Kaplan-Meier in SAS

filename foo url "http://math.unm.edu/~james/small.txt";

data small;
  infile foo firstobs=2;
  input time censor;
run;

proc print data=small;
run;

proc lifetest data=small plots=survival;
  time time*censor(0); *first "time" is a keyword, second "time" is a variable;
run;
Dealing with left-truncated data

The K-M curve allowed for right-censored data. This approach can be modified to allow left-truncated data as well.

The idea is that a patient might not have entered the study until a certain age (for example, only patients who are senior citizens are enrolled, or only patients at a retirement home). This can also be useful to determine the survival curve conditional on already have reached a certain age or conditional on having lived a certain amount of time past diagnosis or treatment.

For this type of data, the $j$th individual has a random age $L_j$ at which the study was joined and a time of death $T_j$. 
Dealing with left-truncated data

For this type of data, \( d_i \) and \( t_i \) are defined as before, with \( d_i \) being the number of deaths observed at time \( i \) and \( t_i \) being the \( i \)th distinct death time, but \( Y_i \) is modified. In particular, \( Y_i \) is the number of individuals satisfying \( L_j < t_i \leq T_j \). In other words,

\[
Y_i = \sum_{j \in \{\text{individuals}\}} I(L_j < t_i \leq T_j)
\]

where \( I(\cdot) \) is an indicator function equal to 1 if the condition is true and 0 otherwise.

Why are the inequalities strict and then not strict?
Dealing with left-truncated data

To understand the inequalities, we presume that someone is alive at the time they are enrolled, so that if $L_j = t_i$, then individual $j$ couldn’t have died at time $t_i$, so we exclude these cases.

Also, if $L_j < t_i$, then if $T_j = t_i$ (which is possible), then they were one of the patients available to die at time $t_i$ who did in fact die at that time.
Dealing with left-truncated data

Essentially, $Y_i$ is defined as before, but we reduce the counts for people available to die at time $t_i$ by not included those individuals who are left-truncated at time $t_i$ – they couldn’t possibly have died by that time because of the way they were recruited into the study.

The product-limit estimator is then given conditional on survival past a certain age $a$ as

$$\hat{S}_a(t) = \prod_{a \leq t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right], \quad t \geq a$$

Given the same sample of data, you could compute, for example

$$\hat{S}_{70}(t)$$

and $\hat{S}_{80}(t)$, which would give different survival curves conditional on having already lived to 70 and 80 years old, respectively. This is deliberately left-truncating your data to determine conditional survival curves.
Example, Channing house data

Data on survival times for patients in a retirement home.

Figure 4.11  Estimated conditional survival functions for Channing house residents. 68 year old females (———); 80 year old females (———); 68 year old males (———); 80 year old males (———).
### Example, psychiatric patients

**TABLE 1.7**

*Survival data for psychiatric inpatients*

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age at Admission</th>
<th>Time of Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>58</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>55</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>30+</td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>28</td>
</tr>
<tr>
<td>Female</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>Female</td>
<td>48</td>
<td>11</td>
</tr>
<tr>
<td>Female</td>
<td>47</td>
<td>14</td>
</tr>
<tr>
<td>Female</td>
<td>25</td>
<td>36+</td>
</tr>
<tr>
<td>Female</td>
<td>31</td>
<td>31+</td>
</tr>
<tr>
<td>Male</td>
<td>24</td>
<td>33+</td>
</tr>
<tr>
<td>Male</td>
<td>25</td>
<td>33+</td>
</tr>
<tr>
<td>Female</td>
<td>30</td>
<td>37+</td>
</tr>
<tr>
<td>Female</td>
<td>33</td>
<td>35+</td>
</tr>
<tr>
<td>Male</td>
<td>36</td>
<td>25</td>
</tr>
<tr>
<td>Male</td>
<td>30</td>
<td>31+</td>
</tr>
<tr>
<td>Male</td>
<td>41</td>
<td>22</td>
</tr>
<tr>
<td>Female</td>
<td>43</td>
<td>26</td>
</tr>
<tr>
<td>Female</td>
<td>45</td>
<td>24</td>
</tr>
<tr>
<td>Female</td>
<td>35</td>
<td>35+</td>
</tr>
<tr>
<td>Male</td>
<td>29</td>
<td>34+</td>
</tr>
<tr>
<td>Male</td>
<td>35</td>
<td>30+</td>
</tr>
<tr>
<td>Male</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>40</td>
</tr>
<tr>
<td>Male</td>
<td>32</td>
<td>39+</td>
</tr>
</tbody>
</table>

* Censored observation
Example, psychiatric patients

Problem 4.8 in the book’s homework problems asks you to plot $Y_i$ as a function of time for this data, treating the data as left-censored based on the entry age of each patient. The problem also asks you to compute the conditional survival function given that the patient entered the hospital at age at least 30. (Note that the data is left-censored because a patient couldn’t have been observed to die earlier than their entrance age, yet from this population, there might have been patients who had died (for example, by suicide) before being admitted.)
Example, psychiatric patients

The unique death times can be found using the `unique()` function in R.

```r
> enter <- c(58,58,59,60,60,61,61,62,62,63,63,64,66,66,67,67,68,69,69,70,70,71,72,72,73,73,73,74,71,72,72,73,73,73,80,74)
> exit <- c(60,63,69,62,65,72,69,73,66,65,68,74,71,68,69,70,77,69,72,79,72,70,76,71,78,79,76,76,73,80,74)
> censor
  [1] 1 1 0 1 1 0 0 0 1 1 1 0 1 1 1 1 1 1 0 1 1 0 1 0 0 1 1 0 1
> length(censor)
[1] 30
> length(enter)
[1] 30
> length(exit)
[1] 30
> ti <- sort(unique(exit[censor==1]))
> ti
  [1]  60  62  63  65  66  68  69  70  71  72  73  74  76  77
```
Example, psychiatric patients

To count $Y_1$, for example, this is the number of patients available to have died at age 60 in the hospital. This includes the one individual who did die at age 60, plus those who were admitted to the hospital at less than 60 years of age but who died (or were right-censored) at age 60 or later. From the table of data, there were only 3 such individuals.

For $Y_2$, the death time is $t_2 = 62$, so we want the number of individuals enrolled at age younger than 62 who were available to die at age 62. There were 7 individuals enrolled before age 62, one of whom died at age 60, so was unavailable. The rest were available, so $Y_2 = 60$. 
Example, psychiatric patients

Some code for computing $Y_i$ from the other data is

```r
> y <- 1:14  # number of unique death times
> for(i in 1:14) {
+ y[i] <- length(enter[enter < ti[i] & ti[i] <= exit])
+ }
> y
[1]  3  6  8 10  8 12 11 10 11 10  9  9  7  5
```

It is typical for left-truncated data that the $Y_i$ terms are small initially, then increase, then decrease again. The truncation can make it so that there aren’t many patients available for earlier times. For data we’ve seen previously that was only right-censored, the $Y_i$ terms were decreasing.
Example, psychiatric patients

Computing $d_i$ from the other data:

```r
> d <- 1:14
> for(i in 1:14) {
+   d[i] <- sum(exit==ti[i] & censor==1)
+ }
> d
[1] 1 1 1 2 1 2 2 2 2 2 2 1 1 1 1
> sum(d)
[1] 20 # check that the total number of deaths is correct
```
Example, psychiatric patients

```r
> km <- 1:14
> for(i in 1:14) {
+ }
> km <- 1:15
> for(i in 1:14) {
+ km[i+1] <- km[i]*(1-d[i]/y[i])
+ }
> km
  [1] 1.00000000 0.66666667 0.55555556 0.48611111 0.38888889 0.34027778
  [7] 0.28356481 0.23200758 0.18560606 0.15185950 0.12148760 0.10798898
  [13] 0.09599021 0.08227732 0.06582185
> ti <- c(58,ti)
> plot(ti,km,type="s",xlab="Age",ylab="Probability of survival",lwd=3,cex.axis=1.3,cex.lab=1.3)
```
Example, psychiatric patients
Example, psychiatric patients

Of course, there is an easier way in R.

```r
km2 <- survfit(Surv(enter,exit,censor)~1,type="kaplan-meier")
plot(km2,xlim=c(58,80),xlab="Age",ylab="Probability of survival",
lwd=c(2,1,1),cex.axis=1.3,cex.lab=1.3)
```
Example, psychiatric patients

```r
> summary(km2)
Call: survfit(formula = Surv(enter, exit, censor) ~ 1, type = "kaplan-meier")

  time n.risk n.event entered censored survival std.err  lower 95% CI
  60    3      1      2      0        0  0.6667  0.2722       0.2995
  62    6      1      3      0        0  0.5556  0.2485       0.2312
  63    8      1      2      0        0  0.4861  0.2269       0.1947
  65   10      2      0      0        0  0.3889  0.1916       0.1480
  66    8      1      2      0        0  0.3403  0.1737       0.1251
  68   12      2      1      0        0  0.2836  0.1493       0.1010
  69   11      2      3      2        0  0.2320  0.1266       0.0796
  70   10      2      3      0        0  0.1856  0.1054       0.0610
  71   11      2      1      0        0  0.1519  0.0889       0.0482
  72   10      2      2      1        0  0.1215  0.0737       0.0370
  73    9      1      2      1        0  0.1080  0.0667       0.0322
  74    9      1      0      1        0  0.0960  0.0604       0.0280
  76    7      1      0      1        0  0.0823  0.0533       0.0231
  77    5      1      0      0        0  0.0658  0.0451       0.0172
```
Example, psychiatric patients
Example, psychiatric patients, default plot with `plot(km2)`
This chapter deals with methods for left-censoring and inter-censoring, as well as right-truncation. In the rare case of having left-censoring but no right-censoring, you can do a trick which is to pick a time $\tau$ which is larger than any observed time (censored or not). Then define all times as $x_i = \tau - t_i$. Then do your analysis on the $x_i$ times, treating observations that were censored as right-censored. The product-limit estimator still works, but then is estimating

$$P[\tau - X > t] = P[X < \tau - t]$$

which is still a decreasing function of $t$. 
We’ll first consider an example with left- and right-censoring.

The example is in section 1.17, on the time to first marijuana use for high school boys in California (based on a study in the 1970s). Boys were asked “When did you first use marijuana”, and ages are given, so data are to the nearest year. If a 15-year old responds “I have never used it?”, how should this be recorded?
Marijuana example

If the boy says he has never used it, then we can treat this as right-censored at age 15 in this case since he might use it in the future. If the boy says he’s been using it at least one year, then it would be left-censored at age 14 since he might have started at age 13 or earlier.
### TABLE 1.8
*Marijuana use in high school boys*

<table>
<thead>
<tr>
<th>Age</th>
<th>Number of Exact Observations</th>
<th>Number Who Have Yet to Smoke Marijuana</th>
<th>Number Who Have Started Smoking at an Earlier Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>19</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>24</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>13</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>&gt;18</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Marijuana example

Having both left- and right-censoring in the same data is a bit tricky to deal with. One approach is an iterative procedure that starts by ignoring the left-censored observations and starts with the product-limit estimator as the initial guess $S_0$ of the survival curve. The curve is then adjusted to deal with the left-censored observations.

For this algorithm, define the times to be $t_i$ (these can include times where only censoring has been observed). Let $c_i$ be the number of left-censored observations for time $t_i$ and let $r_i$ be the number of right-censored observations at time $t_i$. $d_i$ is the number of events at time $t_i$. You can define $Y_i = \sum_{j: j \geq i} d_j + r_j$. Then define $S_0(t)$ as the usual product-limit estimator.
### TABLE 5.1
**Initial Estimate of the Survival Function Formed by Ignoring the Left-Censored Observations**

<table>
<thead>
<tr>
<th>$i$</th>
<th>$t_i$</th>
<th>$c_i$</th>
<th>$d_i$</th>
<th>$r_i$</th>
<th>$Y_i = \sum_{j=i}^{m} d_j + r_j$</th>
<th>$S_0(t_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>179</td>
<td>1.000</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>175</td>
<td>0.978</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>0</td>
<td>19</td>
<td>2</td>
<td>163</td>
<td>0.911</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>1</td>
<td>24</td>
<td>15</td>
<td>142</td>
<td>0.804</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>2</td>
<td>20</td>
<td>24</td>
<td>103</td>
<td>0.669</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>2</td>
<td>13</td>
<td>18</td>
<td>59</td>
<td>0.539</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>3</td>
<td>3</td>
<td>14</td>
<td>28</td>
<td>0.420</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>11</td>
<td>0.375</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0.341</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0.341</td>
</tr>
<tr>
<td>10</td>
<td>&gt;18</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0.000</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>100</td>
<td>79</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Marijuana example

The method is based on taking an initial estimate of the survival function, and using this to estimate the expected number of deaths (or events) at time $t_i$. So given $S_0(t)$, we re-estimate $d_i$ as $\hat{d}_i$ for each $t_i$. The new estimate of the $d_i$ terms leads to another estimate of the survival function, $S_1(t)$. This in tern can be used to re-estimate the $d_i$ terms. Reiterating back and forth, we get $S_1(t), S_2(t), \ldots, S_k(t), \ldots$, a sequence of estimated survival functions.

The sequence of survival functions should converge, so you can stop the algorithm once the estimated survival functions don’t change much. This idea is based on the EM (Expectation-Maximization) algorithm which is used in other areas of statistics as well. The algorithm was formalized in a famous paper in 1977 (Dempster, Laird, and Rubin).
Intuitively, the idea is that if we don’t know how many events occurred at time $t_i$ (due to left-censoring), then we replace $d_i$ with $E[d_i|S_k(t)]$, where $S_k(t)$ itself is estimated, so the expectation isn’t exact. When we use these updated $d_i$ values, we can get an improved estimate $S_{k+1}(t)$ of the survival function, so we can repeat the procedure to get an even better estimate of $d_i$, and so on.
Marijuana example

The algorithm can be described as

1. Let $k = 0$ and estimate the survival function $S_k(t)$ ignoring left-censored observations at times $t_i$.

2. Using the current estimate $S_k(t)$, estimate $p_{ij} = P[t_{j-1} < X \leq t_j| X \leq t_i]$ by $\frac{S_k(t_{j-1}) - S_k(t_j)}{1 - S_k(t_i)}$ for $j \leq i$.

3. Replace the current estimate $d_i$ with $\hat{d}_i = d_i + \sum_{i \geq j} c_i p_{ij}$

4. Compute the survival curve $S_k(t)$ based on only right-censored data but with the updated $d_i$ values. If $S_k(t)$ is close to $S_{k-1}$ (for $k > 1$), then stop. Otherwise, let $k = k + 1$ and go to step 2.
Computing $p_{ij}$

Note that $\sum_{j=1}^{i} p_{ij} = 1$, this gives the probability that the event occurred at time $j$ given that it occurred some time at or before time $j$. So the $p_{ij}$ give the probabilities of the possible earlier times when the event could have occurred. The numerators are successive differences in the survival function, and the denominators don’t depend on $j$.

$$p_{41} = \frac{1.000 - 0.978}{1 - 0.669} = 0.067; \quad p_{42} = \frac{0.978 - 0.911}{1 - 0.669} = 0.202;$$

$$p_{43} = \frac{0.911 - 0.804}{1 - 0.669} = 0.320; \quad p_{44} = \frac{0.804 - 0.669}{1 - 0.669} = 0.410.$$
Computing $p_{ij}$: columns are $i$

**TABLE 5.2**  
*Values of $p_{ij}$ in Step 1*

<table>
<thead>
<tr>
<th>$i/j$</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.067</td>
<td>0.048</td>
<td>0.039</td>
<td>0.036</td>
<td>0.034</td>
<td>0.034</td>
</tr>
<tr>
<td>2</td>
<td>0.202</td>
<td>0.145</td>
<td>0.116</td>
<td>0.107</td>
<td>0.102</td>
<td>0.102</td>
</tr>
<tr>
<td>3</td>
<td>0.320</td>
<td>0.230</td>
<td>0.183</td>
<td>0.170</td>
<td>0.161</td>
<td>0.161</td>
</tr>
<tr>
<td>4</td>
<td>0.410</td>
<td>0.295</td>
<td>0.234</td>
<td>0.218</td>
<td>0.206</td>
<td>0.206</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.281</td>
<td>0.224</td>
<td>0.208</td>
<td>0.197</td>
<td>0.197</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>0.205</td>
<td>0.190</td>
<td>0.180</td>
<td>0.180</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td>0.072</td>
<td>0.068</td>
<td>0.068</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.052</td>
<td>0.052</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
</tr>
</tbody>
</table>
### TABLE 5.3

*First Step of the Self-Consistency Algorithm*

<table>
<thead>
<tr>
<th>$t_i$</th>
<th>$d$</th>
<th>$r_i$</th>
<th>$Y_i$</th>
<th>$S_1(t_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td>10</td>
<td>4.487</td>
<td>0</td>
<td>191.000</td>
<td>0.977</td>
</tr>
<tr>
<td>11</td>
<td>13.461</td>
<td>0</td>
<td>186.513</td>
<td>0.906</td>
</tr>
<tr>
<td>12</td>
<td>21.313</td>
<td>2</td>
<td>173.052</td>
<td>0.794</td>
</tr>
<tr>
<td>13</td>
<td>26.963</td>
<td>15</td>
<td>149.739</td>
<td>0.651</td>
</tr>
<tr>
<td>14</td>
<td>22.437</td>
<td>24</td>
<td>107.775</td>
<td>0.516</td>
</tr>
<tr>
<td>15</td>
<td>14.714</td>
<td>18</td>
<td>61.338</td>
<td>0.392</td>
</tr>
<tr>
<td>16</td>
<td>3.417</td>
<td>14</td>
<td>28.624</td>
<td>0.345</td>
</tr>
<tr>
<td>17</td>
<td>1.207</td>
<td>6</td>
<td>11.207</td>
<td>0.308</td>
</tr>
<tr>
<td>18</td>
<td>0.000</td>
<td>0</td>
<td>4.000</td>
<td>0.308</td>
</tr>
<tr>
<td>&gt;18</td>
<td>4.000</td>
<td>0</td>
<td>4.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
The updated version is fairly close to the estimate ignoring left-censoring. We use the following R code to plot the two estimates simultaneously:

```r
> x <- c(1,.978,.911,.804,.669,.539,.420,.375,.341,.341,0)
> x2 <- c(1,.977,.906,.794,.651,.516,.392,.345,.308,.308,0)
> t <- c(0,10,11,12,13,14,15,16,17,18,19)
> plot(t,x,xlab="Age",ylab="Probability on no marijuana use", cex.axis=1.3,cex.lab=1.3,lwd=3,type="s")
> points(t,x2,lwd=3,type="s",col="blue")
> legend(0,.4,legend=c(expression(S[0](t)),expression(S[1](t))),fill=c("black","blue"),cex=2)
```
Updated survival function

![Graph showing updated survival function with age and probability on no marijuana use.](image)
Repeating the procedure

The book stops at one iteration, and the curves are fairly close, so you might think it is not important to iterate again. We’ll illustrate iterating again since in other data sets it might be more important to iterate more than once. To iterate again, we need to compute a new set of $p_{ij}$ values. For example, we now have

\[
p_{41} = \frac{1.0 - 0.977}{1 - 0.651} = 0.066
\]

\[
p_{42} = \frac{0.977 - 0.906}{1 - 0.651} = 0.203
\]

\[
p_{43} = \frac{0.906 - 0.794}{1 - 0.651} = 0.321
\]

\[
p_{44} = \frac{0.794 - 0.651}{1 - 0.651} = 0.410
\]
Computing $p_{ij}$

To fill in the matrix

```r
> pij <- matrix(ncol=9,nrow=9)
> for(i in 1:9) {
+   for(j in 1:9) {
+     if(j <= i) pij[i,j] <- (S[j]-S[j+1])/(1-S[i+1])
+     else pij[i,j] <- 0
+   }
+ }
```
Repeating the procedure

```r
> options(digits=3)
> t(pij)
[1,] 1.000 0.245 0.112 0.0659 0.0475 0.0378 0.0351 0.0332 0.0332
[2,] 0.000 0.755 0.345 0.2034 0.1467 0.1168 0.1084 0.1026 0.1026
[3,] 0.000 0.000 0.544 0.3209 0.2314 0.1842 0.1710 0.1618 0.1618
[4,] 0.000 0.000 0.000 0.4097 0.2955 0.2352 0.2183 0.2066 0.2066
[5,] 0.000 0.000 0.000 0.0000 0.2789 0.2220 0.2061 0.1951 0.1951
[6,] 0.000 0.000 0.000 0.0000 0.0000 0.2039 0.1893 0.1792 0.1792
[7,] 0.000 0.000 0.000 0.0000 0.0000 0.0000 0.0718 0.0679 0.0679
[8,] 0.000 0.000 0.000 0.0000 0.0000 0.0000 0.0000 0.0535 0.0535
[9,] 0.000 0.000 0.000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
> colSums(t(pij))
[1] 1 1 1 1 1 1 1 1 1
```
Updating the $\hat{d}_i$ values

\begin{verbatim}
> dHatOld <- c(4.487,13.461,21.313,26.963,
22.437,14.714,3.417,1.207,0,4)
> ci <- c(0,0,0,1,2,3,2,3,1) # left-censoring times
  (these don’t change over iterations)
> for(i in 1:9) {
+ dhatNew[i] <- dHatOld[i] + sum(pij[i,]*ci)
+ }
> dHatOld
 [1]  4.49 13.46 21.31  26.96  22.44  14.71  3.42  1.21  0.00  4.00
> dHatNew # oops case-sensitive
Error: object ’dHatNew’ not found
> dhatNew
 [1]  4.49 13.46 21.31 27.37 23.29 16.01  4.76  2.64  1.43  4.00
\end{verbatim}

So the $\hat{d}$ values have changed a little bit.
Computing the survival function

```r
> ri <- c(0,0,2,15,24,18,14,6,0,0)
> Y <- (1:10)*0
> for(i in 1:10) {
+   for(j in i:10) {
+   }
+ }
> Y
 [1] 197.76 193.27 179.81 156.50 114.12  66.83  32.83  14.07  5.43  4.00
```
Computing the survival function

```r
> S2 <- 1:11
> for(i in 2:11) {
+  S2[i] = S2[i-1]*(1-dhatNew[i-1]/Y[i-1])
+  }
> S2
[1] 1.000 0.977 0.909 0.801 0.661 0.526 0.400 0.342 0.278 0.205 0.000
> S1
[1] 1.000 0.977 0.906 0.794 0.651 0.516 0.392 0.345 0.308 0.308 0.000
> S0
[1] 1.000 0.978 0.911 0.804 0.669 0.539 0.420 0.375 0.341 0.341 0.000
```
Updated survival function
Interval-censored data

Interval-censored data can be dealt with in a similar way, making an initial guess about the survival function, and then iteratively computing updated estimated survival curves. The algorithm is outlined below. For the algorithm, an interval-censored observation has the form \((L_i, R_i]\), meaning that the event occurred after time \(L_i\) but before or at time \(R_i\).

To deal with this case, let \(\tau_0 < \tau_1 < \cdots < \tau_m\) be a set of time points that includes all \(L_i\) and \(R_i\) values. Let

\[
\alpha_{ij} = \begin{cases} 
1 & L_i \leq \tau_{j-1} < \tau_j \leq R_i \\
0 & \text{otherwise}
\end{cases}
\]

Then if \(\alpha_{ij} = 1\), then an event with time \((L_i, R_i]\) could have occurred at time \(\tau_j\).
Updated survival function

**Step 1:** Compute the probability of an event’s occurring at time $\tau_j$, $p_j = S(\tau_{j-1}) - S(\tau_j)$, $j = 1, \ldots, m$.

**Step 2:** Estimate the number of events which occurred at $\tau_i$ by

$$d_i = \frac{\sum_{j=1}^{n} \alpha_{ij}p_j}{\sum_k \alpha_{ik}p_k}.$$  

Note the denominator is the total probability assigned to possible event times in the interval $(L_i, R_i]$.

**Step 3:** Compute the estimated number at risk at time $\tau_i$ by $Y_i = \sum_{k=j}^{m} d_k$.

**Step 4:** Compute the updated Product-Limit estimator using the pseudo data found in steps 2 and 3. If the updated estimate of $S$ is close to the old version of $S$ for all $\tau_i$’s, stop the iterative process, otherwise repeat steps 1–3 using the updated estimate of $S$. 