Statistics 582, Problem Set 2 Solutions

Wellner; 1/22/2015

1. Consider nonparametric maximum likelihood estimation of F in the right-censored data problem considered in class, but extend the argument to include ties as follows:
(a) When there are ties, let the distinct Z's be denoted by $T_1 < \ldots < T_k$.Let $m_1, \ldots, m_k$ and $n_1, \ldots, n_k$ be defined by $m_j \equiv \#$ of $Z_i \Delta_i = T_j$, $n_j \equiv \#$ of $Z_i(1-\Delta_i) = T_j$, and let $p_j \equiv \Delta F(T_j) = F(T_j) - F(T_j-)$, $j = 1, \ldots, k$, $p_{k+1} = 1 - F(T_k)$. Show that the likelihood (for F) is

$$L(F|\underline{Z}, \underline{\Delta}) = \prod_{i=1}^{k} p_i^{m_i} (\sum_{j=i+1}^{k+1} p_j)^{n_i}.$$

(b) By defining $\lambda_i = p_i / \sum_{j=i}^{k+1} p_j$ for $i = 1, \ldots, k$ and $\lambda_{k+1} = 1$, and rewriting the likelihood in terms of the λ_i's, show that the likelihood is maximized by

$$\hat{\lambda}_i = m_i / \sum_{j=i}^{k} (m_j + n_j) = \frac{n \Delta \mathbb{H}_n^{uc}(T_i)}{n(1 - \mathbb{H}_n(T_i-))}.$$

and hence that the nonparametric MLE of F is (again) the Kaplan - Meier estimator

$$1 - \widehat{F}_n(t) = \prod_{s \le t} (1 - \Delta \widehat{\Lambda}_n(s)).$$

(c) Compute $1 - \widehat{F}_n$ for the following data (time in days until vaginal cancer in rats, group 1; from Kalbfleisch and Prentice, 1980, page 2):

143, 164, 188, 188, 190, 192, 206, 209, 213, 216,
220, 227, 230, 234, 246, 265, 304, 216+, 244+

here $+$ indicates censoring ($\Delta = 0$).

Solution: (a) When there are ties, let the distinct Z's be denoted by $T_1 < \cdots < T_k$ Let $m_1, \ldots, m_k$ and $n_1, \ldots, n_k$ be defined by $m_j = \#\{i \le n : \ Z_i \Delta_i = T_j\}$, $n_j = \#\{i \le n : \ Z_i(1-\Delta_i) = T_j\}$, and let $p_j \equiv \Delta F(T_j)$, $j = 1, \ldots, k$, $p_{k+1} = 1 - F(T_k-)$. Then the likelihood (for F) is

$$L(F|\underline{Z}, \underline{\delta}) = \prod_{i=1}^{k} p_i^{m_i} \left(\sum_{j=i+1}^{k} p_j \right)^{n_i}.$$

Setting $\lambda_i \equiv p_i / \sum_{j=i}^{k+1} p_j$, $\lambda_{k+1} = 1$ yields

$$\sum_{j=i}^{k+1} p_j = \prod_{j=1}^{i-1} (1 - \lambda_j), \qquad 1 - \lambda_i = \frac{\sum_{j=i+1}^{k+1} p_j}{\sum_{j=i}^{k+1} p_j},$$

and hence

$$\begin{aligned}
L(F|\underline{Z},\underline{\Delta}) &= \prod_{i=1}^{k}\left(\frac{p_i}{\sum_{j=i}^{k+1}p_j}\right)^{m_i}(\sum_{j=i}^{k+1}p_j)^{m_i}\left\{\frac{\sum_{j=i+1}^{k+1}p_j}{\sum_{j=i}^{k+1}p_j}\sum_{j=i}^{k+1}p_j\right\}^{n_i} \\
&= \prod_{i=1}^{k}\lambda_i^{m_i}(1-\lambda_i)^{n_i}\left(\sum_{j=i}^{k+1}p_j\right)^{m_i+n_i} \\
&= \prod_{i=1}^{k}\lambda_i^{m_i}(1-\lambda_i)^{n_i}\left(\prod_{j=1}^{i-1}(1-\lambda_j)\right)^{m_i+n_i} \\
&= \prod_{i=1}^{k}\lambda_i^{m_i}(1-\lambda_i)^{n_i+\sum_{j=i+1}^{k}(m_j+n_j)} \\
&= \prod_{i=1}^{k}\lambda_i^{m_i}(1-\lambda_i)^{r_i-m_i}
\end{aligned}$$

where $r_i \equiv \sum_{j=i}^{k}(m_j+n_j)$.

(b) In view of the binomial form of this expression for each i, we know that it is maximized for each i by

$$\hat{\lambda}_i = \frac{m_i}{r_i} = \frac{m_i}{\sum_{j=i}^{k}(m_j+n_j)} = \frac{n\Delta\mathbb{H}_n^{(uc)}(T_i)}{n(1-\mathbb{H}_n(T_i-))},$$

for $i=1,\ldots,k$. Then

$$\hat{p}_i = \prod_{j=1}^{i-1}(1-\hat{\lambda}_j)\hat{\lambda}_i, \qquad i=1,\ldots,k+1.$$

as before. Note that $\hat{p}_{k+1}>0$ if $n_k>0$. Thus the nonparametric MLE's $\hat{\Lambda}_n$ and $\hat{F}_n$ of Λ and F are the Nelson-Aalen and Kaplan-Meier (or product-limit) estimators

$$\hat{\Lambda}_n(t) = \int_{[0,t]}\frac{d\mathbb{H}_n^{(uc)}(s)}{1-\mathbb{H}_n(s-)}$$

and $1-\hat{F}_n(t) = \prod_{s\le t}(1-\Delta\hat{\Lambda}_n(s))$.

(c) For the given data, there are 17 distinct times T_i: are 143, 164, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 244, 246, 265, 304 with ties at both 188 and 216. If we let $r_i \equiv n(1-\mathbb{H}_n(T_i-))$ and $d_i = n\Delta\mathbb{H}_n^{(uc)}(T_i)$ then we obtain the following table and calculated values of the estimator:

Table 1:

T_i	r_i	d_i	$1-\frac{d_i}{r_i}$	$\prod_{j\le i}(1-\frac{d_i}{r_i})$	$\widehat{Var}(\hat{F})$	$\widehat{Var}_{GW}(\hat{F})$
143	19	1	18/19	.9474	.00249	.00262
164	18	1	17/18	.8947	.00469	.00496
188	17	2	15/17	.7895	.00796	.00875
190	15	1	14/15	.7368	.00935	.01021
192	14	1	13/14	.6842	.01045	.01137
206	13	1	12/13	.6316	.01126	.01225
209	12	1	11/12	.5789	.01179	.01283
213	11	1	10/11	.5263	.01204	.01312
216	10	1	9/10	.4737	.01199	.01312
220	8	1	7/8	.4145	.01187	.01311
227	7	1	6/7	.3553	.01129	.01264
230	6	1	5/6	.2961	.01028	.01170
234	5	1	4/5	.2368	.00882	.01029
244	4	0	1	.1579	.00882	.01029
246	3	1	2/3	.1579	.00669	.00873
265	2	1	1/2	.0789	.00323	.00530
304	1	1	0	.0000		

2. Suppose, as in Example 4.3.10, that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. $\text{Mult}_k(1, \underline{p})$ so that $\underline{N}_n = \sum_{i=1}^n \underline{X}_i \sim \text{Mult}_k(n, \underline{p})$.
(a) Use Jensen's inequality to show that the log-likelihood

$$l_n(\underline{p}|\underline{X}) = \sum_{j=1}^k N_j \log p_j + \sum_{i=1}^n \log\left(\frac{1!}{X_{i1}! \cdots X_{ik}!}\right)$$

is maximized by $\underline{\hat{p}} = \underline{N}_n/n$. [Hint: write the first term of $l_n(\underline{p}|\underline{X})$ as $n\sum_{j=1}^k \hat{p}_j \log p_j$.]
(b) Relate $l_n(\underline{p})$ to $K(\underline{\hat{p}}, \underline{p})$ and hence show again that the maximizing value of $\underline{p}$ is $\hat{p}$.
(c) Use this problem and similar considerations as in the previous problem to formulate a version of Example 6.1 in the lecture notes when ties are present and show that the nonparametric MLE continues to be the empirical measure $\mathbb{P}_n$.

Solution: (a) Our goal is to show that

$$n\sum_{j=1}^k \hat{p}_j \log p_j \le n\sum_{j=1}^k \hat{p}_j \log \hat{p}_j$$

with equality if and only if $\underline{p} = \underline{\hat{p}}$. Subtracting the right side from the left side and dividing by n, we see that we want to show that

$$\sum_{j=1}^k \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) \le 0.$$

But since log is a concave function, Jensen's inequality yields

$$\begin{aligned}\sum_{j=1}^k \hat p_j \log\left(\frac{p_j}{\hat p_j}\right) &\le \log\left(\sum_{j=1}^k \hat p_j \left(\frac{p_j}{\hat p_j}\right)\right)\\ &= \log\left(\sum_{j=1}^k p_j\right) = \log(1) = 0.\end{aligned}$$

(b) Note that in the above argument we have shown that

$$l_n(\underline{p}) - l_n(\underline{\hat p}) = -nK(\underline{\hat p}, \underline{p}) \le 0$$

since $K(P,Q) \ge 0$ for all P, Q. Thus $l_n(\underline{p})$ is maximized by $\underline{p} = \underline{\hat p}$.

3. We showed in class that the nonparametric maximum likelihood estimator of F in the (right) censored data problem, possibly with ties, is the Kaplan-Meier (product limit) estimator $\widehat{F}_n(t)$ given by

$$1 - \widehat{F}_n(t) = \prod_{s \le t}(1 - \Delta\widehat{\Lambda}_n(s))$$

where $\widehat{\Lambda}_n(t)$ is the *Nelson-Aalen* estimator of

$$\Lambda(t) \equiv \Lambda_F(t) \equiv \int_0^t \frac{1}{1 - F_-} dF\,,$$

given by

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{1}{1 - \mathbb{H}_n(s-)} d\mathbb{H}_n^{uc}(s), \qquad 0 \le s \le t\,.$$

Here

$$\mathbb{H}_n^{uc}(t) = \frac{1}{n}\sum_{i=1}^n \Delta_i 1_{[Z_i \le t]}\,, \qquad \mathbb{H}_n(t) = \frac{1}{n}\sum_{i=1}^n 1_{[Z_i \le t]}$$

are the sub-empirical distribution function of the uncensored observations and the marginal empirical distribution of all the Z's, uncensored or censored.

(a) In class I gave a sketch of proof that

$$\sqrt{n}(\widehat{F}_n(t) - F(t)) \Rightarrow (1 - F(t))\mathbb{B}(C(t))$$

as a process uniformly in $t \in [0, \tau]$ for any $\tau < \tau_H$ (i.e. for any τ with $1 - H(\tau) = (1 - F(\tau))(1 - G(\tau)) > 0$, where $\mathbb{B}$ is a standard Brownian motion process and where

$$C(t) \equiv \int_0^t \frac{1}{(1 - H_-(s))^2} dH^{uc}(s)\,, \qquad 0 \le s \le t.$$

Thus we have, for each fixed $t < \tau$,

$$\sqrt{n}(\widehat{F}_n(t) - F(t)) \to_d N(0, (1 - F(t))^2 C(t)).$$

Suggest an estimator of $C(t)$ and hence an estimator of $(1 - F(t))^2 C(t)$.

(b) Show that your estimator of $(1 - F(t))^2 C(t)$ is consistent.

(c) Use the estimator you suggest in (b) to obtain an approximate 90% confidence interval for $F(200)$ based for the data given in problem 1 above.

Solution: (a) In this case there are ties in the data. A table giving the distinct time points T_i together with the numbers at risk and the number of deaths at each time point, together with the successive terms of the product and the resulting Kaplan-Meier estimator was given in problem 1. The last two columns of the table give two variance estimates: column 6 gives the variance estimator from (b) below; column 7 gives the usual Greenwood estimator (cf. part the notes handed out in class on 14 January and Kalbfleisch and Prentice (1980), pages 12 - 14). (b) A natural estimator of

$$C(t) = \int_{[0,t]} \frac{1}{(1-H(s-))^2} dH^{(uc)}(s)$$

is

$$\begin{aligned} \hat{C}_n(t) &= \int_{[0,t]} \frac{1}{(1-\mathbb{H}_n(s-))^2} d\mathbb{H}_n^{(uc)}(s) \\ &= n \int_{[0,t]} \frac{1}{R_n(s)^2} d\left(n\mathbb{H}_n^{(uc)}(s)\right) \end{aligned}$$

where $R_n(s) \equiv n(1-\mathbb{H}_n(s-))$. Note that in the Mathematica program accompanying the solution set the quantity labeled "Cest" is $n^{-1}\hat{C}_n(t) = \int_{[0,t]} R_n(s)^{-2} d(n\mathbb{H}_n^{(uc)}(s))$.

(c) To see that $\hat{C}_n(t) \to_p C(t)$ note that

$$\|\mathbb{H}_n^{(uc)} - H^{(uc)}\|_\infty = \sup_{0<t<\infty} |\mathbb{H}_n^{(uc)}(t) - H^{(uc)}(t)| \to_{a.s.} 0, \quad (1)$$

$$\|\mathbb{H}_n - H\|_\infty = \sup_{0<t<\infty} |\mathbb{H}_n(t) - H(t)| \to_{a.s.} 0 \quad (2)$$

by the Glivenko-Cantelli theorem.

$$\begin{aligned} \hat{C}_n(t) - C(t) &= \int_{[0,t]} \frac{d\mathbb{H}_n^{(uc)}(s)}{(1-\mathbb{H}_n(s-))^2} - \int_{[0,t]} \frac{dH^{(uc)}(s)}{(1-H(s-))^2} \\ &= \int_{[0,t]} \left(\frac{1}{(1-\mathbb{H}_n(s-))^2} - \frac{1}{(1-H(s-))^2} \right) d\mathbb{H}_n^{(uc)}(s) \\ &\quad + \int_{[0,t]} \frac{1}{(1-H(s-))^2} d(\mathbb{H}_n^{(uc)}(s) - H^{(uc)}(s)) \\ &= \int_{[0,t]} \frac{(1-H(s-))^2 - (1-\mathbb{H}_n(s-))^2}{(1-\mathbb{H}_n(s-))^2(1-H(s-))^2} d\mathbb{H}_n^{(uc)}(s) \\ &\quad + \int_{[0,t]} \frac{1}{(1-H(s-))^2} d(\mathbb{H}_n^{(uc)}(s) - H^{(uc)}(s)) \\ &= \int_{[0,t]} \frac{[(1-H(s-)) - (1-\mathbb{H}_n(s-))][(1-H(s-)+(1-\mathbb{H}_n(s-))]}{(1-\mathbb{H}_n(s-))^2(1-H(s-))^2} d\mathbb{H}_n^{(uc)}(s) \\ &\quad + \int_{[0,t]} \frac{1}{(1-H(s-))^2} d(\mathbb{H}_n^{(uc)}(s) - H^{(uc)}(s)) \\ &\equiv I_n(t) + II_n(t) \end{aligned}$$

where

$$
\begin{aligned}
|I_n(t)| &\leq 2\frac{\sup_{0<s\leq t}|\mathbb{H}_n(s-)-H(s-)|}{(1-\mathbb{H}_n(t-))^2(1-H(t-))^2}\int_{[0,t]} d\mathbb{H}_n^{(uc)}(s) \\
&\leq 2\frac{\sup_{0<s\leq t}|\mathbb{H}_n(s-)-H(s-)|}{(1-\mathbb{H}_n(t-))^2(1-H(t-))^2}\cdot 1 \\
&\to_{a.s.} 0\cdot\frac{1}{(1-H(t-))^4}\cdot 1 = 0
\end{aligned}
$$

if $1-H(t-)>0$ by (2). Also,

$$
\begin{aligned}
|II_n(t)| &\leq |\int_{[0,t]}\frac{1}{(1-H(s-))^2}d(\mathbb{H}_n^{(uc)}(s)-H^{(uc)}(s)| \\
&= |n^{-1}\sum_{i=1}^n\left\{\frac{\Delta_i 1_{[0,t]}(Z_i)}{(1-H(Z_i-)^2}-E\left(\frac{\Delta 1_{[0,t]}(Z)}{(1-H(Z-))^2}\right)\right\}| \\
&\to_{a.s.} 0
\end{aligned}
$$

by the strong law of large numbers where we again use $1-H(t-)>0$. Thus $|\hat{C}_n(t)-C(t)|\leq|I_n(t)|+|II_n(t)|\to_{a.s.}0$. Assuming that $1-\hat{F}_n(t)\to_p 1-F(t)$ this yields

$$(1-\hat{F}_n(t))^2\hat{C}_n(t)\to_p(1-F(t))^2C(t).$$

(d) An approximate 90% confidence interval for $F(200)$ is given by

$$\hat{F}_n(200)\pm z_{.95}n^{-1/2}(1-\hat{F}_n(200))\sqrt{\hat{C}_n(200)}.$$

where $P(N(0,1)>z_{.95})=.05$. For the data given I compute $1-\hat{F}_n(200)=0.6842$, $n^{-1}\hat{C}_n(200)=.022323$, and hence an approximate 90% confidence interval for the point estimator $\hat{F}_n(200)=1-.6842=.3158$ is given by

$$
\begin{aligned}
&0.3158\pm 1.64485(.6842)(.022323)^{1/2} \\
&= 0.3158\pm 1.64485(.10223) \\
&= 0.3158\pm 0.16815=(0.14765, 0.48395). \qquad (3)
\end{aligned}
$$

It turns out that the variance estimator based on $\hat{C}_n$ is *not* the usual one for the Kaplan-Meier estimator: instead the usual Greenwood formula for estimation of $C(t)$ is

$$\widehat{C}_n^{GW}(t)=\int_{[0,t]}\frac{d\mathbb{H}_n^{(uc)}(s)}{(1-\mathbb{H}_n(s-)(1-\mathbb{H}_n(s-)-\Delta\mathbb{H}_n^{(uc)}(s))}.$$

This yields $n^{-1}\widehat{C}_n^{GW}(200)=0.024292$ and the resulting value of $\widehat{Var}_{GW}(\hat{F}_n(t))$ at $t=200$ is 0.011372 (rather than $.6842^2\cdot .022323=0.010451$ as in (5)). This leads to the slightly wider confidence interval

$$0.3158\pm 1.64485(.6842)(.024292)^{1/2}=.3158\pm 0.175403=(0.140397, 0.491203). \quad (4)$$

See Kalblfeisch and Prentice page 15 for a brief discussion of alternatives involving transformations to stay in the range [0, 1] and to improve the normal approximation.

4. (Interval censored or current status data). Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables (survival times) with distribution function F as in Example 4.6.5. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables ("observation times") with a distribution function G which are independent of the X_i's. Unfortunately, we cannot observe the X_i's directly but can only observe $(Y_i, 1_{[X_i \le Y_i]}) \equiv (Y_i, \Delta_i)$, $i = 1, \ldots, n$.
(a) Consider the empirical functions

$$\mathbb{G}_n(t) = n^{-1} \sum_{i=1}^{n} 1\{Y_i \le t\} = \mathbb{P}_n 1\{Y \le t\},$$
$$\mathbb{V}_n(t) = n^{-1} \sum_{i=1}^{n} \Delta_i 1\{Y_i \le t\} = \mathbb{P}_n \Delta 1\{Y \le t\}.$$

Show that for each fixed t we have

$$\mathbb{G}_n(t) \to_{a.s.} G(t), \qquad \text{and} \qquad \mathbb{V}_n(t) \to_{a.s.} \int_0^t F dG \equiv V(t).$$

(b) Plot the cumulative sum diagram $\{(n\mathbb{G}_n(T_{(i)}), n\mathbb{V}_n(T_{(i)})) : i = 1, \ldots, n\}$ and the MLE $\hat{F}_n$ of F as described in example 4.6.5, page 38 of the notes, for the following data: $(3.7, 0), (1.3, 1), (5, 5, 1), (6.3, 0), (4.7, 1)$. (c) What would the MLE of F be (at $t = 5$) if we assumed that F is exponential θ distribution (with $1 - F_\theta(x) = \exp(-\theta x)$ for $x > 0$)? Compare with the value of the MLE $\hat{F}_n(5)$.

Solution: (a) By the Glivenko-Cantelli theorem we have $\|\mathbb{G}_n - G\|_\infty = \sup_{t>0} |\mathbb{G}_n(t) - G(t)| \to_{a.s.} 0$, and $\|\mathbb{V}_n - V\|_\infty = \sup_{t>0} |\mathbb{V}_n(t) - V(t)| \to_{a.s.} 0$ where

$$\begin{aligned} V(t) &= E\Delta 1\{Y \le t\} = E\{E[\Delta 1\{Y \le t\}|Y]\} = E\{1\{Y \le t\}E[\Delta|Y]\} \\ &= E\{1\{Y \le t\}F(Y)\} = \int_{[0,t]} F(y) dG(y). \end{aligned}$$

Thus, in particular, the pointwise convergences hold as claimed.
(b) Here is a table of the observed, values, the corresponding cumulative sum diagram, and the estimator at the observed points:

Table 2:

i	1	2	3	4	5
$Y_{(i)}$	1.2	3.5	4.2	5.7	6.1
$\Delta_{(i)}$	1	0	1	1	0
$n\mathbb{G}_n(Y_{(i)})$	1	2	3	4	5
$n\mathbb{V}_n(Y_{(i)})$	1	1	2	3	3
$\hat{P}_i$(left)	1/2	1/2	2/3	2/3	2/3
$\hat{P}_i$(right)	1/2	2/3	2/3	2/3	–

The resulting estimator $\hat{F}_n$ of F is given by

$$\hat{F}_n(t) = \begin{cases} 0, & Y_{(0)} \equiv 0 \le t < 1.2 = Y_{(1)}, \\ 1/2, & Y_{(1)} = 1.2 \le t < 4.2 = Y_{(3)}, \\ 2/3, & Y_{(3)} = 4.2 \le t < 6.1 = Y_{(5)}, \\ 2/3, & Y_{(4)} = 6.1 \le t < \infty. \end{cases}$$

The following figures show the cumulative sum diagram and the resulting estimator of the distribution functions F.

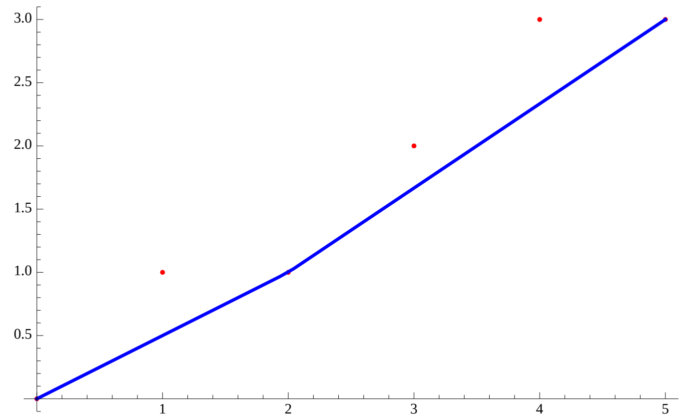

Figure 1: Cumulative Sum Diagram.

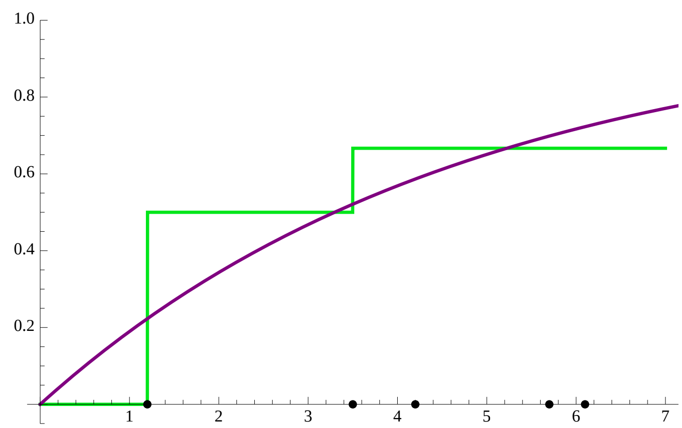

Figure 2: Maximum Likelihood Estimators $\hat{F}_n$ and $\hat{F}_{par}$ of F

(c) If we assume a parametric model for F, namely the exponential distribution $F_\theta(x) = 1 - \exp(-\theta x)$, then the likelihood is

$$\begin{aligned} L(\theta|\underline{Y}, \underline{\Delta}) &= \prod_{i=1}^{n} F_\theta(Y_i)^{\Delta_i}(1 - F_\theta(Y_i))^{1-\Delta_i} g(Y_i) \\ &= \prod_{i=1}^{n} (1 - e^{-\theta Y_i})^{\Delta_i} e^{-\theta Y_i(1-\Delta_i)} \cdot \text{a factor depending on} \quad g \end{aligned}$$

For the given data the likelihood is

$$L(\theta) = (1 - e^{-1.2\theta})(1 - e^{-5.7\theta})(1 - e^{-4.2\theta})e^{-6.1\theta}e^{-3.5\theta}$$

Use of a numerical maximization routine (I used Mathematica) yields $\hat{\theta} = 0.21033$, and this gives $\hat{F}_{par}(5) = 1 - \exp(-5(.21033)) = 0.6506$. This should be compared to the nonparametric estimator at $t = 5$ which is $\hat{F}_n(5) = 2/3$.

5. **Optional bonus problem:** Suppose that F and G are continuous distribution functions.
(a) Show that if $\mathbb{U}$ is a standard Brownian bridge process on $[0,1]$ and $\mathbb{B}$ is a standard Brownian motion process on $[0,\infty)$, then $(1+t)\mathbb{U}(t/(1+t)) \stackrel{d}{=} \mathbb{B}(t)$ as processes on $[0,\infty)$.
(b) Use the result of (a) to show that the limit process for the Kaplan-Meier estimator $(1-F(t))\mathbb{B}(C(t))$ satisfies

$$(1-F(t))\mathbb{B}(C(t)) \stackrel{d}{=} \left(\frac{1-F(t)}{1-K(t)}\right)\mathbb{U}(K(t))$$

as processes on $[0,\tau]$ for any $\tau < \tau_H$ where $C(t) = \int_0^t (1-H(s))^{-2} dH^{(uc)}(s)$ and $K(t) \equiv C(t)/(1+C(t))$.
(c) Show that when there is no censoring (so $G \equiv 0$), $K(t) = F(t)$ for $t < \tau_H$.

Solution: (a) Note that $E\{(1+t)\mathbb{U}(t/(1+t))\} = 0$ for $0 \le t < \infty$; furthermore, for $0 < s < t < \infty$, $s/(1+s) < t/(1+t)$ and hence

$$\begin{aligned} E\left\{(1+s)\mathbb{U}\left(\frac{s}{1+s}\right)(1+t)\mathbb{U}\left(\frac{t}{1+t}\right)\right\} &= (1+s)(1+t)\left\{\frac{s}{1+s} - \frac{s}{1+s}\cdot\frac{t}{1+t}\right\} \\ &= s(1+t) - st = s = s \wedge t. \end{aligned}$$

Since $\{(1+t)\mathbb{U}(t/(1+t)) : 0 \le t < \infty\}$ is a Gaussian process, it follows that $(1+t)\mathbb{U}(t/(1+t)) \stackrel{d}{=} \mathbb{B}(t)$.
(b) By (a) it follows that

$$(1+C(t))\mathbb{U}\left(\frac{C(t)}{1+C(t)}\right) \stackrel{d}{=} \mathbb{B}(C(t))$$

for $0 \le t < \infty$. But $K(t) = C(t)/(1+C(t))$ and $1-K(t) = 1/(1+C(t))$, so it follows that

$$\begin{aligned} (1-F(t))\mathbb{B}(C(t)) &\stackrel{d}{=} (1-F(t))\,(1+C(t))\mathbb{U}\left(\frac{C(t)}{1+C(t)}\right) \\ &= \frac{1-F(t)}{1-K(t)}\mathbb{U}(K(t)). \end{aligned}$$

(c) Suppose that F is continuous and there is no censoring: $G(t) = 0$ for all $t \ge 0$. Then it follows that

$$C(t) = \int_{[0,t]} \frac{1}{(1-F(s))^2} dF(s) = \frac{1}{1-F(t)} - 1 = \frac{F(t)}{1-F(t)}.$$

Then

$$K(t) = \frac{C(t)}{1+C(t)} = \frac{F(t)/(1-F(t))}{1/(1-F(t))} = F(t) \quad \text{for } \ t < \tau_H = \tau_F \wedge \tau_G.$$