SISCER Survival Analysis

Problem Set 1
Answer Key

1. (Life table analysis)

   (a) Three approaches

   - Approach 1

t	n	d	w	$q^r = d/n$	$p^r = 1 - q^r$	$\hat{S}^r = \prod p^r$
0-1	146	27	3	0.185	0.815	0.815
1-2	116	18	10	0.155	0.845	0.689
2-3	88	21	10	0.239	0.761	0.524
3-4	57	9	3	0.158	0.842	0.441
4-5	45	1	3	0.022	0.972	0.432
...						

   - Approach 2

t	n	d	w	$q^l = d/(n-w)$	$p^l = 1 - q^l$	$\hat{S}^l = \prod p^l$
0-1	146	27	3	0.189	0.811	0.811
1-2	116	18	10	0.170	0.830	0.673
2-3	88	21	10	0.269	0.731	0.492
3-4	57	9	3	0.167	0.833	0.410
4-5	45	1	3	0.024	0.977	0.400
...						

   - Approach 3

t	n	d	w	$q = d/(n-w/2)$	$p = 1 - q$	$\hat{S} = \prod p$
0-1	146	27	3	0.187	0.813	0.813
1-2	116	18	10	0.162	0.838	0.681
2-3	88	21	10	0.253	0.747	0.509
3-4	57	9	3	0.162	0.838	0.426
4-5	45	1	3	0.023	0.977	0.417
...						

   (b) Use Greenwoods formula

$$se\left\{\hat{S}(t)\right\} = \hat{S}(t)\left\{\sum_{j=1}^{t} \frac{d_j}{(n_j - w_j/2)(n_j - d_j - w_j/2)}\right\}^{1/2}$$

   and 95% confidence intervals for $S(t)$ is $\hat{S}(t) \pm 1.96 \times se[\hat{S}(t)]$.

2. Exponential distributions.

   (a) Straightforward

   (b) Likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta)^{\Delta_i} S(X_i;\theta)^{1-\Delta_i}$$

   and the MLE can be solved by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta) = \arg\max_{\theta\in\Theta} [\log L(\theta)] = \frac{\sum_i \Delta_i}{\sum_i X_i}.$$

   It is the person-year analysis estimate. A variance estimate is

$$\hat{\text{var}}(\hat{\theta}) = \left[-\frac{\partial^2 \log L(\hat{\theta})}{\partial \theta^2} \right]^{-1} = \frac{\hat{\theta}^2}{\sum_i \Delta_i}$$

   Key assumptions are needed: (1) subjects are independent, (2) non-informative censoring.

3. Use computer to do this exercise

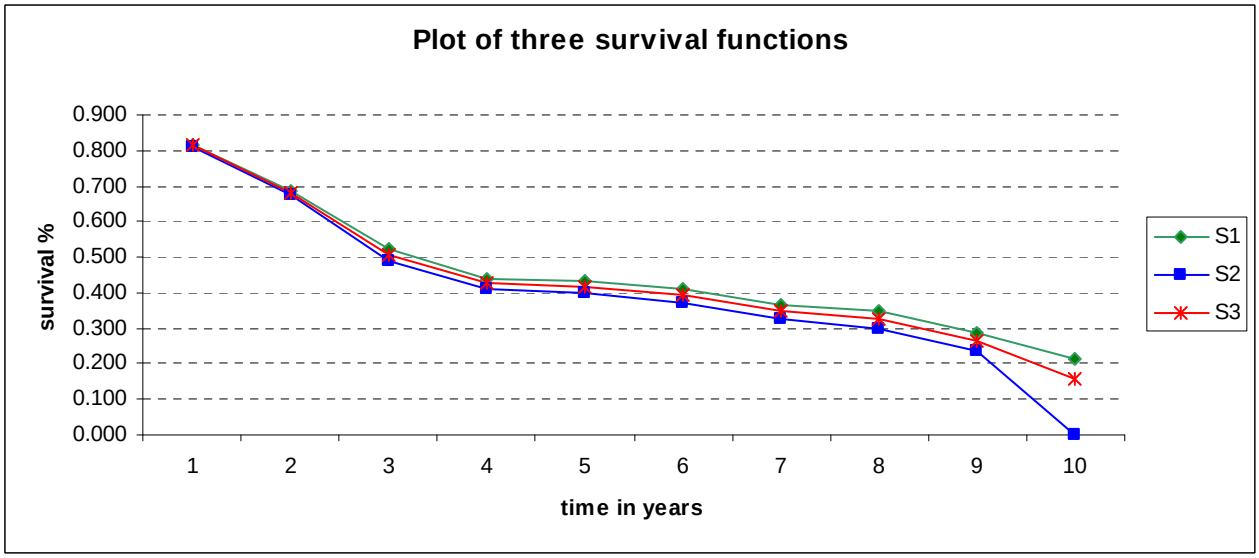
Supplement for Solution of Problem Set 1

Question 1-(b)

95% CI calculation for Approach 3:

Years	$S(t)$	V_j	$\sum_{i=1}^{j} V_j$	$\left(\sum_{i=1}^{j} V_j\right)^{1/2}$	$S(t)\cdot\left(\sum_{i=1}^{j} V_j\right)^{1/2}$	95%CI	
						LB	UB
0-1	0.813	0.0016	0.0016	0.040	0.032	0.750	0.877
1-2	0.681	0.0017	0.0033	0.058	0.039	0.604	0.758
2-3	0.509	0.0041	0.0074	0.086	0.044	0.423	0.595
3-4	0.426	0.0035	0.0109	0.104	0.045	0.339	0.514
4-5	0.417	0.0005	0.0114	0.107	0.045	0.329	0.504
5-6	0.393	0.0017	0.0131	0.115	0.045	0.305	0.481
6-7	0.347	0.0052	0.0184	0.135	0.047	0.255	0.439
7-8	0.325	0.0042	0.0225	0.150	0.049	0.230	0.421
8-9	0.263	0.0224	0.0449	0.212	0.056	0.154	0.373
9-10	0.158	0.1333	0.1783	0.422	0.067	0.027	0.289

$$^{*}\ V_j = \frac{d_j}{\left(n_j - w_j/2\right)\left(n_j - d_j - w_j/2\right)}$$



Three survival estimates are quite similar, however, the first approach produces higher estimates, while the second approach leads to lower estimates which is more conservative. The 95% CIs of the third one cover other two estimates, this approach is widely used in epidemiological studies with right censored data (e.g,, in a longitudinal study, when failure event occurred in the interval between two follow-up visits).

Question 2-(a)

$\because T \sim \exp(\theta)$

$\therefore$ the pdf: $f(t;\theta) = \theta \cdot e^{-\theta t}$ (t>0) (see lecture -1, last two slides)

$\because \lambda(\mathrm{t}) = \theta$

and $\lambda(\mathrm{t}) = \dfrac{f(t)}{S(t)}$

$\therefore$ the survival function: $S(t) = \dfrac{f(t)}{\lambda(t)} = \dfrac{\theta \cdot e^{-\theta t}}{\theta} = e^{-\theta t}$

median of T : $S(t) = e^{-\theta t} = 0.5 \;\Rightarrow\; t = \dfrac{\log 0.5}{-\theta} = \dfrac{\log 2}{\theta}$

Question 2-(b)

(1) the likelihood function:

$$\begin{aligned} \mathrm{L}(\theta) &= \prod_{i=1}^{n} f(X_i;\theta)^{\delta_i} S(X_i;\theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \frac{f(X_i;\theta)^{\delta_i}}{S(X_i;\theta)^{\delta_i}} S(X_i;\theta) \\ &= \prod_{i=1}^{n} \lambda(X_i;\theta)^{\delta_i} S(X_i;\theta) \\ &= \prod_{i=1}^{n} \theta^{\delta_i} \cdot e^{-\theta X_i} \end{aligned}$$

(2) the log-likelihood:

$$\begin{aligned} l(\theta) &= \log\left[\prod_{i=1}^{n} \theta^{\delta_i} \cdot e^{-\theta X_i}\right] \\ &= \sum_{i=1}^{n}\left[\log \theta^{\delta_i} + \log e^{-\theta X_i}\right] \\ &= \sum_{i=1}^{n}\left[\delta_i \log \theta - \theta X_i\right] \end{aligned}$$

(3) take first derivative:

$$\frac{\partial}{\partial\theta}l(\theta)=\sum_{i=1}^{n}\left[\left(\delta_i\log\theta\right)'-\left(\theta X_i\right)'\right]$$
$$=\sum_{i=1}^{n}\left[\delta_i\cdot\frac{1}{\theta}-X_i\right]$$
$$=\frac{\sum_{i=1}^{n}\delta_i}{\theta}-\sum_{i=1}^{n}X_i$$

(4) set $\frac{\partial}{\partial\theta}l(\theta)$ =0, solve for MLE of θ:

$$\frac{\sum_{i=1}^{n}\delta_i}{\theta}-\sum_{i=1}^{n}X_i=0 \quad\Rightarrow\quad \frac{\sum_{i=1}^{n}\delta_i}{\theta}=\sum_{i=1}^{n}X_i \quad\Rightarrow\quad \hat{\theta}_{MLE}=\frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n}X_i}$$

(5) Fisher information (see lecture 1, the last two slides):

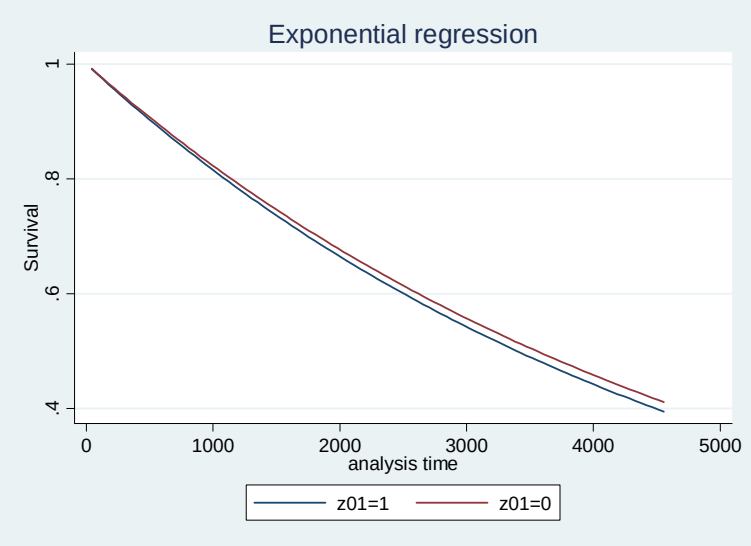
$$\mathrm{I}(\theta)=E\left[-\frac{\partial^2}{\partial\theta^2}\log L(\theta)\right]=E\left[-\frac{\partial}{\partial\theta}\left(\frac{\sum_{i=1}^{n}\delta_i}{\theta}-\sum_{i=1}^{n}X_i\right)\right]$$
$$=E\left[-\left(\frac{\sum_{i=1}^{n}\delta_i}{\theta}\right)'-\left(\sum_{i=1}^{n}X_i\right)'\right] \quad=E\left[-\left(-\frac{\sum_{i=1}^{n}\delta_i}{\theta^2}\right)-0\right]=E\left[\frac{\sum_{i=1}^{n}\delta_i}{\theta^2}\right]=\frac{\sum_{i=1}^{n}\delta_i}{\hat{\theta}_{MLE}^2}$$

(6) $\hat{\mathrm{Var}}\left(\hat{\theta}\right)=\frac{1}{\mathrm{I}(\theta)}=\frac{\hat{\theta}_{MLE}^2}{\sum_{i=1}^{n}\delta_i}$

Key assumption:
(1) subjects are independent;
(2) non-informative censoring (ie, failure time is independent of censoring time).

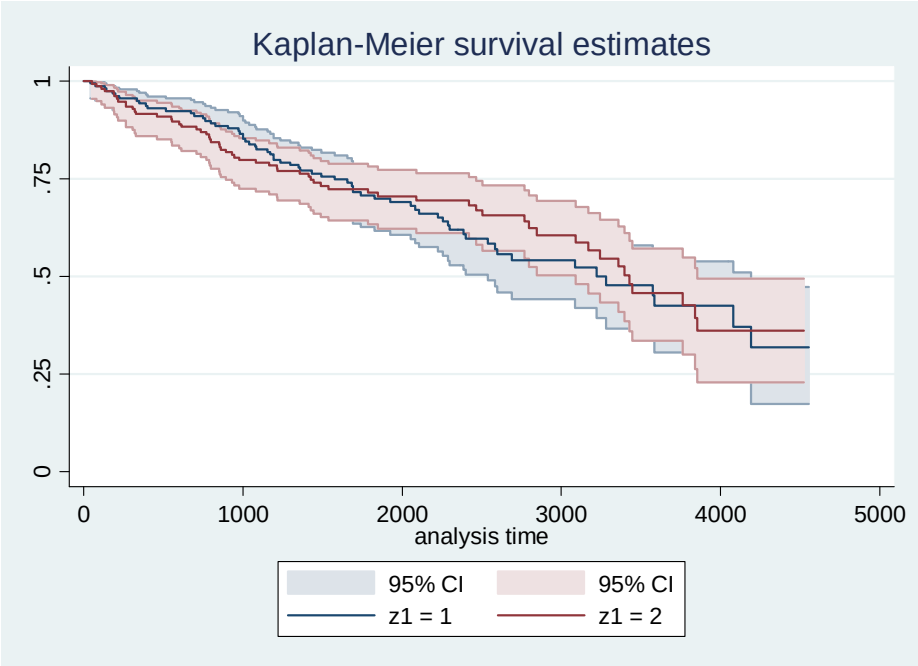
Question 2- (c).

```
. gen z01=z1
. replace z01=0 if z1==2
. streg z01, d(exponential) nohr

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         z01 |   .0450511   .1790287     0.25   0.801    -.3058388    .3959409
       _cons |  -8.541941   .1290994   -66.17   0.000    -8.794971   -8.288911
------------------------------------------------------------------------------

. * hazard rate for two treatment groups
. predictnl haz = predict(hazard), ci(haz_lb haz_ub)
note: Confidence intervals calculated using Z critical values

. list _t haz haz_lb haz_ub z1 in 1/10

     +-------------------------------------------+
     |   _t        haz     haz_lb     haz_ub   z1 |
     |-------------------------------------------|
  1. |  400   .0002041   .0001545   .0002537    1 |
  2. | 4500   .0002041   .0001545   .0002537    1 |
  3. | 1012   .0002041   .0001545   .0002537    1 |
  4. | 1925   .0002041   .0001545   .0002537    1 |
  5. | 1504   .0001951   .0001457   .0002445    2 |
     |-------------------------------------------|
  6. | 2503   .0001951   .0001457   .0002445    2 |
  7. | 1832   .0001951   .0001457   .0002445    2 |
  8. | 2466   .0001951   .0001457   .0002445    2 |
  9. | 2400   .0002041   .0001545   .0002537    1 |
 10. |   51   .0001951   .0001457   .0002445    2 |
     +-------------------------------------------+

. * plot survival functions
. stcurve, survival at1(z01=1) at2(z01=0)
```



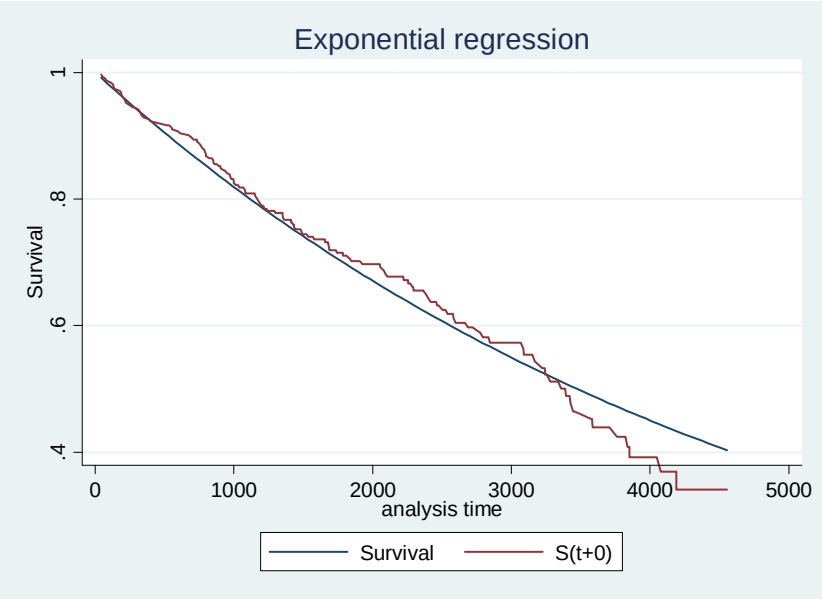
Question 3-a

```
. sts graph, by (z1) gwood
```



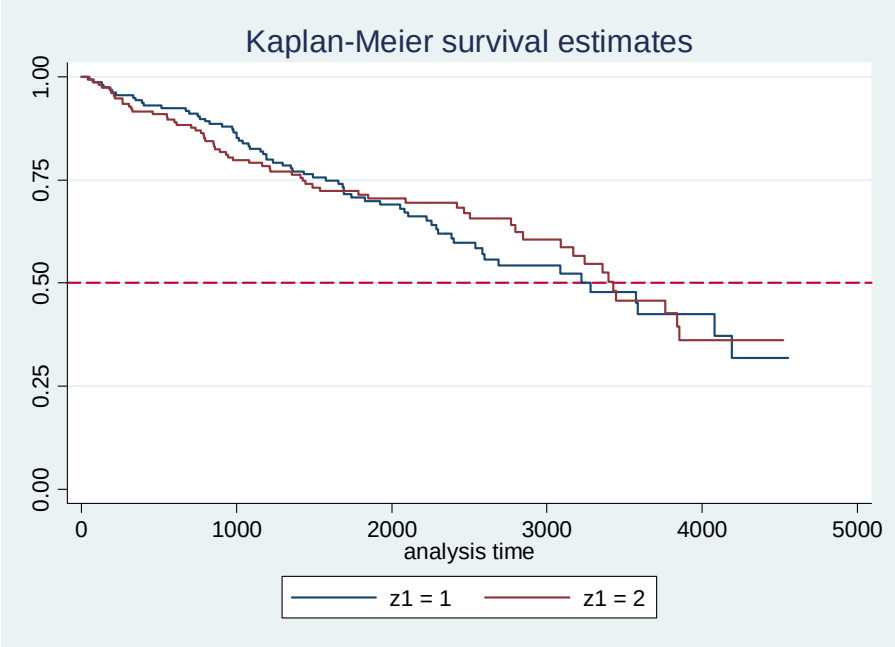
Plot of exponential survival curve versus Kaplan Meier survival curve

```
. * generate Kaplan Meier survival estimate using
. sts gen st= s
. * plot exponential survival curve and Kaplan Meier survival curve in one graph
. stcurve, survival  addplot(line st _t, sort)
```



Question 3-b

```
. sts graph, by (z1)  yline(0.5, lpattern(dash))
```




```
. stci, by(z1)
```

z1	no. of subjects	50%	Std. Err.	[95% Conf.	Interval]
1	158	**3282**	272.7255	**2540**	**4191**
2	154	**3428**	174.1291	**3090**	**3853**
total	312	3395	151.7444	3086	3839

```
. * estimate median time based on exponential regression model
. predict mt,median time
. list mt z1 in 1/10
```

	mt	z1
1.	**3396.08**	**1**
2.	3396.08	1
3.	3396.08	1
4.	3396.08	1
5.	3552.576	2
6.	**3552.576**	**2**
7.	3552.576	2
8.	3552.576	2
9.	3396.08	1
10.	3552.576	2