Tutorial 3: Survival Analysis through R

Miss. Na Zhao

Outline

- Survival Object
- Kaplan-Meier Survival Function
- Cox Model
- RMST Model

Survival Object

Basically observed data in survival analysis is consist of the minimum of survival time and censored time, and one indicator indicating if the observed time is censored or not.

```
# Install the package "survival" in R
# Import the lung cancer dataset and have a look at it
library(survival)
attach(lung)
head(lung)
```

```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1       1       90       100     1175      NA
## 2    3  455      2  68   1       0       90        90     1225      15
## 3    3 1010      1  56   1       0       90        90       NA      15
## 4    5  210      2  57   1       1       90        60     1150      11
## 5    1  883      2  60   1       0      100        90       NA       0
## 6   12 1022      1  74   1       1       50        80      513       0
```

- by help(lung) to see detail of lung dataset
- time: survival time
- status: censoring status 1=censored, 2=dead

Survival Object

```r
#re-order the data by survial time
library(dplyr)
head(arrange(lung, time)) # by default ascending
```

```
##       inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 57     5    5      2  65   2       0      100        80      338       5
## 73     5   11      2  74   1       2       70       100     1175       0
## 79     3   11      2  81   1       0       90        NA      731      15
## 108    1   11      2  67   1       1       90        90      925      NA
## 30     1   12      2  74   1       2       70        50      305      20
## 116    1   13      2  76   1       2       70        70      413      20
```

Survival Object

In R, We need to generate new survival object by function $Surv()$ in "survival" package for further manipulation.

```
# Surv(time, event) creates a survival object for right censored dat
# "time" is the follow up time.
# "event" indicator, normally 0=alive (censored), 1=dead or 1/2 .
Lungsur=Surv(time,status)
head(Lungsur)
```

```
## [1]  306   455  1010+  210   883  1022+
```

```
# pay attention to the meaning of event argument in Surv()
# what if status==1 stands for dead?
Lungsur2=Surv(time,status==1)
head(Lungsur2)
```

```
## [1]  306+  455+ 1010   210+  883+ 1022
```

Kaplan-Meier Survival Function

overall survival curve

- $survfit(formula, \ldots)$ function creates survival curves.
- $formula$ should have the form ” $Surv\ object \sim term1 + \ldots + termk$ ”.
- with term 1, creates the single curve of K-M estimate.

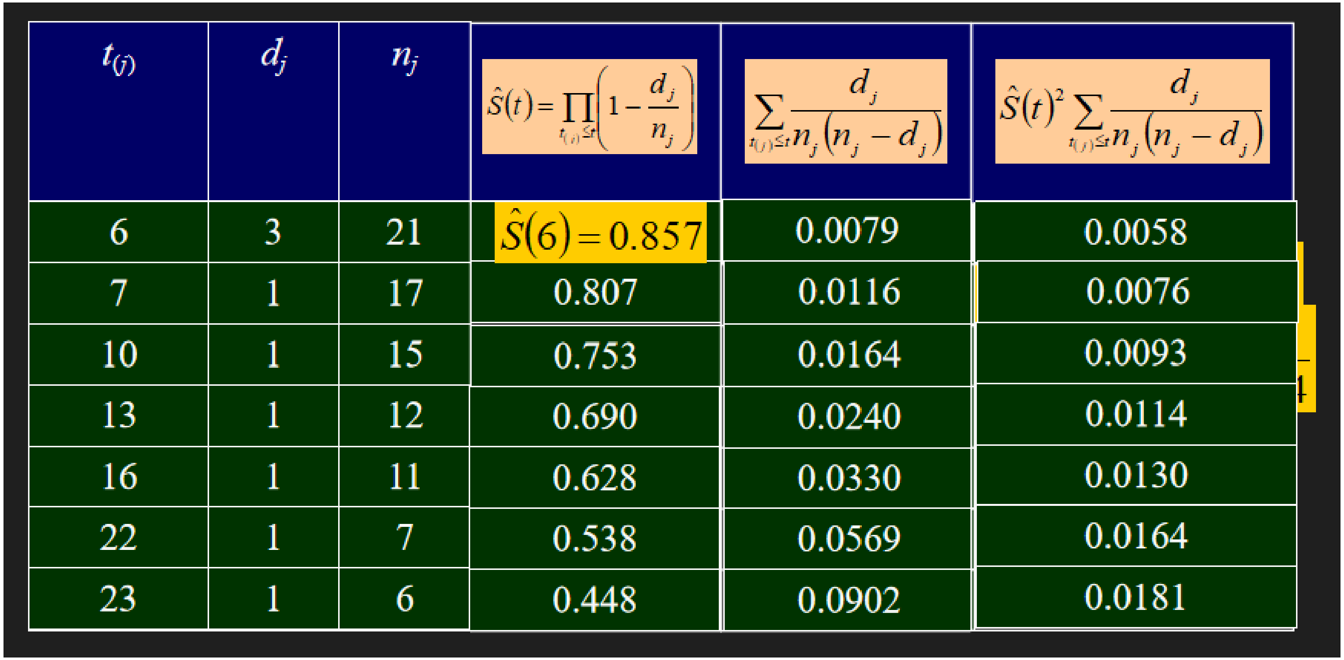
```
fit1=survfit(Lungsur~1,data=lung)
summary(fit1)
```

```
## Call: survfit(formula = Lungsur ~ 1, data = lung)
## 
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     5    228       1   0.9956 0.00438       0.9871        1.000
##    11    227       3   0.9825 0.00869       0.9656        1.000
##    12    224       1   0.9781 0.00970       0.9592        0.997
##    13    223       2   0.9693 0.01142       0.9472        0.992
##    15    221       1   0.9649 0.01219       0.9413        0.989
##    26    220       1   0.9605 0.01290       0.9356        0.986
##    30    219       1   0.9561 0.01356       0.9299        0.983
##    31    218       1   0.9518 0.01419       0.9243        0.980
##    53    217       2   0.9430 0.01536       0.9134        0.974
##    54    215       1   0.9386 0.01590       0.9079        0.970
```

Kaplan-Meier Survival Function

overall survival curve

- You may try to rebuild the K-M estimate as example in p20.
- Construct the same tabel as below, compare it with the previous one.



$t_{(j)}$	d_j	n_j	$\hat{S}(t)=\prod_{t_{(j)}\le t}\left(1-\frac{d_j}{n_j}\right)$	$\sum_{t_{(j)}\le t}\frac{d_j}{n_j\left(n_j-d_j\right)}$	$\hat{S}(t)^2\sum_{t_{(j)}\le t}\frac{d_j}{n_j\left(n_j-d_j\right)}$
6	3	21	$\hat{S}(6)=0.857$	0.0079	0.0058
7	1	17	0.807	0.0116	0.0076
10	1	15	0.753	0.0164	0.0093
13	1	12	0.690	0.0240	0.0114
16	1	11	0.628	0.0330	0.0130
22	1	7	0.538	0.0569	0.0164
23	1	6	0.448	0.0902	0.0181

Kaplan-Meier Survival Function

overall survival curve

- You can choose to plot the confident interval or not.

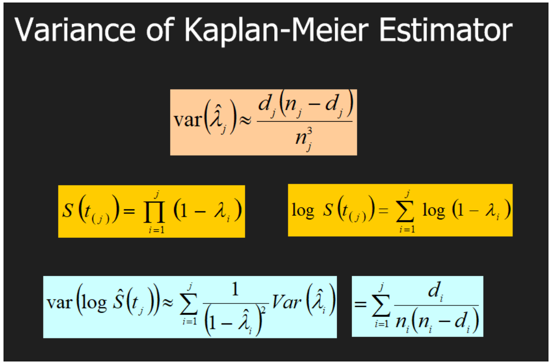
```
plot(fit1,conf.int = FALSE,
   xlab = "Days",
ylab = "Overall survival proba
title("Survival func of Lung c
```

```
plot(fit1,conf.int = TRUE,
xlab = "Days",
ylab = "Overall survival proba
 conf.type = "log")
title("Survival func of Lung c
```

Kaplan-Meier Survival Function

overall survival curve

- $conf.type$ chooses the way how we derive the variance.
- log is the default method, which corresponds to the Greenwood formula.
- Big criticism of Greenwood: can not guarantee the reasonable range of survival function.
- Complimentary log-log transformation can fix this problem.
- The result will be close when sample size is large enough, all based on Delta Method.



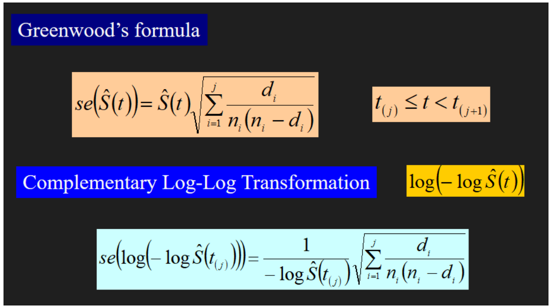
Variance of Kaplan-Meier Estimator

$$\operatorname{var}(\hat{\lambda}_j) \approx \frac{d_j(n_j - d_j)}{n_j^3}$$

$$S(t_{(j)}) = \prod_{i=1}^{j} (1 - \lambda_i)$$

$$\log S(t_{(j)}) = \sum_{i=1}^{j} \log (1 - \lambda_i)$$

$$\operatorname{var}(\log \hat{S}(t_j)) \approx \sum_{i=1}^{j} \frac{1}{(1 - \hat{\lambda}_i)^2} Var(\hat{\lambda}_i) = \sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}$$



Greenwood's formula

$$se(\hat{S}(t)) = \hat{S}(t) \sqrt{\sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}}$$

$$t_{(j)} \leq t < t_{(j+1)}$$

Complementary Log-Log Transformation

$$\log(-\log \hat{S}(t))$$

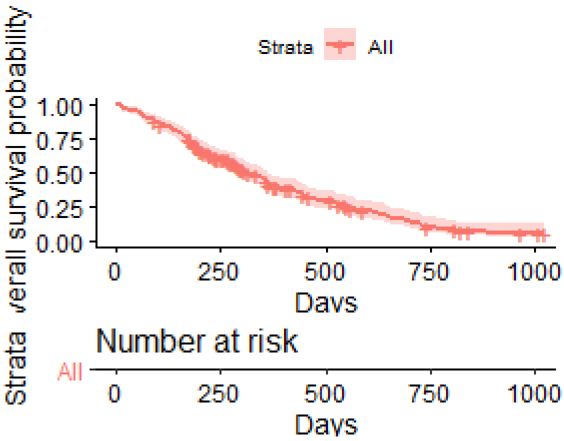
$$se(\log(-\log \hat{S}(t_{(j)}))) = \frac{1}{-\log \hat{S}(t_{(j)})} \sqrt{\sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}}$$

Kaplan-Meier Survival Function

overall survival curve

- Alternatively, *ggsurvplot* function from the *survminer* package is built on *ggplot2*.

```
library(survminer)
ggsurvplot(fit1,data=lung,
           risk.table = TRUE, # show risk table.
           xlab = "Days",ylab = "Overall survival probability")
```



Kaplan-Meier Survival Function

overall survival curve

- Add more information, like median survival time

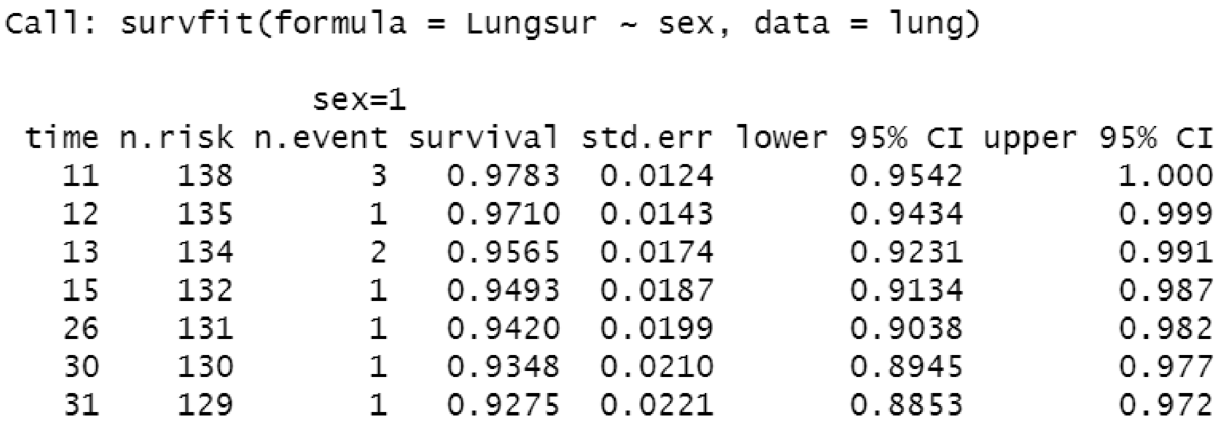
```
ggsurvplot(fit1,  data = lung,
          risk.table = TRUE,
          surv.median.line = "hv",  # add the median survival pointer
        xlab = "Days",ylab = "Overall survival probability")
```

Kaplan-Meier Survival Function

stratified survival curve

- If we want to estimate survival curve for groups.
- You can utilize the *dplyr* package in T2 to select target group and calculate survival function separately.
- Or, directly fitting the curve by some strata variable.

```
fit2=survfit(Lungsur~sex,data=lung)
summary(fit2)
```



```
Call: survfit(formula = Lungsur ~ sex, data = lung)

                sex=1 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   11    138       3   0.9783  0.0124       0.9542        1.000
   12    135       1   0.9710  0.0143       0.9434        0.999
   13    134       2   0.9565  0.0174       0.9231        0.991
   15    132       1   0.9493  0.0187       0.9134        0.987
   26    131       1   0.9420  0.0199       0.9038        0.982
   30    130       1   0.9348  0.0210       0.8945        0.977
   31    129       1   0.9275  0.0221       0.8853        0.972
```

Kaplan-Meier Survival Function

stratified survival curve

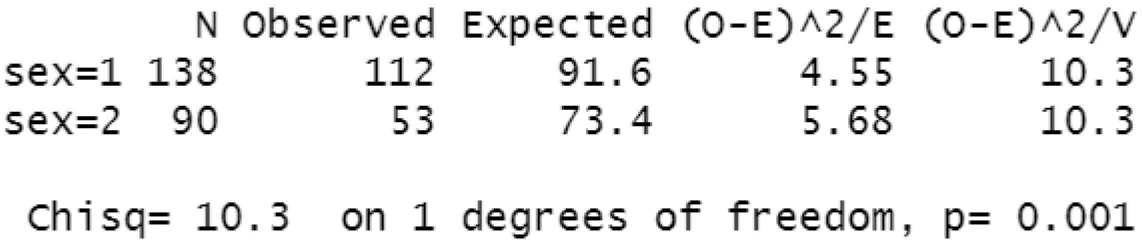
```
ggsurvplot(fit2, data = lung,
           pval = TRUE, #Add p-value
           conf.int = T,
           risk.table = TRUE,        # Add risk table
           legend.labs = c("Male", "Female"),    # Change legend labe
           risk.table.height = 0.25, # Useful to change when you have
           surv.median.line = "hv",  # add the median survival pointe
           ggtheme = theme_bw()      # Change ggplot2 theme
  )
```

Kaplan-Meier Survival Function

stratified survival curve

- In two sample comparison problem, we can also conduct test through $survdiff$ function.
- rho stands for different weighted test.

```
fit3=survdiff(Lungsur~sex,data=lung,
              rho=0 #log-rank test
              )
# log-rank test is default, same as in ggsurvplot.
fit3
```



```
          N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 138      112     91.6      4.55      10.3
sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Cox model

- To quantify an effect size for a single variable.
- Or include more than one variable into a regression model to account for the effects of multiple variables.
- *coxph* to fit a Cox proportional hazards regression model.
- Take gender into the Cox model, coefficient is negative, indicates that Female has higher survival probability compared to Male siginificantly.

```
fit4=coxph(Lungsur ~ sex, data = lung)
fit4
```

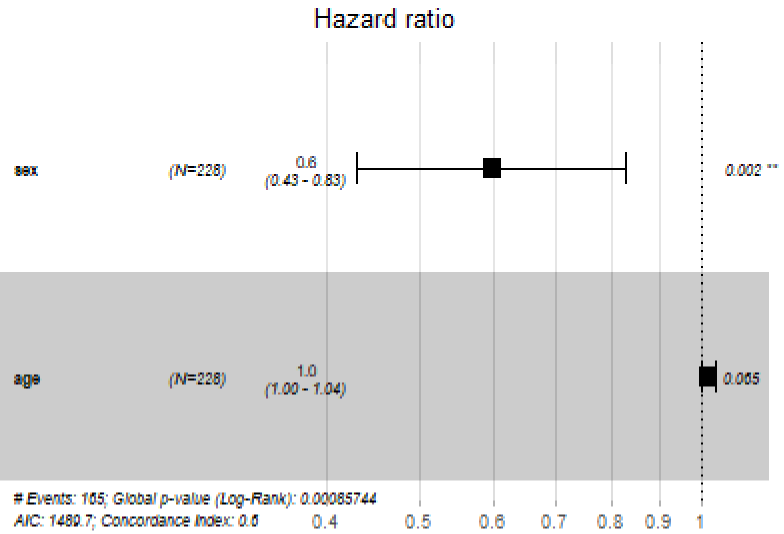
```
## Call:
## coxph(formula = Lungsur ~ sex, data = lung)
## 
##        coef exp(coef) se(coef)      z       p
## sex -0.5310    0.5880   0.1672 -3.176 0.00149
## 
## Likelihood ratio test=10.63  on 1 df, p=0.001111
## n= 228, number of events= 165
```

Cox model

A forest plot for hazard ratio.

- Adding age into the cox model.

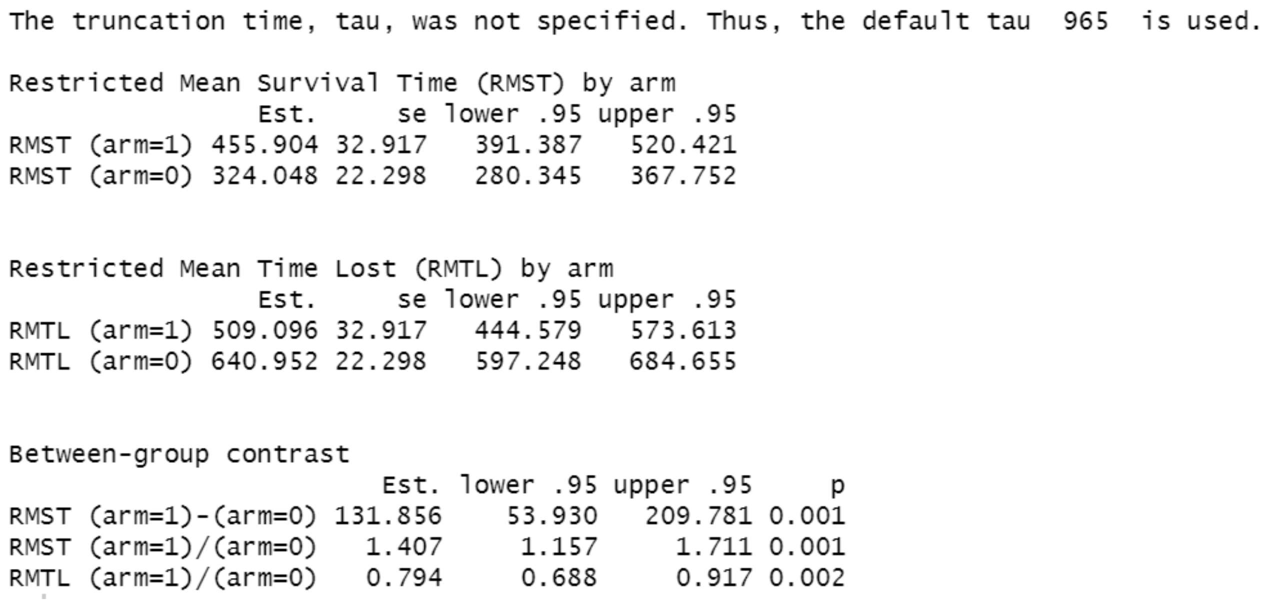
```
fit4_s=coxph(Lungsur ~ sex+age, data = lung)
ggforest(fit4_s)
```



RMST model

- No need for strong model assumption, Model-free and clinically interpretable.
- $rmst2$ function, shoud be careful to each argument definition.

```
library(survRM2)
arm = as.numeric(factor(sex))-1 # 0 for Male and 1 for Female
fit5=rmst2(time, status-1, arm)
fit5
```



```
The truncation time, tau, was not specified. Thus, the default tau  965  is used.

Restricted Mean Survival Time (RMST) by arm
               Est.     se lower .95 upper .95
RMST (arm=1) 455.904 32.917   391.387   520.421
RMST (arm=0) 324.048 22.298   280.345   367.752


Restricted Mean Time Lost (RMTL) by arm
               Est.     se lower .95 upper .95
RMTL (arm=1) 509.096 32.917   444.579   573.613
RMTL (arm=0) 640.952 22.298   597.248   684.655


Between-group contrast
                       Est. lower .95 upper .95     p
RMST (arm=1)-(arm=0) 131.856    53.930   209.781 0.001
RMST (arm=1)/(arm=0)   1.407     1.157     1.711 0.001
RMTL (arm=1)/(arm=0)   0.794     0.688     0.917 0.002
```

RMST model

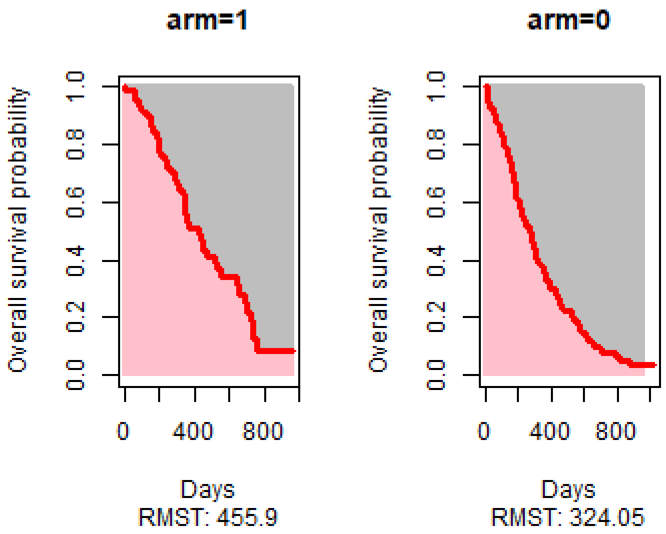
- Adding covariates into model, eg: age.

```
fit6=rmst2(time, status-1, arm,covariates = age)
fit6
```

RMST model

- plot RMST curve

```
plot(fit5,
     col.RMST = "pink",
     col.RMTL = "gray",
     xlab = "Days",ylab = "Overall survival probability")
```



Q&A