Applied Survival Analysis
Solutions to Lab 2: Kaplan-Meier survival estimate

(a) ***Kaplan-Meier Survival Estimate:***

t^+	d_j	c_j	r_j	$(1-d_j/r_j)$	$\hat{S}(t^+)$
1	1	0	12	0.917	0.917
2	2	1	11	0.818	0.750
3	1	0	8	0.875	0.656
5	1	0	7	0.857	0.563
6	1	0	6	0.833	0.469
7	0	1	5	1.000	0.469
8	1	0	4	0.750	0.352
16	0	1	3	1.000	0.352
17	1	0	2	0.500	0.176
34	0	1	1	1.000	0.176

Note that $\hat{S}(t^+)$ stays the same whenever there is a censored observation only.

(b) Use Greenwood's formula to estimate standard error of the estimated survival function: $$\text{var}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j}$$

Since we are evaluating at t^+ (i.e., just after time t), we actually calculate as:

$$\text{var}[\hat{S}(t^+)] = [\hat{S}(t^+)]^2 \sum_{j:\tau_j\le t} \frac{d_j}{(r_j - d_j)r_j}$$

At $t = 1$, $\text{var}[\hat{S}(t^+)] = (0.917)^2 \frac{1}{11\cdot 12} = 0.0064 \Rightarrow se = 0.0798$

At $t = 3$,

$$\text{var}[\hat{S}(t^+)] = (0.656)^2 \left\{ \frac{1}{11\cdot 12} + \frac{2}{9\cdot 11} + \frac{1}{8\cdot 7} \right\} = (0.656)^2 (0.0456) = 0.0196$$

$$\Rightarrow se = 0.1402$$

For STATA confidence intervals:

$$t = 1 \quad \Rightarrow \quad \left[\hat{S}(t^+)^{e^A} \ , \ \hat{S}(t^+)^{e^{-A}} \right] \qquad A = 1.96 \cdot se(\hat{L}(t^+))$$

$$\text{var}\left[\hat{L}(t^+)\right] = \text{var}\left[\log(-\log(\hat{S}(t^+)))\right] = \frac{1}{\left[\log(\hat{S}(t^+))\right]^2}\sum_{j:\tau_j \le t}\frac{d_j}{(r_j - d_j)r_j}$$

$$= \frac{1}{\left[\log(0.917)\right]^2}\cdot\frac{1}{11\cdot 12} = 1.0006$$

$$se\left[\hat{L}(t^+)\right] = 1.0003 \Rightarrow A = 1.9606$$

Lower Bound : $(0.917)^{e^A} = 0.5390$

Upper Bound : $(0.917)^{e^{-A}} = 0.9878$

$t = 3 \quad \Rightarrow \quad \left[0.656^{e^A}\ ,\ 0.656^{e^{-A}}\right]$

$$\text{var}\left[\hat{L}(t^+)\right] = \text{var}\left[\log(-\log(\hat{S}(t^+)))\right] = \frac{1}{\left[\log(0.656)\right]^2}\cdot\{0.0456\} = 0.2569$$

$$se\left[\hat{L}(t^+)\right] = 0.5069 \Rightarrow A = 0.9935$$

Lower Bound : $(0.656)^{e^A} = 0.3204$

Upper Bound : $(0.656)^{e^{-A}} = 0.8557$

In computing confidence intervals, SAS uses the "plain" approach of:
$\hat{S}(t) \pm 1.96 se[\hat{S}(t)]$, where $se[\hat{S}(t)]$ is estimated using Greenwood's formula.
However, this approach does not always yield confidence intervals between [0,1]. This is apparent for `NHLTIME` =1, where the upper bound has been set to 1. In general, the lower bound and upper bound are both shifted higher for the SAS confidence intervals than the STATA ones. On the other hand, STATA uses confidence intervals based on:

$$\left\{[\hat{S}(t)]^{\exp(1.96 se\hat{L}(t))}, [\hat{S}(t)]^{\exp(-1.96 se\hat{L}(t))}\right\}$$

which ensures that the bounds fall into the right range.

(c) Median: survival time (t) such that $\hat{S}(t) \le 0.5 \Rightarrow \hat{S}(6) = 0.469$, so the estimated median survival time is **6**

Lower quartile (25%) : the smallest time (LQ) such that ,
$\hat{S}(LQ) \le 0.75 \Rightarrow \hat{S}(2) = 0.750$, so the estimated 25%-ile survival time is **2**

Upper quartile (75%) : the smallest time (UQ) such that ,
$\hat{S}(UQ) \le 0.25 \Rightarrow \hat{S}(17) = 0.176$, so the estimated 75%-ile survival time is **17** or

`stsum`

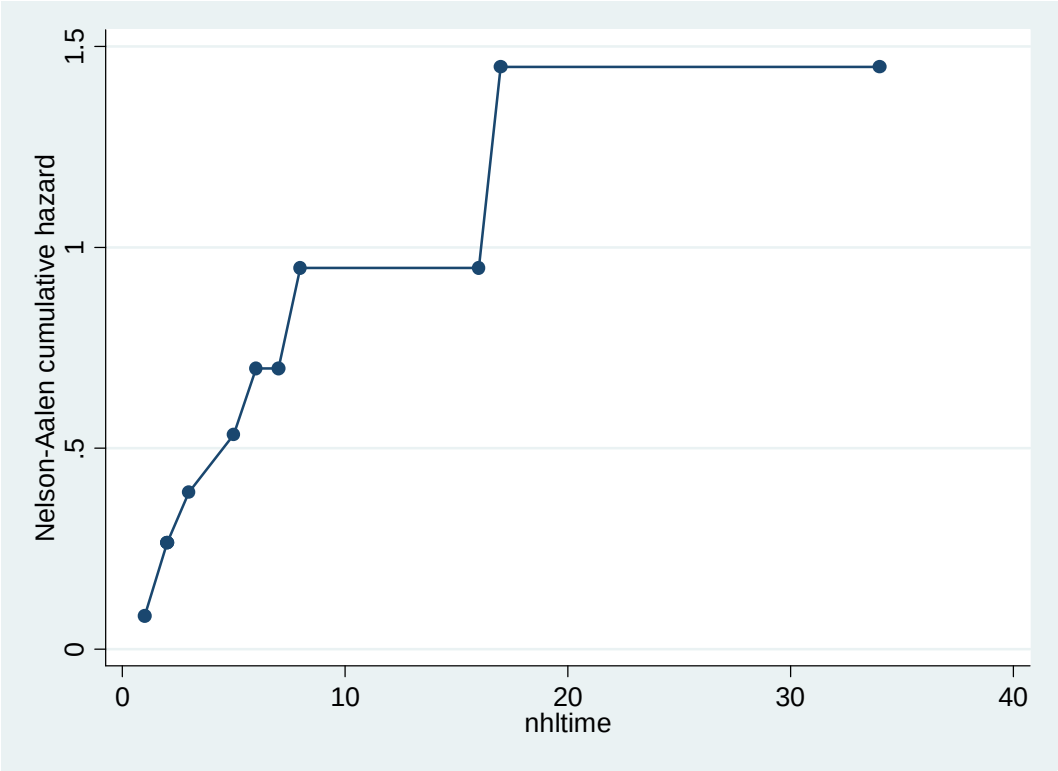
```
          failure _d:  fail
    analysis time _t:  nhltime

          |              incidence       no. of    |------ Survival time -----|
          | time at risk     rate        subjects        25%       50%       75%
----------+-------------------------------------------------------------------
    total |          103   .0776699            12          2         6        17
```

Cumulative Hazard Estimate:

(d) The hazard for relapse or death after a bone marrow transplant for patients wit non-Hodking's lymphoma seems to be increasing over time. Because STATA gives us a plot with steps it is better to construct our own plot, so we have to generate a variable containing the cumulative hazard *nelson.*

`sts generate nelson=na` (**`na`** is an already built-in function)

`scatter nelson nhltime, c(l)`



An exponential model is only appropriate when the hazard is constant over time, which means that we should expect a straight line in the above graph. In the beginning it seems like a straight line but then it curves, so it does not seem likely that an exponential distribution would fit these data well, but we have only 12 observations so it is hard to tell.