2
Probabilistic background

2.1 Preliminaries

Event time data, where one is interested in the time to a specific event occurs, are conveniently studied by the use of certain stochastic processes. The data itself may be described as a so-called counting process, which is simply a random function of time t, $N(t)$. It is zero at time zero and constant over time except that it jumps at each point in time where an event occurs, the jumps being of size 1.

Figure 2.1 shows two counting processes. Figure 2.1 (a) shows a counting process for survival data where one event is observed at time 7 at the time of death for a patient. Figure 2.1 (b) illustrates the counting process for recurrent events data where an event is observed multiple times, such as the times of dental cavities, with events at times 3, 4 and 7.

Why is this useful one could ask. Obviously, it is just a mathematical framework to represent timings of events, but a nice and useful theory has been developed for counting processes. A counting process $N(t)$ can be decomposed into a model part and a random noise part

$$N(t) = \Lambda(t) + M(t),$$

referred to as the compensator $\Lambda(t)$ and the martingale $M(t)$ of the counting process. These two parts are also functions of time and stochastic. The strength of this representation is that a central limit theorem is available for martingales. This in turn makes it possible to derive large sample properties of estimators for rather general nonparametric and semiparametric models

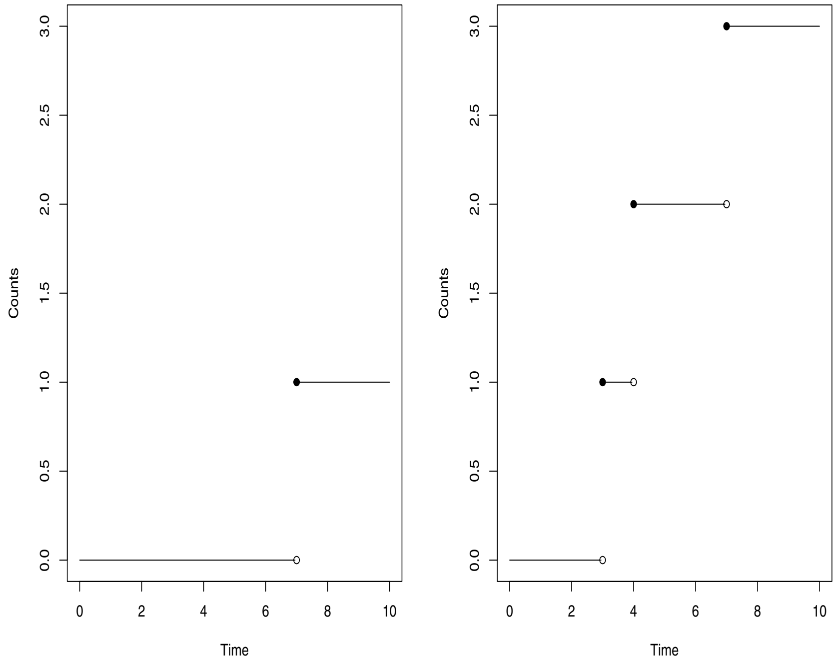

FIGURE 2.1: (a) Counting process for survival data with event time at time 7. (b) Counting process for recurrent events data, with event times at 3,4 and 7.

for such data. One chief example of this is the famous Cox model, which we return to in Section 6.1. To read more about counting processes and their theory we refer to Brémaud (1981), Jacobsen (1982), Fleming & Harrington (1991) and Andersen et al. (1993).

When assumptions are weakened, sometimes the decomposition will not result in an error term that is a martingale but only a zero-mean stochastic process, and in this case asymptotic properties can be developed using empirical process theory; see, for example, van der Vaart & Wellner (1996).

We shall also demonstrate that similar flexible models for longitudinal data may be studied fruitfully by the use of martingale methods. The key to this is that longitudinal data may be represented by a so-called marked point process, a generalization of a counting process. A marked point process is a mathematical construction to represent timing of events and their corresponding marks, and this is precisely the structure of longitudinal data where responses (marks) are collected over time. As for counting processes, a theory has been developed that decomposes a marked point process into a model part (compensator) and a random noise part (martingale). As a consequence of this, the analysis of longitudinal data therefore has many parallels with counting process data, and martingale methods may be invoked when studying large sample properties of concrete estimators. Some key references for additional reading about marked point processes are Brémaud (1981) and Last & Brandt (1995).

Before giving the definitions and properties of counting processes, marked point processes and martingales, we need to introduce some concepts from general stochastic process theory.

Behind all theory to be developed is a measurable space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-field and P is probability measure defined on $\mathcal{F}$. A *stochastic process* is a family of random variables indexed by time $(X(t) : t \geq 0)$. The mapping $t \to X(t, \omega)$, for $\omega \in \Omega$, is called a sample path. The stochastic process X induces a family of increasing sub-σ-fields by

$$\mathcal{F}_t^X = \sigma\{X(s) : 0 \leq s \leq t\}$$

called the *internal history* of X. Often when formulating models we will condition on events that occurred prior in time. We could for example, at time t, condition on the history generated by the process X up to time t. In many applications, however, we will need to condition on further information than that generated by only one stochastic process. To this end we therefore define more generally a *history* $(\mathcal{F}_t;\ t \geq 0)$ as a family of sub-σ-fields such that, for all $s \leq t$, $\mathcal{F}_s \subset \mathcal{F}_t$, which means $A \in \mathcal{F}_s$ implies $A \in \mathcal{F}_t$. A history is also called a *filtration*. Sometimes information (filtrations) are combined and for two filtrations $\mathcal{F}_t^1$ and $\mathcal{F}_t^2$ we let $\mathcal{F}_t^1 \vee \mathcal{F}_t^2$ denote the smallest filtration that contains both $\mathcal{F}_t^1$ and $\mathcal{F}_t^2$. A stochastic process X is *adapted* to a filtration $\mathcal{F}_t$ if, for every $t \geq 0$, $X(t)$ is $\mathcal{F}_t$-measurable, and in this case $\mathcal{F}_t^X \subset \mathcal{F}_t$. We shall often be dealing with stochastic processes with sample paths that, for almost all ω, are right-continuous and with left-hand limits. Such processes are called *cadlag* (continu à droite, limité à gauche). For a function f we define the right-hand limit $f(t+) = lim_{s \to t, s > t} f(s)$ and the left-hand limit $f(t-) = lim_{s \to t, s < t} f(s)$.

A nonnegative random variable T is called a *stopping time* with respect to $\mathcal{F}_t$ if $(T \leq t) \in \mathcal{F}_t$, for all $t \geq 0$. For a stochastic process X and a stopping time T, the stopped process X^T is defined by $X(t) = X(t \wedge T)$, where $a \wedge b$ denotes the minimum of a and b. A localizing sequence is a sequence of stopping times T_n that is nondecreasing and satisfies $T_n \to \infty$ as $n \to \infty$. A property of a stochastic process X is said to hold locally if there exists a localizing sequence (T_n) such that, for each n, the stopped process X^{T_n} has the property.

2.2 Martingales

Martingales play an important role in the statistical applications to be presented in this monograph. Often we shall see that estimating functions (evaluated at true parameter values) and the difference between estimators and true values are (up to a lower-order term) martingales. Owing to the existence of the celebrated central limit theorem for martingales of Rebolledo (1980), there is an elegant and simple approach to derive a complete asymptotic description of the suggested estimators. In the following we give the definition of a martingale.

A martingale with respect to a filtration $\mathcal{F}_t$ is a right-continuous stochastic process M with left-hand limits that, in addition to some technical conditions:

$$\text{(i) } M \text{ is adapted to } \mathcal{F}_t, \text{ and (ii) } E|M(t)| < \infty \text{ for all } t,$$

possesses the key martingale property

$$\text{(iii)} \quad \mathrm{E}(M(t) \,|\, \mathcal{F}_s) = M(s) \quad \text{for all } s \leq t, \tag{2.1}$$

thus stating that the mean of $M(t)$ given information up to time s is $M(s)$ or, equivalently,

$$\mathrm{E}(dM(t) \,|\, \mathcal{F}_{t-}) = 0 \quad \text{for all } t > 0, \tag{2.2}$$

where $\mathcal{F}_{t-}$ is the smallest σ-algebra containing all $\mathcal{F}_s$, $s < t$ and $dM(t) = M((t + dt)-) - M(t-)$. A martingale thus has zero-mean increments given the past, and without conditioning. Condition (ii) above is referred to as M being *integrable*. A martingale may be thought of as an error process in the following sense.

- Since $\mathrm{E}(M(t)) = \mathrm{E}(M(0))$, a martingale has constant mean as a function of time, and if the martingale is zero at time zero (as will be the case in our applications), the mean will be zero. Such a martingale is also called a zero-mean martingale.
- Martingales have uncorrelated increments, that is, for a martingale M it holds that

$$\mathrm{Cov}(M(t) - M(s), M(v) - M(u)) = 0 \tag{2.3}$$

for all $0 \leq s \leq t \leq u \leq v$.

If M satisfies

$$\mathrm{E}(M(t) \,|\, \mathcal{F}_s) \geq M(s) \quad \text{for all } s \leq t, \tag{2.4}$$

instead of condition (2.1), then M is a submartingale. A martingale is called *square integrable* if $\sup_t \mathrm{E}(M(t)^2) < \infty$. A *local martingale* M is a process such that there exist a localizing sequence of stopping times (T_n) such that, for each n, M^{T_n} is a martingale. If, in addition, M^{T_n} is a square integrable martingale, then M is said to be a *local square integrable martingale*.

To be able to formulate the crucial *Doob-Meyer* decomposition we need to introduce the notion of a predictable process. Loosely speaking, a predictable process is a process whose value at any time t is known just before t. Here is one characterization: a process X is predictable if and only if $X(T)$ is $\mathcal{F}_{T-}$-measurable for all stopping times T. The principal class of a predictable processes is the class of $\mathcal{F}_t$-adapted left-continuous processes.

Let X be a cadlag adapted process. Then A is said to be the *compensator* of X if A is a predictable, cadlag and *finite variation* process such that $X-A$ is a local zero-mean martingale. If a compensator exists, it is unique. A process A is said to be of finite variation if for all $t > 0$ (P-a.s.)

$$\int_0^t |dA(s)| = \sup_{\mathcal{D}} \sum_{i=1}^{K} |A(t_i) - A(t_{i-1})| < \infty,$$

where $\mathcal{D}$ ranges over all subdivisions of $[0,t]$: $0 = t_0 < t_1 < \cdots < t_K = t$.

One version of the Doob-Meyer decomposition as formulated in Andersen et al. (1993) is as follows.

Theorem 2.2.1 *The cadlag adapted process X has a compensator if and only if X is the difference of two local submartingales.*

An important simple consequence of the theorem is that, if the cadlag adapted process X is a local submartingale, then it has a compensator since the constant process 0 is a local submartingale.

Let M and $\tilde{M}$ be local square integrable martingales. By Jensen's inequality, M^2 is a local submartingale since

$$\mathrm{E}(M^2(t) \mid \mathcal{F}_s) \geq (\mathrm{E}(M(t) \mid \mathcal{F}_s))^2 = M^2(s)$$

and hence, by the Doob-Meyer decomposition, it has a compensator. This compensator is denoted by $\langle M, M\rangle$, or more compactly $\langle M\rangle$, and is termed the *predictable variation process* of M. By noting that $M\tilde{M} = \frac{1}{4}(M+\tilde{M})^2 - \frac{1}{4}(M-\tilde{M})^2$, it is similarly derived that $M\tilde{M}$ has a compensator, written $\langle M, \tilde{M}\rangle$, and termed the *predictable covariation process* of M and $\tilde{M}$.

The predictable covariation process is symmetric and bilinear like an ordinary covariance. If $\langle M, \tilde{M}\rangle = 0$, then M and $\tilde{M}$ are said to be *orthogonal*. The predictable covariation process is used to identify asymptotic covariances in the statistical applications to follow later on. This is partly explained by the relationship

$$\mathrm{Cov}(M(s), \tilde{M}(t)) = \mathrm{E}(\langle M, \tilde{M}\rangle)(t), \quad s \leq t. \tag{2.5}$$

Estimation of the asymptotic covariances on the other hand may be carried out by use of the quadratic covariation process. This process is defined even when M and $\tilde{M}$ are just local martingales. When M and $\tilde{M}$ further are of finite variation (as will be the case in our applications), the *quadratic covariation process* of M and $\tilde{M}$, denoted by $[M, \tilde{M}]$, has the explicit form

$$[M, \tilde{M}](t) = \sum_{s \le t} \Delta M(s) \Delta \tilde{M}(s). \tag{2.6}$$

In the case where $\tilde{M} = M$, (2.6) is written $[M](t)$ and called the *quadratic variation process* of M. The two processes $[M]$ and $[M, \tilde{M}]$ are also called the *optional variation process* and *optional covariation process*, respectively.

For the process $[M]$, it holds that $M^2 - [M]$ is a local martingale as was also the case with $\langle M \rangle$. An important distinction between the two processes, however, is that $[M]$ may not be predictable; in our applications it will never be! In the applications, the predictable variation process $\langle M \rangle$ will be determined by the model characteristics of the particular model studied while the quadratic variation process $[M]$ may be computed from the data at hand and therefore qualifies as a potential estimator.

Another useful characterization of $[M]$ is the following. When $[M]$ is locally integrable, then M will be locally square integrable and $\langle M \rangle$ will be the compensator of $[M]$! Similarly, $\langle M, \tilde{M} \rangle$ will be the compensator of $[M, \tilde{M}]$. This observation together with (2.6) enable us to compute both the quadratic and predictable covariation process.

In the statistical applications, stochastic integrals will come natural into play. Since we shall be dealing only with stochastic integrals where we integrate with respect to a finite variation process, all the considered stochastic integrals are ordinary pathwise Lebesgue-Stieltjes integrals, see Fleming & Harrington (1991) (Appendix A) for definitions. Of special interest are the integrals where we integrate with respect to a martingale. Such process integrals have nice properties as stated in the following.

Theorem 2.2.2 *Let M and $\tilde{M}$ be finite variation local square integrable martingales, and let H and K be locally bounded predictable processes. Then $\int H\,dM$ and $\int K\,d\tilde{M}$ are local square integrable martingales, and the quadratic and predictable covariation processes are*

$$\Big[\int H\,dM, \int K\,d\tilde{M} \Big] = \int HK\,d[M, \tilde{M}],$$
$$\Big\langle \int H\,dM, \int K\,d\tilde{M} \Big\rangle = \int HK\,d\langle M, \tilde{M} \rangle.$$

The quadratic and predictable variation processes of, for example, $\int H\,dM$ are seen to be

$$\Big[\int H\,dM \Big] = \int H^2\,d[M], \quad \Big\langle \int H\,dM \Big\rangle = \int H^2\,d\langle M \rangle.$$

The matrix versions of the above formulae for the quadratic and predictable covariation processes read

$$[\int H\,dM, \int K\,d\tilde{M}] = \int H\,d[M,\tilde{M}]K^T, \tag{2.7}$$

$$\langle \int H\,dM, \int K\,d\tilde{M}\rangle = \int H\,d\langle M,\tilde{M}\rangle K^T, \tag{2.8}$$

where M and $\tilde{M}$ are two vectors, and H and K are two matrices with dimensions such that the expressions make sense. In this case $[M,\tilde{M}]$ and $\langle M,\tilde{M}\rangle$ should be calculated componentwise.

2.3 Counting processes

Before giving the definition of a counting process we first describe one key example where counting processes have shown their usefulness.

Example 2.3.1 (Right-censored survival data)

Let T^* and C be two nonnegative, independent random variables. The random variable T^* denotes the time to the occurrence of some specific event. It could be time to death of an individual, time to blindness for a diabetic retinopathy patient or time to pregnancy for a couple. In many such studies the exact time T^* may never be observed because it may be censored at time C, that is, one only observes the minimum $T = T^* \wedge C$ of T^* and C, and whether it is the event or the censoring that has occurred, recorded by the indicator variable $\Delta = I(T^* \leq C)$. One simple type of censoring that is often encountered is that a study is closed at some point in time before all subjects have experienced the event of interest. In the counting process formulation the observed data (T, Δ) are replaced with the pair $(N(t), Y(t))$ of functions of time t, where $N(t) = I(T \leq t, \Delta = 1)$ is the counting process jumping at time T^* if $T^* \leq C$ (Figure 2.2), and $Y(t) = I(t \leq T)$ is the so-called *at risk indicator* being one at time t if neither the event nor the censoring has happened before time t. Assume that T^* has density f and let $S(t) = P(T^* > t)$ denote the *survival function*. A key concept in survival analysis is the *hazard function*

$$\alpha(t) = \frac{f(t)}{S(t)} = \lim_{h\downarrow 0} \frac{1}{h} P(t \leq T^* < t + h \,|\, T^* \geq t), \tag{2.9}$$

which may also be interpreted as the instantaneous failure rate. □

The formal definition of a counting process is as follows. A *counting process* $\{N(t)\}$ is stochastic process that is adapted to a filtration $(\mathcal{F}_t)$,

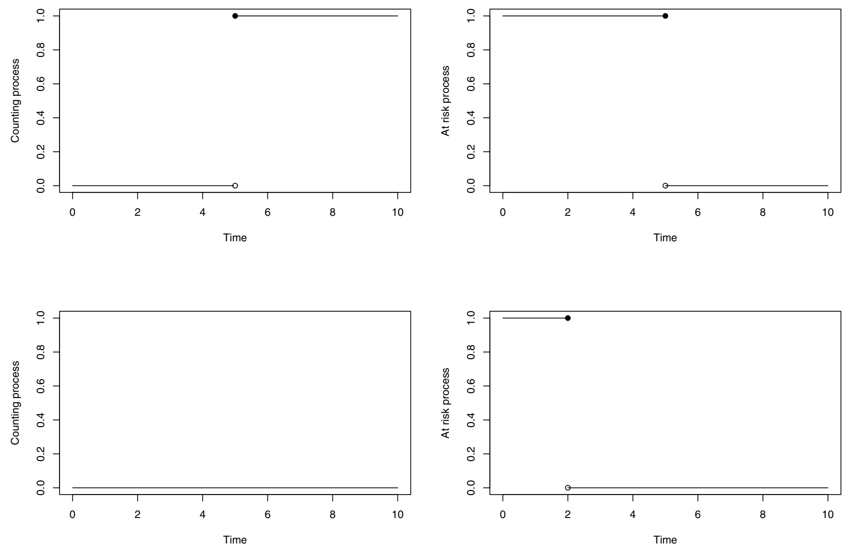

FIGURE 2.2: The counting process $N(t) = I(T \leq t, \Delta = 1)$ with $T^* = 5$ and $C > T^*$ (upper left panel) and corresponding at risk process (upper right panel). The counting process $N(t) = I(T \leq t, \Delta = 1)$ with $T^* > C$ and $C = 2$ (lower left panel) and corresponding at risk process (lower right panel).

cadlag, with $N(0) = 0$ and $N(t) < \infty$ a.s., and whose paths are piecewise constant with jumps of size 1.

A counting process N is a local submartingale and therefore has compensator, Λ, say. The process Λ is nondecreasing and predictable, zero at time zero, and such that

$$M = N - \Lambda$$

is a local martingale with respect to $\mathcal{F}_t$. In fact, M is a local square integrable martingale (Exercise 2.5). It also holds that

$$EN(t) = E\Lambda(t),$$

and further, if $E\Lambda(t) < \infty$, that M is a martingale (Exercise 2.6).

We shall only deal with the so-called *absolute continuous case*, where the above compensator has the special form

$$\Lambda(t) = \int_0^t \lambda(s)\, ds,$$

where the *intensity process* $\lambda(t)$ is a predictable process. The counting process N is then said to have intensity process λ.

By (2.6) is seen that the quadratic variation process of M is

$$[M] = N,$$

and, since it is locally integrable, the predictable variation process of M is

$$\langle M \rangle = \Lambda$$

by the uniqueness of the compensator.

A *multivariate counting process*

$$N = (N_1, \dots N_k)$$

is a vector of counting processes such that no two components jump simultaneously. It follows that

$$\langle M_j, M_{j'} \rangle = [M_j, M_{j'}] = 0, \quad j \neq j',$$

where the M_j's are the associated counting process martingales.

Example 2.3.2 (Continuation of Example 2.3.1)

Let the history be given by

$$\mathcal{F}_t = \sigma\{I(T \leq s, \Delta = 0), I(T \leq s, \Delta = 1) : s \leq t\}.$$

As noted above, the counting process $N(t)$ has a compensator $\Lambda(t)$. It turns out that the compensator is

$$\Lambda(t) = \int_0^t Y(s)\alpha(s)\, ds, \tag{2.10}$$

and hence that $N(t)$ has intensity process

$$\lambda(t) = Y(t)\alpha(t).$$

This may be shown rigorously, see for example Fleming & Harrington (1991). A heuristic proof of the martingale condition is as follows. Since (2.10) is clearly $\mathcal{F}_t$-adapted and left-continuous, it is predictable. By the independence of T^* and C, $dN(t)$ is a Bernoulli variable with conditional probability $Y(t)\alpha(t)\, dt$ of being one given $\mathcal{F}_{t-}$, see also Exercise 2.7. Thus,

$$E(dN(t) \mid \mathcal{F}_{t-}) = Y(t)\alpha(t)\, dt = d\Lambda(t) = E(d\Lambda(t) \mid \mathcal{F}_{t-}),$$

which justify the martingale condition (2.2) for $M = N - \Lambda$. □

Let us see how the decomposition of a counting process into its compensator and martingale parts may be used to construct estimators.

Example 2.3.3 (The Nelson-Aalen estimator)

Let (T_i^*, C_i), $i = 1, \ldots, n$, be n i.i.d. replicates from the model described in Example 2.3.2. Put $N_i(t) = I(T_i \le t, \Delta_i = 1)$ and $Y_i(t) = I(t \le T_i)$ with $T_i = T_i^* \wedge C_i$ and $\Delta_i = I(T_i^* \le C_i)$. Let $\mathcal{F}_t^i$ be defined similarly as $\mathcal{F}_t$ in Example 2.3.1 and 2.3.2 and let $\mathcal{F}_t = \bigvee_i \mathcal{F}_t^i$. Let further

$$N_\cdot(t) = \sum_{i=1}^n N_i(t), \quad Y_\cdot(t) = \sum_{i=1}^n Y_i(t).$$

The counting process $N_\cdot(t)$ is seen to have compensator

$$\Lambda(t) = \int_0^t Y_\cdot(s)\alpha(s)\, ds,$$

and, hence,

$$M_\cdot(t) = N_\cdot(t) - \Lambda(t)$$

is a local square integrable martingale with respect to $\mathcal{F}_t$. In the last display, $M_\cdot(t) = \sum_{i=1}^n M_i(t)$ with $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, \ldots, n$.

Now, decomposing the counting process into its compensator and a martingale term gives

$$N_\cdot(t) = \int_0^t Y_\cdot(s)\alpha(s)\, ds + M_\cdot(t)$$

and since $dM_\cdot(t)$ is a zero-mean process, this motivates the estimating equation

$$Y_\cdot(t) dA(t) = dN_\cdot(t),$$

where $A(t) = \int_0^t \alpha(s)\, ds$. This leads to the *Nelson-Aalen estimator*

$$\hat{A}(t) = \int_0^t \frac{J(s)}{Y_\cdot(s)} dN_\cdot(s) \tag{2.11}$$

of the integrated hazard function $A(t)$, where $J(t) = I(Y_\cdot(t) > 0)$, and where we use the convention that $0/0 = 0$. Notice that the Nelson-Aalen estimator is nothing but a simple sum:

$$\hat{A}(t) = \sum_{T_i \le t} \frac{\Delta_i}{Y_\cdot(T_i)}.$$

One may decompose $\hat{A}(t)$ as

$$\hat{A}(t) = \int_0^t J(s) dA(s) + \int_0^t \frac{J(s)}{Y_\cdot(s)} dM_\cdot(s).$$

By Theorem 2.2.2, it is seen that the second term on the right-hand side of the above decomposition is a local square integrable martingale. Thus, $\hat{A}(t)$ is an unbiased estimator of

$$\int_0^t \alpha(s) P(Y_\bullet(s) > 0)\, ds,$$

which already indicates that the Nelson-Aalen estimator will have sound large-sample properties (under appropriate conditions). One consequence of this is that $E(\hat{A}(t)) \leq A(t)$, and that the estimator will be close to unbiased if there are subjects at risk at all times with high probability.

As we shall see later on, a lot more than (asymptotical) unbiasedness can be said by use of the central limit theorem for martingales.

The Nelson-Aalen estimator may be formulated in the more general context of *multiplicative intensity models* where, for a counting process $N(t)$, it is assumed that the intensity process $\lambda(t)$ has a multiplicative structure

$$\lambda(t) = Y(t)\alpha(t),$$

where $\alpha(t)$ is a nonnegative deterministic function (being locally integrable) while $Y(t)$ is a locally bounded predictable process. The extension thus allows $Y(t)$ to be something else than an at risk indicator and is useful to deal with a number of different situations. The Nelson-Aalen estimator is then

$$\hat{A}(t) = \int_0^t \frac{J(s)}{Y(s)} dN(s),$$

where $J(t) = I(Y(t) > 0)$. The estimator $\hat{A}(t)$ was introduced for counting process models by Aalen (1975, 1978b) and it generalizes the estimator proposed by Nelson (1969, 1972). □

The concept of a filtration $\mathcal{F}_t$ may seem rather technical. It is important, however, as it corresponds to what information we are given, which in turn is used when specifying models. Sometimes we may be interested in conditioning on more information than that carried by $\mathcal{F}_t$. This additional information may give rise to a new filtration, $\mathcal{G}_t$ say, such that $\mathcal{F}_t \subseteq \mathcal{G}_t$, for all t. Assume that the counting process $N(t)$ is adapted to both $\mathcal{F}_t$ and $\mathcal{G}_t$, and that $N(t)$ has intensity $\lambda(t)$ with respect to $\mathcal{G}_t$. The intensity with respect to the smaller filtration $\mathcal{F}_t$ is then

$$\tilde{\lambda}(t) = \mathrm{E}(\lambda(t) \mid \mathcal{F}_{t-}), \tag{2.12}$$

which will generally be different from $\lambda(t)$ as we condition on less information. The above result is the so-called *innovation theorem*.

The following two examples of counting process models, illustrates how the innovation theorem can be used to adjust models to the amount of available information.

Example 2.3.4 (Clustered survival data)

Consider the situation where we are interested in studying the time to the occurrence of some event. Suppose in addition that there is some cluster structure in the data. An example could be the time to onset of blindness in patients with diabetic retinopathy, see Lin (1994). Patients were followed over several years and the pair of waiting times to blindness in the left and right eyes, respectively, were observed. In such a study one should expect some correlation between the waiting times within clusters. One approach to model such data is to use a random effects model, where the random effect accounts for possible (positive) correlation within the clusters. For ease of notation we describe the model in the situation where there is no censoring. Let T_{ik} denote the ith waiting time in the kth cluster, and put $N_{ik}(t) = I(T_{ik} \leq t)$, $Y_{ik}(t) = I(t \leq T_{ik})$, $i = 1, \ldots, n$, $k = 1, \ldots, K$. Assume that $T_k = (T_{ik}, \ldots, T_{nk})$, $k = 1, \ldots, K$ are i.i.d. random variables such that T_{ik} and T_{jk} $(i \neq j)$ are independent given the random effect Z_k. Let $\mathcal{F}_t^{ik}$ be the internal history of N_{ik}, $\mathcal{F}_t^k = \bigvee_i \mathcal{F}_t^{ik}$ and $\mathcal{F}_t = \bigvee_k \mathcal{F}_t^k$. The Clayton-Oakes model (Clayton (1978); Oakes (1982)) is obtained by assuming that $N_{ik}(t)$ has intensity

$$\lambda_{ik}(t) = Y_{ik}(t) Z_k \alpha(t)$$

with respect to the enlarged filtration $\mathcal{G}_t$, where

$$\mathcal{G}_t = \bigvee_k \mathcal{G}_t^k; \quad \mathcal{G}_t^k = \mathcal{F}_t^k \vee \sigma(Z_k);$$

and by assuming that the Z_k's are i.i.d. gamma distributed with expectation 1 and variance η^{-1}. The random effect Z_k is also referred to as a frailty variable, see Chapter 9. Besides carrying the information generated by the counting processes, $\mathcal{G}_t$ also holds the information generated by the random effects. The filtration $\mathcal{G}_t$ is not fully observed due to the unobserved random effects. The observed filtration is $\mathcal{F}_t$, and we now find the $\mathcal{F}_t$-intensities using the innovation theorem. One may show that

$$\mathrm{E}(Z_k \mid \mathcal{F}_{t-}) = \frac{\eta + N_{\bullet k}(t-)}{\eta + \int_0^t Y_{\bullet k}(s)\alpha(s)\,ds},$$

where $N_{\bullet k}(t) = \sum_{i=1}^n N_{ik}(t)$ and $Y_{\bullet k}(t) = \sum_{i=1}^n Y_{ik}(t)$, $k = 1 \ldots, K$. The $\mathcal{F}_t$-intensity is hence

$$\tilde{\lambda}_{ik}(t) = Y_{ik}(t)\left(\frac{\eta + N_{\bullet k}(t-)}{\eta + \int_0^t Y_{\bullet k}(s)\alpha(s)\,ds}\right)\alpha(t).$$

Estimation of $A(t) = \int_0^t \alpha(s)\,ds$ in this context may be carried out by use of the EM-algorithm, which was originally suggested by Gill (1985) and further developed by Klein (1992) and Nielsen et al. (1992), see also Andersen et al. (1993).

The above approach could be called *conditional* in the sense that the intensity of $N_{ik}(t)$ is modeled conditional on Z_k. An alternative approach that avoids joint modeling of data is the so-called *marginal approach* where the intensity of $N_{ik}(t)$ is only specified with respect to the marginal filtration $\mathcal{F}_t^{ik}$. It is assumed that $N_{ik}(t)$ has $\mathcal{F}_t^{ik}$-intensity

$$Y_{ik}(t)\alpha(t), \tag{2.13}$$

whereas it is *not* assumed that the $\mathcal{F}_t$-intensity is governed by (2.13) because that would correspond to assuming independence between subjects within each cluster, which obviously would be wrong with data like those mentioned in the beginning of this example. Estimation of $A(t)$ using the marginal approach is done by the usual Nelson-Aalen estimator ignoring the cluster structure of the data. Standard error estimates, however, should be computed differently. We return to clustered survival data in Chapter 9.

□

Example 2.3.5 (The additive hazards model and filtrations)

Consider the survival of a subject with covariates $X = (X_1, ..., X_p, X_{p+1})$ and assume that the corresponding counting process of the subject, $N(t)$, has intensity on the additive hazards form

$$\lambda^{p+1}(t) = Y(t)\left(\sum_{j=1}^{p+1} X_j\alpha_j(t)\right)$$

with respect to the history $\mathcal{F}_t^N \vee \sigma(p+1)$, where $\mathcal{F}_t^N$ is the internal history of N and $\sigma(i) = \sigma(X_1, ..., X_i)$ for $i = 1, ..., p+1$ the σ-fields generated by different sets of the covariates, In the above display, $Y(t)$ is an at risk indicator and $\alpha_j(t)$, $j = 1, ..., p+1$, are locally integrable deterministic unknown functions.

If only the p first covariates are known, or used, in the model the intensity changes, by the innovation theorem, to

$$\begin{aligned}\lambda^p(t) &= E(\lambda^{p+1}(t)|\mathcal{F}_t^N \vee \sigma(p)) \\ &= \sum_{i=1}^{p} Y(t)X_j\alpha_j(t) + Y(t)\alpha_{p+1}(t)E(X_{p+1}|Y(t) = 1, X_1, ..., X_p).\end{aligned}$$

The last conditional mean of X_{p+1} given that the subject is at risk (has survived), and the observed covariates can be computed (under regularity conditions) to be minus the derivative of $\log(f(t))$, where

$$f(t) = E(\exp(-\int_0^t \alpha_{p+1}(s)dsX_{p+1})|X_1, ..., X_p)$$

is the conditional Laplace transform of X_{p+1} evaluated at $\int \alpha_{p+1}$. Under certain assumptions, such as independence between the covariates, it is seen that the additive structure of the intensity is preserved, see Exercise 5.1. This example was given by Aalen (1989) □

2.4 Marked point processes

Later on we shall describe how nonparametric and semiparametric models for regression data and longitudinal data may be analyzed fruitfully by the use of martingale calculus. A key notion in this treatment is a generalization of counting processes, or point processes, to marked point processes, which will be introduced in the following. To a large extent we follow the exposition of marked point processes given by Brémaud (1981), see also the recent Last & Brandt (1995).

The idea is that instead of just recording the time points T_k at which specific events occur (as for the counting processes) we also observe an additional variable Z_k (the response variable in the longitudinal data setting) at each time point T_k. To make things precise we fix a measurable space $(E, \mathcal{E})$, called the mark space, and assume that

(i) $(Z_k, k \geq 1)$ is a sequence of random variables in E,

(ii) the sequence $(T_k, k \geq 1)$ constitutes a counting process

$$N(t) = \sum_k I(T_k \leq t).$$

The double sequence (T_k, Z_k) is called a *marked point process* with (Z_k) being the marks. To each $A \in \mathcal{E}$ is associated a counting process

$$N_t(A) = \sum_k I(T_k \leq t) I(Z_k \in A),$$

that counts the number of jumps before time t with marks in A. The marked point process is also identified with its associated counting measure defined by

$$p((0,t] \times A) = N_t(A), \qquad t > 0,\ A \in \mathcal{E}.$$

A marked point process counting measure thus accumulates information over time about the jump times and marks just as in the simpler counting process situation where there are no marks. Just as for counting processes it is also useful to consider integrals with respect to the marked point process. A *marked point process integral* has the following simple interpretation:

$$\int_0^t \int_E H(s,z) p(ds \times dz) = \sum_k H(T_k, Z_k) I(T_k \leq t).$$

The internal history of the marked point process is defined by

$$\mathcal{F}_t^p = \sigma(N_s(A) : 0 \leq s \leq t, A \in \mathcal{E}),$$

and we let $\mathcal{F}_t$ be any history of p, that is, $\mathcal{F}_t^p \subset \mathcal{F}_t$. If, for each $A \in \mathcal{E}$, $N_t(A)$ has intensity $\lambda_t(A)$ (predictable with respect to $\mathcal{F}_t$), we then say that $p(dt \times dz)$ admits the *intensity kernel* $\lambda_t(dz)$. We let $\lambda_t = \lambda_t(E)$ and assume that λ_t is locally integrable. A probability measure on $(E, \mathcal{E})$ is then defined by

$$F_t(dz) = \frac{\lambda_t(dz)}{\lambda_t}.$$

The pair $(\lambda_t, F_t(dz))$ is called the *local characteristics* of $p(dt \times dz)$. If the history $\mathcal{F}_t$ has the special form $\mathcal{F}_t = \mathcal{F}_0 \vee \mathcal{F}_t^p$, we have the following

$$F_{T_k}(A) = P(Z_k \in A \,|\, \mathcal{F}_{T_k-}) \quad \text{on } \{T_k < \infty\},$$

where

$$\mathcal{F}_{T_k-} = \sigma(T_j, Z_j; 1 \leq k-1; T_k)$$

is the history generated by the occurrence times and marks obtained before time T_k, and by T_k itself. The important above characterization of the second term of the local characteristics as the distribution of the current mark given past history and the time of the current mark is proved in Brémaud (1981).

Let $\mathcal{F}_t$ be a history of $p(dt \times dz)$ and let $\tilde{\mathcal{P}}(\mathcal{F}_t)$ be the history generated by the mappings

$$H(t, z) = C(t) 1_A(z),$$

where C is a $\mathcal{F}_t$-predictable process and $1_A(z)$ is the indicator of z being in A, $A \in \mathcal{E}$. Any mapping $H : (0, \infty) \times \Omega \times E \to \mathbf{R}$, which is $\tilde{\mathcal{P}}(\mathcal{F}_t)$-measurable is called an $\mathcal{F}_t$-predictable process indexed by E. Let p have intensity kernel $\lambda_t(dz)$ and let H be a $\mathcal{F}_t$-predictable process indexed by E. We shall now consider the measure

$$q(dt \times dz) = p(dt \times dz) - \lambda_t(dz)dt \tag{2.14}$$

obtained by compensating the marked point process measure by its intensity kernel. One may show, for all $t \geq 0$, that

$$M(t) = \int_0^t \int_E H(s, z) q(ds \times dz) \tag{2.15}$$

is a locally square integrable martingale (with respect to $\mathcal{F}_t$) if and only if

$$\int_0^t \int_E H^2(s, z) \lambda_s(dz) ds < \infty \qquad P-a.s.$$

We now turn to the computation of the quadratic variation and predictable variation process of M given by (2.15). Since (2.15) is of finite variation, the optional variation process is

$$[M](t) = \sum_{s \le t} \Delta M(s)^2 = \int_0^t \int_E H^2(s,z) p(ds \times dz).$$

The predictable variation process $\langle M \rangle$ is the compensator of $[M]$, and by the uniqueness of the compensator, we hence have

$$\langle M \rangle(t) = \int_0^t \int_E H^2(s,z) \lambda_s(dz) ds.$$

Let $p_1(dt \times dz)$ and $p_2(dt \times dz)$ be two marked point processes with intensity kernels $\lambda_t(dz)$ and $\mu_t(dz)$, respectively. Let H_j, $j = 1,2$, be $\mathcal{F}_t$-predictable processes indexed by E where $\mathcal{F}_t \supset \mathcal{F}_t^{p_1} \vee \mathcal{F}_t^{p_2}$, and assume that

$$\int_0^t \int_E H_1^2(s,z) \lambda_s(dz) ds < \infty, \qquad \int_0^t \int_E H_2^2(s,z) \mu_s(dz) ds < \infty \qquad P - a.s.$$

Write $M_j(t) = \int_0^t \int_E H_j(s,z) q_j(ds \times dz)$, $j = 1,2$. Assume that the two induced counting process, $N_1(t)$ and $N_2(t)$, have no jumps in common. Proceeding as above one may then derive that $[M_1, M_2] = 0$ and hence $\langle M_1, M_2 \rangle = 0$. Also,

$$[\int_0^t \int_E H_1 q_1(ds \times dz), \int_0^t \int_E H_2 q_1(ds \times dz)] = \int_0^t \int_E H_1 H_2 p_1(ds \times dz),$$

and

$$\langle \int_0^t \int_E H_1 q_1(ds \times dz), \int_0^t \int_E H_2 q_1(ds \times dz) \rangle = \int_0^t \int_E H_1 H_2 \lambda_s(dz) ds,$$

where the dependence of the integrands on s and z has been suppressed.

The following example illustrates how i.i.d. regression data may be put into the marked point process framework. Note how the techniques in the example closely parallel the similar development of the Nelson-Aalen estimator in the counting process setup.

Example 2.4.1 (Regression data)

Consider a sample (T_i, Z_i), $i = 1, \ldots, n$, of n i.i.d. regression data with Z_i being the (one-dimensional) response and T_i the (one-dimensional) regressor. Let

$$\mathrm{E}(Z_i \mid T_i = t) = \phi(t)$$

and assume that T_i has an absolute continuous distribution on $[0,\infty)$ with hazard function $\alpha(t)$. For simplicity we further assume for the moment that this distribution is known, that is, the hazard function is assumed to be known. Assume also that $\int_0^t \alpha(s)\phi(s)\,ds < \infty$ for all t. Each (T_i, Z_i) constitutes a marked point process p_i and with

$$\int_0^t \int_E z p_i(ds \times dz) = Z_i I(T_i \leq t),$$

we have the decomposition

$$\int_0^t \int_E z p_i(ds \times dz) = \int_0^t Y_i(s)\alpha(s)\phi(s)\,ds + \int_0^t \int_E z q_i(ds \times dz),$$

where the second term on the right-hand side of this display is a martingale with respect to the internal filtration $\mathcal{F}_t^{p_i}$. Writing the above equation in differential form and summing over all subjects gives

$$\sum_{i=1}^n \int_E z p_i(dt \times dz) = Y_\bullet(t)\alpha(t)\,d\Phi(t) + \sum_{i=1}^n \int_E z q_i(dt \times dz), \tag{2.16}$$

where $Y_\bullet(t) = \sum_{i=1}^n Y_i(t)$ and $\Phi(t) = \int_0^t \phi(s)\,ds$. Assume that $\inf_t \alpha(t) > 0$. Since α is known, (2.16) suggests the following estimator of $\Phi(t)$:

$$\begin{aligned}\hat{\Phi}(t) &= \sum_{i=1}^n \int_0^t \int_E \frac{z}{Y_\bullet(s)\alpha(s)} p_i(ds \times dz) \\ &= \sum_{i=1}^n \frac{Z_i}{Y_\bullet(T_i)\alpha(T_i)} I(T_i \leq t). \qquad (2.17)\end{aligned}$$

For this estimator we have

$$\hat{\Phi}(t) = \int_0^t J(s)d\Phi(s) + M(t),$$

where $J(t) = I(Y_\bullet(t) > 0)$ and

$$M(t) = \sum_{i=1}^n \int_0^t \int_E \frac{J(s)z}{Y_\bullet(s)\alpha(s)} q_i(ds \times dz),$$

which is seen to be a martingale with respect to the filtration spanned by all the $\mathcal{F}_t^{p_i}$'s. This implies that

$$E(\hat{\Phi}(t)) = \int_0^t P(Y_\bullet(t) > 0)d\Phi(s)$$

just as in the Nelson-Aalen estimator case. The estimator will thus be close to unbiased if there is a high probability that subjects are at risk at all times. If $\phi(t)$ is positive, then $E(\hat{\Phi}(t)) \leq \Phi(t)$. □

2.5 Large-sample results

As mentioned earlier, one of the strengths of representing the data as either a counting process or a marked point process is that we get martingales into play and that a central limit theorem for martingales is available. This theorem will be the main tool when we derive asymptotic results for concrete estimators. The theorem is stated below.

We shall consider a sequence of $\mathbb{R}^k$-valued local square integrable martingales $(M^{(n)}(t) : t \in \mathcal{T})$ with either

$$\mathcal{T} = [0, \infty) \text{ or } \mathcal{T} = [0, \tau]$$

with $\tau < \infty$. For $\epsilon > 0$, we let $M_\epsilon^{(n)}$ be the $\mathbb{R}^k$-valued local square integrable martingale where $M_{\epsilon l}^{(n)}$ contains all the jumps of $M_l^{(n)}$ larger in absolute value than ϵ, $l = 1, \ldots k$, i.e.,

$$M_{\epsilon l}^{(n)}(t) = \sum_{s \le t} \Delta M_l^{(n)}(s) I(|\Delta M_l^{(n)}(s)| > \epsilon), \quad l = 1, \ldots, k.$$

Note, that for counting process martingales of the form

$$\tilde{M}(t) = \int_0^t H(s) dM(s)$$

with $M(t) = N(t) - \Lambda(t)$ then

$$\tilde{M}_{\epsilon j}(t) = \sum_l \int_0^t H_{jl}(s) I(|H_{jl}(s)| > \epsilon) dM_l(s).$$

A *Gaussian martingale* is an $\mathbb{R}^k$-valued martingale U such that $U(0) = 0$ and the distribution of any finite family $(U(t_1), \ldots, U(t_j))$ is Gaussian. Write $V(t)$ for the variance-covariance matrix of $U(t)$. It follows that

(i) $\langle U \rangle(t) = V(t)$ for $t \ge 0$;

(ii) $V(t) - V(s)$ is positive semidefinite for $s \le t$;

(iii) $U(t) - U(s)$ is independent of $(U(r); r \le s)$ for $s \le t$.

A stochastic process U with the only requirement that is has continuous sample paths and normal distributed finite dimensional distributions is said to be a *Gaussian process.*

We may then state one form of the martingale central limit theorem.

Theorem 2.5.1 *(CLT for martingales). Let $(M^{(n)}(t) : t \in \mathcal{T})$ be a sequence of $\mathbb{R}^k$-valued local square integrable martingales. Assume that*

$$\langle M^{(n)} \rangle(t) \xrightarrow{P} V(t) \qquad \text{for all } t \in \mathcal{T} \text{ as } n \to \infty, \tag{2.18}$$

$$\langle M_{\epsilon l}^{(n)} \rangle(t) \xrightarrow{P} 0 \qquad \text{for all } t \in \mathcal{T},\ l \text{ and } \epsilon > 0 \text{ as } n \to \infty. \tag{2.19}$$

Then

$$M^{(n)} \xrightarrow{\mathcal{D}} U \text{ in } (D(\mathcal{T}))^k \text{ as } n \to \infty, \tag{2.20}$$

where U is a Gaussian martingale with covariance function V. Moreover, $\langle M^{(n)} \rangle$ and $[M^{(n)}]$ converge uniformly on compact subsets of $\mathcal{T}$, in probability, to V.

The theorem is due to Rebolledo (1980). The result (2.20) says that we have weak convergence of the process $M^{(n)}$ to U on the space $(D(\mathcal{T}))^k$ that consists of cadlag functions on $\mathcal{T}$ into $\mathbb{R}^k$ and is endowed with the Skorokhod topology, see e.g. Fleming & Harrington (1991) for definitions.

The condition (2.19) states that the jumps of $M^{(n)}$ should become negligible as $n \to \infty$ (see (2.25)), which makes sense if $M^{(n)}$ shall converge towards a process with continuous sample paths. Condition (2.18) says that the (predictable) variation process of $M^{(n)}$ becomes deterministic and approaches the variance function of the limit process as $n \to \infty$, which also makes sense in light of (2.5).

To illustrate the use of the martingale central limit theorem, we consider the Nelson-Aalen estimator (Example 2.3.3), and the i.i.d. regression set-up (Example 2.4.1).

Example 2.5.1 (The Nelson-Aalen estimator)

Consider the situation with n possibly right-censored survival times as described in Example 2.3.3. It was seen there that the Nelson-Aalen estimator of the cumulative hazard function $A(t) = \int_0^t \alpha(s)\,ds$ takes the form

$$\hat{A}(t) = \int_0^t \frac{J(s)}{Y_\bullet(s)} dN_\bullet(s),$$

where $J(t) = I(Y_\bullet(t) > 0)$,

$$N_\bullet(t) = \sum_{i=1}^n N_i(t), \quad Y_\bullet(t) = \sum_{i=1}^n Y_i(t),$$

with $N_i(t) = I(T_i \le t, \Delta_i = 1)$ and $Y_i(t) = I(t \le T_i)$, $i = 1, \dots, n$, the basic counting processes and the at risk indicators, respectively. With

$$A^*(t) = \int_0^t J(s) dA(s),$$

it was also seen that

$$\hat{A}(t) - A^*(t) = \int_0^t \frac{J(s)}{Y_\bullet(s)} dM_\bullet(s)$$

is a local square integrable martingale. Recall that $M_\bullet(t) = \sum_{i=1}^n M_i(t)$ with $M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha(s)\,ds$. By writing

$$n^{1/2}(\hat{A}(t) - A(t)) = n^{1/2}\big((A^*(t) - A(t)) + (\hat{A}(t) - A^*(t))\big)$$
$$= n^{1/2}\int_0^t (J(s)-1)\alpha(s)\,ds + n^{1/2}\int_0^t \frac{J(s)}{Y_\bullet(s)}dM_\bullet(s), \tag{2.21}$$

we see that under regularity conditions the asymptotic distribution of the Nelson-Aalen estimator on $[0,t]$, $t \in \mathcal{T}$ is a Gaussian martingale if it can be shown that the second term in (2.21) converges to a Gaussian martingale and that the first term in (2.21) converges to zero uniformly in probability.

We assume that $\int_0^t \alpha(s)\,ds < \infty$ for all $t \in \mathcal{T}$, and that there exists a function $y(s)$ such that

$$\sup_{s\in[0,t]} |n^{-1}Y_\bullet(s) - y(s)| \xrightarrow{P} 0; \qquad \inf_{s\in[0,t]} y(s) > 0. \tag{2.22}$$

It may now be shown that (Exercise 2.8)

$$\sup_{s\in[0,t]} |n^{1/2}\int_0^s (J(u)-1)\alpha(u)\,du| \xrightarrow{P} 0,$$

and we may hence concentrate on the martingale term

$$M(s) = n^{1/2}(\hat{A}(s) - A^*(s).$$

We see that, for $s \le t$,

$$\langle M\rangle(s) = \int_0^s \frac{J(u)}{n^{-1}Y_\bullet(u)}\alpha(u)\,du \xrightarrow{P} \int_0^s \frac{\alpha(u)}{y(u)}\,du$$

and

$$\langle M_\epsilon\rangle(s) = \int_0^s \frac{J(u)}{n^{-1}Y_\bullet(u)}\alpha(u)I\left(n^{1/2}\frac{J(u)}{Y_\bullet(u)} > \epsilon\right)du \xrightarrow{P} 0$$

(Exercise 2.8). Thus,

$$n^{1/2}(\hat{A}(s) - A(s)) \xrightarrow{\mathcal{D}} U(s)$$

in $\mathcal{D}[0,t]$, $t \in \mathcal{T}$, where U is a Gaussian martingale with variance function

$$V(s) = \int_0^s \frac{\alpha(u)}{y(u)}\,du.$$

Moreover, a uniformly consistent estimator of the variance function is given by the quadratic variation process

$$[M](s) = n\int_0^s \frac{J(u)}{(Y_\bullet(u))^2}dN_\bullet(u).$$

In the case of simple random censorship, that is, the C_i's are i.i.d. with distribution function F_C, say, (2.22) is fulfilled provided that $F_C(t-) < 1$, which says that the censoring must not be too heavy. In this case, $y(s) = (1 - F_{T*}(s))(1 - F_C(s))$, where F_{T*} denotes the distribution function of the survival times. □

Example 2.5.2 (Regression data)

Consider the i.i.d. regression setup of Example 2.4.1 where we observe i.i.d. regression data where

$$Z_i = \phi(T_i) + e_i$$

and the residual terms $e_1, \ldots, e_n$ are independent with zero mean such that $E(Z_i|T_i = t) = \phi(t)$. As noted there an estimator of $\Phi(t) = \int_0^t \phi(s)ds$ was given by

$$\hat{\Phi}(t) = \sum_{i=1}^n \int_0^t \int_E \frac{z}{Y_\bullet(s)\alpha(s)} p_i(ds \times dz),$$

which may be rewritten as

$$\hat{\Phi}(t) = \int_0^t J(s)d\Phi(s) + M(t),$$

where $J(t) = I(Y_\bullet(t) > 0)$ and

$$M(t) = \sum_{i=1}^n \int_0^t \int_E \frac{J(s)z}{Y_\bullet(s)\alpha(s)} q_i(ds \times dz),$$

the latter being a martingale with respect to the filtration spanned by all the $\mathcal{F}_t^{p_i}$'s. By imposing appropriate conditions we may show that

$$n^{1/2}(\hat{\Phi}(t) - \Phi(t)) = n^{1/2}M(t) + o_p(1),$$

uniformly in t, and the asymptotic distribution of $\hat{\Phi}(t)$ may hence be derived by use of the martingale central limit theorem. We have, for all $s \le t$,

$$\langle n^{1/2}M\rangle(s) = \int_0^s \frac{J(u)\psi(u)}{n^{-1}Y_\bullet(u)\alpha(u)}\, du \xrightarrow{P} \int_0^s \frac{\psi(u)}{y(u)\alpha(u)}\, du,$$

where

$$\psi(s) = \mathrm{E}(Z_i^2 \,|\, T_i = s)$$

and $y(s)$ is the limit in probability of $n^{-1}Y_\bullet(t)$ assuming that $\inf_{t\in\mathcal{T}} y(t) > 0$. Assume also that $\psi(t) < \infty$ for all t. The martingale containing the jumps of absolute size larger than ϵ is

$$(n^{1/2}M)_\epsilon(s) = n^{1/2} \sum_{i=1}^n \int_0^s \int_E \frac{J(u)|z|}{Y_\bullet(u)\alpha(u)} I\Bigl(n^{1/2}\frac{J(u)|z|}{Y_\bullet(u)\alpha(u)} > \epsilon\Bigr) q_i(du \times dz)$$

and hence

$$\langle M_\epsilon \rangle(s) = \int_0^s \frac{J(u)}{n^{-1}Y_\bullet(u)\alpha(u)} E\left(Z^2 I\left(n^{1/2}\frac{J(u)|Z|}{Y_\bullet(u)\alpha(u)} > \epsilon\right) | T = u\right) du \xrightarrow{P} 0,$$

Exercise 2.11. Thus,

$$n^{1/2}(\hat{\Phi}(s) - \Phi(s)) \xrightarrow{\mathcal{D}} U(s)$$

in $\mathcal{D}[0,t]$, $t > 0$, where U is a Gaussian martingale with variance function

$$V(s) = \int_0^s \frac{\psi(u)}{y(u)\alpha(u)} du.$$

A uniformly consistent estimator of the variance function is given by the quadratic variation process

$$\begin{aligned}[n^{1/2}M](s) &= n \sum_{i=1}^n \int_0^s \int_E \frac{J(u)z^2}{(Y_\bullet(u)\alpha(u))^2} p_i(du \times dz) \\ &= n \sum_{i=1}^n \frac{J(T_i)Z_i^2}{(Y_\bullet(T_i)\alpha(T_i))^2} I(T_i \le s).\end{aligned}$$

□

Once we have established convergence of our estimator as in the previous two examples, we can use their large-sample properties for hypothesis testing and construction of confidence bands and intervals. Consider, for example, the estimator $\hat{\Phi}(t)$ in the previous example that converged towards a Gaussian martingale $U(t)$. Suppose that the limit process $U(t)$ is $\mathbb{R}$-valued and has variance process $V(t)$. Then a $(1-\alpha)$ pointwise confidence interval for $\Phi(t) = \int_0^t \phi(s)ds$, for fixed t, is

$$\left[\hat{\Phi}(t) - c_{\alpha/2}\hat{\Sigma}(t)^{1/2}, \hat{\Phi}(t) + c_{\alpha/2}\hat{\Sigma}(t)^{1/2}\right]$$

where $n\hat{\Sigma}(t)$ is an (uniformly consistent) estimator of $V(t)$, like the one based on the quadratic variation process, and $c_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. Since we often will be interested in the behavior of $\phi(t)$, or, $\Phi(t)$, as function of t, inferences based on confidence bands may be more informative than pointwise confidence limits. One type of such confidence bands are the so-called Hall-Wellner bands (Hall & Wellner, 1980). These bands are uniform for some interval of interest that we here denote as $[0,\tau]$. Since

$$U(t)V(\tau)^{1/2}[V(\tau) + V(t)]^{-1}$$

is distributed as

$$B^0\left(\frac{V(t)}{V(\tau) + V(t)}\right),$$

where B^0 is the standard Brownian bridge (see Exercise 2.2), it follows that approximate $100(1-\alpha)\%$ confidence bands for $\Phi(t)$ are given by

$$\left[\hat{\Phi}(t) - d_\alpha \hat{\Sigma}(\tau)^{1/2}\left(1+\frac{\hat{\Sigma}(t)}{\hat{\Sigma}(\tau)}\right), \hat{\Phi}(t) + d_\alpha \hat{\Sigma}(\tau)^{1/2}\left(1+\frac{\hat{\Sigma}(t)}{\hat{\Sigma}(\tau)}\right)\right],$$

where d_α is the $(1-\alpha)$-quantile in the distribution of

$$\sup_{t\in[0,1/2]} |B^0(t)|,$$

see also Exercise 2.3. Tables of d_α may be found in Schumacher (1984); here we list some of the most used ones: $d_{0.01} = 1.55$, $d_{0.05} = 1.27$ and $d_{0.1} = 1.13$. Likewise, the hypothesis

$$\mathrm{H}_0: \ \phi(t) = \phi_0(t) \quad \text{for all } t$$

may be tested by use of a *Kolmogorov-Smirnov test* that rejects at level α if

$$\sup_{t\le\tau} |(\hat{\Phi}(t) - \Phi_0(t))\hat{\Sigma}(\tau)^{1/2}[\hat{\Sigma}(\tau) + \hat{\Sigma}(t)]^{-1}| \ge d_\alpha, \tag{2.23}$$

where $\Phi_0(t) = \int_0^t \phi_0(u)\,du$. The *Cramér-von Mises test* rejects at level α if

$$\int_0^\tau \left(\frac{(\hat{\Phi}(t) - \Phi_0(t))/\hat{\Sigma}^{1/2}(\tau)}{1+\hat{\Gamma}(t)}\right)^2 d\left(\frac{\hat{\Gamma}(t)}{1+\hat{\Gamma}(t)}\right) \ge e_\alpha \tag{2.24}$$

where e_α is the (1-α)-quantile in the distribution of $\int_0^{1/2} B^0(u)^2\,du$ and $\hat{\Gamma}(t) = \hat{\Sigma}(t)/\hat{\Sigma}(\tau)$. For reference: $e_{0.01} = 0.42$, $e_{0.05} = 0.25$ and $e_{0.1} = 0.19$; a detailed table of e_α may be found in Schumacher (1984).

Example 2.5.3

We here present a small simulation study to illustrate the use of confidence bands and the performance of the Kolmogorov-Smirnov and Cramér-von Mises tests. We generated data from the model described in Example 2.4.1 with T being exponential with mean one. The response variable is normal distributed with mean $\phi(t)$ and standard deviation 1/3. The true regression function is $\phi(t) = 1/(1+t)$ resulting in the cumulative regression function $\Phi(t) = \log(1+t)$. The sample size was first set to $n = 100$ and we then generated 500 datasets. Figure 2.3 (a) shows the true $\Phi(t)$ (thick full line), a randomly chosen estimate (thin dotted line) and the average of the 500 estimators (thick dotted line), which is almost indistinguishable from the true $\Phi(t)$. A slight bias is seen towards the end of the shown interval, which has upper limit equal to 4.6 corresponding to the 99%-quantile of the exponential distribution with mean one. According to the derived formulae

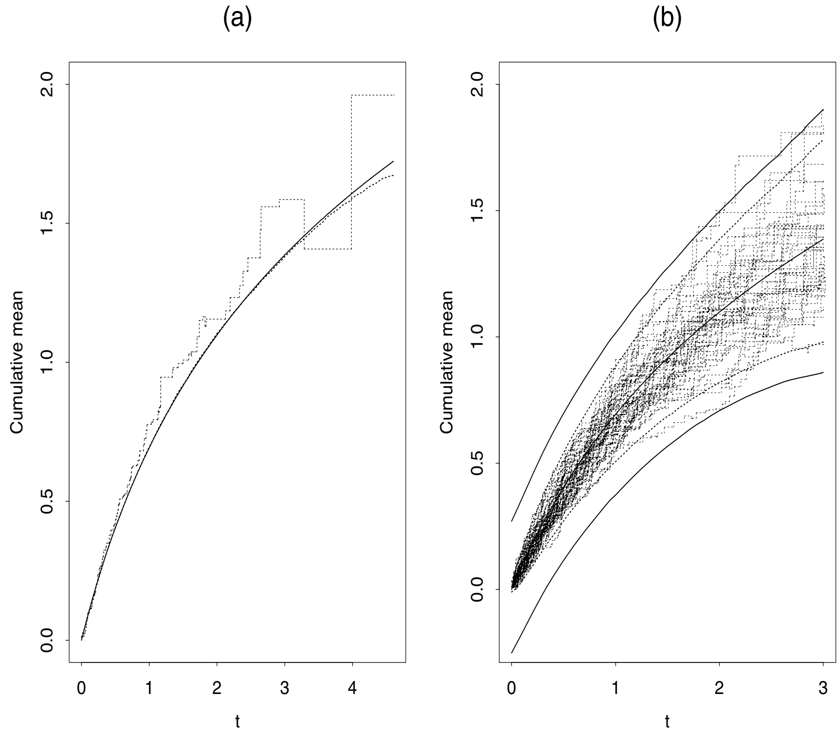

FIGURE 2.3: (a) True cumulative regression function $\Phi(t)$ (thick full line); average of 500 estimates of the cumulative regression function (thick dotted line); a typical estimator of the cumulative regression function (thin dotted line). (b) True cumulative regression function (thick full line) together with 95% pointwise confidence limits (thick dotted lines) and 95% Hall-Wellner confidence bands (thick full lines); and 40 randomly chosen estimates of the cumulative regression function (thin dotted lines).

this bias is due to the probability of being at risk towards the end of the interval deviating slightly from 1. We notice that the estimator $\hat{\hat{\Phi}}(t)$, which is given by (2.17), is a step function (like the Nelson-Aalen estimator) with jumps at the observed values of t. Figure 2.3 (b) shows the true $\Phi(t)$ (thick full line), 40 randomly chosen estimates (thin dotted lines), 95% pointwise confidence limits (thick dotted lines) and 95% Hall-Wellner bands (thick full lines) with $\tau = 3$, which corresponds to 95% quantile of the considered exponential distribution. We see that the estimators are contained within the confidence bands with the exception of one or two estimators.

We also look at the performance of the Kolmogorov-Smirnov test and the Cramér-von Mises test under the null. We generated data as described above with sample size $n = 100, 400$ and computed the rejection probabilities for the two tests at level $\alpha = 1\%, 5\%, 10\%$. These are shown in Table 2.1, where each entry is based on 10000 repetitions.

n	Test statistic	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
100	KS	0.9	3.3	6.8
	CM	1.0	5.1	10.2
400	KS	0.8	3.9	8.1
	CM	0.9	4.9	10.6

TABLE 2.1: Rejection probabilities for the Kolmogorov-Smirnov test (KS) and the Cramér-von Mises test (CM) computed at levels $\alpha = 1\%, 5\%, 10\%$

It is seen from Table 2.1 that the Cramér-von Mises test has the correct level already at sample size $n = 100$. The Kolmogorov-Smirnov test is somewhat conservative for $n = 100$ but approaches the correct level for $n = 400$. □

A useful result is the so-called *Lenglart's inequality*, see Andersen et al. (1993), which, in the special case of a local square integrable martingale M, says that

$$P(\sup_{[0,\tau]} |M| > \eta) \leq \frac{\delta}{\eta^2} + P(\langle M\rangle(\tau) > \delta) \tag{2.25}$$

for any $\eta > 0$ and $\delta > 0$. Hence $\sup_{[0,\tau]} |M| \xrightarrow{P} 0$ if $\langle M\rangle(\tau) \xrightarrow{P} 0$. A typical application of (2.25) is the following. Suppose that H_n is a sequence locally bounded predictable stochastic processes such that

$$\sup_{[0,\tau]} |H_n| \xrightarrow{P} 0,$$

and that M_n is a sequence of local square integrable martingales such that $\langle M_n\rangle(t) = O_p(1)$. We then have

$$\sup_{[0,\tau]} |\int_0^t H_n dM_n| \xrightarrow{P} 0, \tag{2.26}$$

since

$$\langle \int H_n dM_n\rangle(\tau) \xrightarrow{P} 0.$$

In some applications, however, we may not have that the H_n's are predictable. A useful result, due to Spiekerman & Lin (1998), says that (2.26) is still true provided that

$$\int_0^\tau |dH_n(t)| = O_p(1),$$

that is, H_n is of bounded variation. The result can be further relaxed by noticing that the proof of Spiekerman & Lin (1998) remains valid if M_n is some process that converges in distribution to some zero-mean process with continuous limits M. This extended version does not require any martingales, and is used in a couple of places in the proofs and is referred to as the Lemma by Spiekerman & Lin (1998); see also Lin et al. (2000) and Lin & Ying (2001).

Often we wish to conclude that

$$\int_0^t X^{(n)}(s)\,ds \xrightarrow{\text{P}} \int_0^t f(s)\,ds \quad \text{as } n \to \infty, \tag{2.27}$$

where we know that $X^{(n)}(t) \xrightarrow{\text{P}} f(t)$ for almost all $t \in [0, \tau]$ and $\int_0^\tau |f(t)|\,dt < \infty$. A result by Gill (1983) says that (2.27) holds uniformly in t if, for all $\delta > 0$, there exists a k_δ with $\int_0^\tau k_\delta(t)\,dt < \infty$ such that

$$\lim_{n\to\infty} \inf P(|X^{(n)}(s)| \le k_\delta(s) \text{ for all } s) \ge 1 - \delta. \tag{2.28}$$

We refer to (2.28) as *Gill's condition*.

A related dominated convergence theorem says that with $0 \le X_n(s) \le Y_n(s)$ for $s \in [0, \tau]$ and with ν a measure such that

$$Y_n(s) \xrightarrow{\text{P}} Y(s), \quad X_n(s) \xrightarrow{\text{P}} Y(s)$$

for ν almost all s and

$$\int Y_n(s) d\nu \xrightarrow{\text{P}} \int Y(s) d\nu < \infty \ (a.e)$$

then

$$\int X_n(s) d\nu \xrightarrow{\text{P}} \int X(s) d\nu.$$

The *delta-method* and its equivalent functional version are very useful for deriving the asymptotic distribution in the case where a function (functional) is applied to a random-vector (process) that converges in distribution.

The simple version states that if the p-dimensional random vector's X_n, X and fixed μ satisfy that

$$n^{1/2}(X_n - \mu) \xrightarrow{\mathcal{D}} X,$$

then if f is differentiable ($f : \mathbb{R}^p \to \mathbb{R}^q$) at μ with derivative $\dot{f}(\mu)$ (a $p \times q$ matrix function), it follows that

$$n^{1/2}(f(X_n) - f(\mu)) \xrightarrow{\mathcal{D}} \dot{f}(\mu) X.$$

This can be extended to functional spaces by the concept of Hadamard differentiability (Andersen et al., 1993). Consider the functional spaces $B =$

$D[0,\tau]^p$ and $B' = D[0,\tau]^q$ and let $f : B \to B'$ with derivative $\dot{f}(\mu)$ at μ (a continuous linear map, $\dot{f}(\mu) : B \to B'$) such that

$$a_n(f(\mu + a_n^{-1}h_n) - f(\mu)) \to \dot{f}(\mu) \cdot h$$

for all real sequences $a_n \to \infty$ and all convergent sequences $h_n \to h$ in B. The mapping f is then said to be Hadamard differentiable at μ. If X_n and X are processes in B, μ is a fixed point in B and f is Hadamard differentiable at μ, it then follows that

$$n^{1/2}(f(X_n) - f(\mu)) \xrightarrow{\mathcal{D}} \dot{f}(\mu) \cdot X.$$

The *functional delta theorem* can obviously be defined for all Banach spaces and one typical application is one where the p-dimensional process B_n and the q-dimensional vector θ_n jointly converge such that

$$n^{1/2}(B_n - b, \theta_n - \mu) \xrightarrow{\mathcal{D}} (X_1, X_2)$$

and then

$$n^{1/2}(f(B_n, \theta_n) - f(b, \mu)) \xrightarrow{\mathcal{D}} \dot{f}(b, \mu) \cdot (X_1, X_2)$$

for differentiable f.

We close this section by briefly mentioning the *conditional multiplier central limit theorem*. Suppose that $X_1, \cdots, X_n$ are i.i.d. real-valued random variables and $G_1, \cdots, G_n$ are independent standard normals independent of $X_1, \cdots, X_n$. Then if

$$n^{-1/2} \sum_{i=1}^{n} X_i \xrightarrow{\mathcal{D}} U$$

it follows from the *conditional multiplier central limit theorem* that also

$$n^{-1/2} \sum_{i=1}^{n} G_i X_i \xrightarrow{\mathcal{D}} U,$$

under suitably conditions (van der Vaart & Wellner, 1996) given almost every sequence of $X_1, \cdots, X_n$.

One practical use of this is that when X_i are the residuals from some regression model then it will often also be true that

$$n^{-1/2} \sum_{i=1}^{n} G_i \hat{X}_i \xrightarrow{\mathcal{D}} U,$$

where $\hat{X}_i$ are estimated based on the data, and this result can also be expanded to functional cases where for example X_i is a residual process on $D[0,\tau]$. We will use this approach to approximate the asymptotic distribution for many estimators as suggested in the counting process situation by Lin et al. (1993).

2.6 Exercises

2.1 (Poisson process) A Poisson process $N(t)$ with intensity $\lambda(t)$ is a counting process with

- independent increments and such that
- $N(t) - N(s)$ follows a Poisson distribution with parameter $\int_s^t \lambda(u)\,du$ for all $0 \leq s \leq t$.

(a) Find the compensator Λ of N and put $M = N - \Lambda$. Show by a direct calculation that $E(M(t) \mid \mathcal{F}_s) = M(s)$, where $\mathcal{F}_t$ is the internal history N. Is M a local square integrable martingale?

(b) Find the compensator of M^2.

2.2 (Brownian motion and Brownian bridge) The Brownian motion or the Wiener process is the Gaussian process B such that $\mathrm{E}B(t) = 0$ and $\mathrm{Cov}(B(s), B(t)) = s \wedge t$ for $s, t \geq 0$.

(a) Show that B has independent increments. Show that B is a martingale and find the compensator of B^2.

The Brownian bridge (tied down Wiener process) $B^0(t)$ with $t \in [0, 1]$ is the Gaussian process such that $\mathrm{E}B^0(t) = 0$ and $\mathrm{Cov}(B^0(s), B^0(t)) = s(1 - t)$ for $0 \leq s \leq t \leq 1$.

(b) Show that the processes $B^0(t)$ and $B(t) - tB(1)$ have the same distribution on $[0, 1]$.

(c) Show that the processes $B(t)$ and $(1 + t)B^0(t/(1 + t))$ have the same distribution on $[0, \infty)$.

2.3 (Hall-Wellner bands) Consider the time interval $[0, \tau]$. Let $U(t)$ be a Gaussian martingale with covariance process $V(t)$, $t \in [0, \tau]$. Show that

$$U(t)V(\tau)^{1/2}[V(\tau) + V(t)]^{-1}$$

has the same distribution as

$$B^0\left(\frac{V(t)}{V(\tau) + V(t)}\right),$$

where B^0 is the standard Brownian bridge.

2.4 Let M_1 and M_2 be the martingales associated with the components of the multivariate counting process $N = (N_1, N_2)$ with continuous compensators. Show that

$$\langle M_1, M_2 \rangle = [M_1, M_2] = 0.$$

2.5 Let $M = N - \Lambda$ be the counting process local martingale. It may be shown that Λ is locally bounded Meyer (1976), Theorem IV.12.

(a) Show that N is a local submartingale with localizing sequence

$$T_n = n \wedge \sup\{t : N(t) < n\}.$$

(b) Show that M is a local square integrable martingale using the below cited optional stopping theorem.

Theorem. *Let M be a $\mathcal{F}_t$-martingale and let T be an $\mathcal{F}_t$-stopping time. Then $(M(t \wedge T) : t \geq 0)$ is a martingale.*

2.6 Let $M = N - \Lambda$ be the counting process local martingale.

(a) Show that $\mathrm{E}N(t) = \mathrm{E}\Lambda(t)$ (hint: use the monotone convergence theorem).

(b) If $\mathrm{E}\Lambda(t) < \infty$, then show that M is a martingale by verifying the martingale conditions.

(c) If $\sup_t \mathrm{E}\Lambda(t) < \infty$, then show that M is a square integrable martingale.

2.7 Let $N(t) = (N_1(t), \ldots, N_k(t))$, $t \in [0, \tau]$, be a multivariate counting process with respect to $\mathcal{F}_t$. It holds that the intensity

$$\lambda(t) = (\lambda_1(t), \ldots, \lambda_k(t))$$

of $N(t)$ is given (heuristically) as

$$\lambda_h(t) = P(dN_h(t) = 1 \mid \mathcal{F}_{t-}), \tag{2.29}$$

where $dN_h(t) = N_h((t+dt)-) - N_h(t-)$ is the change in N_h over the small time interval $[t, t+dt)$.

(a) Let T^* be a lifetime with hazard $\alpha(t)$ and define $N(t) = I(T^* \leq t)$. Use the above (2.29) to show that the intensity of $N(t)$ with respect to the history $\sigma\{N(s) : s \leq t\}$ is

$$\lambda(t) = I(t \leq T^*)\alpha(t).$$

(b) Let T^* be a lifetime with hazard $\alpha(t)$ that may be right-censored at time C. We assume that T^* and C are independent. Let $T = T^* \wedge C$, $\Delta = I(T^* \leq C)$ and $N(t) = I(T \leq t, \Delta = 1)$. Use the above (2.29) to show that the intensity of $N(t)$ with respect to the history

$$\sigma\{I(T \leq s, \Delta = 0), I(T \leq s, \Delta = 1) : s \leq t\}$$

is

$$\lambda(t) = I(t \leq T)\alpha(t).$$

2.8 Let $M(s)$ and $M_\epsilon(s)$ denote the martingales introduced in Example 2.5.1.

(a) Verify the expressions for $\langle M \rangle(s)$, $[M](s)$ and $\langle M_\epsilon \rangle(s)$ given in that example and show that they converge in probability as $n \to \infty$ verifying Gill's condition (2.28).

(b) From the same example, show that:

$$\sup_{s \in [0,t]} |n^{1/2} \int_0^s (J(u) - 1)\alpha(u)\, du| \xrightarrow{P} 0.$$

2.9 (Asymptotic results for the Nelson-Aalen estimator) Let $N^{(n)}(t)$ be a counting process satisfying the multiplicative intensity structure $\lambda(t) = Y^{(n)}(t)\alpha(t)$ with $\alpha(t)$ being locally integrable. The Nelson-Aalen estimator of $\int_0^t \alpha(s)\, ds$ is

$$\hat{A}^{(n)}(t) = \int \frac{1}{Y^{(n)}(s)} dN^{(n)}(s).$$

Define $A^*(t) = \int_0^t J^{(n)}(s)\alpha(s)\, ds$ where $J^{(n)}(t) = I(Y^{(n)}(t) > 0)$.

(a) Show that $A^{(n)}(t) - A^*(t)$ is a local square integrable martingale.

(b) Show that, as $n \to \infty$

$$\sup_{s \leq t} |\hat{A}^{(n)}(t) - A(t)| \xrightarrow{P} 0$$

provided that

$$\int_0^t \frac{J^{(n)}(s)}{Y^{(n)}(s)}\alpha(s)\, ds \xrightarrow{P} 0 \quad \text{and} \quad \int_0^t (1 - J^{(n)}(s))\alpha(s)\, ds \xrightarrow{P} 0,$$

as $n \to \infty$.

(c) Show that the two conditions given in (b) are satisfied provided that

$$\inf_{s \le t} Y^{(n)}(t) \xrightarrow{P} \infty, \quad \text{as } n \to \infty.$$

Define $\sigma^2(s) = \int_0^s \frac{\alpha(u)}{y(u)}\, du$, where y is a non-negative function so that α/y is integrable over $[0,t]$.

(d) Let $n \to \infty$. If, for all $\epsilon > 0$,

$$n \int_0^s \frac{J^{(n)}(u)}{Y^{(n)}(u)} \alpha(u) I\left(\left| n^{1/2} \frac{J^{(n)}(u)}{Y^{(n)}(u)} \right| > \epsilon\right) du \xrightarrow{P} 0,$$

$$n^{1/2} \int_0^s (1 - J^{(n)}(u))\alpha(u)\, du \xrightarrow{P} 0 \text{ and } n \int_0^s \frac{J^{(n)}(u)}{Y^{(n)}(u)} \alpha(u)\, du \xrightarrow{P} \sigma^2(s)$$

for all $s \le t$, then show that

$$n^{1/2}(\hat{A}^{(n)} - A) \xrightarrow{\mathcal{D}} U$$

on $D[0,t]$, where U is a Gaussian martingale with variance function σ^2.

2.10 (Right-censoring by the same stochastic variable) Let $T_1^*, \ldots, T_n^*$ be n i.i.d. positive stochastic variables with hazard function $\alpha(t)$. The observed data consist of $(T_i, \Delta_i)_{i=1,\ldots,n}$, where $T_i = T_i^* \wedge U$, $\Delta_i = I(T_i = T_i^*)$. Here, U is a positive stochastic variable with hazard function $\mu(t)$, and assumed independent of the T_i^*'s. Define

$$N_\bullet(t) = \sum_{i=1}^n N_i(t), \quad Y_\bullet(t) = \sum_{i=1}^n Y_i(t)$$

with $N_i(t) = I(T_i \le t, \Delta_i = 1)$ and $Y_i(t) = I(t \le T_i)$, $i = 1, \ldots, n$.

(a) Show that $\hat{A}(t) - A^*(t)$ is a martingale, where

$$\hat{A}(t) = \int_0^t \frac{1}{Y_\bullet(s)} dN_\bullet(s), \quad A^*(t) = \int_0^t J(s)\alpha(s)\, ds.$$

(b) Show that

$$\sup_{s \le t} |\hat{A}(s) - A^*(s)| \xrightarrow{P} 0$$

if $P(T_i \le t) > 0$.

(c) Is it also true that $\hat{A}(t) - A(t) \xrightarrow{P} 0$?

2.11 Consider again Example 2.5.2.

(a) Verify the expressions for $(n^{1/2}M)_\epsilon(s)$ and $\langle M_\epsilon \rangle(s)$.

(b) Show that $\langle M_\epsilon \rangle(s) \xrightarrow{P} 0$ using Gill's condition and that

$$\lim_{n\to\infty} \int_{A_n} X dP = 0,$$

where X is a random variable with $E|X| < \infty$, A_n is measurable and $A_n \searrow \emptyset$.

2.12 (Simulations from Example 2.5.3**)** Consider the simulations in Example 2.5.3. Work out the asymptotic bias for the simulations as a function of time and compare with Figure 2.3.

2.13 (Counting process with discrete compensator**)** Let N be a counting process with compensator Λ that may have jumps. Put $M = N - \Lambda$.

(a) Show by a direct calculation that

$$[M](t) = N(t) - 2\int_0^t \Delta\Lambda(s) dN(s) + \int_0^t \Delta\Lambda(s) d\Lambda(s),$$

where $\Delta\Lambda(t)$ denotes the jumps of $\Lambda(t)$.

(b) Show that

$$\langle M \rangle(t) = \Lambda(t) - \int_0^t \Delta\Lambda(s) d\Lambda(s).$$

Dynamic Regression Models for Survival Data
Martinussen, T.; Scheike, Th.H.
2006, XIV, 470 p., Hardcover
ISBN: 978-0-387-20274-7