C.3 Modern Survival Problem sheet 3: Estimating quantiles and excess mortality

(1) Show that Duhamel's equation (6.5) holds at a point s where S_1 or S_2 is discontinuous.
At a discontinuity point of S_1 or S_2

$$\begin{aligned} d\frac{S_1(s)}{S_2(s)} &= \frac{S_1(s)}{S_2(s)} - \frac{S_1(s-)}{S_2(s-)} \\ &= \frac{dS_1(s)}{S_2(s)} + S_1(s-)\left(\frac{1}{S_2(s)} - \frac{1}{S_2(s-)}\right) \\ &= \frac{S_1(s-)}{S_2(s)} \cdot \frac{dS_1(s)}{S_1(s-)} - \frac{S_1(s-)}{S_2(s)}\left(\frac{S_2(s)}{S_2(s-)} - 1\right) \\ &= \frac{S_1(s-)}{S_2(s)}\left(\frac{dS_1(s)}{S_1(s-)} - \frac{dS_2(s)}{S_2(s-)}\right). \end{aligned}$$

(2) (a) Explain why this is a reasonable estimator for the quantiles of the survival function.
This code computes two vectors `xpless` and `xpmore`, corresponding to the event times in the data. These are lower and upper confidence limits for the cumulative hazard, according to the Nelson–Aalen estimator. The p-th quantile of survival is the time when the cumulative hazard crosses $-\log p$. Thus, we define two indices `upper` and `lower`, to be the last time when the upper confidence bound is below $-\log p$, and the first time when the upper confidence bound is above $-\log p$ respectively.

(b) Use this to compute a 95% confidence interval for median survival in the `ovarian` data set (ignore treatment type);

```
 library(survival)
data(ovarian) # load the ovarian dataset into the workspace

ov.surv <- Surv(ovarian$futime, ovarian$fustat)
ov.fit <- survfit( ov.surv~1 )
quantileCI( ov.fit, 0.5 )
[1] 431  NA
>
> #
> # It actually makes more sense to look at something like 80th percentile of
> # survival, since survival barely gets down below .5 in the study time
> # (Hence the NA for the upper end of the confidence interval)
>
> quantileCI(ov.fit, 0.8)
[1] 115 563
```

(c) Use this to compute a 95% confidence interval for median survival in a collection of data simulated from an exponential distribution with parameter 1, available on the course web site.

```
> load("expdata.dat")
>
> exp.surv <- Surv(t,ev)
> exp.fit  <- survfit(exp.surv~1)
>
> quantileCI( exp.fit , 0.5 )
[1] 0.5 0.7
```

(d) Suppose we ignored the censoring, so only included the uncensored times. What would you estimate for the median survival time?

```
T_new <- t[ev] #only use the events

exp.surv_new <- Surv(T_new, rep(TRUE, sum(ev)) )
exp.fit_new  <- survfit(exp.surv_new~1)

quantileCI( exp.fit_new, 0.5 )
[1] 0.3 0.5
```

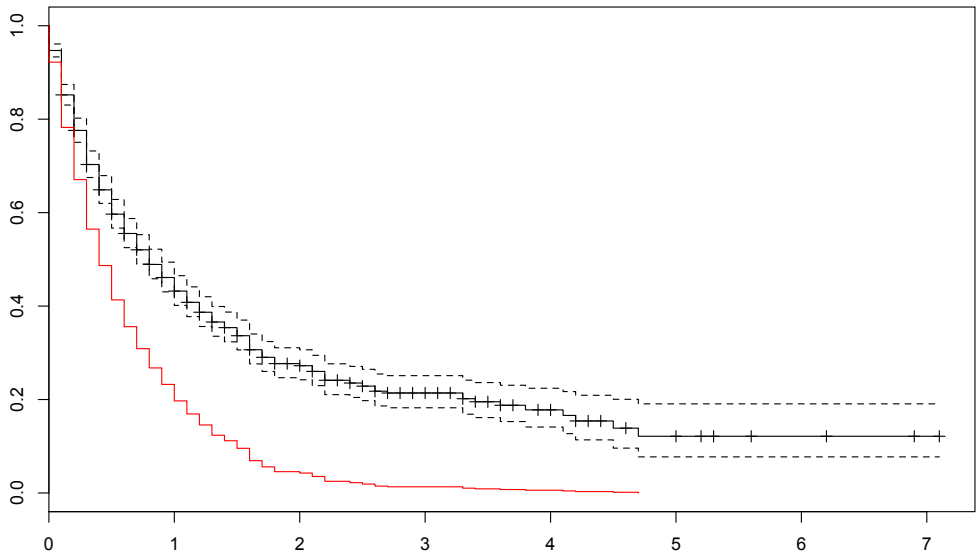

Figure C.1: Plot of Kaplan–Meier estimator with 95% confidence interval for the exponential simulation (black), and the false plot based on ignoring censored data (red).

(3) Look back to the derivation in the lecture notes section 7.2 of the estimator $\hat{\Gamma}(t)$ for cumulative excess mortality in the two-sample setting. Think of $k_c(t)$ as arbitrary predictable random variables.

(a) Construct a martingale to show that the estimator (7.5) for excess mortality in the two-sample case is unbiased for appropriate choice of $k_c(t)$. What conditions must $k_c(t)$ satisfy?
For each c,

$$N(c,0;t) - \int_0^t \alpha(c;s)Y(c,0;s)ds \text{ and}$$
$$N(c,1;t) - \int_0^t (\alpha(c;s)+\gamma(s))Y(c,1;s)ds$$

are both martingales. Dividing the increments by Y, we see that

$$\int_0^t \frac{dN(c,0;s)}{Y(c,0;s)} - \alpha(c,0;s)ds \text{ and}$$
$$\int_0^t \frac{dN(c,1;s)}{Y(c,1;s)} - \int_0^t \alpha(c,1;s)ds - \Gamma(t)$$

are martingales. Thus, the difference

$$\int_0^t \left(\frac{dN(c,1;s)}{Y(c,1;s)} - \frac{dN(c,0;s)}{Y(c,0;s)}\right) - \Gamma(t)$$

is also a martingale.
Thus, if k_c is predictable with $\sum k_c = 1$, then

$$M(t) := \sum_c \left(\int_0^t k_c(t)\left(\frac{dN(c,1;s)}{Y(c,1;s)} - \frac{dN(c,0;s)}{Y(c,0;s)}\right) - \Gamma(t)\right) = \hat{\Gamma}(t) - \Gamma(t)$$

is a martingale, and its expectation is 0 for any t, implying that $\hat{\Gamma}(t)$ is an unbiased estimator for $\Gamma(t)$.

(b) Find an expression for estimating the variance of $\hat{\Gamma}$.
The optional variation of M is

$$[M](t) = \sum_c \sum_{t_i^{(c,1)} \le t} \frac{k_c(t_i)^2}{Y(c,1;t_i^{(c,1)})^2} + \sum_{t_i^{(c,0)} \le t} \frac{k_c(t_i)^2}{Y(c,0;t_i^{(c,0)})^2}\Big)$$
$$= \sum_{t_i \le t} \frac{k_{c_i}(t_i)^2}{Y(c_i,G_i;t_i)^2}.$$

(c) Show that for the particular choice (7.4) the bound (7.6)

$$\mathrm{Var}(\hat{\Gamma}(t)) \approx \sum_{t_i \le t} \left(\sum_c Y(c, -; t_i) \right)^{-2}$$

is a conservative estimate for the variance of the estimator. That is, it is a good estimate for large samples, and tends not to underestimate the variance.

We need to show that the optional variation is bounded by

$$\sum_{t_i \le t} \left(\sum_c Y(c, -; t_i) \right)^{-2}$$

when we take

$$k_c(t) = \frac{Y(c, -; t)}{\sum_{c'} Y(c', -; t)} \, .$$

This follows immediately from the above formula, since $Y(c_i, -; t_i) \le Y(c_i, G_i; t_i)$, so

$$\frac{k_{c_i}(t_i)^2}{Y(c_i, G_i; t_i)^2} \le \Big(\sum_c Y(c, -; t_i) \Big)^2 .$$

(d) Since any choice of k_c yields an estimator, we are free to make a convenient choice. Why is the choice (7.4) a good one?

A sensible strategy is to minimise the expected next variance increment. Since the next individual to have an event is approximately uniformly chosen from the available individuals, hence proportional to $Y(c, G; t)$, we want to minimise

$$\sum Y(c, G; t) \frac{k_c^2}{Y(c, G; t)^2}$$

subject to $\sum k_c = 1$, leading us to choose k_c proportional to $Y(c, G; t)$. However, k_c can only depend on c, not on G, so we take $k_c(t) = Y(c, -; t) / \sum_{c'} Y(c', -; t)$. (Otherwise, it wouldn't be predictable.) In that case, we get the variance bound

$$\sum_{t_i \le t} \left(\sum_c Y(c, -; t) \right)^{-2} \tag{C.1}$$

If we said instead we wanted to minimise the maximum of the next squared increment, we would also choose $k_c(t)$ proportional to $Y(c, -; t)$.

(e) Supposing the groups to be of approximately equal size, what will the relation be between the variance of our estimator for the cumulative excess mortality, and the variance we would estimate for the difference in the cumulative hazards between the groups $\{G_i = 0\}$ and $\{G_i = 1\}$, ignoring the classification c_i.

If we ignore the classification, we will estimate the difference in cumulative hazards as

$$\sum_{t_i \le t} \frac{(-1)^{G_i}}{\sum_c Y(c, G_i; t_i)},$$

with variance estimated by

$$\sum_{t_i \le t} \frac{1}{\sum_c Y(c, G_i; t_i)^2}.$$

The bound we have just derived replaces $\sum_c Y(c, G_i; t_i)$ by $\sum_c Y(c, -; t_i)$. Thus, the variance estimate when we stratify by category is always larger, but they differ only to the extent that the numbers at risk in the two groups differ.