# P8110: Applied Regression II
# Homework #2 [15 points]

**NOTE: For the hand calculation questions, please only keep the first three decimals. Show all work.**

1. Listed below are values of survival time in years for 10 patients, with the censored observations denoted by a "+" as a superscript: 0.4, 1.2, $1.2^+$, $3.4^+$, 4.3, 5.0, 5.0, $5.0^+$, $6.1^+$, 7.1. The Kaplan-Meier estimates and 95% CIs of the survival functions are:

| $j$ | $t_j$ | For $t$ in | $n_j$ | $d_j$ | $1-\frac{d_j}{n_j}$ | $\hat{S}(t_j)$ | 95% CI |
|---|---|---|---|---|---|---|---|
| 0 | 0.0 | [0,0.4) | 10 | 0 | 1 | 1 | (1.0, 1.0) |
| 1 | 0.4 | [0.4,1.2) | 10 | 1 | 0.9 | 0.9 | (0.473, 0.985) |
| 2 | 1.2 | [1.2,4.3) | 9 | 1 | 0.889 | 0.8 | (0.409, 0.946) |
| 3 | 4.3 | [4.3,5.0) | 6 | 1 | 0.833 | 0.667 | (0.272, 0.882) |
| 4 | 5.0 | [5.0,7.1) | 5 | 2 | 0.6 | 0.4 | (0.097, 0.698) |
| 5 | 7.1 | [7.1,$\infty$) | 1 | 1 | 0 | 0 | (0, 0) |

   (a) Use SAS to re-generate the estimates of survival functions and 95% CIs in the table above. Show the SAS codes and outputs (only cut and paste the relevant SAS outputs) [2 points]

   <u>*SAS code:*</u>

```
data HW2;
input time status;
datalines;
0.4 1
1.2 1
1.2 0
3.4 0
4.3 1
5.0 1
5.0 1
5.0 0
6.1 0
7.1 1
;
run;
```

```
proc lifetest data=HW2 method=KM alpha=0.05 outsurv=A stderr;
time time*status(0);
run;

proc print data = A; run;
```

proc print data = A; run;

| Obs | time | _CENSOR_ | SURVIVAL | SDF_STDERR | SDF_LCL | SDF_UCL |
|---|---|---|---|---|---|---|
| 1 | 0.0 | . | 1.00000 | 0.00000 | 1.00000 | 1.00000 |
| 2 | 0.4 | 0 | 0.90000 | 0.09487 | 0.47301 | 0.98528 |
| 3 | 1.2 | 0 | 0.80000 | 0.12649 | 0.40869 | 0.94587 |
| 4 | 1.2 | 1 | 0.80000 | . | . | . |
| 5 | 3.4 | 1 | 0.80000 | . | . | . |
| 6 | 4.3 | 0 | 0.66667 | 0.16102 | 0.27168 | 0.88147 |
| 7 | 5.0 | 0 | 0.40000 | 0.17512 | 0.09658 | 0.69824 |
| 8 | 5.0 | 1 | 0.40000 | . | . | . |
| 9 | 6.1 | 1 | 0.40000 | . | . | . |
| 10 | 7.1 | 0 | 0.00000 | . | 0.00000 | 0.00000 |

(b) Hand calculate the median survival time and its 95% confidence interval. Interpret the results. [4 points]

*The median survival time is*

$$\hat{t}_{50} = \min\{t : \hat{S}(t) < 0.5\} = 5,$$

*The 95% CI for the median survival time is* $[0.4, 7.1)$*, because 0.5 lies within the 95% CI of* $S(t)$ *for any value of* $t$ *in the time intervals* $[0.4, 1.2)$*,* $[1.2, 4.3)$*,* $[4.3, 5.0)$*, and* $[5.0, 7.1)$*.*

*Interpretation: We estimate that half of the patients die before 5 years after having acute myocardial infarction. We are 95% confident that the median survival time lies between 0.4 years and 7.1 years.*

(c) Hand calculate the mean survival time estimate $\hat{u}_1$ and the variance estimate $\widehat{Var}(\hat{u}_1)$. [3 points]

*The number of unique time points with events is* $J = 6$ *and the last observed survival time is 7.1. The* $\hat{u}_1$ *can be estimated as the mean* $\hat{\mu}(t_6 = 7.1)$ *based on the observed* *range of event times, as following:*

$$\begin{aligned}
\hat{u}_1 &= \hat{\mu}(7.1) = \sum_{j=1}^{5} \hat{S}(t_{j-1})(t_j - t_{j-1}) \\
&= 1 \times (0.4 - 0) + 0.9 \times (1.2 - 0.4) + 0.8 \times (4.3 - 1.2) \\
&\quad +0.6667 \times (5 - 4.3) + 0.4 \times (7.1 - 5) \\
&= 4.91
\end{aligned}$$

*The variance of this estimate can be calculated as follows:*

$$Var(\hat{u}_1) = \frac{n_d}{n_d - 1} \sum_{j=1}^{4} \frac{A_j^2 d_j}{n_j(n_j - d_j)}$$

$$\text{for} \quad n_d = \sum_{j=1}^{5} d_j = 6 \quad \text{and} \quad A_j = \sum_{k=j+1}^{J} \hat{S}(t_{j-1})(t_j - t_{j-1})$$

*First calculate the* $A_1, ..., A_4$

$\mathrm{A}_1 = \hat{S}(0.4) \times (1.2 - 0.4) + \hat{S}(1.2) \times (4.3 - 1.2) + \hat{S}(4.3) \times (5 - 4.3) + \hat{S}(5) \times (7.1 - 5) = 4.51$

$\mathrm{A}_2 = \hat{S}(1.2) \times (4.3 - 1.2) + \hat{S}(4.3) \times (5 - 4.3) + S\hat{(5)} \times (7.1 - 5) = 3.79$

$\mathrm{A}_3 = \hat{S}(4.3) \times (5 - 4.3) + \hat{S}(5) \times (7.1 - 5) = 1.31$

$\mathrm{A}_4 = \hat{S}(5) \times (7.1 - 5) = 0.84$

*Then the variance can be calculated as,*

$$Var(\hat{u_1}) = \frac{6}{6-1}\left[\frac{4.51^2 \times 1}{10(10-1)} + \frac{3.79^2 \times 1}{9(9-1)} + \frac{1.31^2 \times 1}{6(6-1)} + \frac{0.84^2 \times 2}{5(5-2)}\right] = 0.692$$

(d) Let $\hat{u}_2$ be the mean survival time estimate if all the censored observations were events. Without calculating $\hat{\mu}_2$, do you know which one of $\hat{u}_1$ and $\hat{u}_2$ is bigger? Briefly explain. Use SAS to estimate $\hat{\mu}_2$. [4 points]

$\hat{u}_1 > \hat{u}_2$, *because the actual survival time for the censored subjects is longer than their censored time, and the estimation of* $\hat{u}_2$ *is based on treating the censored observations as if they were events, while the estimation of* $\hat{u}_1$ *is based on the original censoring and event status.*

*SAS code:*

```
data HW2;
input time status;
datalines;
0.4 1
1.2 1
1.2 1
3.4 1
4.3 1
5.0 1
5.0 1
5.0 1
6.1 1
7.1 1
;
run;

proc lifetest data=HW2 method=KM alpha=0.05 outsurv=A stderr;
time time*status(0);
run;
```

*The estimate of* $\hat{u}_2$ *is 3.87*

(e) Repeat the analyses in part (b) and (c) using SAS. Show the SAS codes and outputs (only cut and paste the relevant SAS outputs). [2 points]

*The median survival time is 5 with 95% CI (0.4,7.1) and the mean is 4.91 with variance* $0.83^2 = 0.69$

*SAS code:*

```
data HW2;
input time status;
datalines;
0.4 1
1.2 1
1.2 0
3.4 0
4.3 1
5.0 1
5.0 1
5.0 0
```

```
6.1 0
7.1 1
;
run;

proc lifetest data=HW2 method=KM alpha=0.05 outsurv=A stderr;
time time*status(0);
run;
```

**Summary Statistics for Time Variable time**

| Quartile Estimates | | | | |
|---|---|---|---|---|
| | | 95% Confidence Interval | | |
| Percent | Point Estimate | Transform | [Lower | Upper) |
| 75 | 7.10000 | LOGLOG | 5.00000 | 7.10000 |
| 50 | 5.00000 | LOGLOG | 0.40000 | 7.10000 |
| 25 | 4.30000 | LOGLOG | 0.40000 | 5.00000 |

| Mean | Standard Error |
|---|---|
| 4.90667 | 0.83125 |