Survival models and health sequences

Walter Dempsey

University of Michigan

July 27, 2015

Survival Data

- Survival data is commonplace in medical studies, consisting of failure time information for each patient i, T_i.
- Many studies require patient monitoring, generating a series of measurements of a health process, Y_i.

A complete uncensored observation for one patient consists of the following triple:

$$(Y, T, \mathbf{t})$$

- $Y \sim$ health process (ex. CD4 cell count, Prothrombin Index, Quality of Life Index)
- $T \sim$ failure time as measured from recruitment
- $\mathbf{t} \sim$ appointment schedule

Joint modeling of the repeated measurements and survival data is necessary in order to gauge the effect of covariates on the response.

Survival Process

Survival Process

A stochastic process $Y_i(t) : \mathbb{R} \to \mathcal{R}$ where $\mathcal{R}$ is the state space.

- Typically $Y_i(t)$ is state of health or quality of life of patient i at time t, typically measured since recruitment
  - ex. Simple Survival Process, state space is $\mathcal{R} = \{0, 1\}$ is sufficient

Survival Process

Survival Process

A stochastic process $Y_i(t) : \mathbb{R} \to \mathcal{R}$ where $\mathcal{R}$ is the state space.

- Typically $Y_i(t)$ is state of health or quality of life of patient i at time t, typically measured since recruitment
  - ex. Simple Survival Process, state space is $\mathcal{R} = \{0, 1\}$ is sufficient

Flatlining

An absorbing state $\flat \in \mathcal{R}$ such that $Y(t) = \flat \Rightarrow Y(t') = \flat$ for all $t' > t$.
The survival time, T, is the time to failure:

$$T_i = \sup\{t : Y_i(t) \neq \flat\}$$

Appointment Schedule

The **appointment schedule** is a random subset $\mathbf{t} \subset [0, T)$, which is informative for survival regardless of health trajectory

Sequential Conditional Independence

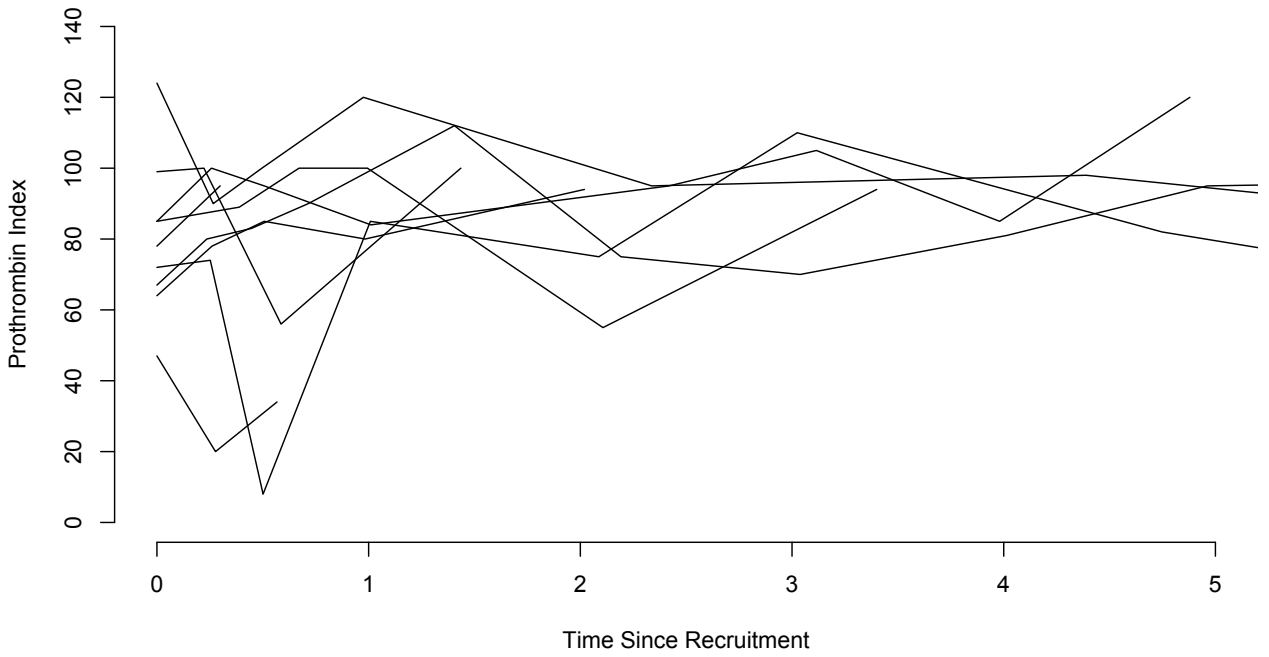
Consider patient with k appointments $\mathbf{t}^{(k)} = (t_0 < \ldots < t_{k-1})$. The sequence $Y[\mathbf{t}^{(k)}]$ may affect the scheduled appointment date t_k. The *sequential conditional independence* assumption states

$$t_k \perp\!\!\!\perp Y \mid (T, \mathbf{t}^{(k)}, Y[\mathbf{t}^{(k)}]). \tag{1}$$

(Example) Half of patients had appointment within the last three weeks of life $\rightarrow$ unclear if assumption is violated by patient-initiated appointments (71 within 10 days prior to death).

Motivating Example : Prednisone Case Study

- From 1962 to 1969, 488 patients with histologically verified liver cirrhosis at hospitals in Copenhagen were randomly assigned to the two treatment arms.
- 251 and 237 patients in the prednisone and placebo treatment groups respectively
  - 292 uncensored records $\Rightarrow$ 40% patients have censored records (45% of total measurements correspond to censored patient records)
- The purpose of the trial was to ascertain whether prednisone prolonged survival for patients with cirrhosis.



Motivating Example : Prednisone Case Study

Table 1 shows average Y-values by T and t. The cell averages 8–266 non-independent, highly variable measurements.

Table 1 : Average prothrombin levels indexed by T and t

Survival time (T)	Time t after recruitment (yrs)								
	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8+
0–1	58.0								
1–2	72.5	66.4							
2–3	72.6	73.2	66.0						
3–4	69.8	71.2	68.5	54.2					
4–5	68.5	75.7	72.5	74.6	57.7				
5–6	70.5	77.3	73.5	57.1	64.5	60.9			
6–7	81.8	73.6	81.1	80.6	79.4	75.5	75.8		
7–8	84.4	88.8	88.1	92.1	85.2	81.2	84.3	88.1	
8+	77.3	73.6	87.0	74.1	92.0	80.3	89.2	79.4	84.7

Suggests *reverse-time* is a more effective way of organizing the data to display main trends in the mean response.

Section 2

Revival Process

The Revival Process

Assuming the survival time is *finite with probability one*, we can define the **Revival Process**

$$Z_i(s) = Y_i(T_i - s)$$

- The revival process is the re-aligned health process.
- This is well-defined conditional on the survival time and provides an invertible mapping from the survival process to the joint survival time and revival process.

$$Y \leftrightarrow (Z, T)$$

- We consider this joint process in lieu of the survival process as we hope it provides effective alignment of patient records for comparison and signal extraction.

The Revival Process

Table 2 confirms there is considerable excess variation associated with rows (116.8) and with the reverse-time factor (77.8) but not so much with columns.

Table 2 : ANOVA decomposition for Table 1

Source	$\mathcal{U}/\mathcal{V}$	$\|P_{\mathcal{U}}Y\|^2 - \|P_{\mathcal{V}}Y\|^2$	d.f.	M.S.
Diagonal	$(R+C+D)/(R+C)$	544.3	7	77.8
Column	$(R+C+D)/(R+D)$	237.9	7	34.0
Row	$(R+C+D)/(C+D)$	817.3	7	116.8
Residual	$RC/(R+C+D)$	497.2	21	23.7

$\Rightarrow$ Time reversal yields effective alignment of patient records for comparison and signal extraction.

Covariates

Definition (Temporal and time-evolving variable)

A *temporal variable* x is a function defined for every $t \geq 0$.

- It is a covariate if is a function on the units (entire function is determined at baseline)
- Typically implies x constant, but there are exceptions (eg. patient's age)

A *time-evolving* variable x' is a temporal variable that is not a function on the units (eg. marital status, quality of life, and air quality)

Every time-evolving variable is necessarily part of the response process

- The joint distribution of time-evolving variables and survival time may be used to predict survival time beyond t whose status is known at times $\mathbf{t}$ prior to t.
- Probabilistic prediction is not possible without the requisite mathematical structure of σ-fields.

Treatment effect: definition and estimation

- For patient i, let $a_i(t)$ be the treatment arm scheduled for patient i at time t.
- *Null* level required for times at and before recruitment, $t \leq 0$.
- Entire temporal trajectory, $a_i(t)$, for $t \geq 0$ is specified at baseline and determined by randomization (ie. it is a time-dependent covariate).

Let $\bar{a}_i(s) = a_i(T - s)$ be the treatment arm expressed in revival time.

- $Z \perp\!\!\!\perp T \mid \bar{a}$ because T is a function of $\bar{a}$
- *Lack of interference (I)* : the treatment assigned to one individual has no effect on the response distribution for other individuals.
- *Lack of interference (II)* : the treatment protocol at one time has no effect on the response distribution at other times.

$$Z[\mathbf{s}] \perp\!\!\!\perp \bar{a} \mid \bar{a}[\mathbf{s}]$$

Treatment effect: definition and estimation

Consider two patients, one in each treatment arm,

$$a_i(t) = \bar{a}_i(T_i - t) = 1, \qquad a_j(t) = \bar{a}_j(T_j - t) = 0$$

such that $x_i = x_j$.

The conventional treatment definitions are

$$\gamma_{10}(t) = E(Y_i(t) - E(Y_j(t)) \quad \text{or} \quad \gamma'_{10}(t) = E(Y_i(t) - E(Y_j(t) \mid T_i, T_j > t)$$

Supposing the dependence on T is additive, the difference of means at revival time s

$$E(Z_i(s) \mid T) - E(Z_j(s) \mid T) = \tau_{10}(s) + \gamma(T_i) - \gamma(T_j)$$

contains both a treatment effect and an effect due to the difference in survival times.

Conditional Distribution

The joint density of $(T, \mathbf{t}^{(k)}, Y[\mathbf{t}^{(k)}]$ at $(t, \mathbf{t}^{(k)}, y)$ is a product of three factors

$$\begin{aligned} & f(t) \times \prod_{j<k} p(t_j, y_j \mid \mathcal{H}_j, T = t) \\ = \quad & f(t) \times \prod_{j<k} p(y_j \mid \mathcal{H}_j, T = t) \times \prod_{j<k} p(t_j \mid \mathcal{H}_j, T = t) \\ = \quad & f(t) \times g_k(y; t - \mathbf{t}^{(k)} \mid t) \times \prod_{j<k} p(t_j \mid \mathcal{H}_j, T = t), \end{aligned} \tag{2}$$

where $f = F'$ is the survival density, and $\mathcal{H}_j$ is the observed history $(\mathbf{t}^{(j)}, Y[\mathbf{t}^{(j)}])$ at time t_{j-1}.

We assume the appointment schedule is uninformative for prediction in the sense that

$$p(t_k \mid \mathcal{H}_k, T = t) = p(t_k \mid \mathcal{H}_k, T = \infty) \tag{3}$$

for $t_{k-1} < t_k < t$.

Likelihood Factorization : Survival distribution specification

Re-alignment by the survival time *requires* the survival time to be finite with probability one.

$$\mathrm{pr}(T < \infty) = 1$$

Likelihood Factorization : Survival distribution specification

Re-alignment by the survival time *requires* the survival time to be finite with probability one.

$$\mathrm{pr}(T < \infty) = 1$$

Harmonic Process

The *harmonic process* is an exchangeable survival process defined by two non-negative parameters (ν, ρ).

$$T_i \sim \mathrm{Exp}\left(\nu(\psi(1+\rho) - \psi(\rho))\right),$$

where $\psi(\cdot)$ is the derivative of the log-gamma function.

Given unique survival times $T_1 < \ldots < T_k$ the conditional hazard is the product of a continuous and discrete component. The continuous component is

$$H(t) = \sum_{i:T_i \leq t} \nu \frac{T_i - T_{i-1}}{R^\sharp(T_{i-1}) + \rho} + \nu \frac{t - T_j}{R^\sharp(T_j) + \rho},$$

where $R^\sharp(t) = \#\{i : T_i < t\}$. The discrete component is

$$\prod_{j:T_j \leq t} \frac{r_j + \rho}{r_j + d_j + \rho}$$

Incomplete Records

On the assumption that censoring is uninformative, the joint density is

$$\int_{t \geq c} f(t) p(\mathbf{t}_c \mid t) g(y; t - \mathbf{t}_t) dt$$

where $\mathbf{t}_c = \mathbf{t} \cap [0, c]$ and $Y[\mathbf{t}_c] = y$.

Incomplete Records

On the assumption that censoring is uninformative, the joint density is

$$\int_{t \geq c} f(t) p(\mathbf{t}_c \mid t) g(y; t - \mathbf{t}_t) dt$$

where $\mathbf{t}_c = \mathbf{t} \cap [0, c]$ and $Y[\mathbf{t}_c] = y$.

Imputation

Impute survival times T' using the conditional survival distribution

$$f\left(T \mid Y[\mathbf{t}_c], \mathbf{t}_c, T > c; \hat{\psi}_u, \hat{\theta}\right) \propto f(T; \hat{\theta}) g(Y[\mathbf{t}_c]; T - \mathbf{t}_c; \hat{\psi}_u) \mathbf{1}[T > c]$$

The log-likelihood component associated with the imputed, uncensored record is given by

$$\log g(y; T' - \mathbf{t}_c; \psi) + \log f(T'; \theta)$$

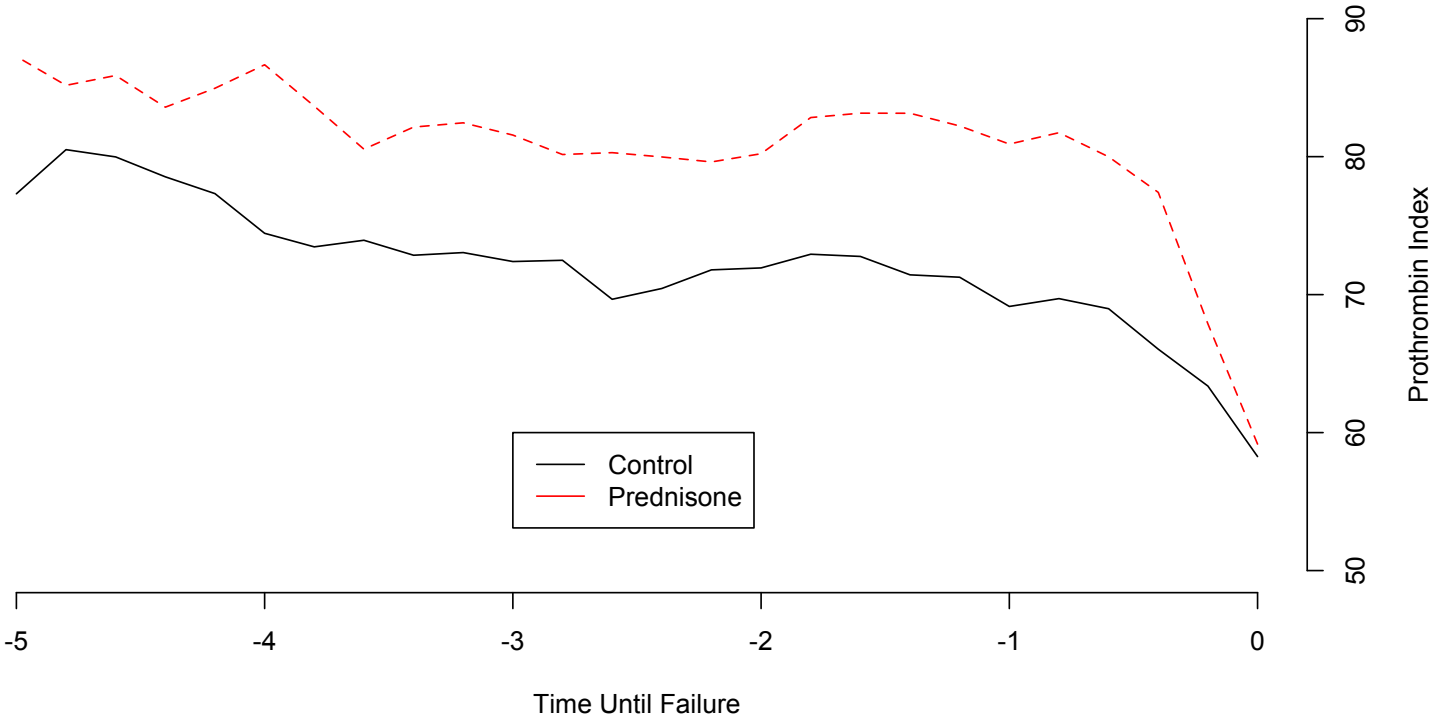
so parameter estimation after imputation is straightforward.

Section 5

A worked example : cirrhosis study

Motivating Example: Continued

Figure 1 : Smoothed Mean Curve for Prothrombin Index Aligned by Time until Failure



Motivating Example: Continued

We suggest the following model based on Figure 1,

$$\mathrm{E}(Z_i(s)|\,T) = \alpha + \tau_{\bar{a}_i(s)} + \beta_0 T_i + \beta_1 s + \beta_s \log(s+\delta)$$

$$\mathrm{cov}(Z_i(s), Z_j(s')|\,T) = \sigma_1^2 \delta_{ij} K_1(s, s'; \lambda) + \sigma_2^2 \delta_{ij} + \sigma_3^2 \delta_{ij} \delta_{ss'}$$

$$K_1(s, s'; \lambda) = \exp\left(-\frac{|s - s'|}{\lambda}\right)$$

Table 3 : Coefficients for revival model

	Censored records			Uncensored records			
Covariate	Coef.	S.E.	Ratio	Coef.	S.E.	Ratio	$T^\star$
Null Treatment	0.00	-	-	0.00	-	-	-
Control	4.13	1.84	2.3	2.41	1.43	1.7	0.94
Prednizone	11.56	1.75	6.6	13.55	1.47	9.2	−1.14
Survival (T)	2.65	0.39	6.9	1.75	0.47	3.7	2.79
Revival (s)	−2.78	0.49	−5.7	−2.11	0.47	−4.5	−1.72
$\log(s+\delta)$	3.74	2.68	1.4	4.66	0.41	11.5	−0.46
λ				0.164			

Motivating Example: Continued

Table 4 : Variance components

		Censored records		Uncensored records	
		Coefficient	S.E.	Coefficient	S.E.
AR1	σ_1^2	166.27	29.79	209.95	29.54
Patient	σ_2^2	155.84	31.02	206.82	34.48
White Noise	σ_3^2	223.69	17.30	179.59	12.90

Effect of prothrombin on prognosis

Over a period of 5 years and one month following recruitment, patient u had eight appointments with prothrombin values as follows:

$\mathbf{t}_u$ (days)	0	126	226	392	770	1127	1631	1855
$Y_u[\mathbf{t}_u]$	49	93	122	120	110	100	72	59

- This is the record for patient 402 who was assigned to prednisone and was subsequently censored at 2661 days.
- The log density at y_u is a quadratic form

$$h(t, y_u) = \text{const} - (y_u - \mu)' \Sigma^{-1} (y_u - \mu)/2$$

depending on t only through μ.

Effect of prothrombin on prognosis

This estimated factor is shown in Fig. 2a for three versions of the record in which the final prothrombin value is 59, 69 or 79.

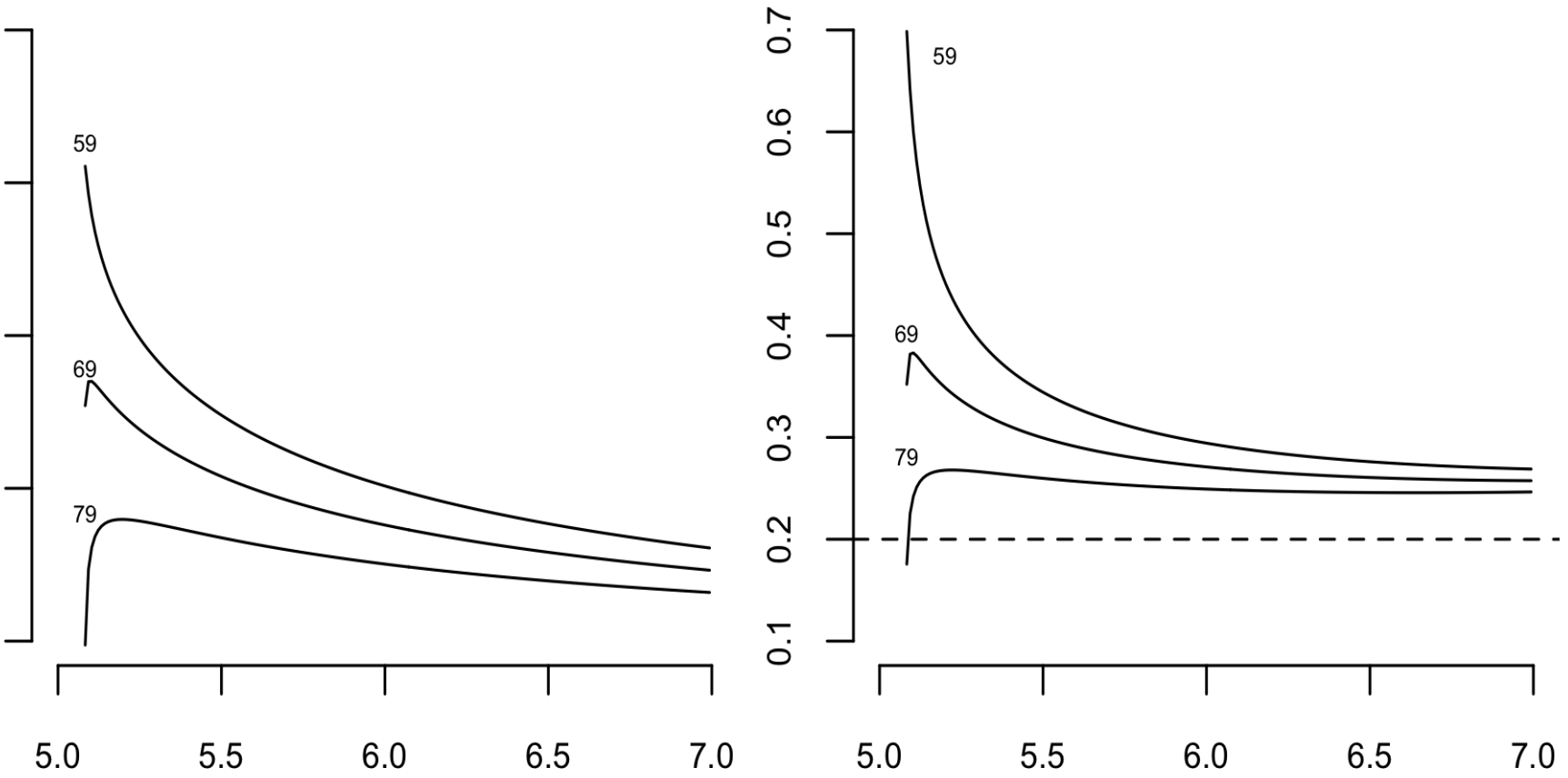


Figure 2 : Three versions of the record for patient 402: log modification factors for the predictive survival density (left panel) and hazard functions (right panel).

Summary

- The health sequence is regarded as a random process in its own right, not as a time-dependent covariate governing survival.
- To a substantial extent, the model for survival time is decoupled from the revival model for the behaviour of the health sequence in reverse time.
- Realignment implies that value $Y_i(0)$ at recruitment must not be treated as a covariate, but as an integral part of the response sequence. If they were available, values prior to recruitment could also be used.
- The definition of a treatment effect is not the usual one because the natural way to compare the records for two individuals is not at a fixed time following recruitment, but at a fixed revival time. The treatment value need not be constant in revival time.
- The predictive value of a partial health sequence for subsequent survival emerges naturally from the joint survival-revival distribution. In particular, the conditional hazard given the finite sequence of earlier values is typically not constant during the subsequent inter-appointment period.

Summary

- Records cannot be aligned until the patient dies, which means that the revival process is not observable component-wise until T is known. As a result, the likelihood analysis for incomplete records is technically more complicated.
- The omission of incomplete records from the revival likelihood does not lead to bias in estimation, but it does lead to inefficiency, which could be substantial if the majority of records are incomplete.
- The principal assumption, that appointment dates be uninformative for subsequent survival, does not affect likelihood calculations, but it does affect prognosis calculations for individual patients. For that reason, it is advisable to label all appointments as scheduled or unscheduled.

Thank you