P8131_hw10

Shan Jiang

Problem 1

```r
# load data
##############################################
time  = c(4, 2, 12, 6, 15, 8, 21, 10, 23, 19)
pair = c(1, 1, 2, 2, 3, 3, 4, 4, 5, 5)
cens = c(0, 0, 1, 1, 0, 1, 1, 0, 0, 0)
treat = rep(c(1, 2), 5 )

df = data.frame(pair, time, cens, treat)
head(df)
```

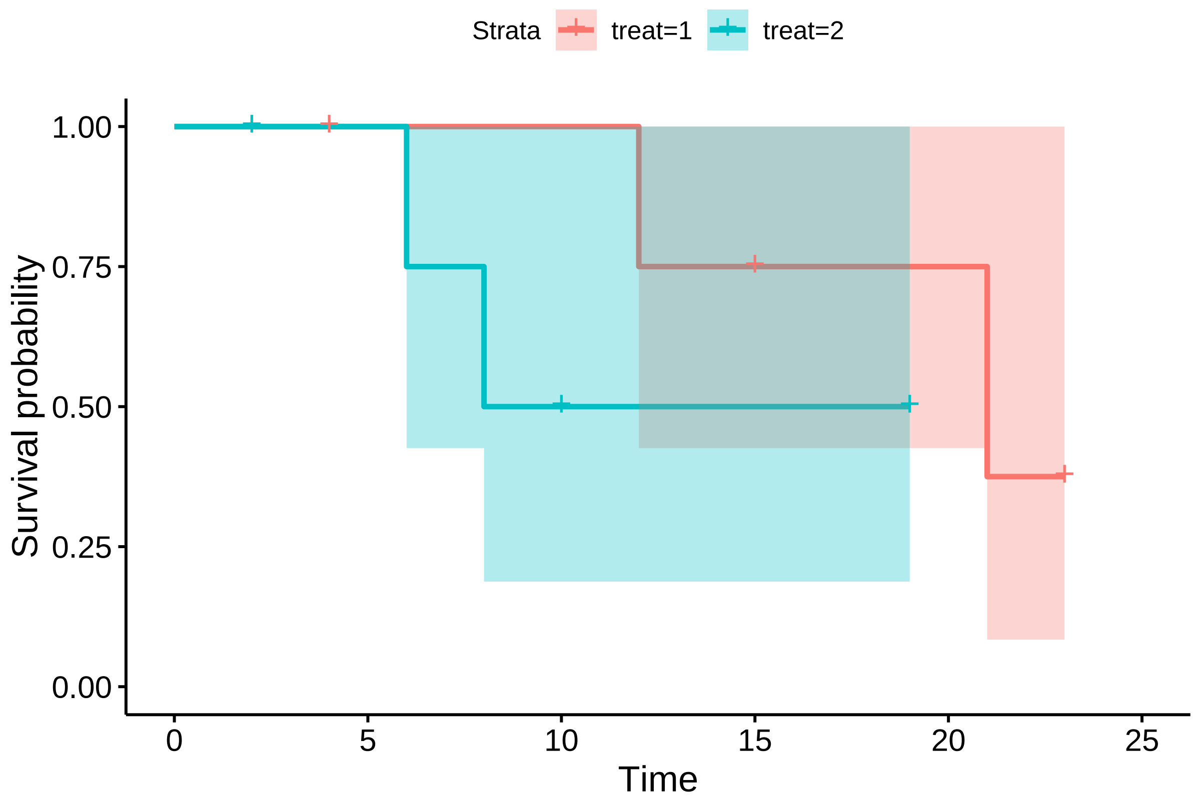
```
##   pair time cens treat
## 1    1    4    0     1
## 2    1    2    0     2
## 3    2   12    1     1
## 4    2    6    1     2
## 5    3   15    0     1
## 6    3    8    1     2
```

$H_0 : h_1(t) = h_2(t)$ for all t $H_1 : h_1(t) \neq h_2(t)$

```r
## Log Rank test
survdiff(Surv(time,cens)~treat, data = df) # log rank test
```

```
## Call:
## survdiff(formula = Surv(time, cens) ~ treat, data = df)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## treat=1 5        2     2.87     0.264      1.16
## treat=2 5        2     1.13     0.673      1.16
##
##  Chisq= 1.2  on 1 degrees of freedom, p= 0.3
```

```r
ggsurvplot( survfit(Surv(time, cens) ~ treat, data = df), conf.int = TRUE)
```



Interpretation:

- The log rank test: at the significance of 0.05, we cannot reject the null hypothesis as p > 0.3.So, we cannot say there is significant difference for hazard function between these two groups.

Problem 2

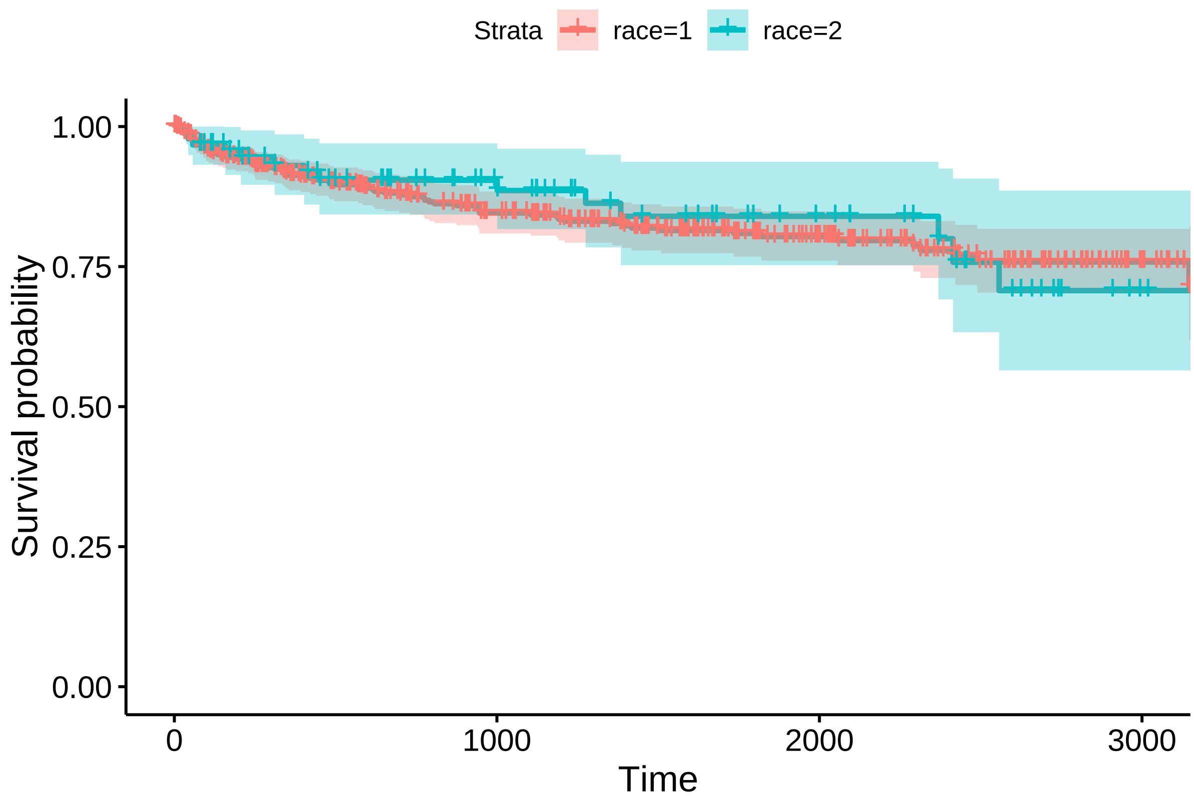
```
library(KMsurv)
data(kidtran)
```

- gender: 1 = Male, 2 = Female;
- race: 1 = white, 2 = black;
- delta: Death indicator (0 = alive, 1 = dead);

Two gender group data: visual comparison

Group 1: Male(gender = 1)

```
## gender: Male
ggsurvplot(survfit(Surv(time, delta) ~  race, data = subset(kidtran, gender == "1")), conf.int=TRUE)
```



More formal way to compare intra-group differences, is by using the log-rank test to test hypotheses.

```
# Male
survdiff(Surv(time,delta) ~ race, data = subset(kidtran, gender == "1"))
```

```
## Call:
## survdiff(formula = Surv(time, delta) ~ race, data = subset(kidtran,
##     gender == "1"))
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## race=1 432       73     71.9    0.0168     0.097
## race=2  92       14     15.1    0.0801     0.097
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.8
```

In the log-rank test, we fail to reject the null hypothesis and conclude that there is no significant difference in two races groups for t defined.

```
fit = coxph(Surv(time, delta) ~  gender + race,
            data = kidtran)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ gender + race, data = kidtran)
```

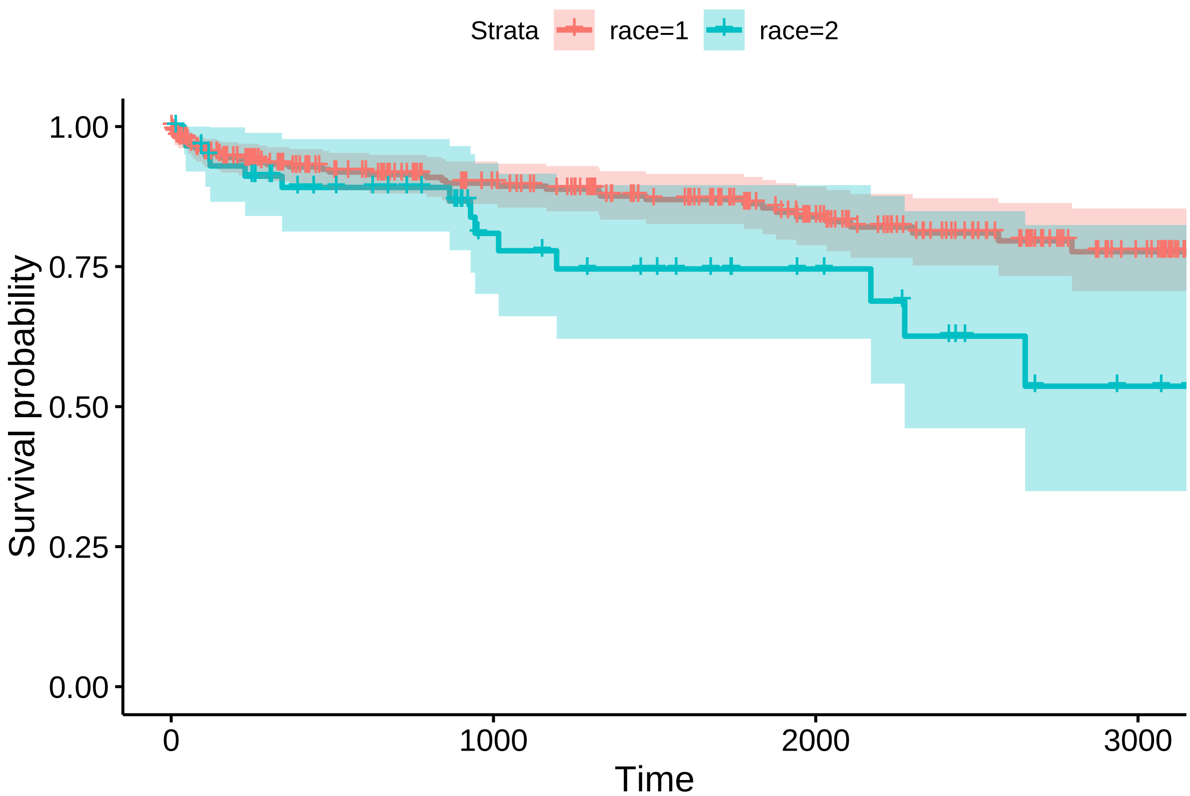
```
##
##   n= 863, number of events= 140
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## gender -0.09347   0.91077  0.17441 -0.536    0.592
## race    0.21992   1.24597  0.21148  1.040    0.298
##
##        exp(coef) exp(-coef) lower .95 upper .95
## gender    0.9108     1.0980    0.6471     1.282
## race      1.2460     0.8026    0.8232     1.886
##
## Concordance= 0.526  (se = 0.023 )
## Rsquare= 0.002   (max possible= 0.87 )
## Likelihood ratio test= 1.35  on 2 df,   p=0.5
## Wald test            = 1.39  on 2 df,   p=0.5
## Score (logrank) test = 1.4  on 2 df,   p=0.5
```

```r
## In male group
fit_male = coxph(Surv(time, delta) ~ race,
            data = subset(kidtran, gender == "1"))
summary(fit_male)
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ race, data = subset(kidtran,
##     gender == "1"))
##
##   n= 524, number of events= 87
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## race -0.0908    0.9132   0.2918 -0.311    0.756
##
##      exp(coef) exp(-coef) lower .95 upper .95
## race    0.9132      1.095    0.5154     1.618
##
## Concordance= 0.513  (se = 0.02 )
## Rsquare= 0   (max possible= 0.854 )
## Likelihood ratio test= 0.1  on 1 df,   p=0.8
## Wald test            = 0.1  on 1 df,   p=0.8
## Score (logrank) test = 0.1  on 1 df,   p=0.8
```

Group 2: Female(gender = 2)

```r
## gender: Female
ggsurvplot(survfit(Surv(time, delta) ~  race, data = subset(kidtran, gender == "2")), conf.int=TRUE)
```




```r
# Female
survdiff(Surv(time,delta) ~ race, data = subset(kidtran, gender == "2"))
```

```
## Call:
## survdiff(formula = Surv(time, delta) ~ race, data = subset(kidtran,
##     gender == "2"))
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## race=1 280       39    44.79     0.748      4.85
## race=2  59       14     8.21     4.076      4.85
##
##  Chisq= 4.8  on 1 degrees of freedom, p= 0.03
```

For female subjects, there is a significant difference in hazards ratio ($p < 0.03$) for Black and white people.

```r
## Female(gender = 2)
fit_male = coxph(Surv(time, delta) ~ age + race,
            data = subset(kidtran, gender == "2"))
summary(fit)
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ gender + race, data = kidtran)
##
##   n= 863, number of events= 140
```

```
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## gender -0.09347   0.91077  0.17441 -0.536    0.592
## race    0.21992   1.24597  0.21148  1.040    0.298
##
##        exp(coef) exp(-coef) lower .95 upper .95
## gender    0.9108     1.0980    0.6471     1.282
## race      1.2460     0.8026    0.8232     1.886
##
## Concordance= 0.526  (se = 0.023 )
## Rsquare= 0.002   (max possible= 0.87 )
## Likelihood ratio test= 1.35  on 2 df,   p=0.5
## Wald test            = 1.39  on 2 df,   p=0.5
## Score (logrank) test = 1.4  on 2 df,   p=0.5
```

Problem 3

```r
data(larynx)

larynx = larynx %>%
            mutate( Z1 = ifelse(stage == 2, 1, 0)) %>%
            mutate( Z2 = ifelse(stage == 3, 1, 0)) %>%
                mutate( Z3 = ifelse(stage == 4, 1, 0)) %>%
                mutate( Z4 = age) %>%
                mutate( Z1_Z4 =  Z1 * Z4) # ADD interaction term Z1 * Z

head(larynx)
```

```
##   stage time age diagyr delta Z1 Z2 Z3 Z4 Z1_Z4
## 1     1  0.6  77     76     1  0  0  0 77     0
## 2     1  1.3  53     71     1  0  0  0 53     0
## 3     1  2.4  45     71     1  0  0  0 45     0
## 4     1  2.5  57     78     0  0  0  0 57     0
## 5     1  3.2  58     74     1  0  0  0 58     0
## 6     1  3.2  51     77     0  0  0  0 51     0
```

Model building

$h(t) = h_0(t)exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_1 \times Z_4$

```r
fit_3 = coxph(Surv(time, delta) ~  age + Z1 + Z2 + Z3 + Z4 + Z1_Z4,
            data = larynx, ties = 'breslow')

summary(fit_3)
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ age + Z1 + Z2 + Z3 + Z4 +
##     Z1_Z4, data = larynx, ties = "breslow")
##
##   n= 90, number of events= 50
```

```
##
##              coef  exp(coef)   se(coef)      z Pr(>|z|)
## age    0.0059729  1.0059908  0.0148792  0.401   0.6881
## Z1    -7.3820143  0.0006223  3.4027542 -2.169   0.0301 *
## Z2     0.6218044  1.8622853  0.3558078  1.748   0.0805 .
## Z3     1.7534270  5.7743576  0.4239595  4.136 3.54e-05 ***
## Z4            NA         NA  0.0000000     NA       NA
## Z1_Z4  0.1116674  1.1181409  0.0476728  2.342   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## age   1.0059908     0.9940 9.771e-01    1.0358
## Z1    0.0006223  1606.8231 7.900e-07    0.4903
## Z2    1.8622853     0.5370 9.272e-01    3.7403
## Z3    5.7743576     0.1732 2.516e+00   13.2550
## Z4           NA         NA        NA        NA
## Z1_Z4 1.1181409     0.8943 1.018e+00    1.2277
##
## Concordance= 0.682  (se = 0.04 )
## Rsquare= 0.235   (max possible= 0.988 )
## Likelihood ratio test= 24.11  on 5 df,   p=2e-04
## Wald test            = 23.77  on 5 df,   p=2e-04
## Score (logrank) test = 27.98  on 5 df,   p=4e-05
```

Interpretation of results:

- $\beta_1 + \beta_5$: The log hazard ratio for subjects in stage II versus stage I is (-7.382 + 0.112 $\times$age) given of the same age.
- β_2: The log hazards ratio for subjects in stage III versus stage I is 0.621 given the same age.
- β_3: The log hazards ratio for subjects in stage IV versus stage I is 1.753 given they have same age.
- β_4: Compared to non-Stage II, the log hazards ratio for subjects with one unit changes in age is 0.006 given they are in same stage.
- $\beta_4 + \beta_5$: In the stage II for subjects, the log hazards ratio for subjects with one unit changes in age is 0.118.

Relative risk:

- For the hazards of dying for a stage II patient of age 50 is:

$$h_2(t) = h_0(t) \times exp(7.382»1 + 0.111»1»5)$$

- For the hazard of dying for a stage I patient of age 50 is

$$h_1(t) = h_0(t) exp(0)$$

- *So* the hazard ratio: $HR(t) = \frac{h_2(t)}{h_1(t)} = 0.16$