ADA Homework 11
Wenxin Liang wl2455

Consider the Mayo Clinic Lung Cancer Data in R package survival: data (lung) or data (cancer): including the variables
inst: Institution code
time: Survival time in days
status: censoring status 1=censored, 2=dead
age: Age in years
sex: Male=1 Female=2, etc.

Define 'AGE GROUP' as 'YOUNG" if 'age < 65", and OLD, otherwise.

1. Using a Cox proportional hazards model, estimate the hazard rate for old relative to young
As R code showed,

```
> fitcox_old_young
Call:
coxph(formula = Surv(data3$time, data3$status) ~ +factor(age,
    levels = c("Young", "Old")), data = data3)


                                         coef exp(coef) se(coef)    z     p
factor(age, levels = c("Young", "Old"))Old 0.298      1.35    0.156 1.91 0.056

Likelihood ratio test=3.62  on 1 df, p=0.057  n= 228, number of events= 165
> fitcox_young_old
Call:
coxph(formula = Surv(data3$time, data3$status) ~ +factor(age,
    levels = c("Old", "Young")), data = data3)


                                            coef exp(coef) se(coef)     z     p
factor(age, levels = c("Old", "Young"))Young -0.298     0.742    0.156 -1.91 0.056

Likelihood ratio test=3.62  on 1 df, p=0.057  n= 228, number of events= 165
```

Also based on R, for old relative to young, which may be old over young.

```
> young_index <- which(data$age <65)
> young_index
  [1]   3   4   5   9  10  11  14  15  18  19  20  22  23  24  27  29  33  34  35  36  40  43
 [23]  44  48  50  52  54  55  56  58  59  60  62  68  71  72  74  75  77  78  81  82  83  84
 [45]  85  87  88  89  91  94  99 101 103 104 105 107 109 112 115 117 122 127 128 130 131 132
 [67] 133 134 137 138 139 140 145 146 148 150 152 153 154 155 158 159 160 161 162 164 165 166
 [89] 167 168 169 171 172 173 174 175 176 177 178 179 181 182 185 186 188 189 190 192 193 194
[111] 197 198 199 200 201 203 204 205 206 207 208 210 211 214 217 220 225 228
> young_time <- data$time[young_index]
> mean(young_time)
[1] 320.2266
> old_index2 <- which (data$age >= 65)
> old_index2
  [1]   1   2   6   7   8  12  13  16  17  21  25  26  28  30  31  32  37  38  39  41  42  45
 [23]  46  47  49  51  53  57  61  63  64  65  66  67  69  70  73  76  79  80  86  90  92  93
 [45]  95  96  97  98 100 102 106 108 110 111 113 114 116 118 119 120 121 123 124 125 126 129
 [67] 135 136 141 142 143 144 147 149 151 156 157 163 170 180 183 184 187 191 195 196 202 209
 [89] 212 213 215 216 218 219 221 222 223 224 226 227
> old_time <- data$time[old_index2]
> mean(old_time)
[1] 286.04
```
The mean of time for age young is greater than the mean of time for age old then we conclude that the hazard ratio for age old relative to age young is 0.742.

2. Assess the validity of the proportional hazards assumption in (1)

For the proportional hazards assumption we test whether the ratio of hazards $\frac{h(t;\beta_1,x_1=1)}{h(t;\beta_1,x_1=0)} = e^{\beta_1}$ is independent of t. Suppose x_1 is an indicator variable, taking the values 0 or 1.

Defining the auxiliary variable may test this,

$$x_2(t)\begin{cases} log\ t - \bar{t}, if\ young \\ 0, if\ old \end{cases}$$

Then the hazard ratio becomes,

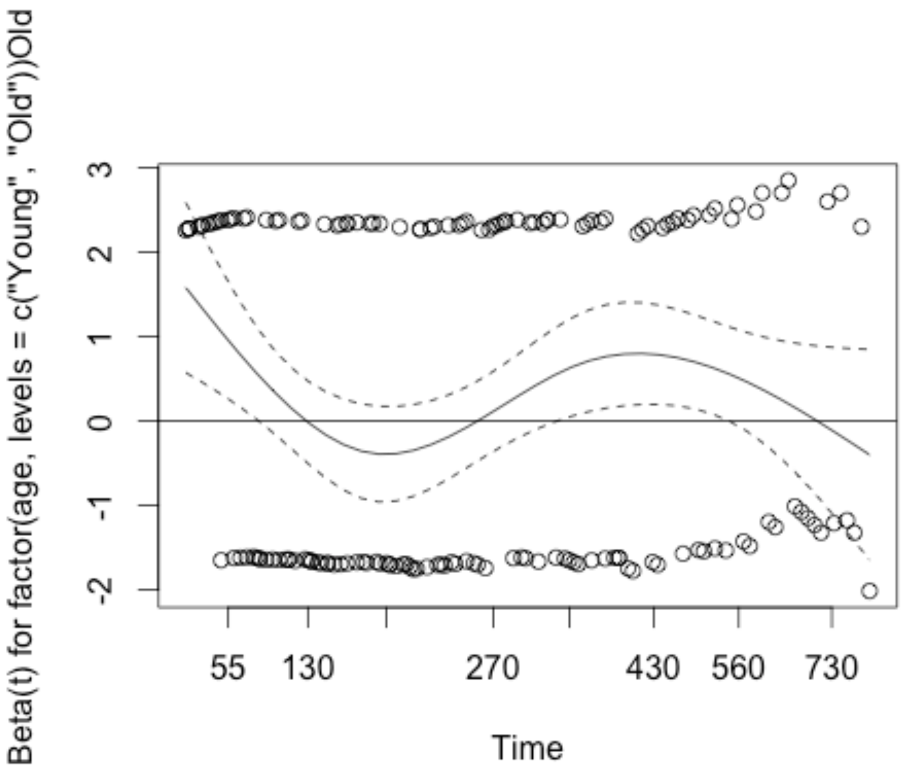
$$e^{\beta_1+\beta_2(lnt-\bar{t})} = t^{\beta_2}e^{\beta_1-\beta_2\bar{t}}$$

Thus, the hypotheses are followed,

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

Based on R, the result and the plot showed,

```
> prop <- cox.zph(fitcox_old_young)
> prop
                                        rho  chisq     p
factor(age, levels = c("Young", "Old"))Old -0.0211 0.0727 0.787
```
Since p-value >0.05 we fail to reject the null hypothesis then we conclude that the proportional hazards assumption is true. There is evidence of proportional hazard for old relative to young.



The plot showed the scaled Schoenfeld residuals against transformed time for the covariate, young, in the model fit to the data. The solid line is a smoothing-spline fit to the plot, with the broken lines representing a ± 2-standard-error band around the fit. Since the plot is not systematic departures from a horizontal line, then we can conclude that the proportional hazard assumption is true.

3. Repeat 1, adjusting "Sex".

Based on R,

```
> fitcox_male_female <- coxph(Surv(data3$time,data3$status)~+factor(sex,levels=c('Female','Male'))
data=data3)
> fitcox_female_male <- coxph(Surv(data3$time,data3$status)~+factor(sex,levels=c('Male','Female'))
data=data3)
> fitcox_male_female
Call:
coxph(formula = Surv(data3$time, data3$status) ~ +factor(sex,
    levels = c("Female", "Male")), data = data3)


                                           coef exp(coef) se(coef)    z      p
factor(sex, levels = c("Female", "Male"))Male 0.531       1.7    0.167 3.18 0.0015

Likelihood ratio test=10.6  on 1 df, p=0.00111  n= 228, number of events= 165
> fitcox_female_male
Call:
coxph(formula = Surv(data3$time, data3$status) ~ +factor(sex,
    levels = c("Male", "Female")), data = data3)


                                             coef exp(coef) se(coef)     z      p
factor(sex, levels = c("Male", "Female"))Female -0.531     0.588    0.167 -3.18 0.0015

Likelihood ratio test=10.6  on 1 df, p=0.00111  n= 228, number of events= 165
```

Also we calculate the mean of the time for female and male, as followed,

```
> male_index <- which(data$sex== 1)
> male_time <- data$time[male_index]
> mean(male_time)
[1] 283.2319
> female_index <-which(data$sex == 2)
> female_time <- data$time[female_index]
> mean(female_time)
[1] 338.9667
```

The mean of time for female is greater than the mean of time for male then we conclude that the hazard ratio for age old relative to age young is 0.588.

Following code,

```
#ADA Homework 11
library(survival)
data = lung

# Question 1
data3 <- data
data3$age[which(data3$age<65)]    <- 1
data3$age[which(data3$age>65)]    <- 0
data3$age[which(data3$age==65)] <- 0
data3$age[which(data3$age == 0)] <- "Old"
data3$age[which(data3$age == "1")] <- "Young"
fitcox_old_young                                                                                                         <-
coxph(Surv(data3$time,data3$status)~+factor(age,levels=c('Young','Old')),data=data3)
fitcox_young_old                                                                                                         <-
coxph(Surv(data3$time,data3$status)~+factor(age,levels=c('Old','Young')),data=data3)
young_index <- which(data$age <65)
young_index
young_time <- data$time[young_index]
mean(young_time)
old_index2 <- which (data$age >= 65)
old_index2
old_time <- data$time[old_index2]
mean(old_time)
# young >old
# choose smaller one
# old/young
# 0.7419

# Question 2
prop <- cox.zph(fitcox_old_young)
plot(prop)
score <- resid(fitcox_old_young,type="score")
score_frame <- data.frame(score)
plot(data$age,score_frame[,1],ylab="Score Residuals",xlab ="AGE")

# Question 3
data3$sex[which(data3$sex == 1)] <- "Male"
data3$sex[which(data3$sex == "2")] <- "Female"
fitcox_male_female                                                                                                       <-
coxph(Surv(data3$time,data3$status)~+factor(sex,levels=c('Female','Male')),data=data3
)
fitcox_female_male                                                                                                       <-
coxph(Surv(data3$time,data3$status)~+factor(sex,levels=c('Male','Female')),data=data3
```

```
)

fitcox_male_female1 <- coxph(Surv(data$time,data$status)~sex,data=data)
summary(fitcox_male_female1)

male_index <- which(data$sex== 1)
male_time <- data$time[male_index]
mean(male_time)
female_index <-which(data$sex == 2)
female_time <- data$time[female_index]
mean(female_time)
```