Lecture 23
Introduction to Modern Survival Models

Economic Motivation. This Section was inspired by a talk by Gregory Kordas in April, 2001 based on work of Kordas and Deltas. An interesting analogue to survival models arises in discrete demand analysis. Consider a commodity like MP3 music players for which consumers buy either one or zero units. The willingness to pay of consumer i, say w_i, may be viewed as a survival time. The willingness to pay may be viewed as the interval of prices from zero to w_i such that the consumer is willing to purchase the good. Typically, in survey market research a sample of prospective consumers are asked, e.g., "would you be willing to purchase a player at price v_i?" Responses may be viewed as censored observations on the actual w_i's. It is standard practice to analyze such data by expressing w_i as a function of covariates, plus a random component, e.g.,

$$w_i = x_i'\beta + u_i$$

and then assuming the u_i's are iid normal, the probability of a "yes" is

$$\begin{aligned} P(w_i > v_i) &= P(-u_i < x_i'\beta - v_i) \\ &= P(u_i > -(x_i'\beta - v_i)) \\ &= 1 - \Phi(-(x_i'\beta - v_i)) \\ &= \Phi(x_i'\beta - v_i). \end{aligned}$$

So we model $\pi_i = P(y_i = 1|x_i)$ as $\Phi(x_i'\beta - v_i)$, yielding a probit model that includes the x_i's and v_i as covariates, that is we estimate,

$$\Phi^{-1}(\pi_i) = x_i\beta + \alpha v_i$$

We expect $\alpha < 0$, and then *impose* $\alpha = -1$ on the final estimate of β so

$$\hat{\beta} = -\hat{\gamma}/\hat{\alpha}.$$

Of course the iid assumption may be poor.

The usual survival analysis, say á la Kaplan-Meier, isn't much help here since we *never* observe an uncensored "event." Kordas and Deltas consider an alternative estimator $\hat{\beta}(\tau)$ analogous to quantile for this problem:

$$\min \sum \rho_\tau(y_i - I(x_i\beta - v_i > 0)).$$

A somewhat more sophisticated design that is often used in contingent valuation settings involves two questions. If the respondent says "no" to the first question, the interviewer lowers the price and asks once more. If "yes", then the price is raised. This yields a data structure

somewhat analogous to the interval censored data of survival analysis. Let U and V denote the upper and lower prices respectively and W be the value assigned by the subject. We see

$$\begin{aligned} \delta &= I(W \leq U) \\ \gamma &= I(U < W \leq V) \end{aligned}$$

If we have a model like

$$\text{logit } F(t|z) = \text{logit } F_0(t) + x'\beta,$$

we can form a likelihood

$$\mathcal{L} = \prod F(u_i)^{\delta_i}(F(v_i) - F(u_i))^{\gamma_i}(1 - F(v_i))^{1-\delta_i-\gamma_i}$$

This model is investigated in Huang and Rossini (1997, *JASA*), in the context of survival analysis it is usually called interval censoring.

The Relationship between Survival and Binary Response Models

This section is based mainly on Doksum and Gasko (1990, Intl Stat Review). We can think of the usual binary response model as a survival model in which we fix the time of survival and ask, what is the probability of surviving up to time t. For example, in the 472 problem set on quit behavior of Western Electric workers, we can ask what is the probability of not quitting up to time 6 months. By then varying t we get a nice 1-1 correspondence between the two classes of models. We can specify the general failure-time distribution,

$$F(t|x) = P(T < t|x)$$

and fixed t so we are simply modeling a survival probability, say $S(t|x) = 1 - F(t|x)$ which depends on covariates. We will consider two leading examples to illustrate this, the logit model, and the Cox proportional hazard model.

Logit

In the logit model we have,

$$\text{logit } (S(t|x)) = \log(S(t|x)/(1 - S(t|x)) = x'\beta$$

where $F(z) = (1 + e^{-z})^{-1}$ is the df of the logistic distribution. In survival analysis this would correspond to the model

$$\text{logit } (S(t|x)) = x'\beta + \log \Gamma(t)$$

where $\Gamma(t)$ is a baseline odds function which satisfies the restriction that $\Gamma(0) = 0$, and $\Gamma(\infty) = \infty$. For fixed t we can simply absorb $\Gamma(t)$ into the intercept of $x'\beta$. This is the proportional-odds model. Let

$$\Gamma(t|x) = S(t|x)/(1 - S(t|x)) = \Gamma(t) \exp\{x'\beta\}$$

and by analogy with other logit type models we can characterize the model as possessing the property that the ratio of the odds-on-survival at any time t don't depend upon t, i.e.

$$\Gamma(t|x_1)/\Gamma(t|x_2) = \exp(x_1'\beta)/\exp(x_2'\beta).$$

Now choosing some explicit functional form for $\Gamma(t)$ for example $\log \Gamma(t) = \gamma \log(t)$, ie. $\Gamma(t) = t^\gamma$, gives the survival model introduced by Bennett (1983).

Proportional Hazard Model

One can, of course, model not S, as above, but some other aspect of S which contains equivalent information, like the hazard function,

$$\lambda(t|x) = f(t|x)/(1 - F(t|x))$$

or the cumulative hazard,

$$\Lambda(t|x) = -\log(1 - F(t|x)).$$

In the Cox model we take

$$\lambda(t|x) = \lambda(t)e^{x'\beta},$$

so

$$\Lambda(t|x) = \Lambda(t)e^{x'\beta},$$

which is equivalent to

$$\log(-\log(1 - F(t|x)) = x'\beta + \log \Lambda(t).$$

This looks rather similar to the the logit form,

$$\text{logit } (F(t|x)) = x'\beta + \log \Gamma(t).$$

but it is obviously different. This form of the proportional hazard model could also be written as,

$$F(t|x) = \Psi(x'\beta + \log \Lambda(t)).$$

where $\Psi(z) = 1 - e^{-e^z}$ is the Type I extreme value distribution. For fixed t we can again absorb the $\log \Lambda(t)$ term into the intercept of the $x'\beta$ contribution and we have the formulation,

$$\log(-\log(1 - \theta(x))) = x'\beta$$

this is sometimes called the complementary $\log - \log$ model in the binary response literature. So this would provide a binary response model which would be consistent with the Cox proportional hazard specification of the survival version of the model. In general, this strategy provides a useful way to go back and forth between binary response and full-blown survival models.

Accelerated Failure Time Model

A third alternative, which also plays an important role in the analysis of failure time data is the accelerated failure time (AFT) model, where we have

$$\log(T) = x'\beta + u$$

with the distribution of u unspecified, but typically assumed to be iid. A special case of this model is the Cox model with Weibull baseline hazard, but in general we have

$$P(T > t) = P(e^u > te^{-x'\beta}) = 1 - F(te^{-x'\beta})$$

where F denotes the df of e^u and therefore in this model,

$$\lambda(t|x) = \lambda_0(te^{-x'\beta}e^{x'\beta})$$

where λ_0 denotes the hazard function corresponding to F. In effect the covariates are seen to simply rescale time in this model. An interesting extension of this model is to write,

$$Q_{h(T)}(\tau|x) = x'\beta(\tau)$$

and consider a family of quantile regression models. This allows the covariates to act rather flexibly with respect to the shape of the survival distribution.

Consider the general form of the binary response form of the survival model

$$G^{-1}(S(t|x)) = h(t) + x'\beta$$

Here we are saying that there is a transformation G^{-1} of the conditional survival probability such that the resulting $G^{-1}(\pi)$ is additively separable in t and $x'\beta$. This is strong, but subsumes some interesting models. Then,

$$\begin{aligned} P(h(T) > t|\lambda) &= P(T > h^{-1}(t)|x) \\ &= S(h^{-1}(t)|x) \\ &= G(h(h^{-1}(t))|x) \\ &= G(t - x\beta) \end{aligned}$$

This says that,

$$\begin{aligned} P(h(T) < t) &= 1 - G(t - x\beta) \\ P(h(T) < t + x'\beta) &= 1 - G(t) \\ P(h(T) - x'\beta < t) &= 1 - G(t) \end{aligned}$$

so $h(T) - x'\beta$ is iid with df $1 - G$ and hence, $h(T) = x'\beta + v$ with v iid with df $(1 - G)$.

Example: Cheng, Ying and Wei, (1995) *Biometrika*, 82, 435-45.

Consider the transformation model

$$h(T) = X'\beta + U$$

where U is iid F. If, for example, F is extreme value $F(s) = 1 - e^{-e^s}$, then we have the proportional odds models. It is of considerable interest to explore the general problem of estimating semiparametric models and methods for this model especially if they can be adapted to censoring.

Suppose we have data $\{Y_i, \delta_i\}$ where $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \leq C_i)$ as usual. If h is strictly increasing, then the ranks of $\{h(T_i)\}$ are the same as the ranks of $\{T_i\}$ so Chen, Ying and Wei suggest that using the marginal ranks to make inference about β might be appropriate. However, this is difficult, so they adopt the following simpler approach

Consider an estimator based on the following observation;

$$\begin{aligned} E(I(T_i \geq T_j)|X_i, X_j) &= P(h(T_i) \geq h(T_j)|X_i, X_j) \\ &= P(U_i - U_j \geq X_{ij}'\beta) \\ &= \xi(X_{ij}'\beta) \end{aligned}$$

where $X_{ij} = X_i - X_j$,
and

$$\xi(s) = \int_{-\infty}^{\infty} (1 - F(t+s))dF(t).$$

How does this yield an estimator? We have a "moment condition" and we can look for a β to solve the nonlinear equations

$$U(\beta) = \sum_i \sum_j X_{ij}(I(T_i \geq T_j) - \xi(X_{ij}'\beta))$$

Extensions

1. We can add weights: MLE type weights would be $w_i = \xi'(\cdot)/(\xi(\cdot)(1 - \xi(\cdot)))$.
2. If there is censoring, then we don't always observe $I(T_i \geq T_j)$, but

$$\begin{aligned} E\left(\frac{\delta_j I(Y_i \geq Y_j)}{G^2(Y_j)}\middle| X_i, X_j\right) &= E\left[E\left(\frac{I(T_i \geq T_j)I(\min(C_i, C_j) \geq T_j)}{G^2(T_j)}\middle| T_j, X_i, X_j\right)\right] \\ &= E(I(h(T_i) \geq h(T_j))|X_i, X_j) \\ &= \xi(X_{ij}'\beta_0) \end{aligned}$$

so we have

$$\tilde{U}(\beta) = \sum_i \sum_j X_{ij}\left(\frac{\delta_j I(Y_i \geq Y_j)}{G^2(Y_j)} - \xi(X_{ij}'\beta)\right)$$

where $G(\cdot$ is the survival function of the censoring n. variable C, and can be estimated. Horowitz (1998) also discusses this model, and mentions the Cheng, Ying, Wei estimator.

The Nelson-Aalen estimator of $\Lambda(t)$.

We have already introduced the Kaplan Meier estimator of the survival function $S(t)$. In this section we consider an alternative strategy in which we estimate instead the cumulative hazard function $\Lambda(t)$ for iid observations. This enables us to introduce some basic concepts. The discussion follows Therneau and Grambsch (2000) and is intended to introduce the some aspects of the approach I would like to pursue in my Fall 2001, 478 topics course.

Consider the Nelson (1969) fan data, 70 observations on the failure times of diesel generator fans. The data is heavily censored – only 12 of the 70 observations represents an actual failure, the others are all censoring times. The data are conveniently printed in R and S using the convention of Therneau's Survival5 package that censored event times are indicated by a + sign. The event times are given in thousands of hours.

```
> Surv(fans,cens)

 [1]   4.5    4.6+  11.5   11.5   15.6+  16.0   16.6+  18.5+  18.5+  18.5+
[11]  18.5+  18.5+  20.3+  20.3+  20.3+  20.7   20.7   20.8   22.0+  30.0+
[21]  30.0+  30.0+  30.0+  31.0   32.0+  34.5   37.5+  37.5+  41.5+  41.5+
[31]  41.5+  41.5+  43.0+  43.0+  43.0+  43.0+  46.0   48.5+  48.5+  48.5+
[41]  48.5+  50.0+  50.0+  50.0+  61.0+  61.0   61.0+  61.0+  63.0+  64.5+
[51]  64.5+  67.0+  74.5+  78.0+  78.0+  81.0+  81.0+  82.0+  85.0+  85.0+
[61]  85.0+  87.5+  87.5   87.5+  94.0+  99.0+ 101.0+ 101.0+ 101.0+ 115.0+
```

The practical question of interest is quite simple: Is the failure rate, i.e., the hazard function, increasing decreasing or constant? It might be conjectured that the fans were heterogeneous and that after the demise of a few "bad apples " the remaining fans would appear quite robust. This would suggest decreasing hazard. Alternatively, we might have parts that gradually wore out, which would suggest increasing hazard. An answer to this question would be an important piece of the more complicated problem of designing a good maintenance/replacement policy, see Rust (1987) for an extended analysis of this sort.

We will assume that the failure times T_i^* are iid with *df* F, density f and hazard, λ. To formalize the heterogeneity hypothesis we may view F, its associated density function f and the associated survival function S as mixtures of some underlying set of fundamental fan "types," but unless we are able to untangle these types with some covariate, for example, the day of the week of manufacture, we are just as well to consider them as iid provided we permit a flexible form for their distribution. We may view this as an early example of a frailty model.

As usual, we observe $T_i = \min\{T_i^*, C_i^*\}$, where C_i^* denotes a censoring time for the i^{th} observation.

Counting Process Formulation

Let $Y_i(t) = I(\{T_i \geq t\})$ so $Y_i(t)$ is 1 until failure, T_i^*, or censoring C_i^*, whichever comes first. The *counting process*, $N_i(t)$ associated with $Y_i(t)$ is simply the number of observed events in $[0, t]$ for unit i. In our fan example, $N_i(t)$ is 0 up to T_i and 1 thereafter, but the formalism obviously accommodates multiple events. Thus

$$\begin{aligned} N_i(t) &= I(\{T_i \leq t\}, \{\delta_i = 1\}) \\ Y_i(t) &= I(\{T_i \geq t\}) \end{aligned}$$

Note the right continuity of $Y_i(t)$ and the left continuity of $N_i(t)$; this is quite crucial. We may designate $Y_i(t)$ as a predictable process – if we need to know $Y_i(t)$, then we are assured that it is sufficient to know $Y_i(t-)$. In gambling, $Y_i(t)$ might indicate subject i's wealth at time t and something about his bets "at risk". This can depend upon the past in a complicated fashion, but it is *known* at t.

Now consider the aggregated processes

$$\begin{aligned} \bar{Y}(t) &= \sum Y_i(t) \quad \text{\# at risk at } t \\ \bar{N}(t) &= \sum N_i(t) \quad \text{\# of failures up to } t \end{aligned}$$

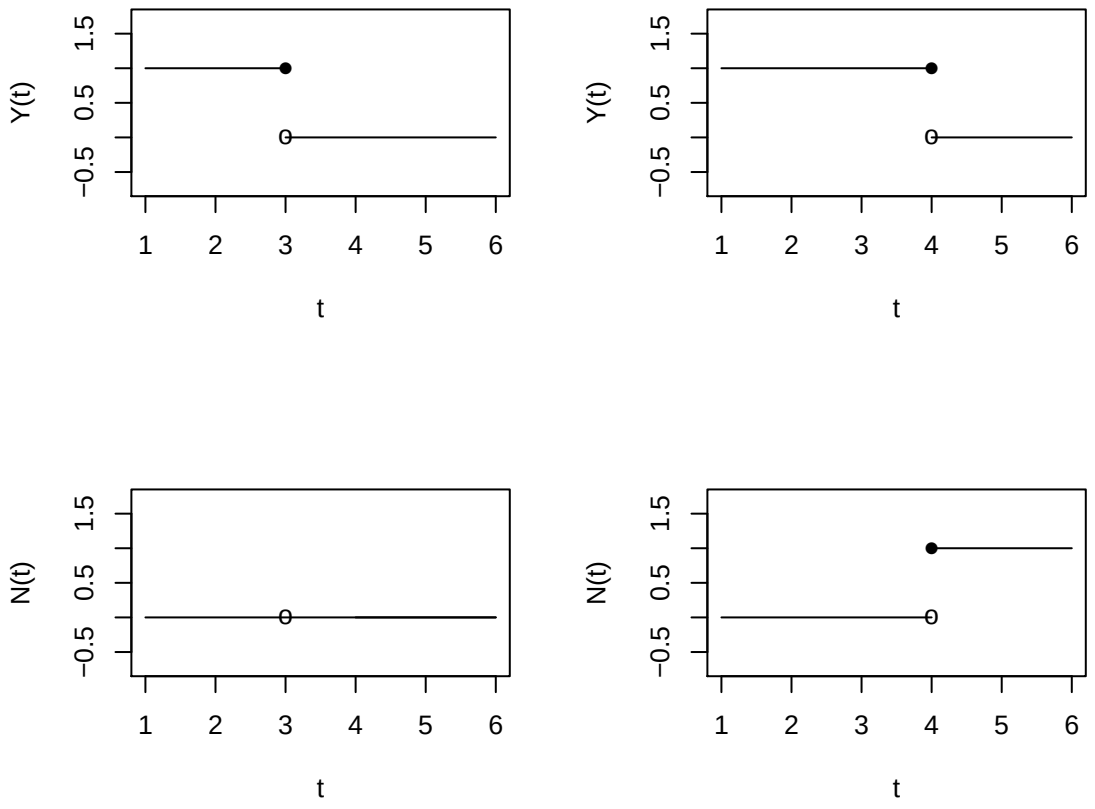

Figure 1: This figure illustrates two examples of the N and Y processes. In the two left panels we illustrate a censored observation with event time 3. In the two right panels we illustrate an uncensored observation with event time 4.

Since

$$\Lambda(s+h) - \Lambda(s) \approx \lambda(s)h$$

it is natural to estimate this by the number of events occurring in $[s, s+h]$ divided by the number of subjects at risk at s, i.e., by $(\bar{N}(s+h) - \bar{N}(s))/\bar{Y}(s)$. Summing over all of $[0,t]$ we have

$$\hat{\Lambda}(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)}.$$

We need to be careful about the notation. In principal $d\bar{N}(s)$ can accommodate both discrete and continuous components of the counting process, i.e.,

$$d\bar{N}(t) = \Delta\bar{N}(t) + n(t)dt$$

where $\Delta\bar{N}(t)$ denotes the discrete component and the $n(t)dt$ denotes a component with density with respect to Lebesgue measure. The quantity,

$$\Delta\bar{N}(t) = \bar{N}(t) - \bar{(}N(t-)$$

is the number of events occurring at precisely t. Since counting processes are pure jump processes the continuous part is unnecessary, so we may rewrite the Nelson-Aalen estimator in somewhat less intimidating, but fully equivalent fashion as

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)}$$

Note carefully the $t_i \leq t$!

Interpretation: Two versions

1. $\hat{\Lambda}(t)$ estimates the average number of failures up to time t, so, for example, up to $t = 87,500$ hours $\hat{\Lambda}(t) = .3368$ there would be about $\frac{1}{3}$ of the installed fans failing. This needs to be carefully interpreted – think of a repair policy that minimally repaired each of the fans so they were not "good as new" at failure time, but "good as at time of failure". This *policy* yields $\approx \frac{1}{3}$ of the installed fans failing.

2. The slope of $\hat{\Lambda}(t)$ is $\hat{\lambda}(t)$. Constant slope indicates exponential hazard, the mle of λ for the exponential model is # of failures /"total time on trial"

$$\begin{aligned} \hat{\lambda} &= \sum \delta_i / \sum T_i \\ &= \frac{12}{3443} = 0.0035 \end{aligned}$$

   and this fits quite well.

The Nelson-Aalen estimate of the cumulative hazard function is illustrated in Figure 2 with the exponential fit superimposed as the dotted line. This reproduces Figure 2.1 of TG] the R code to produce this figure is given below.

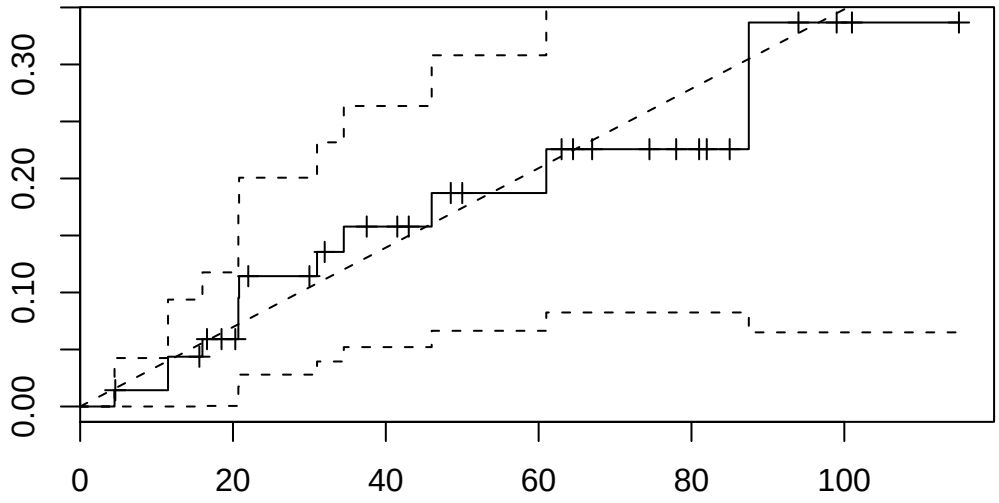

Figure 2: This figure illustrates the Nelson Aalen estimate of the cumulative hazard function for the fan data. The dotted line represents the fit of the exponential model.

```
#Reproduce Fig 2.1 from Therneau and Grambsch
d <- read.table("fans.dat",header=T)
fans <- d[,1]
cens <- d[,2]
ps.options(paper="special",width=6,height=4,horizontal=F)
postscript(file="fans1.ps")
plot(survfit(Surv(fans,cens),type="fleming-harrington"),fun="cumhaz")
#exponential fit
lambda <- sum(cens)/sum(fans)
abline(coef=c(0,lambda),lty=2)
frame()
```

It is important to assess the precision of $\hat{\Lambda}(t)$. A crucial virtue of the counting process formulation is that it reveals clearly that $N(t)$ can be modeled as a Poisson process, at least locally. Thus, the number of events in a small interval $[t, t+h], \bar{N}(t+h) - N(t) = \Delta_h N(t)$ is approximately Poisson with some intensity or rate parameter,

$$\int_t^{t+h} \bar{Y}(s)\lambda(s)ds \approx \bar{Y}(t)\lambda(t)h.$$

Conditional on the past,

$$E(\Delta_h \bar{N}(t)/\bar{Y}(t)) \approx \lambda(t)h$$

and

$$V(\Delta_h \bar{N}(t)/\bar{Y}(t)) \approx \lambda(t)h/\bar{Y}(t)$$

Note that we use the predictability of $\bar{Y}(t)$ here in a crucial way. The variance of the estimate can be estimated quite easily. Since $E(\Delta_h \bar{N}(t)) \approx \bar{Y}(t)\lambda(t)h$ and $V(\Delta_h \bar{N}(t)) \approx \bar{Y}(t)\lambda(t)h$ we have,

$$V(\Delta_h \bar{N}(t)/\bar{Y}(t)) \approx \lambda(t)h/\bar{Y}(t)$$

and thus,

$$V(\Lambda(t)) = \int_0^t \bar{Y}(s)^{-2} d\bar{N}(s) = \sum_{i:t_i \leq t} \Delta N(t_i)/\bar{Y}^2(t_i)$$

Variability for Poisson models is sometimes assessed on the log scale. Using the δ-method we have

$$V(\log \hat{\Lambda}(t)) \approx V(\hat{\Lambda}(t))/\hat{\Lambda}^2(t)$$

This for $\log \Lambda(t)$ we have the confidence interval

$$\log \Lambda(t) \in \log \hat{\Lambda}(t) \pm z_\alpha \hat{\sigma}/\hat{\Lambda}(t)$$

where $\hat{\sigma}^2$ is the estimate of $V(\hat{\Lambda}(t))$ described above. Or we have

$$\Lambda(t) \in \hat{\Lambda}(t) \exp\{\pm z_\alpha \hat{\sigma}/\hat{\Lambda}(t)\}$$

as an alternative interval for $\Lambda(t)$ on the original scale.

Nelson-Aalen vs. Kaplan-Meier

There is a very close connection between the Nelson-Aalen estimator of $\Lambda(t)$ and the Kaplan-Meier estimator of $S(t)$. To explore this recall that

$$\Lambda(t) = \int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1-F(s)}ds = -\log(1-F(t)) = -\log S(t)$$

so

$$d\Lambda(s) = \frac{dF(s)}{1-F(s-)}$$

and

$$F(t) = \int_0^t dF(s) = \int_0^t (1-F(s-))d\Lambda(s).$$

Thus, we can recursively define the estimator,

$$\hat{S}(t) = 1 - \int_0^t S(s-)d\hat{\Lambda}(s)$$

and since $\hat{S}(t-) - \hat{S}(t) = -\Delta\hat{S}(t) = \hat{S}(t-)\frac{\Delta N(t)}{Y(t)}$ we have,

$$\hat{S}(t) = \hat{S}(t-)\left(1 - \frac{\Delta N(t)}{Y(t)}\right) = \prod_{s<t}\left(1 - \frac{\Delta N(s)}{Y(s)}\right)$$

which is recognizable as the Kaplan-Meier estimator.

It is also informative to contrast the usual estimates of the precision of the two estimators. In the notation of L22, we can write the Kaplan-Meier estimator

$$\hat{S}(t) = \prod\left(1 - \frac{d_j}{n_j}\right) \equiv \prod(1 - h_j)$$

The product is rather awkward, but we can easily consider

$$\begin{aligned}
\text{Var }(\log \hat{S}(t)) &\cong \sum \text{ Var }(\log(1 - h_j)) \\
&\approx \sum(1 - h_j)^{-2} \text{ Var }(h_j) \\
&\approx \sum(1 - h_j)^{-2}\frac{h_j(1 - h_j)}{n_j} \\
&\approx \sum \frac{d_j}{n_j(n_j - d_j)}
\end{aligned}$$

So, again using delta method,

$$\text{Var }(S(t)) \approx S(t)^2 \sum \frac{d_j}{n_j(n_j - d_j)}$$

which is known as Greenwood's (1926) formula.

This is almost the same formula we found for the Nelson-Aalen estimator variance

$$\text{Var }(\hat{\Lambda}(t)) = \sum \Delta N(t_i)/Y^2(t_i) = \sum \frac{d_j}{n_j^2},$$

except for the slight modification of the denominator that obviously causes a serious problem when the last $d_j = 1$.

Introduction to Martingales for Survival Analysis One natural property of the Nelson Aalen estimator is that

$$(*) \qquad \sum_{i=1}^{n} \hat{\Lambda}(T_i) = \sum_{i=1}^{n} N_i(T_i)$$

If we observe that (why?)

$$\hat{\Lambda}(T_i) = \int_0^\infty Y_i(s)d\hat{\Lambda}(s)$$

and (why?)

$$N_i(T_i) = N_i(\infty)$$

we can write (*) as

$$\sum_{i=1}^{n}(N_i(\infty) - \int_0^\infty Y_i(s)d\hat{\Lambda}(s)) = 0$$

we can view the summands as a useful residual derivable from martingale properties of the $\hat{\Lambda}(s)$ estimator.

Some Basic Concepts

This will be a very brief sketch of some basics of martingales. In order to be explicit about conditioning we need some fundamental notion of the history of the process up to time t on which we will condition.

This is usually called the filtration

$$\mathcal{F}_t = \sigma((N_i(s), Y_i(s+), X_i(s)) : i = 1, \ldots, n, \;\; 0 < s < t)$$

where $\sigma(A)$ denotes the σ-algebra comprising A. Clearly, for $s < t, \mathcal{F}_s \subseteq \mathcal{F}_t$, so information (history) accumulates as time passes. You can think of the filtration as any conceivable way of packaging the history of the process up the the present. For a counting process $N_i(t)$, the increments $dN_i(t)$ over the interval $[t, t + dt)$ satisfies

$$E(dN_i(t)|\mathcal{F}_{t-}) = Y_i(t)\lambda(t)dt$$

Presuming that the observations are $\perp\!\!\!\perp$ over i,

$$\begin{aligned} E(dN_i(t)|\mathcal{F}_{t-}) &= P(dN_i(t) = 1|\mathcal{F}_{t-}) \\ &= P(dN_i(t) = 1|Y_i(t)) \end{aligned}$$

If $Y_i(t) = 0$, then the failure has already occurred and the conditional probability is 1, otherwise

$$P(dN_i(t) = 1|Y_i(t) = 1) = P(T_i^* \in [t, t + dt)|t \leq T_i^*, t \leq C_i^*)$$

Now assuming the $\perp\!\!\!\perp$ of T_i^* and C_i^* the conditioning on C_i^* is irrelevant so

$$P(dN_i(t) = 1|Y_i(t) = 1) = P(T_i^* \in [t, t + dt)|t \leq T_i^*) = \lambda(t)dt$$

or

$$P(dN_i(t) = 1|Y_i(t)) = Y_i(t)\lambda(t)dt.$$

Consider the process

$$(*) \qquad M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s)ds$$

Definition: A sequence of rv's $X_1, X_2, \ldots$ on a P-space $(\Omega, \mathcal{A}, P)$ that is adapted to an increasing sequence of σ-fields $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ is called a martingale if $E|X_i| < \infty$ for all i and

$$E(X_i|\mathcal{F}_j) = X_j \quad \text{for all} \quad j \leq i$$

Extending this definition to processes we may write this as

$$E(M(t)|\mathcal{F}_s) = M(s) \quad \text{for} \quad 0 \leq s < t$$

We may also write it in terms of increments as

$$E(dM(t)|\mathcal{F}_{t-}) = 0$$

So martingale increments have mean 0 and martingale increments are uncorrelated (though they are not necessarily $\perp\!\!\!\perp$)

$$\begin{aligned} \text{Cov } (M(t), M(t+u) - M(t)) &= 0 \\ \text{Cov } (M(t) - M(t-s), M(t+u) - M(t)) &= 0 \end{aligned}$$

Counting processes are examples of submartingales, processes for which $E|N(t)| < \infty$ and

$$E(N(t)|\mathcal{F}_s) \geq N(s) \quad \text{for} \quad 0 \leq s < t$$

Since our $N(t)$ process is monotone increasing this is particularly clear in this case. In general, submartingales need not be increasing, only the expectation condition is needed. For example, by Jensen's inequality if $M(t)$ is a martingale $M^2(t)$ is a submartingale; it has the same filtration as $M(\cdot)$, but its conditional expectation is strictly greater than $M^2(s)$ for $0 \leq s < t$.

The Doob-Meyer Decomposition

Submartingales may be decomposed into a predictable component, usually called "their compensator" and a martingale component. This is a crucial result due initially to Doob and extended by Meyer and others. An example is (*) where

$$C_i(t) = \int_0^t Y_i(s)\lambda(s)ds$$

is the compensator, $M_i(t)$ is a martingale and $C_i(t)$ subtracts off the conditional expectation of $N_i(t)$, given the history of the process up to time t.

It is (perhaps) useful to think about this in terms gambling strategies. You might, in a more general setting, think of the $Y_i(t)$ process as the process that generates bets at time t, it can depend in quite complicated ways on the whole history up to time t, including the present wealth of the bettor, but the crucial thing is that it is left continuous so we can think of bets being placed just prior to time t.

The counting process $N_i(t)$ is the returns process, it is right continuous so at time t it generates outcomes of the gambles based on bets placed at t, represented by $Y_i(t)$. At each point, t, we may assume (optimistically!) that the conditional expectation of returns $N_i(t)$ given the past exceeds the current value $N_i(t-)$. This is clear for the counting process, which can only move from 0 to 1 and then stay there, but in general we may even consider returns processes which may fall but still rise *in expectation*.

For such processes, submartingales, we may compute the conditional expectation of $N_i(t)$ given the past and then subtract that from $N_i(t)$. This yields a predictable process representing the conditional mean and the difference between $N_i(t)$ and its conditional mean is a process that has conditional expectation zero. This process is a martingale and this property yields a powerful source of implications for the behavior of the original process.

In the case of the counting process we have

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s)ds$$

where

$$E(M(t)|\mathcal{F}_s) = M(s)$$

or, apparently more generally,

$$E(M(t)|M(u) : 0 < u < s) = M(s)$$

Many nice features follow from the decomposition and the martingale structure of the process. Further discussion of these properties will have to be deferred to later however.