\left[\mathbf{Z}_i'(u)\hat{\boldsymbol{\beta}}\right]\mathrm{Z}_{im}(u)}{\sum_{i=1}^{n} \mathrm{Y}_i(u)\exp\left[\mathbf{Z}_i'(u)\hat{\boldsymbol{\beta}}\right]}. \tag{6.65}$$

Equation (6.65) is the weighted mean of the covariate over the risk set at time u, as indicated in Chapter 5 and in Section 6.1.

Given the specification of martingale residuals, Equation (6.64) can be rewritten as

$$\begin{aligned}\left[\frac{\partial \log L_p}{\partial \beta_m}\right]_{\boldsymbol{\beta}=\boldsymbol{b}} &= \sum \int_0^\infty \left[\mathrm{Z}_{im}(u) - \bar{\mathrm{Z}}_m(b,u)\right]\mathrm{dM}_i(u) \\ &\equiv \sum \mathrm{L}_{im}(b,\infty).\end{aligned} \tag{6.66}$$

The first equality in Equation (6.66) follows the specification of the score function when evaluated at $\boldsymbol{\beta} = \boldsymbol{b}$. As a result, the score process for individual i at time t can be formally defined by the vector

$$\mathbf{L}_i(\boldsymbol{\beta},t) = \int_0^t \left[\mathbf{Z}_i(u) - \bar{\mathbf{Z}}(\boldsymbol{\beta},u)\right]\mathrm{dM}_i(\boldsymbol{\beta},u). \tag{6.67}$$

The vector $\hat{\mathbf{L}}_i \equiv \mathbf{L}_i\left(\hat{\boldsymbol{\beta}},\infty\right)$, the first partial derivatives of the log partial likelihood function, defines the score residual for individual i.

In particular, assuming covariate x_m to be time independent, the score residual at a specific survival time is

$$\begin{aligned} \mathrm{L}_{im}(\boldsymbol{\beta},t) &= \delta_i \mathrm{Y}_i(t)\left[\mathrm{Z}_{im} - \bar{\mathrm{Z}}_m(t_i)\right] \\ &\quad - \sum_{t_b \le t}\left[\mathrm{Z}_{im} - \bar{\mathrm{Z}}_m(t_b)\right]\mathrm{Y}_i(t_b)\exp\left(\mathbf{Z}_i'\hat{\boldsymbol{\beta}}\right)\left[\hat{\Lambda}_0(t_b) - \hat{\Lambda}_0(t_{b-1})\right], \end{aligned} \tag{6.68}$$

where $t_b = t_0 < t_1 < \cdots < t_d$ are the ordered actual survival times. Consequently, for individual i ($i = 1, 2, \cdots, n$), the score process for the mth covariate is specified.

The score residuals are useful to evaluate the influence of each individual on each individual parameter estimate. The overall score residual for an observation can also be obtained by summing the component residuals within the observation. Technically, the score residuals are simply the modification of the Schoenfeld residuals by applying the martingale theory. As the development of this method is based on martingales, the score residuals are generally viewed as a martingale-based transform. The reader interested in learning more details of this residual type might want to read Barlow and Prentice (1988).

6.2.5 Deviance residuals

Although the martingale residuals are useful for providing information on model adequacy, one inherent limitation in the specification is the marked skewness of random departures. Given the 0–1 observed value for the outcome variable in the Cox model, the martingale residual does not follow the theory of normality, therefore causing difficulty in analyzing its impact. As indicated above, this is a common problem of specifying random disturbances encountered in all generalized linear models. One popular perspective to handle skewness in qualitative outcome data is to transform a nonnormal distribution to a distribution 'as normal as possible' (Box and Cox, 1964; McCullagh and Nelder, 1989; Sakia, 1992).

Therneau, Grambsch, and Fleming (1990) use the *deviance score*, a statistic widely used in econometrics, to measure residuals. Specifically, the deviance is defined as

$$\tilde{D} = 2\left[\log L(\text{saturated}) - \log L\left(\hat{\boldsymbol{\beta}}\right)\right]. \tag{6.69}$$

where the saturated model indicates a perfect regression model that has no random errors (each individual has his or her own $\hat{\boldsymbol{\beta}}$ vector). While $\tilde{D}$ is distributed as χ^2 with $(n - M)$ degrees of freedom, its square root approximates a normal distribution, thereby possessing potentials to derive a more efficient residual type.

The nuisance parameter $h_0(t)$ in the Cox model is assumed to be constant across the saturated model and the reduced proportional hazard model with $\hat{\boldsymbol{\beta}}$. Let $\tilde{\boldsymbol{g}}_i$ be the individual-specific estimate of $\boldsymbol{\beta}$. The deviance with $\mathbf{Z}_i$ (including time-independent covariates only) and known Λ_0 is extended as

$$\begin{aligned} \tilde{D} = 2\sup\sum\Big\{ &\int\left[\log\exp(\mathbf{Z}_i'\tilde{\boldsymbol{g}}_i) - \log\exp\left(\mathbf{Z}_i'\hat{\boldsymbol{\beta}}\right)\right]d\mathrm{N}_i(u) \\ &-\int \mathrm{Y}_i(u)\left[\exp(\mathbf{Z}_i'\tilde{\boldsymbol{g}}_i) - \exp\left(\mathbf{Z}_i'\hat{\boldsymbol{\beta}}\right)\right]d\Lambda_0(u)\Big\}. \end{aligned} \tag{6.70}$$

After some transformation, Equation (6.70) leads to the estimator of the deviance residual $\tilde{D}_i$ as a transform of the martingale residual:

$$\tilde{D}_i = \operatorname{sign}\left(\hat{\mathrm{M}}_i\right)\sqrt{2\left\{-\hat{\mathrm{M}}_i - \mathrm{N}_i(\infty)\log\left[\frac{\mathrm{N}_i(\infty)-\hat{\mathrm{M}}_i}{\mathrm{N}_i(\infty)}\right]\right\}}, \tag{6.71}$$

where sign(·) is the sign function. Mathematically, the use of the sign function guarantees $\tilde{D}_i$ to take the same sign as the martingale residual, the square root shrinks large martingale residuals, and the logarithmic transformation makes martingale residuals close to 0 (Fleming and Harrington, 1991). Consequently, the deviance residuals, while taking the same signs as martingale residuals, are more symmetrically distributed around 0 than the martingale residuals, providing tremendous convenience for assessing the model adequacy. In the presence of light censoring, the deviance scores against the linear predictor $\mathbf{Z}_i'\hat{\boldsymbol{\beta}}$ should approximate a normal distribution.

For the Cox model, the deviance residual reduces to

$$\tilde{D}_i^{\mathrm{Cox}} = \operatorname{sign}\left(\hat{\mathrm{M}}_i\right)\sqrt{2\left[-\hat{\mathrm{M}}_i - \delta_i\log\left(\delta_i - \hat{\mathrm{M}}_i\right)\right]}. \tag{6.72}$$

Equation (6.72) indicates that the deviance residual is a transform of the martingale residuals; therefore it is also classified as a martingale-based statistic.

6.2.6 Illustration: Residual analysis on the Cox model of smoking cigarettes and the mortality of older Americans

In the present illustration, I calculate residuals from results of the Cox model on the association between smoking cigarettes and the mortality of older Americans, using the same survival information as presented in Subsection 5.7.3. The purpose of this diagnostic analysis is to assess departures of model estimates from the observed survival data, thereby generating information on the adequacy of the Cox model. Specifically, I obtain four types of residuals from analytic results of the Cox model on smoking cigarettes and mortality – the martingale, the deviance, the score, and the Schoenfeld residuals. In the Cox model, the four covariates – smoking cigarettes, age, gender, and educational attainment – are measured as centered variables, with names given as, respectively, ‘Smoking_mean,’ ‘Age_mean,’ ‘Female_mean,’ and ‘Educ_mean.’

The PROC PHREG procedure is applied to generate those residuals. First, I request SAS to include the four residuals in the OUTPUT statement by specifying an OUT = OUT_RES temporary dataset. Then I ask SAS to plot those residual scores by using the PROC SGPLOT procedure. Below is the SAS program for the work.

SAS Program 6.1:

```
......

proc phreg data = new noprint ;
  model duration*Status(0) = smoking_mean age_mean female_mean educ_mean ;
  output out = out_res XBETA = XB RESMART = Mart RESDEV = Dev RESSCO =
   Scosmoking_mean Scoage_mean Scofemale_mean Scoeduc_mean RESSCH =
   Schsmoking_mean Schage_mean Schfemale_mean Scheduc_mean;
run;

Title "Figure 6.1a. Martingale residuals";
proc sgplot data = out_res ;
  yaxis grid;
  refline 0 / axis = y ;
  scatter y = Mart x = xb ;
run;

Title "Figure 6.1b. Deviance residuals";
proc sgplot data = out_res ;
  yaxis grid;
  refline 0 / axis = y ;
  scatter y = Dev x = xb ;
run;

Title "Figure 6.1c. Score residuals against Age_mean";
proc sgplot data = out_res ;
  yaxis grid;
  refline 0 / axis = y ;
  scatter y = Scoage_mean x = duration ;
run;

Title "Figure 6.1d. Schonfeld residuals against Age_mean";
proc sgplot data = out_res ;
  yaxis grid;
  refline 0 / axis = y ;
  scatter y = Schage_mean x = duration ;
run;
```

SAS Program 6.1 specifies a classical PROC PHREG model by adding an OUTPUT statement. The keywords XBETA, RESMART, RESDEV, RESSCO, and RESSCH specify the linear predictor, the martingale, the deviance, the score, and the Schoenfeld residuals, respectively. The first three new variables are identified as XB, MART, and DEV. As the score and the Schoenfeld residuals are designed to display a series of individual-specific residual values corresponding to each covariate, the keyword RESSCO or RESSCH is followed by the variable names of the four covariates considered in the Cox model. Given the nature of the illustration, in this example I only plot the score and the Schoenfeld residuals corresponding to the covariate Age_mean. As residuals on the entire model, the martingale and the deviance residuals are plotted against the linear predictor XB for displaying their distributions. The score and the Schoenfeld residuals, on the other hand, are plotted against the rank order of survival times to demonstrate the time trend of proportionality for Age_mean. These four residuals are then output into the temporary data file OUT_RES. The Breslow approximation method, as default in SAS, is used to handle survival data with tied survival times. The four residual plots are generated by four PROC SGPLOT steps with each defining a specific type of residual, as shown in Figure 6.1, which includes four residual plots against the linear predictor $\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ for the martingale and the deviance residuals or against the rank order of survival times for the score and the Schoenfeld residuals, from which the pattern of residuals can be evaluated. The first plot, Figure 6.1a, shows the martingale residuals. The linear predictor has a range between −3 and 2 since all covariates in the Cox model are centered about means. As expected, the martingale residuals are skewed given the single event setting and heavy right censoring in the dataset. Nevertheless, these residuals are concentrated around zero, highlighting a fairly decent fit of the model. In this plot, an outlier can be easily identified with a value below −3.

Figure 6.1b demonstrates deviance residuals against the linear predictor. Here, the logarithms, as discussed in Subsection 6.2.4, derive impact on narrowing the range of residuals, thereby making the plot look more balanced. Compared to the martingale residual, the deviance residuals are more intensely scattered around zero, with the outlier, identified in Figure 6.1a, nearly vanishing. Because of heavy censoring, however, a large quantity of residuals are clustered near zero, thus disturbing the expected normal approximation.

The score residuals corresponding to the covariate Age_mean, as another martingale transform and displayed in Figure 6.1c, look randomly distributed against the rank order of survival times, with the vast majority of the scores narrowly scattered around zero. Most significantly, no distinct outliers can be identified in this plot.

The last plot, Figure 6.1d, presents the Schoenfeld residuals corresponding to the covariate Age_mean. Obviously, these age-specific residuals are independent of survival times as they are scattered randomly around zero without displaying any systematic pattern. As a result, the effect of age is not shown to depart from the proportional hazards assumption in the Cox model. Because the Schoenfeld residuals are not defined for censored observations, Figure 6.1d only plots residuals for those who died in the observation period. Therefore, as a modification of the Schoenfeld residual, the score residuals are obviously preferable to show a complete, more refined set of residuals.

From the above evaluation of four residual plots, it may be appropriate to conclude that there is no evidence of model misspecification on individual observations, both overall and with specific regard to age. All four residual types behave as expected, without revealing a distinct trend of model inadequacy. Additionally, only one distinct outlier can be identified. Therefore, I have reason to believe that the underlying Cox model is adequately applied for analyzing the association between smoking cigarettes and the mortality of older Americans. Even so, in the presence of heavy censoring, it seems very difficult to find a highly efficient residual transform that can display an explicit and normally distributed random process as generated in linear regression models.

6.3 Assessment of proportional hazards assumption

In Chapter 5, I discussed the use of some simple graphical checks on the proportional hazards assumption. Specifically, if two or more stratum-specific log–log survival curves, with all covariates set at zero, are approximately parallel, it can be inferred that the baseline hazard rates across strata tend to be proportional. If not, the effects of the stratification factor on the hazard function are probably not multiplicative over time, thereby displaying violation of the proportionality hypothesis. Such graphical checks, however, do not provide sufficient information on nonproportionality because they are not associated with an explicit and unambiguous statistical criterion on observed deviations from a proportional effect. Some statisticians have developed more refined techniques for checking the proportionality assumption of the Cox model, based on various methodological perspectives (Andersen, 1982; Arjas, 1988; Gill and Schumacher, 1987; Lagakos and Schoenfeld, 1984; Lin, Wei, and Ying, 1993; Storer and Crowley, 1985; Struthers and Kalbfleisch, 1986).

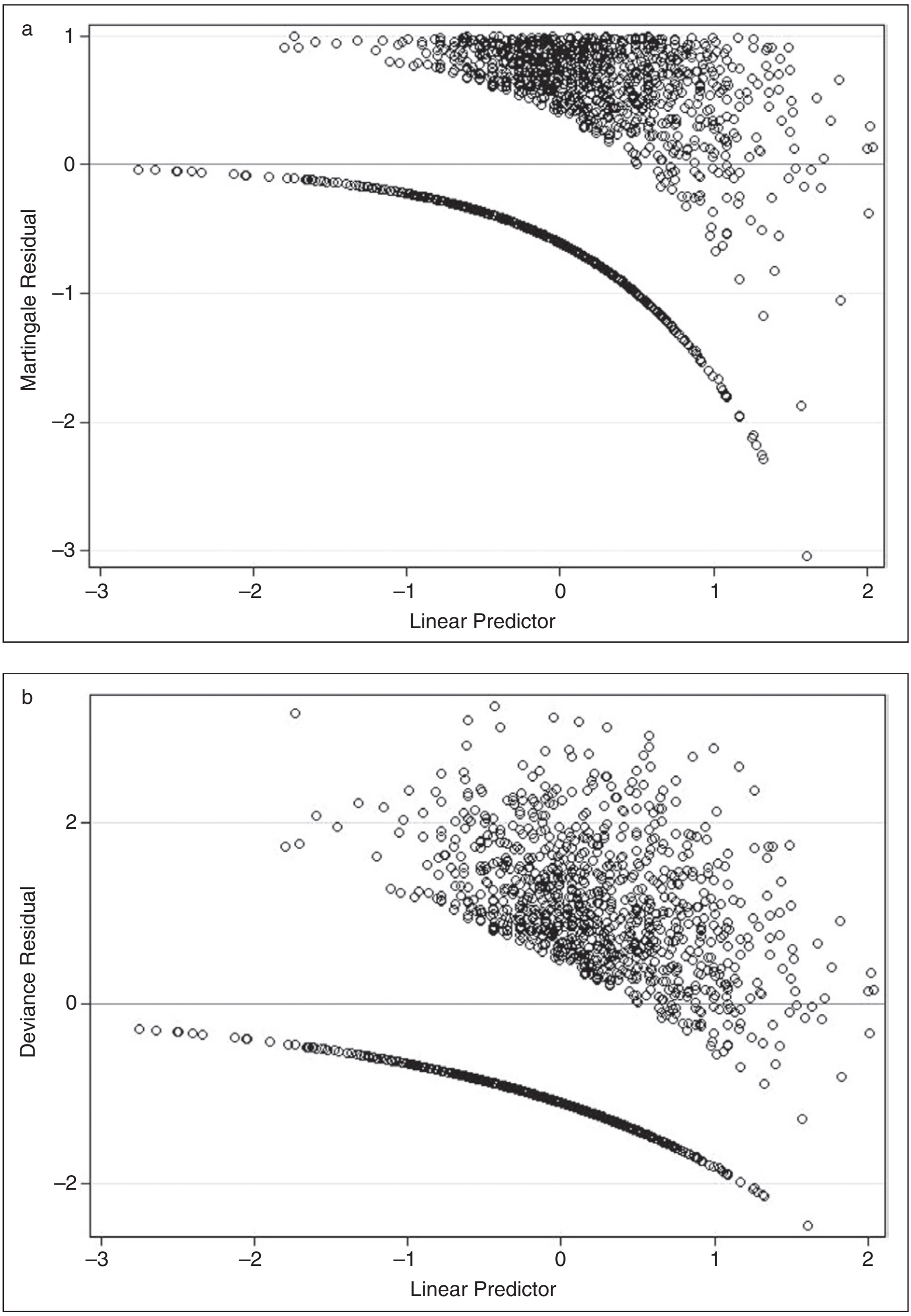


Figure 6.1 a Martingale residuals. b Deviance residuals.

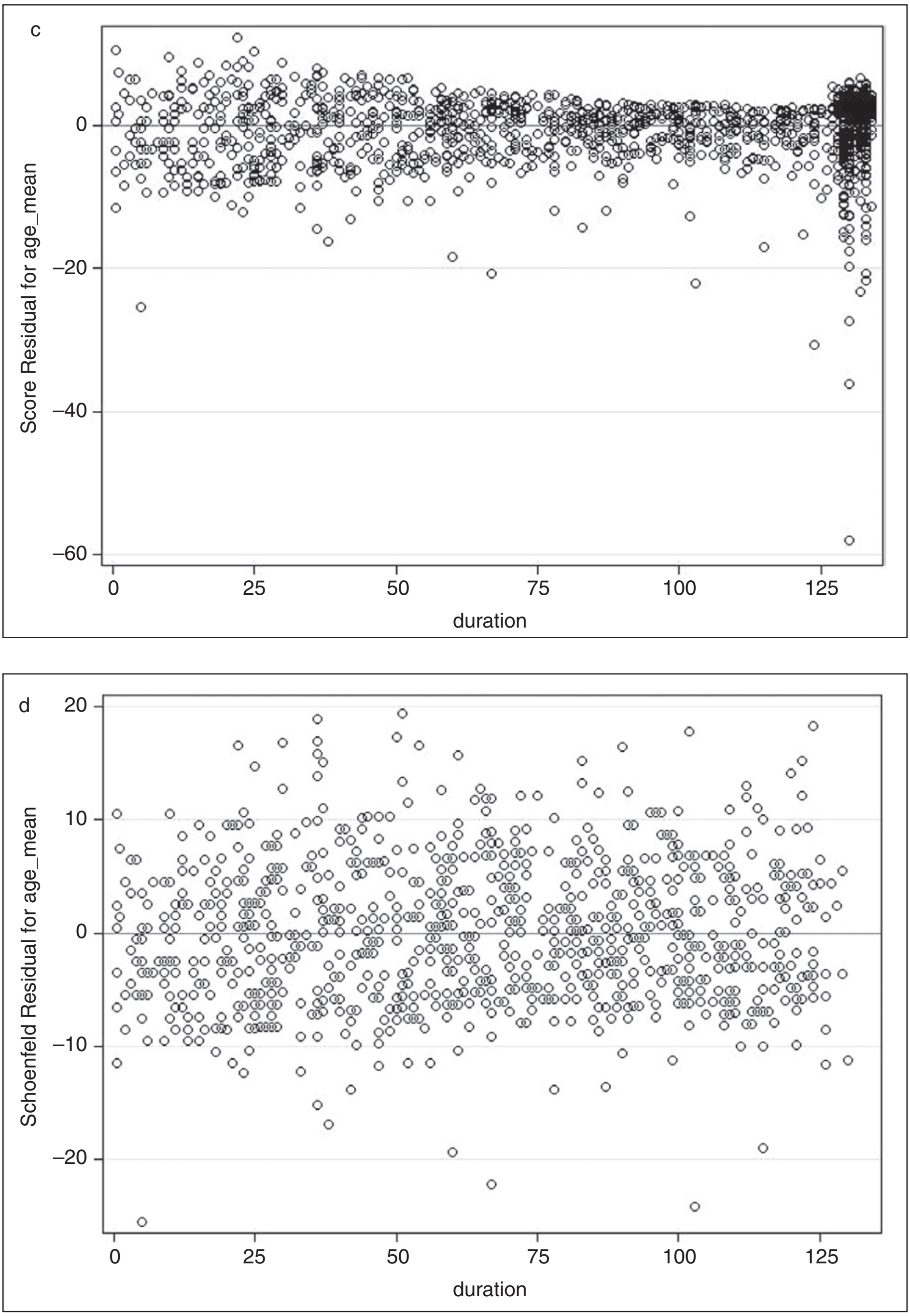


Figure 6.1 (Continued) c Score residuals against Age_mean. d Schonfeld residuals against Age_mean.

In this section, I introduce five methods in this regard: (1) checking proportionality by adding a time-dependent variable, (2) the Andersen plots, (3) checking proportionality with the scaled Schoenfeld residuals, (4) the Arjas plots, and (5) checking the proportional hazards assumption with cumulative sums of martingale-based residuals. Lastly, I provide an empirical illustration on testing the proportionality assumption in the Cox model, using survival data of older Americans.

6.3.1 Checking proportionality by adding a time-dependent variable

The Cox model is based on the assumption that the hazard ratio of a given covariate x_m is independent of time. This assumption is violated if the hazard ratio changes significantly over time, thus highlighting the absence of a constant multiplicative effect. From this logic, a straightforward approach to check the validity of the proportionality assumption is to compare two Cox models, one specifying x_m as a single covariate with time-independent effects and one adding a time-dependent interaction term ($x_m \times t$) with the assumption that the effect of x_m varies over time. While the first regression is a standard Cox model, the second model can be written by

$$h(t|x_m, \boldsymbol{x}_r) = h_0(t)\exp[x_m\beta_1 + (x_m \times t)\beta_2 + \boldsymbol{x}_r'\boldsymbol{\beta}_r], \tag{6.73}$$

where β_2 measures the time-dependent effect of x_m and $\boldsymbol{x}_r$ is the vector of other covariates, used as controls.

In practice, the estimation of Equation (6.73) may encounter some specification problems, so some functional adjustments are needed. First, the distribution of t may not necessarily be linearly associated with log hazards, thus causing some numeric instability. As widely applied in survival and longitudinal data analyses, using $\log t$ to replace t can considerably improve stability in the estimation process (Collett, 2003). Second, the two covariates, x_m and $x_m \times t$, come from the same source of variability; therefore they tend to be highly correlated, thus affecting the efficiency of the model fit. Using the centered log time for t, defined as $[\log t - \text{mean}(\log t)]$, can substantially reduce multicollinearity in longitudinal data (Hedeker and Gibbons, 2006). Accordingly, Equation (6.73) can be adjusted by

$$h(t|x_m, \boldsymbol{x}_r) = h_0(t)\exp\{x_m\beta_1 + x_m[\log t - \text{mean}(\log t)]\beta_2 + \boldsymbol{x}_r'\boldsymbol{\beta}_r\}. \tag{6.74}$$

By performing a significance test on the time-dependent component of x_m, the validity of the proportionality assumption on x_m can be statistically assessed. In particular, the Wald test can be used for testing the null hypothesis H_0: $\beta_2 = 0$, given a chi-square distribution. If β_2 is not statistically significant, the addition of the time-dependent component $\{x_m[\log t - \text{mean}(\log t)]\}$ should be viewed as redundant. In this situation, as the multiplicative effect of x_m does not depend on time, the 1-unit hazard ratio of x_m is $\exp(\beta_1)$.

If it is statistically significant, however, the value of β_2 reflects the extent to which the hazard ratio of x_m increases ($\beta_2 > 0$) or decreases ($\beta_2 < 0$) over time. Then, the proportional hazards assumption on the effects of x_m should be considered violated. The 1-unit hazard ratio of x_m then becomes

$$\begin{aligned} HR(x_m|t) &= \frac{h_0(t)\exp\left\{(x_{m0}+1)\beta_1 + (x_{m0}+1)[\log t - \text{mean }(\log t)]\beta_2 + \boldsymbol{x}_r'\hat{\boldsymbol{\beta}}_r\right\}}{h_0(t)\exp\left\{x_{m0}\hat{\beta}_1 + x_{m0}[\log t - \text{mean}(\log t)]\hat{\beta}_2 + \boldsymbol{x}_r'\hat{\boldsymbol{\beta}}_r\right\}} \\ &= \frac{\exp\left\{(x_{m0}+1)\hat{\beta}_1 + (x_{m0}+1)[\log t - \text{mean }(\log t)]\hat{\beta}_2\right\}}{\exp\left\{x_{m0}\hat{\beta}_1 + x_{m0}[\log t - \text{mean }(\log t)]\hat{\beta}_2\right\}} \\ &= \exp\left[\hat{\beta}_1 - \text{mean }(\log t)\hat{\beta}_2\right] t^{\hat{\beta}_2}, \end{aligned} \tag{6.75}$$

where $\exp\left(\hat{\beta}_1\right)$ is the hazard ratio at the mean survival time because $\log t$ is centered. At other time points, the hazard ratio varies significantly if the value of β_2 is sizable, thereby highlighting misspecification of the proportionality assumption in the Cox model.

More formally, statistical testing on the proportional hazards assumption can be executed by using the partial likelihood ratio test. With respect to the above-mentioned two Cox models, the score of the log partial likelihood ratio test is

$$G_{\beta_2} = -2\times\left[\log L_p\left(\hat{\beta}_1, \hat{\beta}_2, \hat{\boldsymbol{\beta}}_r; t, x_m, \boldsymbol{x}_r\right) - \log L_p\left(\hat{\beta}_1, \hat{\boldsymbol{\beta}}_r; x_m, \boldsymbol{x}_r\right)\right], \tag{6.76}$$

where G_{β_2} is the partial likelihood ratio statistic for the time-dependent effect of x_m, distributed as χ^2 with one degree of freedom. In the bracket on the right of Equation (6.76), the first term is the log partial likelihood ratio statistic for the time-dependent model, whereas the second is for the standard Cox model. If $G_{\beta_2} < \chi^2_{(1-\alpha;1)}$, the model is not statistically improved by adding the time-dependent component $\{x_m[\log t - \text{mean}(\log t)]\}$; hence the proportional hazards assumption is not violated. If $G_{\beta_2} > \chi^2_{(1-\alpha;1)}$, the addition of the time-dependent variable significantly increases the quality of the overall fit, so the null hypothesis on the proportional effects of covariate x_m should be rejected.

On most occasions, I expect the above two test statistics, the Wald and the partial likelihood ratio tests, to yield the same result on the significance of the time-dependent effect. Specifically, the Wald statistic checks the statistical significance of a single parameter estimate, whereas the partial likelihood ratio test is based on the entire model fit. In certain circumstances, however, the two tests can generate statistics that designate different test results. These test statistics, in turn, would provide ambiguous information about statistical significance of nonproportionality in the Cox model. For example, one statistic is associated with a p-value greater than α, whereas the other is less than α. If such a situation happens, the results generated from the partial likelihood ratio test are recommended as the final criterion.

When applying this numeric method, however, we must consider the potential problems in the specification of a time-dependent covariate. As stated in Chapter 5, without extensive knowledge about the mechanisms involved in a time-dependent process, the estimated effect of a time-dependent covariate can imply other interrelationships, thereby leading to misleading conclusions. Additionally, a sizable effect on the hazard ratio may not necessarily translate into strong effects on the hazard rate itself and, consequently, some graphical checks are useful to aid in the interpretation of numeric results.

6.3.2 The Andersen plots for checking proportionality

The Andersen (1982) plots are somewhat similar to the graphical checks described in Chapter 5 (Section 5.5). Suppose that $M + 1$ covariates are considered and that $\boldsymbol{x} = (x_1, \ldots, x_M)'$ is included in a proportional hazard model. The covariate x_{M+1} is the independent variable under assessment concerning whether its effects on the hazard rate are proportional over time. Given such, the null hypothesis for a statistical test is

$$\mathrm{H}_0 : h_0(t; x_{M+1}) \exp(\boldsymbol{x}'\boldsymbol{\beta}) = h_0(t) \exp(\boldsymbol{x}'\boldsymbol{\beta} + x_{M+1}\beta_{M+1}). \tag{6.77}$$

As discussed in Chapter 5, checking the proportionality assumption for a single covariate can be performed by stratifying on this covariate, fitting a proportional hazard model for each stratum, and then combining them by assuming a common coefficient vector $\boldsymbol{\beta}$. In stratum k, where $k = 1, \ldots, K$, the proportional hazard model with covariate vector $\boldsymbol{x}$ is

$$h_k(t; \boldsymbol{x}_k) = h_{0k}(t) \exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{6.78}$$

where h_{0k} is the baseline hazard function for stratum k. Graphical checks can be made to evaluate whether the K baseline hazard functions are proportional to each other. As the log–log survival function is actually the log transformation of the cumulative hazard rate, it is informative to compare stratum-specific log–log survival curves with all covariates set at zero. In Chapter 5, I provided an empirical illustration on how to perform this method.

Based on the graphical check on a plot of $\log \hat{H}_k(t)$ versus t, Andersen (1982) proposes a unique graphical checking method for the proportionality assumption in the Cox model. Let $K = 2$ be a binary variable for variable x_{M+1}, observed at time points $t_1, \ldots, t_n$. Then, if the proportional hazards assumption holds, a plot of $\hat{H}_2(t_1), \ldots, \hat{H}_2(t_n)$ versus $\hat{H}_1(t_1), \ldots, \hat{H}_1(t_n)$ should be approximately a straight line through the origin. Accordingly, the slope of the line should approximate the regression coefficient of x_{M+1} if the Cox model is valid. In contrast, considerable deviations of this plot from a straight line would suggest that the proportionality assumption may be violated. For $K > 2$ in x_{M+1}, each pair of $\hat{H}_k(t_1), \ldots, \hat{H}_k(t_n)$ versus $\hat{H}_1(t_1), \ldots, \hat{H}_1(t_n)$, where $k \neq 1$ may be plotted to assess the adequacy of the proportional hazards assumption on this added covariate. Such graphical checks are known as the Andersen plots. As Gill and Schumacher (1987) summarize, the shape of such a plot demonstrates the direction of deviations from the proportionality assumption. If the hazard ratio increases over time, the Andersen plot should appear convex; if the hazard ratio decreases over time, the plot should appear concave.

Andersen (1982) also attempts to develop some numeric tests to support the results of the graphical checks, using the stratification perspective. Specifically, t is divided into a number of time intervals and the stratum-specific baseline hazard function is assumed to be constant in each of those intervals. Given those additional assumptions, statistical tests on the proportionality assumption in the Cox model become very tedious and can easily come across some specification problems. Therefore, I do not present this numeric method in this text. The interested reader on this numeric method is referred to Andersen's (1982) article.

6.3.3 Checking proportionality with scaled Schoenfeld residuals

Grambsch and Therneau (1994) proposed the use of scaled Schoenfeld residuals to check the proportional hazards assumption in the Cox model. First, for a single covariate x_m, they expand the proportional hazards by adding a time-varying coefficient, given by

$$\beta_m(t) \equiv \beta_m + \beta_m g_m(t), \tag{6.79}$$

where $g_m(t)$ is defined as a predictable process, which can take various functional forms. Accordingly, the ith scaled Schoenfeld residual corresponding to x_m, represented by Equation (6.56), is also expanded about $\beta_m(t_i) = \beta_m$, given by

$$\tilde{r}_{im}^{*,\text{Schoen}} = \frac{\tilde{r}_{im}^{\text{Schoen}}(\beta_m)}{V(\beta_m, t_i)}. \tag{6.80}$$

Therefore, within this context, the scaled Schoenfeld residual is evaluated with respect to $\beta_m(t_i)$ rather than to β_m itself.

Assuming g to vary about 0 and $\boldsymbol{G}_i = \boldsymbol{G}(t_i)$ to be an $M \times M$ diagonal matrix with the mmth element being $g_m(t_i)$, the expected value of the Schoenfeld residual can be expressed by

$$\mathrm{E}\left[\tilde{\boldsymbol{r}}_i^{\text{Schoen}}(\boldsymbol{\beta})\right] \approx V(\boldsymbol{\beta}, t_i)\boldsymbol{G}(t_i)\bar{\boldsymbol{\theta}}. \tag{6.81}$$

Because the scaled Schoenfeld residual is defined as $\tilde{\boldsymbol{r}}_i^{*,\text{Schoen}} = \tilde{\boldsymbol{r}}_i^{*,\text{Schoen}}(\boldsymbol{\beta}) = V^{-1}(\boldsymbol{\beta}, t_i)\tilde{\boldsymbol{r}}_i^{\text{Schoen}}(\boldsymbol{\beta})$, the expected value of the scaled Schoenfeld residual is

$$\mathrm{E}\left(\tilde{\boldsymbol{r}}_i^{*,\text{Schoen}}\right) \approx \boldsymbol{G}_i\bar{\boldsymbol{\theta}}, \tag{6.82}$$

with variance

$$V\left(\tilde{\boldsymbol{r}}_i^{*,\text{Schoen}}\right) = V^{-1}(\boldsymbol{\beta}, t_i). \tag{6.83}$$

The score test can be performed on the null hypothesis that H_0: $\beta_m(t_i) = \beta_m$. As $\bar{\theta}_m$ can be viewed as the time-dependent component of $\beta_m(t)$, given the functional form of $g_m(t)$, Equation (6.82) can be more conventionally written as

$$\mathrm{E}\left(\tilde{r}_{im}^{*,\text{Schoen}}\right) \approx \beta_m(t_i) - \beta_m, \tag{6.84}$$

where $\beta_m(t_i)$ is the regression coefficient of covariate x_m at observed time t_i and $\hat{\beta}_m$ is the estimate of β_m from the Cox model.

Grambsch and Therneau (1994) suggest that as the variance matrix of $\boldsymbol{x}(t)$ is stable over time (also discussed in Section 6.2), a smoothed scatter plot of the values of $\tilde{r}_{im}^{*,\text{Schoen}} + \beta_m$ against t_i can be used to track the nonproportionality of $\beta_m(t)$. Specifically, a horizontal line of $\tilde{r}_{im}^{*,\text{Schoen}} + \beta_m$ versus t_i suggests the hazard ratio of x_m to be constant, in turn reflecting the validity of the proportional hazards assumption. By contrast, if the line of $\tilde{r}_{im}^{*,\text{Schoen}} + \beta_m$ versus t_i deviates considerably from horizontality, the Cox proportional hazard model may be misspecified. As also suggested by Grambsch and Therneau (1994), a smoothed line can be drawn to supplement the interpretation of analytic results.

This graphical check has some advantages for use. It is intimately linked to the scaled Schoenfeld residual, so that computation of this graphical check is handy and the resulting plot is easy to comprehend. In particular, a plot can be readily drawn with the estimated regression coefficient of a given covariate plus the scaled Schoenfeld residuals. As a coarse graphical approach, however, this test does not make distinct improvements in checking the proportionality assumption compared to other traditional graphical checking methods. Additionally, it is not associated with a numeric statistic that can be used to perform more rigorous statistical testing on the proportionality hypothesis.

6.3.4 The Arjas plots

Another popular graphical method for checking the proportional hazards assumption in the Cox model is using the Arjas (1988) plots. Specifically, the Arjas plots are designed to make direct comparisons between observed and estimated event frequencies without adding a time-dependent variable. Therefore, this method is not based on the estimation of alternative models and only involves parameter estimates already derived from the partial likelihood procedure.

According to Arjas (1988), the application of the stratified Cox model is subject to two types of defects: (1) an influential covariate may be deleted from the model (this defect has been discussed in Section 5.5 of this book) and (2) the stratified Cox model is based on the assumption of a common baseline hazard for all individuals, so that the individuals are stratified according to the baseline hazard. These two defects can seriously influence the efficiency of the Cox model, thus making it difficult to perform a graphical check correctly on the validity of the proportionality hypothesis. Accordingly, he proposes to test the proportionality assumption directly from the proportional hazard model including all $(M + 1)$ covariates.

Practically, deriving the Arjas plots can be performed by taking the following steps. First, divide n individuals into K strata of the $(M + 1)$th covariate according to the research interest of a particular study or previous findings. If the $(M + 1)$th covariate is a continuous variable, classify the sample respondents into a few categories according to an existing theory or results from a previous empirical analysis. Second, calculate the estimated cumulative hazard rate at each observed survival time for each stratum using the parameter estimates obtained from the Cox model. Third, compute the cumulative number of actual events at each survival time for each stratum. Fourth, plot the estimated cumulative hazard rate at each actual survival time along the y axis against the corresponding observed cumulative number of events on the x axis for each stratum. Eventually, discrepancies between the estimated cumulative hazard rate and the empirical data can display whether the estimated hazard rates of those strata are scattered randomly or systematically too high or too low.

In particular, if the proportionality assumption for the $(M + 1)$th covariate holds, the stratum-specific plots should be approximately linear with various slopes, so that the discrepancy between the estimated and the observed should be a martingale. If all stratum-specific plots are closely clustered around a 45-degree line, then adding the $(M + 1)$th covariate to the Cox model is unnecessary because the variable does not contribute additional information into the Cox model. Likewise, if those stratum-specific curves are separated in a nonlinear fashion, it can be inferred that the proportional hazards assumption for the

($M + 1$)th covariate is violated. Between two stratum-specific curves, if the curve for the first stratum is concave and the second curve convex, the hazard ratio between the two strata would increase over time. Obviously, this checking approach is somewhat similar to the graphical check using the scaled Schoenfeld residuals.

The Arjas plots have the advantage that the plots are derived from an integrated proportional hazard model, rather than from several stratified models. Therefore, this graphical approach increases the statistical power of the assessment. These plots are particularly useful when the baseline hazard function tends to be common across all strata. If there is strong evidence on distinct differences in the baseline hazard function, however, using this graphical check is inappropriate; under such circumstances, the Andersen plots are preferred.

6.3.5 Checking proportionality with cumulative sums of martingale-based residuals

This graphical and numeric method is based on the martingale theory, so I use the counting processes terminology for its description. In principle, the application of the martingale residuals and their transforms, described in Section 6.1, can detect departures from the proportionality assumption by plotting the score process versus follow-up times. Just looking at such residuals, however, is not clear enough to make a firm conclusion about the validity of the proportional hazards assumption because considerable residual deviations can occur even when the model is correctly specified. Accordingly, it is essential to develop a martingale transform that can be used both for a graphical check and as a summary statistic with a known distribution, thereby providing additional information on whether the null hypothesis on the proportional hazards should be accepted or rejected.

Lin, Wei, and Ying (1993) propose a combined method for the assessment of the Cox model based on the cumulative sums of martingale residuals and the transforms. The rationale of this method is that the failure of the proportionality assumption in the Cox model would be reflected by deviations of observed martingale residuals from some standardized martingale transforms with a known distribution.

Specifically, they first group the martingale-based residuals cumulatively with respect to follow-up times and/or covariate values. Then they develop the following two classes of multiparameter Wiener stochastic processes:

$$\mathrm{W}_{\mathbf{z}}(t,\mathbf{z}) = \sum_{i=1}^{n} \mathrm{f}(\mathbf{Z}_i)\mathrm{I}(\mathbf{Z}_i \leq \mathbf{z})\hat{\mathrm{M}}_i(t), \tag{6.85}$$

$$\mathrm{W}_{\mathrm{r}}(t,\mathrm{r}) = \sum_{i=1}^{n} \mathrm{f}(\mathbf{Z}_i)\mathrm{I}(\mathbf{Z}_i'\boldsymbol{\beta} \leq \mathrm{r})\hat{\mathrm{M}}_i(t), \tag{6.86}$$

where $\mathrm{f}(\cdot)$ is a known smooth function, $\mathbf{z} = (\mathrm{z1}, \ldots, \mathrm{z}_M)' \in \mathcal{R}^M$, and $(\mathcal{Z}_i \leq \mathcal{z})$ means that all the M covariates in $\mathbf{Z}_i$ are not larger than the respective components of $\mathbf{z}$. The distributions of these two stochastic processes under the proportionality hypothesis can be approximated by the distributions of certain zero-mean Gaussian processes, assuming a known stochastic structure of the martingale process $\mathrm{M}_i(t)$. In particular, a standardized process $\mathrm{N}_i(u)\mathrm{G}_i$ is recommended given $\langle \mathrm{M}_i \rangle(t) = \mathrm{E}[\mathrm{N}_i(t)]$ (Fleming and Harrington, 1991), where $(\mathrm{G}_1, \ldots, \mathrm{G}_n)$ are the standard normal variables that are independent of the triple $(T_i, \delta_i, \mathbf{Z}_i)$. Given a zero-mean Gaussian distribution, these two standardized processes should fluctuate randomly around zero. Consequently, each observed martingale process can be plotted along with a number of realizations from simulation using an estimator of W_z. Observed patterns of residuals can then be compared, both graphically and numerically, under the null distribution. The resulting graphical plots and the attached numeric statistics, in turn, enable the researcher to assess more objectively whether the observed residual pattern deviates significantly from random fluctuations.

In particular, given the definition of the martingale residuals, formulated by Equation (6.62), the empirical score process $\tilde{\boldsymbol{U}}\left(\hat{\boldsymbol{\beta}},t\right)=\left[\tilde{U}_1\left(\hat{\boldsymbol{\beta}},t\right),\ldots,\tilde{U}_M\left(\hat{\boldsymbol{\beta}},t\right)\right]$ can be viewed as a transform of the martingale residuals, given by

$$\tilde{\boldsymbol{U}}\left(\hat{\boldsymbol{\beta}},t\right)=\sum_{i=1}^{n}\mathbf{Z}_i\hat{\mathrm{M}}_i(t). \tag{6.87}$$

The standardized empirical score process for the mth component of $\mathbf{Z}$, denoted by $\tilde{U}_m^*(t)$, is

$$\tilde{U}_m^*(t)=\sup\left\{\left[\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right)_{mm}\right]^{1/2}\tilde{U}_m^*\left(\hat{\boldsymbol{\beta}},t\right)\right\},\quad m=1,\ldots,M. \tag{6.88}$$

where $\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right)_{mm}$ represents the diagonal elements in the inverse of the observed information matrix.

This standardized test statistic $\tilde{U}_m^*(t)$, under the null hypothesis that the proportional hazards assumption holds, is a special case of $W_{\mathbf{z}}(t, \mathbf{z})$ with $\mathbf{z} = \infty$ and $f(\cdot) = \cdot$. Given the Taylor series expansion, it can be approximated by

$$\begin{aligned}\tilde{U}_m^*(t)=\left[\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right)_{mm}\right]^{1/2}&\left\{\sum_{l=1}^{n}\mathrm{I}(T_l\le t)\delta_l\left[\mathbf{Z}_{ml}-\bar{\mathbf{Z}}_m\left(\hat{\boldsymbol{\beta}},t\right)\right]\mathrm{G}_l\right.\\&-\sum_{k=1}^{n}\int_0^t\mathrm{Y}_k(u)\exp\left(\mathbf{Z}_k'\hat{\boldsymbol{\beta}}\right)\mathbf{Z}_{mk}\left[\mathbf{Z}_k-\bar{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}},u\right)\right]'\mathrm{d}\hat{\Lambda}_0(u)\\&\left.\times\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right)\sum_{l=1}^{n}\delta_l\left[\mathbf{Z}_l-\bar{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}},T_l\right)\right]\mathrm{G}_l\right\},\end{aligned} \tag{6.89}$$

where $\bar{\mathbf{Z}}_m\left(\hat{\boldsymbol{\beta}},t\right)$ is the mth component of $\bar{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}},t\right)$. As the standardized score converges to a zero-mean Gaussian process, the resulting p-values are valid asymptotically regardless of the covariance structure.

Given the above empirical score process, the proportional hazards assumption for the mth covariate, according to Lin, Wei, and Ying (1993), can be assessed by plotting a dozen or so realizations of the simulated $\tilde{U}_m^*(t)$ on the same graph as the observed $\tilde{U}_m^*(t)$, thereby checking whether the observed scores fit in the null distribution samples. Additionally, given Equation (6.89), this graphical method can be supplemented by applying a Kolmogorov-type supremum test. The test statistic is

$$\sup_t\left\|\tilde{U}_m^*\left(\hat{\boldsymbol{\beta}},t\right)\right\|$$

or

$$\sup_t \sum_{m=1}^{m} \left[\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right)_{mm} \right]^{1/2} \left| \tilde{U}_m^*\left(\hat{\boldsymbol{\beta}}, t\right) \right|.$$

Lin, Wei, and Ying (1993) contend that such test scores are consistent against the nonproportional hazards alternative, in which the effects of one or more covariates are not time independent. Given the value of α, the researcher can make a decision on whether or not the proportional hazards assumption in the Cox model is valid. Specifically, if the p-value of the Kolmogorov-type supremum test is smaller than α on a given covariate, it is appropriate to conclude with sufficient confidence that the proportionality assumption for this covariate is invalid.

This checking method is appealing in several aspects. First, it uses a standardized martingale transform that is distributed asymptotically as a zero-mean Gaussian process, so that it can be applied effectively to compare the expected and the observed martingales with a known distribution. Second, such a plot of martingale-typed residuals is linked with a Kolmogorov-type supremum test score, thus making a graphical check on the proportionality assumption in conjunction with a routine statistical test with a known distribution. Third, the development of this method provides a solid theoretical foundation for further refinements for handling other statistical issues encountered in survival analysis, as will be described in the next section and in Chapter 7.

6.3.6 Illustration: Checking the proportionality assumption in the Cox model for the effect of age on the mortality of older Americans

In Section 5.5, I displayed three log–log survival curves for three age groups: 70–74 years, 75–84 years, and 85 years or over. These survival curves present distinct separations and appear approximately parallel over time. Therefore, this graph provides some evidence that the effects of age on the mortality of older Americans are approximately proportional and thus can be considered a covariate in the Cox model. In Section 6.2, a plot of the Schoenfeld residuals corresponding to age also displays the effect of age to be independent of time. Those graphical checks, however, are crude in several aspects. First, in the stratified Cox model age is roughly divided into three groups, so that the proportionality within each age group cannot be assessed. Second, the stratified Cox model using three age groups as strata is poorly fitted and thus the graph derived from its results is not highly efficient (this issue is also discussed in Subsection 6.3.4). Third, it is difficult to conclude with sufficient confidence that the approximate parallel of the three log–log curves is statistically significant because a separation between two curves can also come from sampling errors. Lastly, the Schoenfeld residuals do not specify residuals for censored observations. Given these limitations, more refined methods should be applied to check the proportional hazards assumption further on the effects of age.

In the present illustration, I check the proportionality assumption on the effects of age by using two refined methods: (1) the checking approach with the addition of a time-dependent variable and (2) the graphical and numeric method using cumulative sums of the martingale-based residuals developed by Lin, Wei, and Ying (1993). The observation period is still from the baseline survey to the end of the year 2004. Three mean-centered variables – ‘Age_mean,’ ‘Female_mean,’ and ‘Educ_mean’ – are used as covariates with their values fixed at baseline.

As the first step, I illustrate the application of the first approach, which adds a time-dependent component to the effect of age. Specifically, a time-dependent variable is created – {Age_mean × [log t – mean(log t)]} – for checking the statistical significance of the time-dependent component on the effect of age. The working hypothesis on this numeric test is that the estimated regression coefficient of the time-dependent component of age is not statistically significant, thus validating the effect of age to be multiplicatively constant. The estimation procedure is exactly the same as described in Subsection 5.3.3, except for the addition of a time-dependent variable.

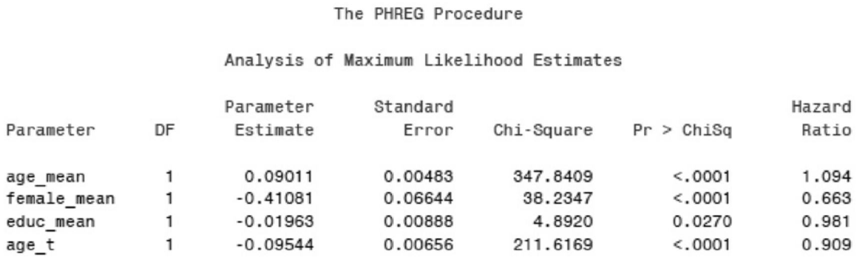
The SAS program for estimating this hazard model is given below.

SAS Program 6.2:

```
……

log_t = log(duration);

……

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,
    log_t - mean(log_t) as logt_mean
  from new;
quit;

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

In SAS Program 6.2, I first create a variable log(t) to increase numeric stability in estimating the Cox model. Then, in the PROC SQL procedure, a mean-centered variable of log(t) is constructed for reducing collinearity. In the MODEL statement, the time-dependent variable, subsequently defined and named as 'Age_t,' is included in the regression, so that at each survival time, the individuals exposed to the risk of dying just before t are subject to two age dimensions. The Breslow method is applied again to handle tied observations. The following SAS output table displays parameter estimates.

SAS Program Output 6.1:



```
                              The PHREG Procedure

                     Analysis of Maximum Likelihood Estimates

                       Parameter    Standard                                 Hazard
Parameter      DF       Estimate       Error    Chi-Square    Pr > ChiSq      Ratio

age_mean        1        0.09011     0.00483      347.8409        <.0001      1.094
female_mean     1       -0.41081     0.06644       38.2347        <.0001      0.663
educ_mean       1       -0.01963     0.00888        4.8920        0.0270      0.981
age_t           1       -0.09544     0.00656      211.6169        <.0001      0.909
```

In SAS Program Output 6.1, the estimated regression coefficients of all four covariates are statistically significant. In particular, the regression coefficient of the time-dependent variable ‘Age_t’ is −0.0954, with the Wald chi-square statistic very strongly statistically significant ($p < 0.0001$). The difference in the chi-square of the likelihood ratio test scores with or without the time-dependent variable, not presented here, is as high as 154.57 (470.56–315.99), also very strongly significant given one degree of freedom ($p < 0.0001$). Therefore, from the results of this test, the effect of age on the hazard function is shown to be a function of time. It is interesting to note that the main effect of age is positive, whereas the regression coefficient of the time-dependent component is negative, which, combined, highlights a decreasing trend over time in the hazard ratio of age. If such a decrease in the hazard ratio is sizable, it can be inferred that the proportional hazards assumption on age is invalid.

Closer examination on the hazard ratio of the time-dependent variable, however, suggests some caution in rejecting the null hypothesis on the proportional hazards of age. As the interaction of time is measured as log time, a hazard ratio of 0.91, though statistically significant, may not necessarily translate into a strong offsetting effect on proportional hazards. Given this concern, it is informative to display a plot that supports the absence of proportionality along the life course. Accordingly, next I illustrate Lin, Wei, and Ying’s (1993) method with cumulative sums of the martingale-based residuals on the same data.

Below is the SAS program for using this method to check the proportionality assumption on the effects of age.

SAS Program 6.3:

```
......
ods graphics on ;
proc phreg data = new ;
  model duration*Status(0) = age_mean female_mean educ_mean /  ties = BRESLOW ;
  assess ph / resample seed = 25 ;
run;
ods html close ;
```

In SAS Program 6.3, the MODEL statement does not include the time-dependent variable Age_t because this method is based on the martingale residual and its transforms, rather than on specifying an additional time-dependent covariate. The ASSESS statement tells SAS that the graphical and numeric methods of Lin, Wei, and Ying (1993) should be performed for checking the proportional hazards assumption and, as will be presented in next section, the adequacy of some other specifications in the Cox model. In particular, the PH option requests SAS to check the proportional hazards assumption for all covariates in the model. For each covariate, the observed score process component is plotted versus follow-up times along with 20 simulated patterns. The RESAMPLE option requests SAS to compute the Kolmogorov-type supremum test on 1000 simulated patterns. The last option in the ASSESS statement, the SEED = 25, specifies the number used to create simulated realizations for plots and the Kolmogorov-type supremum tests.

The plot in Figure 6.2 displays the graphical results of the proportional hazards assumption check for the covariate Age_mean. The standardized and the observed score processes are shown for covariate Age_mean, suggesting that in the early stage of the life course, the observed scores are consistently below zero, thus showing some systematic variability. Overall, however, the observed process tends to fluctuate randomly around zero, particularly at later survival

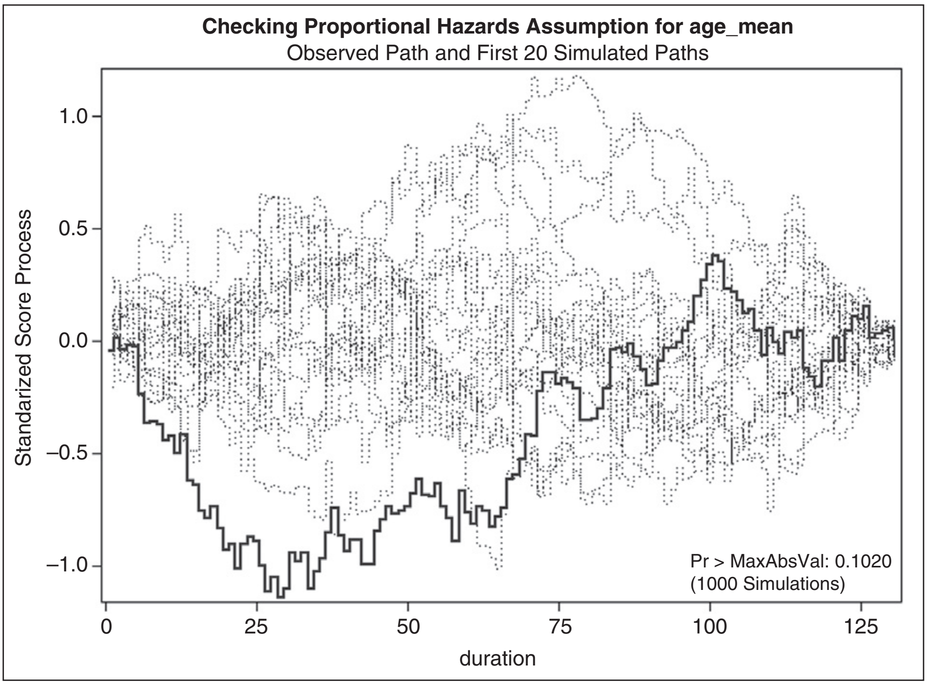
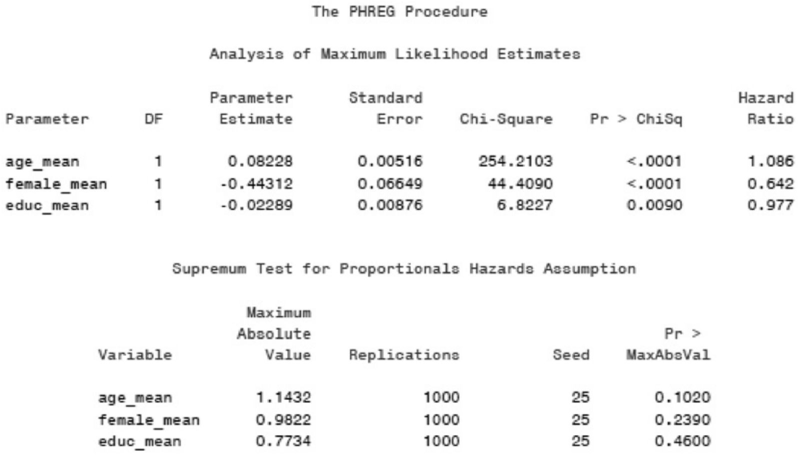


Figure 6.2 Checking the proportional hazards assumption for age.

times. Therefore, it seems that the null hypothesis on the proportionality assumption about age cannot be rejected. SAS Program 6.3 also generates graphical results for Female_mean and Educ_mean, not presented here, with both plots revealing proportional hazards.

The following SAS output tables present the numeric results of the regression coefficients and the Kolmogorov-type supremum tests for all the three covariates, also produced from SAS Program 6.3.

SAS Program Output 6.2:



```
                              The PHREG Procedure

                     Analysis of Maximum Likelihood Estimates

                        Parameter      Standard                                 Hazard
Parameter       DF       Estimate         Error    Chi-Square    Pr > ChiSq      Ratio

age_mean         1        0.08228       0.00516      254.2103        <.0001      1.086
female_mean      1       -0.44312       0.06649       44.4090        <.0001      0.642
educ_mean        1       -0.02289       0.00876        6.8227        0.0090      0.977


                 Supremum Test for Proportionals Hazards Assumption

                       Maximum
                      Absolute                                      Pr >
         Variable        Value    Replications        Seed    MaxAbsVal

         age_mean       1.1432            1000          25       0.1020
         female_mean    0.9822            1000          25       0.2390
         educ_mean      0.7734            1000          25       0.4600
```

Not surprisingly, the estimated regression coefficients of all three covariates on the hazard rate are statistically significant, with all p-values smaller than 0.01, consistent with the results previously reported. From the results of the Kolmogorov-type supremum tests on all three covariates, the proportional hazards assumption on each of them is not violated, with all p-values of those tests greater than 0.10. Obviously, the results derived from this method contradict those obtained from the method adding a time-dependent component.

6.4 Checking the functional form of a covariate

In Section 6.3, I introduce several refined graphical and numeric methods for checking the proportional hazard assumption in the Cox model. Sometimes, researchers are also concerned with the functional form of a covariate. As widely discussed in the literature on regression modeling, a continuous variable x_m may take many functional forms, such as $\log(x_m)$, $(x_m)^2$, $\sqrt{x_m}$, with each implying a unique distribution. In certain circumstances, x_m can be partitioned into two or more subgroups for capturing nonlinearity of an association. In survival analysis, misuse of a functional form for a covariate may lead to misspecification of model parameters, thereby yielding misleading and erroneous analytic results.

In this section, I introduce two statistical methods for checking the functional form of a covariate – the checking method comparing the model fit of different link functions in terms of a specific covariate and the graphical and numeric method using cumulative sums of the martingle-based residuals. An illustration is provided for checking the function form of age in the Cox model on the mortality of older Americans.

6.4.1 Checking model fit statistics for different link functions

Statistically, this method is simple, straightforward, and consistent with the corresponding approaches used for other linear or nonlinear regressions. Suppose we want to check the functional form of covariate x_m. First, the covariate vector $\boldsymbol{x}$ in the Cox model can be partitioned into two parts – x_m, the covariate under assessment, and $\boldsymbol{x}_r$, the vector of other covariates in $\boldsymbol{x}$. Assuming the proportional hazards for $\boldsymbol{x}_r$ to be valid and letting the true functional form of x_m be denoted by $\mathrm{f}(x_m)$, the correct Cox model should be written by

$$h(t|x_m, \boldsymbol{x}) = h_0(t)\exp[\mathrm{f}(x_m)\beta_1 + \boldsymbol{x}_r'\boldsymbol{\beta}_r], \tag{6.90}$$

where β_1 is the regression coefficient of $\mathrm{f}(x_m)$. On most occasions, $\mathrm{f}(x_m)$ is simply approximated by taking its natural form x_m or its mean-centered form (x_m minus mean(x_m)), assuming the variable to be linearly associated with the log hazards. If there is strong evidence against the assumption of such linearity, however, exploring an appropriate functional form of x_m becomes necessary.

Practically, we may create $\acute{n}$ regression models by using $\acute{n}$ different link functions of x_m for checking which functional form fits most closely with the log hazards. Suppose that x_m is independent of other covariates; an appropriate functional form of x_m can be determined by comparing the results of the Wald test statistic for those $\acute{n}$ Cox models. The regression having the highest Wald test score on the estimated regression coefficient among a variety of functional forms of x_m should be regarded as the most appropriate Cox model. Therefore, that specific functional form of x_m can be taken as the one closest to the true function f(x_m).

Alternatively, the statistical testing on the functional form of x_m can be performed by using the partial likelihood ratio test. In particular, for two successive Cox models that specify two different functional forms of x_m, denoted by $\mathrm{f}^{(\gamma-1)}$ and by $\mathrm{f}^{(\gamma)}$, respectively, the partial likelihood ratio test score is given by

$$G_{\beta_1(\beta-1,\beta)} = -2\times\left\{\log L_p\left[\hat{\beta}_1^{(\gamma-1)}, \hat{\boldsymbol{\beta}}_r\right] - \log L_p\left[\hat{\beta}_1^{(\gamma)}, \hat{\boldsymbol{\beta}}_r\right]\right\}, \quad \gamma = 1, \ldots \dot{w}, \tag{6.91}$$

where $\hat{\beta}_1^{(\gamma-1)}$ and $\hat{\beta}_1^{(\gamma)}$ are the estimated regression coefficients of x_m transforms from models with functional forms $\mathrm{f}^{(\gamma-1)}$ and $\mathrm{f}^{(\gamma)}$, respectively, and $G_{\beta_1(\gamma-1,\gamma)}$ is the partial likelihood ratio test statistic that reflects whether or not the functional form $\mathrm{f}^{(\gamma)}$ gains statistical information as compared to the functional form $\mathrm{f}^{(\gamma-1)}$, distributed as χ^2 with one degree of freedom. Within the brace on the right of Equation (6.91), the first term is the log partial likelihood ratio statistic for the model with functional form $\mathrm{f}^{(\gamma-1)}$, whereas the second is the same statistic for the Cox model with functional form $\mathrm{f}^{(\gamma)}$. If $G_{\beta_1(\gamma-1,\gamma)} < \chi^2_{(1-\alpha;1)}$, the specification of $\mathrm{f}^{(\gamma)}$ does not improve the quality of the model fit, thereby suggesting that $\mathrm{f}^{(\gamma)}$ should be dropped from further comparison and the function $\mathrm{f}^{(\gamma-1)}$ should be retained. If $G_{\beta_1(\gamma-1,\gamma)} > \chi^2_{(1-\alpha;1)}$, the specification of the functional form $\mathrm{f}^{(\gamma)}$ predicts the hazard rate statistically better than does the functional form $\mathrm{f}^{(\gamma-1)}$; therefore this functional form should be retained for further comparison and, accordingly, $\mathrm{f}^{(\gamma-1)}$ be dropped. Eventually, the most appropriate functional form of x_m can be determined statistically from those $\acute{n}$ candidates ordered by the level of complexity.

Sometimes, it is technically difficult to judge the validity of a specific functional form of a covariate simply from the above procedure. For example, other than several frequently used functional forms (e.g., $\log x_m$, $(x_m)^2$, $\sqrt{x_m}$), some high-order polynomial functions are occasionally used. Given the problem of correlation among the linear, quadratic, and high-order terms, centering x_m is necessary to reduce multicollinearity, thereby complicating the selection of an appropriate function form. Correlation between x_m and other covariates must also be considered, which can significantly complicate the selection process. Specifically, the existence of complex interrelationships among covariates can make the test results dubious. Perhaps due to these reasons, some statisticians recommend computing and plotting the martingale residuals to find a functional form that is closest to f (Klein and Moeschberger, 2003; Therneau, Grambsch, and Fleming, 1990). If the martingale plot with x_m as a covariate appears linear, no transformation of x_m is needed; if the plot appears nonlinear, then a transformation of x_m may be necessary. In the presence of heavy censoring, however, it is very infrequent to find a plot of the martingale residual or its transforms to be distributed linearly, as evidenced in Subsection 6.3.6.

6.4.2 Checking the functional form with cumulative sums of martingale-based residuals

This method for checking the functional form of a covariate is an integral part of the graphical and numeric approach described in Subsection 6.3.5. As previously discussed, it is difficult to derive a convincing conclusion about the validity of the proportional hazards assumption just by examining the martingale residuals and their transforms. The same problem exists for checking the functional form of a covariate. Lin, Wei, and Ying (1993) provide a less subjective approach by examining the two stochastic processes with a known distribution, specified by Equations (6.85) and (6.86). In terms of checking the functional form of covariate x_m, they propose to plot the partial-sum processes of the martingale residuals, written as

$$\mathrm{W}_m(\mathrm{z}) = \sum_{i=1}^{n} \mathrm{I}(\mathrm{Z}_{im} \le \mathrm{z})\hat{\mathrm{M}}_i, \quad m = 1, \ldots, M, \tag{6.92}$$

where $\mathrm{W}_m(\mathrm{z})$ is a special case of $\mathrm{W}_{\mathrm{z}}(t, \mathrm{z})$, specified in Equation (6.85), with $\mathrm{f}(\cdot) = 1$ and $z_{m'} = \infty (m' \ne m)$.

As indicated above, the null distribution of $\mathrm{W}_m(\cdot)$ can be approximated through simulating the corresponding zero-mean Gaussian process $\hat{\mathrm{W}}_m(\cdot)$ along with a dozen or so realizations. In the case of checking the functional form of covariate m, $\mathrm{W}_m(\mathrm{z})$ can be approximated by

$$\begin{aligned}\hat{\mathrm{W}}_m(\mathrm{z}) = \sum_{l=1}^{n} \delta_l & \left\{ \mathrm{I}(\mathrm{Z}_{lm} \le \mathrm{z}) - \frac{\sum_{i=1}^{n} \mathrm{Y}_i(T_l)\exp\left(\mathbf{Z}_i'\hat{\boldsymbol{\beta}}\right)\mathrm{I}(\mathrm{Z}_{im} \le \mathrm{z})}{\sum_{i=1}^{n} Y_i(\mathrm{T}_l)\exp\left(\mathbf{Z}_i'\hat{\boldsymbol{\beta}}\right)} \right\} G_l \\ & - \sum_{\bar{k}=1}^{n} \int_0^t Y_{\bar{k}}(u)\exp\left(\mathbf{Z}_{\bar{k}}'\hat{\boldsymbol{\beta}}\right)\mathrm{I}(\mathrm{Z}_{\bar{k}m} \le \mathrm{z})\left[\mathbf{Z}_{\bar{k}} - \bar{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}}, u\right)\right]' \mathrm{d}\hat{\Lambda}_0(u) \\ & \times \boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}}\right) \sum_{l=1}^{n} \delta_l \left[\mathbf{Z}_l - \bar{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}}, T_l\right)\right]\mathrm{G}_l \Bigg\}.\end{aligned} \tag{6.93}$$

Consequently, if the null hypothesis about the functional form of x_m holds, $\hat{\mathrm{W}}_m(\cdot)$ should fluctuate randomly around zero. Accordingly, a natural numeric measure can be created:

$$\tilde{\mathrm{s}}_m = \sup_{\mathrm{z}} \left| \mathrm{w}_m(\mathrm{z}) \right|, \tag{6.94}$$

where $\mathrm{w}_m(\cdot)$ is the observed value of the Gaussian approximate $\mathrm{w}_m(\cdot)$. This numeric score, $\tilde{\mathrm{s}}_m$, can be used to display whether or not the functional form of x_m is statistically suitable. Specifically, if the value of $\tilde{\mathrm{s}}_m$ is beyond a critical point, the underlying functional form of x_m is questionable; therefore, using another transform of x_m in the Cox model may be necessary. The p-value for the distribution of $\tilde{\mathrm{s}}_m$, under the null hypothesis, can be approximated by $\Pr\left(\hat{\tilde{\mathrm{S}}}_m \ge \tilde{\mathrm{s}}_m\right)$, where

$$\hat{\tilde{\mathrm{S}}}_m = \sup_{\mathrm{z}} \left| \hat{\mathrm{W}}_m(\mathrm{z}) \right|.$$

The calculation of $\Pr\left(\hat{\tilde{\mathrm{S}}}_m \geq \tilde{\mathrm{s}}_m\right)$ is conditional on the triple (T_i, δ_i, $\mathbf{Z}_{\ i}$). Lin, Wei, and Ying (1993) showed that $\Pr\left(\hat{\tilde{\mathrm{S}}}_m \geq \tilde{\mathrm{s}}_m\right)$ converges to $\Pr\left(\tilde{\mathrm{S}}_m \geq \tilde{\mathrm{s}}_m\right)$ as n tends to ∞, consistent against incorrect functional forms for x_m if there is no additional model misspecification and if x_m is independent of other covariates.

6.4.3 Illustration: Checking the functional form of age in the Cox model on the mortality of older Americans

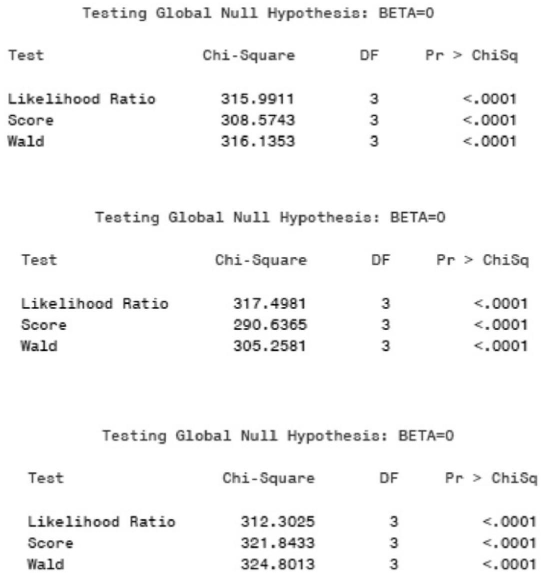
In this illustration, I check the functional form of age by using the two methods described above: (1) the method based on the partial likelihood ratio test and (2) the graphical and numeric approach using cumulative sums of the martingale-based residuals developed by Lin, Wei, and Ying (1993). The observation period is still the same as indicated in the previous example (from the outset of the baseline survey to the end of the year 2004). Three mean-centered variables – ‘Age_mean,’ ‘Female_mean,’ and ‘Educ_mean’ – are used as time-independent covariates. For applying the first method, I focus on comparing three functional forms of age – Age_mean (age – mean[age]), log(age), and age × age. Specifically, I want to know whether the second and the third functional forms significantly improve the overall fit of the Cox model by examining the partial likelihood ratio test statistics. The SAS program for estimating the three models are displayed below.

SAS Program 6.4:

```
……
proc SQL;
  create table new as
  select *, age - mean(age) as age_mean,
    log(age) as log_age,
    (age)**2 as age_2,
    female - mean(female) as female_mean,
    educ - mean(educ) as educ_mean
  from new;
quit;

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

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

proc phreg data = new ;
  model duration*Status(0) = age_2 female_mean educ_mean /  ties = BRESLOW ;
run;
```

In SAS Program 6.4, I create two additional functional forms of age – log(age) and (age)^2 – using the PROC SQL procedure. Then, three Cox models are specified using those three different functional forms. In the MODEL statement of each model, age is considered in the regression with a unique functional form. The Breslow method is applied again to handle tied observations. The following output table displays the results of the overall test statistics on the three models.

SAS Program Output 6.3:



```
              Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio       315.9911        3         <.0001
Score                  308.5743        3         <.0001
Wald                   316.1353        3         <.0001


              Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio       317.4981        3         <.0001
Score                  290.6365        3         <.0001
Wald                   305.2581        3         <.0001


              Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio       312.3025        3         <.0001
Score                  321.8433        3         <.0001
Wald                   324.8013        3         <.0001
```

As shown in SAS Program Output 6.3, differences in the chi-square value of the likelihood ratio test score between the three models do not support the proposition that the specification of the second and the third functional forms of age significantly improve the overall fit of the Cox model. Changes in this test score are not statistically significant ($p > 0.05$), thus suggesting that the two additional functional forms of age do not have a better link with the hazard function. Among the three functional forms, the centered variable Age_mean is the natural and the most parsimonious transform of age.

As discussed above, however, this method does not necessarily provide sufficient information for deriving a convincing conclusion. Like the case in checking the proportionality assumption in the Cox model, a statistical method combining both graphical and numeric checks provides more insights for an appropriate functional form of a covariate. For this reason, next I apply Lin, Wei, and Ying's (1993) approach with cumulative sums of the martingale-based residuals. In particular, I start by using [age – mean(age)], represented by covariate Age_mean, as the functional form. Below is the SAS program for fitting this model.

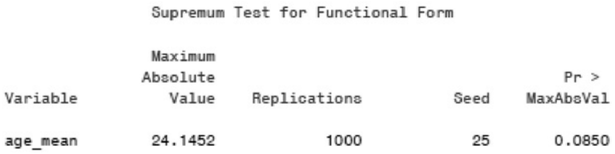
SAS Program 6.5:

```
……
ods graphics on ;
proc phreg data = new ;
 model duration*Status(0) = age_mean female_mean educ_mean /  ties = BRESLOW ;
 assess var = (age_mean) / resample seed = 25 ;
run;
ods html close ;
```

In SAS Program 6.5, I fit the Cox model with the covariate 'Age_mean,' using 'Female_mean' and Educ_mean' as controls. The Breslow method is used again to handle tied observations. As mentioned in Subsection 6.3.6, the ASSESS statement can be used to check the adequacy of some other specifications in the Cox model, including the capability of checking the functional form of a covariate. Accordingly, I use the VAR=(AGE_MEAN) option to create a plot of the cumulative martingale residuals against values of the covariate Age_mean. From this option, the functional form of this age transform can be assessed both visually and analytically. The RESAMPLE and SEED options are explained in Section 6.3.

SAS Program 6.5 yields an output table and a graph. I first review the result of the supremum test for this functional form of age, shown below.

SAS Program Output 6.4:



```
                    Supremum Test for Functional Form

                    Maximum
                    Absolute                                      Pr >
Variable               Value    Replications          Seed    MaxAbsVal

age_mean             24.1452            1000            25       0.0850
```

SAS Program Output 6.4 displays that the *p*-value of Age_mean from the Kolmogorov-type supremum test, based on 1000 simulations, is 0.0850. Statistically, this *p*-value is marginal: if the value of α is set at 0.05, it can be said that the observed martingale process does not deviate significantly from the simulated realizations, so that using another transform of age seems unnecessary. If $\alpha = 0.10$, the observed process would be considered atypical compared to the normal simulations, which, in turn, suggests that a more appropriate functional form of age should be used to replace centered age. Therefore, a graphical check may be helpful to draw a more confident conclusion. The graph generated from SAS Program 6.5 is shown in Figure 6.3, which displays the plot of the observed cumulative martingale residual process for Age_mean, along with 20 simulated realizations from the null

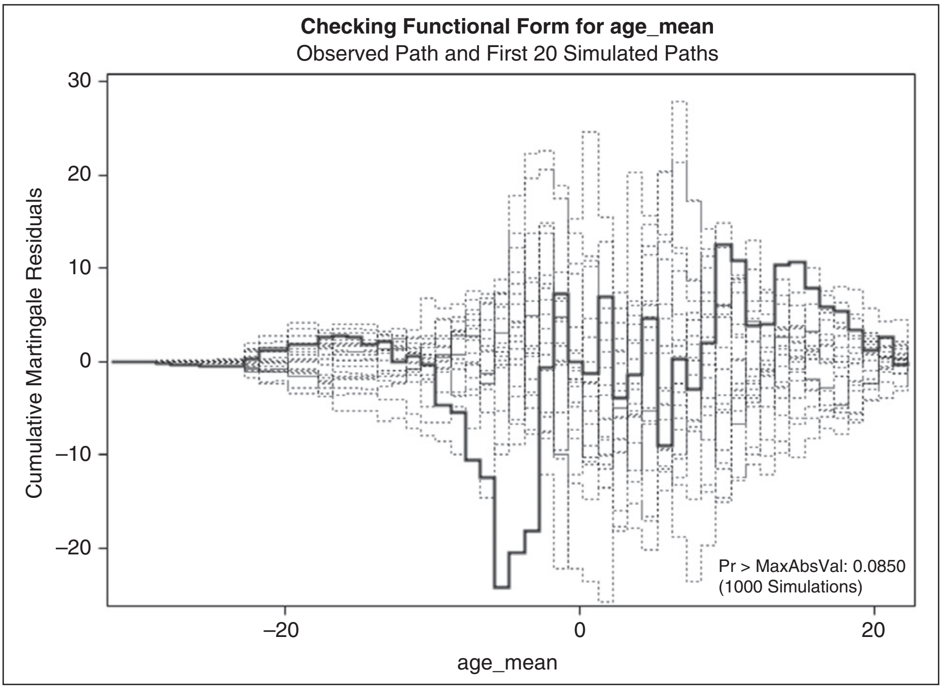


Figure 6.3 Checking the functional form of the covariate 'age_mean'.

distribution. According to this plot, the observed martingale residuals do not seem to fall off systematically from the null distribution, except at some survival times located in the middle range of Age_mean; therefore, using another transform of age seems unnecessary.

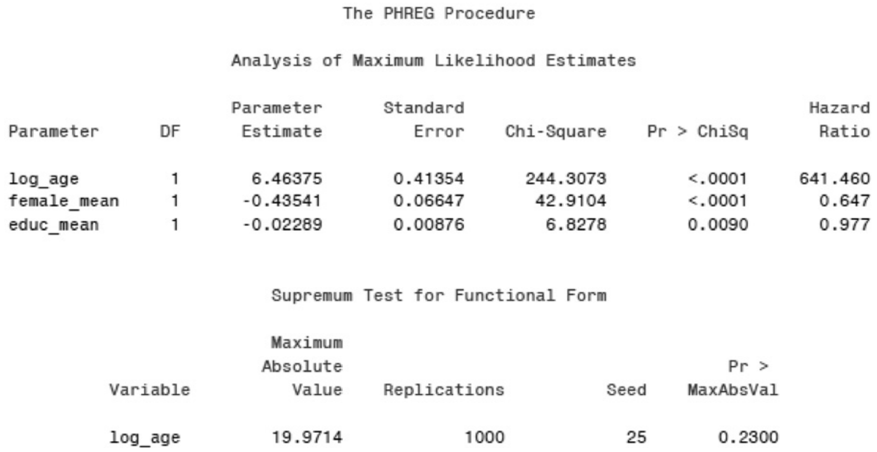
To ensure that [age – mean(age)] is the appropriate functional form of age in the Cox model, I perform some additional analyses by replacing this variable with the functional form log(age) in the Cox model. In particular, I want to check whether the cumulative martingale residual plot, obtained from the model using the new covariate Log_age as the transform of age, differs significantly from Figure 6.3. Below is the SAS program for this step.

SAS Program 6.6:

```
……
ods graphics on ;
proc phreg data = new ;
  model duration*Status(0) = log_age female_mean educ_mean /  ties = BRESLOW ;
  assess var = (log_age) / crpanel resample seed = 25 ;
run;
ods html close ;
```

SAS Program 6.6 considers log(age) as the functional form of age. In the ASSESS statement, the CRPANEL option is added to request a panel of four plots for extensive checks, with each plotting the observed cumulative martingale residual process along with two simulated realizations. The following SAS output table presents the estimated regression coefficients of the three covariates and the result of the supremum test on covariate Log_age.

SAS Program Output 6.5:



```
                              The PHREG Procedure

                    Analysis of Maximum Likelihood Estimates

                     Parameter      Standard                                Hazard
Parameter      DF     Estimate         Error    Chi-Square    Pr > ChiSq     Ratio

log_age         1      6.46375       0.41354      244.3073        <.0001   641.460
female_mean     1     -0.43541       0.06647       42.9104        <.0001     0.647
educ_mean       1     -0.02289       0.00876        6.8278        0.0090     0.977


                       Supremum Test for Functional Form

                       Maximum
                      Absolute                                    Pr >
          Variable       Value    Replications       Seed    MaxAbsVal

          log_age      19.9714            1000         25       0.2300
```

In the above SAS output table, the estimated regression coefficients of all three covariates are very strongly statistically significant, consistent with the results shown previously. The *p*-value for the Kolmogorov-type supremum test based on 1000 simulations is now 0.23,

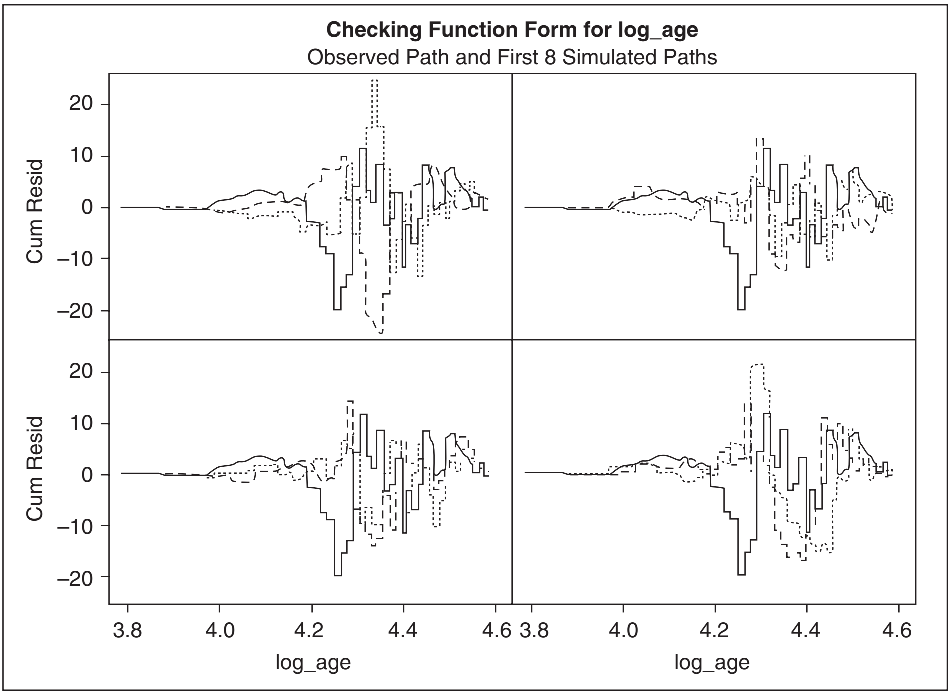


Figure 6.4 Panel plot of cumulative martingale residuals against log (age).

indicating that using additional functional forms of age is unnecessary for estimating the effects of age.

The plot in Figure 6.4 displays the panel with four plots, where the observed cumulative martingale process agrees nicely with each set of two realizations of the null distributions. The next graph further demonstrates this consistency, where Figure 6.5 displays a summary plot of the cumulative martingale residuals against log(age) generated from the VAR= option in the ASSESS statement. This graph plots the observed martingale score process for log(age) along with 20 realizations from the null distribution. In general, this graph appears consistent both with Figure 6.3 and with those in the panel plot. Obviously, either using Age_mean or using Log_age yields the same results on the overall fit of the Cox model and on the distribution of the martingale residuals. As a result, I now have sufficient confidence to conclude that using covariate Age_mean in the Cox model is most appropriate given the simplicity and parsimony attaching to this functional form. As in the survival data of older Americans, an older person's age is not highly correlated with gender and educational attainment, this conclusion seems valid and reliable. This conclusion is also in accordance with the result obtained from the method using the partial likelihood ratio test statistic.

6.5 Identification of influential observations in the Cox model

Another important aspect of regression diagnostics on the Cox model is identification of influential observations. In the literature of regression diagnostics, it is an essential statistical

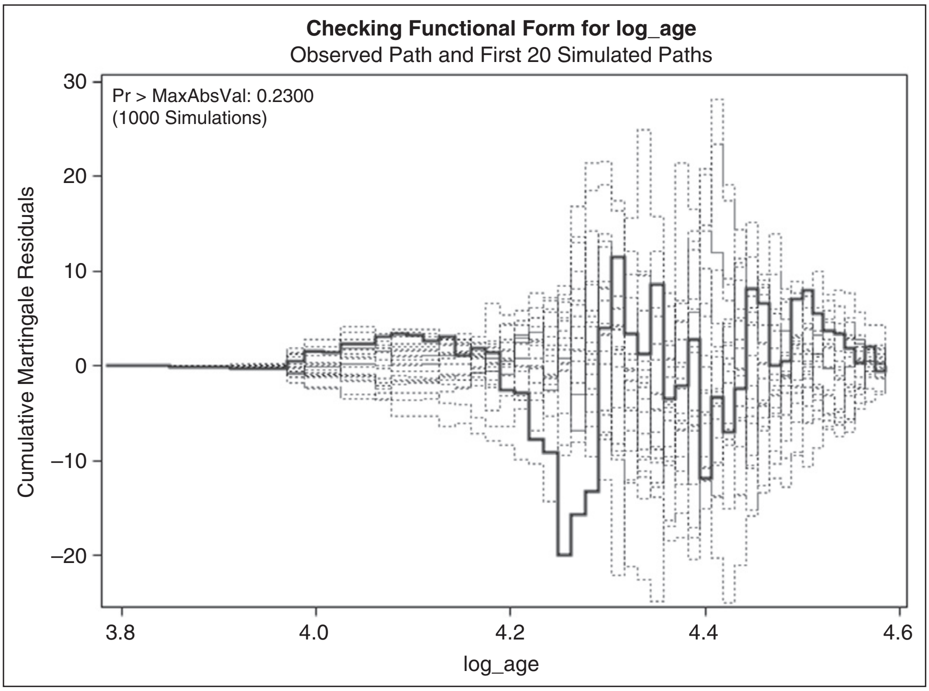


Figure 6.5 Cumulative martingale residuals against log(age).

step to ascertain particular observations that have extraordinary influences on analytic results of a linear or a nonlinear regression model. With regard to the Cox model, removal of such influential cases from the regression should be followed by substantially increased or decreased hazard rates. Identification of influential observations in survival data, however, is not an easy undertaking and differs markedly from conventional perspectives in several ways. Most significantly, the diagnostic techniques on survival data involve individuals at many survival times, rather than at a single data point. As a result, elimination of one individual may affect a series of risk sets, in turn magnifying its influence on parameter estimates. This unique impact is especially strong for those who experience a particular event late in an observation period. Given such characteristics, some of the conventional diagnostic measures, like the Cook's distance (Fox, 1991), would not perform appropriately in the Cox model. Identification of influential observations for the proportional hazard model thus calls for the development of more refined techniques.

This section describes two popular diagnostic methods for identifying influential observations in the Cox model – the likelihood displacement score (the *LD* statistic) and the *LMAX* standardized statistic. An illustration is provided for checking whether there are any influential observations in the Cox model on the mortality of older Americans.

6.5.1 The likelihood displacement statistic approximation

The Cox model is a log-linear regression model and therefore, like other relevant techniques involving a linear predictor, some observations can have an unduly impact on inferential procedures for deriving parameter estimates. Theoretically, such influential cases can be identified by changes in the estimated regression coefficients after deleting each observation in a sequence.

Let $\hat{\boldsymbol{\beta}}$ be the value of $\boldsymbol{\beta}$ that maximizes the log partial likelihood function and $\hat{\boldsymbol{\beta}}_{(-i)}$ be the same estimate of $\boldsymbol{\beta}$ when individual i is eliminated. For covariate x_m, the distance in the estimated regression coefficient after removing individual i, denoted $\bar{D}_{mi}$, is given by

$$\bar{D}_{mi} = \hat{\beta}_m - \hat{\beta}_{m(-i)}. \tag{6.95}$$

The above equation provides an exact measure for the absolute influence of deleting individual i from a regression model on the estimates of regression coefficients. By applying the likelihood ratio test with one degree of freedom, the significance of this influence can be statistically assessed. This likelihood ratio test is referred as the *likelihood displacement* (*LD*) statistic, given by

$$LD_i = 2\log L\left(\hat{\boldsymbol{\beta}}\right) - 2\log L\left[\hat{\boldsymbol{\beta}}_{(-i)}\right]. \tag{6.96}$$

The likelihood displacement statistic is distributed as chi-square under the null hypothesis that $\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}}$. Therefore, the observations having a strong impact on $\hat{\boldsymbol{\beta}}$ can be statistically identified, thus helping the researcher decide whether or not they should be removed from an estimation process.

When analyzing survival data, however, the use of this exact effect is not realistic. First, when the sample size is large, the checking process becomes extremely tedious and time-consuming. Consider, for example, a sample of 2000 observations: the analyst needs to create 2001 Cox models to identify which case or cases have an exceptionally strong impact on the estimation of regression coefficients. Second, unlike other types of generalized linear regression models, eliminating one observation from the Cox model would affect a series of risk sets. In particular, for those who experience a particular event early, less weight should be considered in the estimating process because their contributions to the partial likelihood function are relatively limited. In contrast, for individuals who have the event late, they are involved in more risk sets than others in likelihoods and thus should be given more weight. Hence, identifying influential observations in the Cox model involves a much more complicated procedure than in a common generalized linear regression model. If the distance score $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}$ can be statistically approximated by a scalar measure in the Cox model, influential observations can be identified without removing each case in sequence from the estimation process.

Cain and Lange (1984) developed a method to approximate Equation (6.95) by introducing weights into the partial likelihood function. Suppose that a weighted analysis is needed to identify influential observations in the Cox model, with individual i assigned weight w_i and all other observations given the weight 1. The approximation to Equation (6.95), based on the first-order Taylor series expansion about $w_i = 1$, is

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} \cong \frac{\partial \hat{\boldsymbol{\beta}}}{\partial w_i}. \tag{6.97}$$

This approximation is called the *infinitesimal jackknife approach* (Pettitt and Bin Raud, 1989). This derivative can be evaluated in terms of the score vector $\tilde{\boldsymbol{U}}\left[\hat{\boldsymbol{\beta}}(w_i), w_i\right]$.

The derivative of the log partial likelihood function with respect to $\boldsymbol{\beta}$ can be written as

$$\tilde{\boldsymbol{U}}\left(\hat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{d} \tilde{\boldsymbol{U}}_i = \sum_{i=1}^{d} \left[w_i \boldsymbol{x}_i - w_i \hat{\mathrm{E}}\left(\boldsymbol{x} | \mathcal{R}_i\right) \right], \tag{6.98}$$

where

$$\hat{\mathrm{E}}\left(\boldsymbol{x} | \mathcal{R}_i\right) = \frac{\sum\limits_{l \in \mathcal{R}(t_i)} w_l \boldsymbol{x}_l \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right)}{\sum\limits_{l \in \mathcal{R}(t_i)} \mathrm{w}_l \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right)}.$$

Let D_i be the set of individuals who experience a particular event before or at time t_i. After some algebra, the derivative of $\tilde{\boldsymbol{U}}$ with respect to w_i is

$$\frac{\partial \tilde{\boldsymbol{U}}}{\partial w_i} = \delta_i \left[\boldsymbol{x}_i - \hat{\mathrm{E}}\left(\boldsymbol{x} | \mathcal{R}_i\right) \right] - \sum_{i=1}^{D_i} \frac{w_i \exp\left(\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}\right)}{\sum\limits_{l \in \mathcal{R}_i} w_l \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right)} \left[\boldsymbol{x}_i - \hat{\mathrm{E}}\left(\boldsymbol{x} | \mathcal{R}_i\right) \right]. \tag{6.99}$$

Equation (6.99) shows that the score vector $\tilde{\boldsymbol{U}}$ with respect to changes in w_i can be decomposed into two components. The first component is the Schoenfeld (1982) residual, defined as the difference between the covariate vector for individual i at time t and the expected values of the covariate vector at the same time. The second component measures the impact that changes in w_i have on all the risk sets including individual i. The second component increases in absolute magnitude with t because it is the sum of an increasing number of terms. Consequently, the second term plays an increasingly important role in estimating the regression coefficients as time progresses.

In Equation (6.99), D_i is fixed for censored patients between event times, so can be written as

$$\frac{\partial \tilde{\boldsymbol{U}}}{\partial w_i} = \delta_i \left[\boldsymbol{x}_i - \frac{\sum\limits_{l \in \mathcal{R}_{tt}} \boldsymbol{x}_l \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right)}{\sum\limits_{l \in \mathcal{R}_{tt}} \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right)} \right] - \exp\left(\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}\right) \left\{ \frac{\boldsymbol{x}_i}{\sum\limits_{l \in \mathcal{R}_{tt}} \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right)} - \frac{\sum\limits_{l \in \mathcal{R}_{tt}} \boldsymbol{x}_l \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right)}{\left[\sum\limits_{l \in \mathcal{R}_{tt}} \exp\left(\boldsymbol{x}_l' \hat{\boldsymbol{\beta}}\right) \right]^2} \right\}. \tag{6.100}$$

Clearly, the above approximation evades the specification of w_i. As a result, such simplification facilitates the development of an approximation for the likelihood displacement statistic.

Consequently, when $w_i = 0$ and all w_l are unity ($l \neq i$), the regular Cox model can be applied to approximate the likelihood displacement statistic without fitting numerous regressions and deleting each individual in sequence. Cain and Lange (1984) suggest that with information of the triple (t, δ, $\boldsymbol{x}$) for each individual as well as of $\hat{\boldsymbol{\beta}}$ and the observed information matrix, an M vector of estimated influence for each individual can be produced.

Assuming the elimination of individual i not to change the values of the observed information matrix, Pettitt and Bin Raud (1989) proposed the following approximation of the likelihood displacement statistic for individual i:

$$\boldsymbol{LD}_i \approx \hat{\mathbf{L}}_i' \hat{\boldsymbol{I}}^{-1}\left(\hat{\boldsymbol{\beta}}\right) \hat{\mathbf{L}}_i, \tag{6.101}$$

where $\hat{\mathbf{L}}_i$ is the score residual vector of individual i, described in Subsection 6.2.4. The mth element of the LD_i vector, referred to as the *delta-beta statistic*, approximates the case influence on the estimated regression coefficient of covariate x_m after deleting individual i. More important, from the LD_i scores, a single measure of the LD statistic can be created by summing all component score residuals within an individual. This global influence statistic reflects the case influence on the overall fit of the Cox model.

The above LD statistic approximates the change in the estimated regression coefficients after deleting each individual from estimation of the Cox model, thus evading elimination of useful survival data. Empirically, both Collett (2003) and Klein and Moeschberger (2003) display remarkable agreement between the exact and the approximate LD statistics using some survival data with a small sample size.

6.5.2 LMAX statistic for identification of influential observations

Another innovative technique for identifying influential observations in the Cox model is the *LMAX* statistic, originally developed by Cook (1986) as a standardized diagnostic method for general regression modeling and later introduced and advanced into survival analysis by Pettitt and Bin Raud (1989). Specifically, this method maximizes the approximation to the $LD(w)$ statistic for changes standardized to the unit length.

Cook (1986) suggests the use of a standardized likelihood displacement statistic for more accurately measuring influences of particular cases in estimating a regression model. Given the likelihood displacement (LD) statistic, represented by Equation (6.96), Cook first defines a symmetric matrix $\mathbf{B}$, given by

$$\mathbf{B} = \mathbf{L}'\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)^{-1}\mathbf{L}, \tag{6.102}$$

where $\mathbf{L}$ is the matrix with rows containing the score residual vector $\hat{\mathbf{L}}_i$. With $\mathbf{B}$ defined, Cook considers the direction of the $n \times 1$ vector $\tilde{\mathbf{l}}$ that maximizes $\tilde{\mathbf{l}}'\mathbf{B}\tilde{\mathbf{l}}$, and $\tilde{\mathbf{l}}$ is standardized to have unit length. Because the $M \times M$ matrix $\hat{\boldsymbol{I}}\left(\hat{\boldsymbol{\beta}}\right)^{-1}$ is positive definite, the $n \times n$ symmetric matrix $\mathbf{B}$ is positive semi-definite with rank no more than M. The statistic $\tilde{\mathbf{l}}_{\max}$ corresponds to the unit length eigenvector of $\mathbf{B}$ that has the largest eigenvalue $\tilde{\gamma}_{\max}$. Then, $\tilde{\mathbf{l}}'_{\max}\mathbf{B}\tilde{\mathbf{l}}_{\max}$ maximizes $\tilde{\mathbf{l}}'\mathbf{B}\tilde{\mathbf{l}}$ and satisfies the equation

$$\mathbf{B}\tilde{\mathbf{l}}_{\max} = \ddot{\lambda}_{\max}\tilde{\mathbf{l}}_{\max} \quad \text{and} \quad \tilde{\mathbf{l}}'_{\max}\tilde{\mathbf{l}}_{\max} = 1,$$

where $\ddot{\lambda}_{\max}$ is the largest eigenvalue of $\mathbf{B}$ and $\tilde{\mathbf{l}}_{\max}$ is the eigenvector associated with $\ddot{\lambda}_{\max}$. The elements of $\tilde{\mathbf{l}}_{\max}$, standardized to unit length, measure the sensitivity of the model fit to each observation of the data. The absolute value of $\tilde{\mathbf{l}}_i$, the ith element in $\tilde{\mathbf{l}}_{\max}$, is used as the *LMAX* score for individual i. Given the unit length, the expected value of the squared *LMAX* statistic for each observation is $1/n$, where n is sample size, so that a value significantly greater than this expected figure indicates a strong influence on the overall fit of a regression model thereby being identified. If $M = 1$, the *LMAX* statistic is proportional to $\mathbf{L}'$ and $\ddot{\lambda}_{\max} = \boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)^{-1}\|\mathbf{L}\|$. When $M > 1$, an advantage of looking at elements of the *LMAX* statistic, rather than at the delta-beta statistics, is that each case has a single summary measure of influence.

As the *LMAX* score is a standardized statistic, it is highly useful to plot elements of *LMAX* scores against survival times and/or values of covariates. The standardization of $\tilde{\mathbf{l}}_{\text{max}}$ to unit length means that the squares of the elements in $\tilde{\mathbf{l}}_{\text{max}}$ sum to unity, so that signs of the elements of $\tilde{\mathbf{l}}_{\text{max}}$ are not of concern. Accordingly, for $\tilde{\mathbf{l}}_i$, only the absolute value needs to be plotted. Observations that have most unduly influence on parameter estimates and the model fit can then be identified by examining the relative influence of the elements in $\tilde{\mathbf{l}}_{\text{max}}$. If none of the observations has an undue impact on model inference, the plot of the elements of the *LMAX* scores should approximate a horizontal line.

6.5.3 Illustration: Checking influential observations in the Cox model on the mortality of older Americans

In the present illustration, I extend the example presented in Subsection 6.4.3. Specifically, I want to identify whether there are influential cases in terms of the overall fit of the Cox model on the mortality of older Americans, using the centered variables Age_mean, Female_mean, and Educ_mean as covariates. Both the *LD* and the *LMAX* statistics are applied for such identification, as against the rank order of survival times. Here, only the case influence on the overall model fit is considered, given the aforementioned advantage of using a single summary measure. Therefore, I do not examine the delta-beta statistic for each covariate. Additionally, I plot the elements of the *LMAX* statistic against values of a covariate – age, as recommended by Collett (2003) and Pattitt and Bin Raud (1989). The SAS program for generating the three *LD* and *LMAX* plots are displayed below.

SAS Program 6.7:

```
......

proc phreg data = new noprint ;
  model duration*Status(0) = age_mean female_mean educ_mean /  ties = BRESLOW ;
  output out = out_case LD = LD LMAX = LMAX ;
run;

Title "Figure 6.6. Case influence against duration based on LD";
proc sgplot data = out_case ;
  yaxis grid;
  refline 0 / axis = y ;
  scatter y = LD x = duration ;
run;

Title "Figure 6.7. Case influence against duration based on LMAX";
proc sgplot data = out_case ;
  yaxis grid;
  refline 0 / axis = y ;
  scatter y = LMAX x = duration ;
run;

Title "Figure 6.8. Case influence against age based on LMAX";
proc sgplot data = out_case ;
  yaxis grid;
  refline 0 / axis = y ;
  scatter y = LMAX x = age ;
run;
```

In SAS Program 6.7, I first specify the Cox model on the mortality of older Americans, with Age_mean, Female_mean, and Educ_mean as covariates. In the PROC PHREG procedure, an OUTPUT statement is added for saving information needed for creating the three plots. The keyword LD specifies the approximate likelihood displacement for each individual. Similarly, the keyword LMAX tells SAS to derive the score of relative influence on the overall fit of the Cox model, standardized to unit length. Those two keywords identify two new variables – *LD* and *LMAX* – and they are then output into the temporary data file OUT_CASE. Lastly, I use the PROC SGPLOT procedure to generate the three plots, specifying different variables for axis *y* and axis *x* in each graph.

The first plot, displaying the likelihood displacement scores against the rank order of survival times, is displayed in Figure 6.6, where numerous values of the *LD* statistic are plotted against the duration of survival times, actual or censored. As can be easily identified, there is an outstanding case located near the end of the observation period. Additionally, there are some other observations obviously deviating markedly from the vast majority of the cases, though distant from the most outstanding case. In total, however, the influences of those observations on the overall fit of the Cox model seem small, considering the limited range of values of the *LD* statistics. Even for the most influential case, the value of the *LD* statistic is less than 0.10. As indicated earlier, when sample size is large, such an *LD* approximate is supposed to agree closely with the exact *LD* statistic. At least, I do not need to create $n + 1$ Cox models to identify those influential observations.

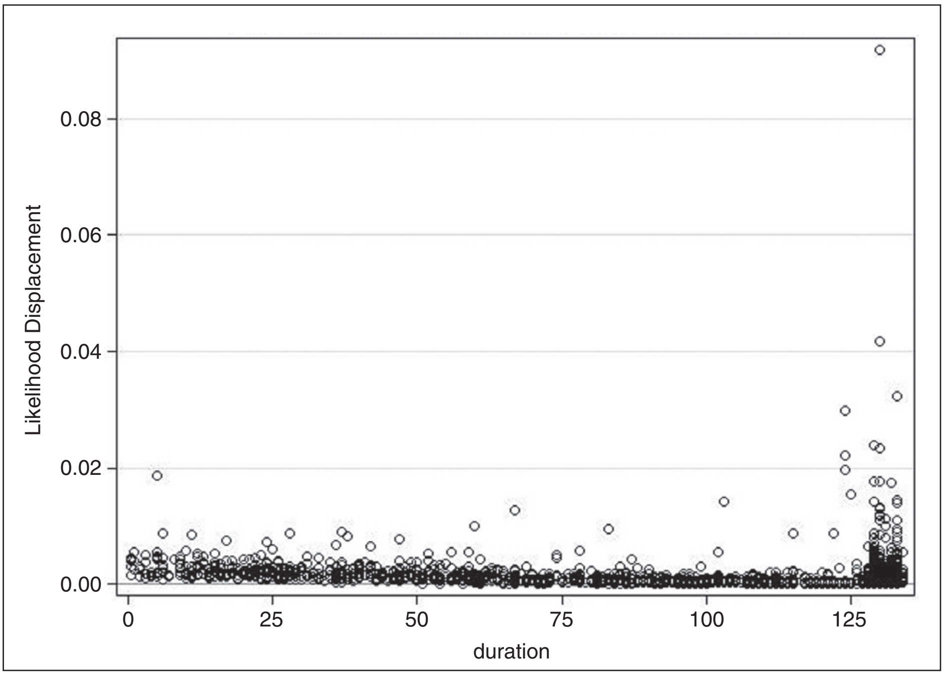


Figure 6.6 Case influence against duration based on LD.

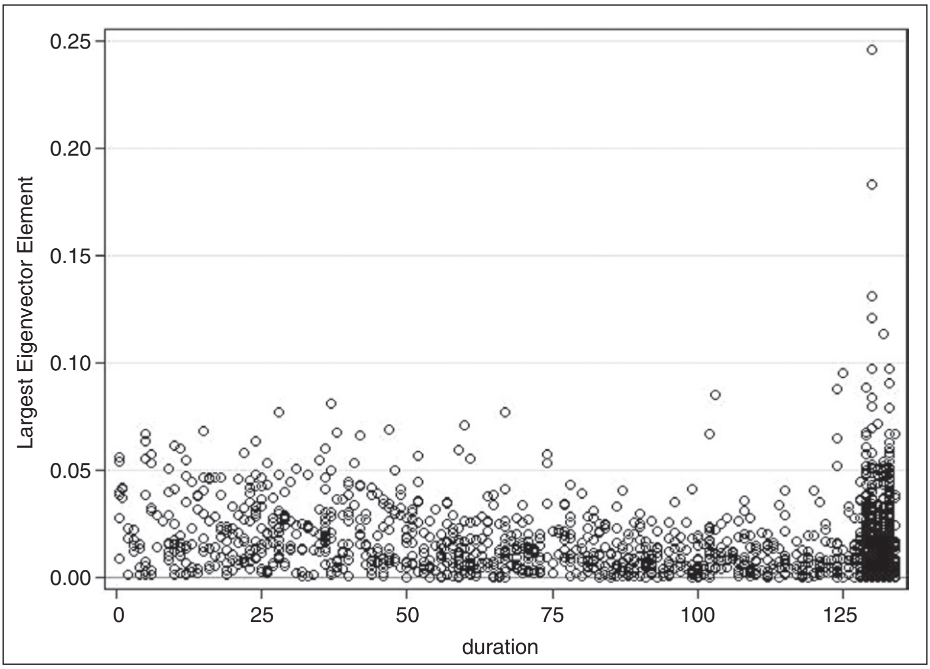


Figure 6.7 Case influence against duration based on LMAX.

Just looking at the *LD* approximate statistic, however, is not enough to derive sufficient confidence for making the decision that the null hypothesis $\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}}$ cannot be rejected. Given this reservation, the case influence on the overall fit of the Cox model is further examined by a plot of the *LMAX* statistics, standardized to the unit length of the total sum and also against the rank order of survival times.

Figure 6.7 displays the same pattern of case influences as shown in Figure 6.6. There are two outstanding cases with *LMAX* scores that have much higher scales than those of the others, both located near the ending limit of the observation period. As the square of an element in $\tilde{\mathbf{l}}_{\max}$ measures the proportion of the total sum of squares of unity, the influence of each of those two cases can be assessed by checking its proportional contribution to the unit length. For example, the square of the *LMAX* statistic for the most influential case is about 0.06 (0.25 × 0.25), indicating a considerable contribution of about 6 % to the overall fit of the Cox model. Considering a sample size of 2000 in this analysis, this case is shown to have a very significant relative influence. Likewise, for the second most outstanding case, the *LMAX* statistic is about 0.18, contributing about 3 % in the overall fit of the Cox model. Nevertheless, the high value of such a relative influence may not necessarily lead to a substantial change in the estimates of regression coefficients and standard errors, especially when the sample size is large. Therefore, the exact influences of those two most outstanding cases need to be checked before a firm conclusion can be reached. For this reason, those two observations need to be identified exactly.

It is useful to plot the elements of the *LMAX* scores against the value of a covariate for identifying the two most influential observations. I will display the third plot specified in SAS Program 6.7, a plot of the *LMAX* statistic against age, which can help us identify those two individuals.

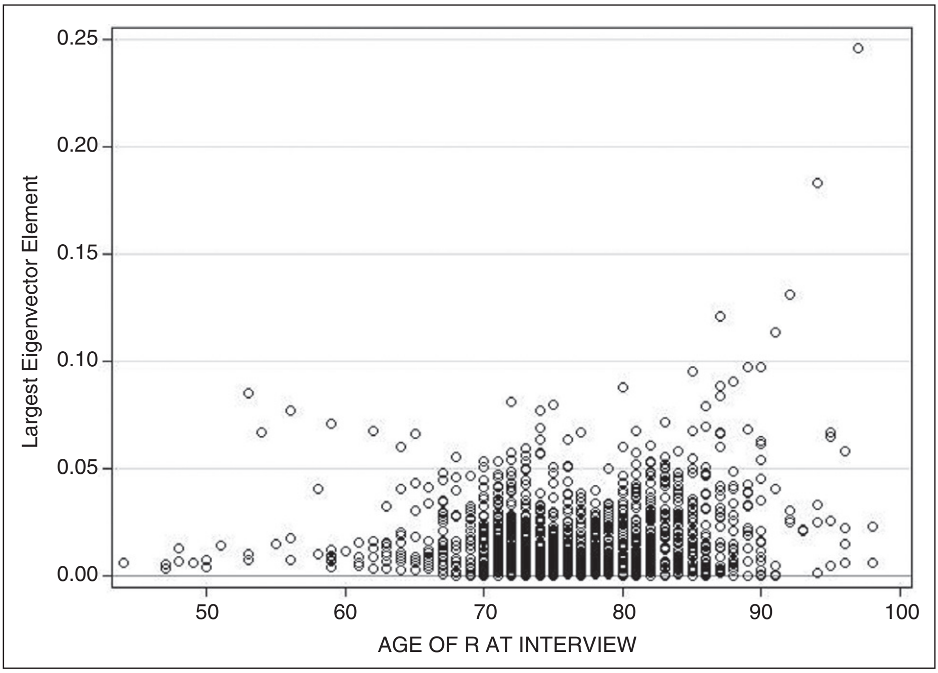


Figure 6.8 Case influence against age based on LMAX.

Figure 6.8 shows that the most influential cases are two individuals who are over 90 years of age at the baseline survey. After some additional graphical and numeric checks, those two observations are further identified. In particular, the most influential case is aged 97 years and the other one is aged 94, and both are female and right censored. With regard to educational attainment, one has 11 years of education and one has 13 years; thus both have education around the level of high school graduation. It can be inferred that compared to other cases, these two observations are so highly influential on the overall fit of the Cox model because both are expected to die early in the observation period at such old ages. In other words, it is their unexpected survival throughout the entire observation that yields very strong statistical impact on the inferential procedures.

Given the high proportional contribution of those two influential cases to the overall fit of the Cox model, it is necessary to check the exact likelihood displacement for both. Accordingly, I create two additional Cox models with each deleting one of those influential cases. The SAS program for this step is not presented here because the two models follow the standard procedures as previously exhibited, except removal of a single observation. Table 6.1 summarizes the results.

In Table 6.1, the first Cox model uses full data, with results previously reported. The second Cox model is fitted after removing the most influential case. The exact *LD* statistic, obtained from the formula $2\log L\left(\hat{\boldsymbol{\beta}}\right) - 2\log L\left[\hat{\boldsymbol{\beta}}_{(-i)}\right]$, is statistically significant with one degree of freedom and at $\alpha = 0.05$ ($LD = 6.15$, $p < 0.05$), indicating the most influential observation to make a strong statistical impact on the overall fit of the first Cox model. Likewise, the third Cox model is fitted after deleting the second most influential case, with the exact likelihood displacement statistic also statistically significant at the same criterion

Table 6.1 Maximum likelihood estimates and the likelihood displacement statistic for three Cox models.

Explanatory variable	Parameter estimate	Standard error	Chi-square	*p*-value	Hazard ratio
Cox model with full data (–2 LL = 13583.84; *p* < 0.0001)					
Age_mean	0.0823	0.0052	254.2103	<0.0001	1.086
Female_mean	–0.4431	0.0665	44.4090	<0.0001	0.642
Educ_mean	–0.0229	0.0088	6.8227	0.0090	0.977
Cox model deleting most influential case (–2 LL = 13577.69; *p* < 0.0001)					
Age_mean	0.0838	0.0052	257.6550	<0.0001	1.087
Female_mean	–0.4390	0.0665	43.6150	<0.0001	0.645
Educ_mean	–0.0224	0.0088	6.5178	0.0107	0.978
LD statistic	6.15 (*df* = 1; *p* < 0.05)				
Cox model deleting second most influential case (–2 LL = 13579.28; *p* < 0.0001)					
Age_mean	0.0833	0.0052	257.1688	<0.0001	1.087
Female_mean	–0.4399	0.0665	43.7917	<0.0001	0.654
Educ_mean	–0.0220	0.0088	6.4085	0.0114	0.978
LD statistic	4.56 (*df* = 1; *p* < 0.05)				

($LD = 4.56$, $p < 0.05$). In both the second and the third models, however, the parameter estimates, including the regression coefficients and the standard errors, do not vary noticeably at all after removing each of those two cases. The three sets of hazard ratios are almost identical. Obviously, deleting those two cases makes no genuine impact on the results of the model fit. Given such remarkable similarities of the estimated regression coefficients, I do not see any reason that the two influential cases should be eliminated from the Cox model, albeit their exceptionally strong statistical contributions. Indeed, for large samples, a strong relative influence of a few particular cases can be easily averaged out by the effects of the vast majority of the normal observations in the estimation process. Consequently, deleting any influential observation can hardly make an actual impact on the estimates of regression coefficients and standard errors.

6.6 Summary

In survival analysis, one of the most remarkable progresses in the past two decades is the application of counting processes, martingales in continuous time, and stochastic integration for the development of refined techniques. Attaching to this work, the martingale central limit theorem provides a strong theoretical foundation for verifying the efficiency and robustness of various statistical models based on counting processes. In this chapter, therefore, I first describe basic specifications of the counting process system, the martingale theory, and the stochastic integrated function as martingale transforms. The martingale central limit theorems are also presented. Additionally, given the close connection between counting processes and the partial likelihood perspective, I consider it helpful to respecify the Cox model as a stochastic counting process so that the reader can have a better comprehension of this popular regression model. It is worth noting here that in using this highly flexible counting system for developing advanced techniques, the score function, regarded as a martingale transform, plays an extremely important role in specifying various complex functions and distributions.

I perform the residuals analysis on the survival data of older Americans, describing five types of residuals widely used in survival analysis. Not surprisingly, these residuals do not display any signs of model inadequacy, as they are all scattered around zero, displaying a regular distributional pattern as shown by some other studies of this kind (Grambsch and Therneau, 1994; Schoenfeld, 1982; Therneau, Grambsch, and Fleming, 1990). Compared to linear regression models, however, these residuals do not behave explicitly enough to generate deterministic implications on the adequacy of the Cox model (Fleming and Harrington, 1991). Given the unobservable nature of the hazard function and the regular existence of heavy censoring, it is difficult to develop a residual that follows an unambiguous and known distribution. Thus, developing more efficient residuals in survival analysis remains a challenge to statisticians and other quantitative methodologists.

Compared to the residual-related diagnostic methods, the techniques for identification of influential observations on the overall fit of the Cox model are more mature. In Section 6.5, I described two popular approaches in this area, and interesting results are displayed in an empirical illustration. For large samples, the actual impact of influential cases is often found to be very limited, even though they may statistically affect the summary measure of the Cox model. The techniques for identifying the case influence are particularly useful for clinical trials and the observational studies characterized with a small sample size. Here, I recommend the following steps for identifying influential cases in survival analysis with small samples. First, identify the most influential observations by using the *LD* and *LMAX* approximations. Then examine the exact changes in the estimated regression coefficients and the overall model fit of the Cox model after deleting each of those cases. Following this strategy, a decision can be made regarding whether those cases should be removed in fitting the Cox model.