Proj4_partB

Wenbo Fei

4/8/2021

Part B: Analyzing menopause_age

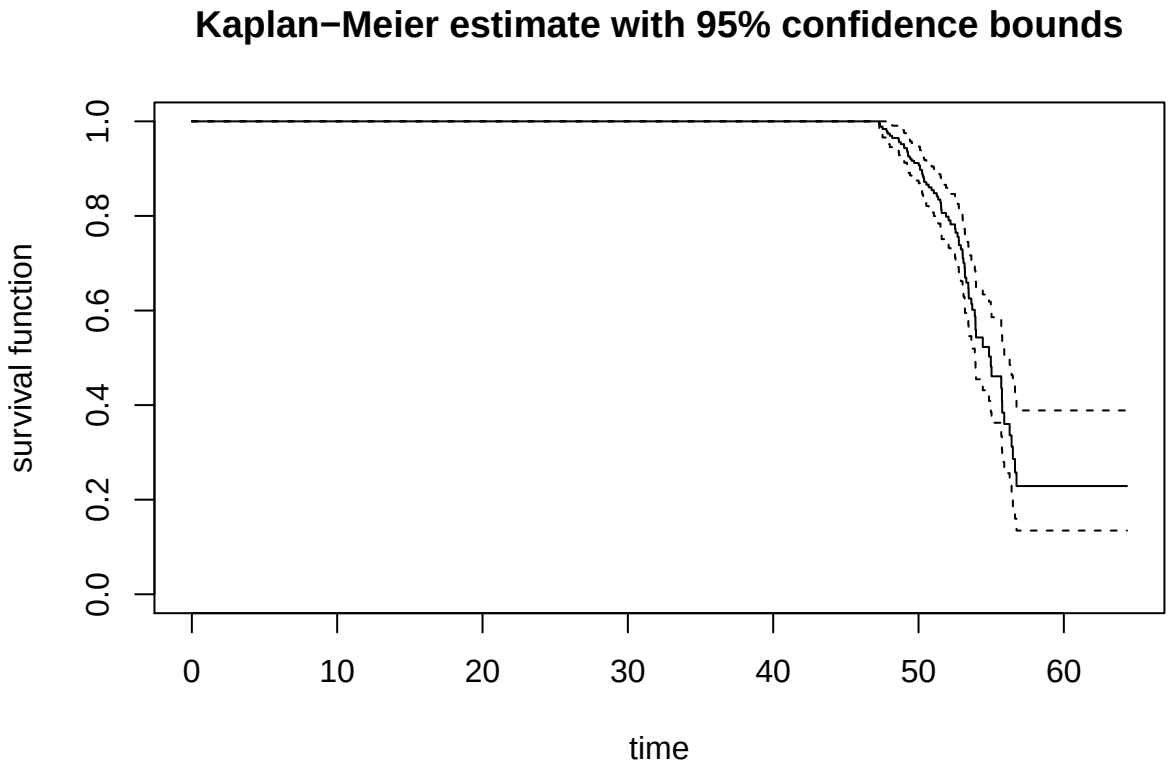
Since menopause_age is always no less than intake_age for every patient due to the sampling design, so this is a left truncated, right cencered data.

```
data = read.table("Menopause.dat")
colnames(data) = c("Patient_ID","Intake_Age","Menopause_Age",
                   "Menopause","Race","Education")
data = data %>%
  mutate(Race = ifelse(Race == 0,"White",ifelse(Race == 1,"Black","Other")),
         Education = ifelse(Education == 0,"Post-Grad",
                            ifelse(Education == 1,"College Grad",
                                   ifelse(Education == 2,"Some College","HS or less"))),
         Race = factor(Race, ordered = FALSE, levels = c("White","Black","Other")),
         Education = factor(Education, ordered = FALSE, levels = c("HS or less","Some College", "Colleg
```

III. Compute a nonparametric estimate for the survival function

The Kaplan-Meier estimate is a nonparametric maximum likelihood estimate (MLE) of the survival function, S(t).

```
my.surv.object <- Surv(data$Intake_Age, data$Menopause_Age, data$Menopause)
my.fit <- survfit(my.surv.object~1) #
plot(my.fit, main="Kaplan-Meier estimate with 95% confidence bounds",  xlab="time", ylab="survival funct
```




```
a = summary(my.fit)
table3 = round(a$surv,4)
names(table3) = round(a$time,2)# length(table3) # output 75 = 5*15 table?
table3
```

```
##   47.3  47.35  47.52  47.79  47.87  48.02  48.18  48.63  48.65  48.78     49
## 0.9945 0.9890 0.9838 0.9790 0.9741 0.9694 0.9649 0.9605 0.9561 0.9518 0.9476
##  49.02  49.21  49.23  49.28  49.28  49.41  49.52   49.7  50.01   50.1   50.1
## 0.9433 0.9389 0.9345 0.9301 0.9257 0.9212 0.9167 0.9119 0.9071 0.9022 0.8974
##  50.25  50.25  50.32  50.36  50.39  50.54   50.7   50.9  51.06  51.26  51.35
## 0.8924 0.8874 0.8822 0.8769 0.8716 0.8661 0.8604 0.8544 0.8480 0.8413 0.8346
##  51.51  51.55  51.56  51.59  51.89  52.07   52.2  52.51  52.54  52.69  52.77
## 0.8275 0.8205 0.8135 0.8063 0.7985 0.7905 0.7821 0.7734 0.7645 0.7557 0.7471
##  52.78  52.91  53.03  53.04  53.09  53.17  53.19  53.19  53.29  53.43  53.44
## 0.7384 0.7294 0.7199 0.7103 0.7004 0.6903 0.6800 0.6697 0.6590 0.6479 0.6367
##  53.46  53.63  53.69  53.87  53.92  53.92  53.96  54.42  54.85  54.98  55.02
## 0.6255 0.6137 0.6015 0.5875 0.5728 0.5581 0.5430 0.5229 0.5028 0.4818 0.4609
##  55.69  55.74  55.76   55.9  56.27   56.4  56.49  56.64  56.74
## 0.4353 0.4097 0.3841 0.3601 0.3361 0.3121 0.2861 0.2574 0.2288
```

```
survfit(my.surv.object~1, conf.type = "log")
```

```
## Call: survfit(formula = my.surv.object ~ 1, conf.type = "log")
##
## records   n.max n.start  events  median 0.95LCL 0.95UCL
##   380.0   224.0    18.0    75.0    55.0    53.9    56.3
```

```
fit3.2 = flexsurv::flexsurvreg(my.surv.object~1,
                               data = data,dist = "exponential")
summary(fit3.2, type = "median")
```

```
##
```

```
##        est      lcl      ucl
## 1 12.20345 9.748101 15.38844
```

From the table, median survival age of the KM estimator is 55.0.

The results here represents that the median survival age of not having menopause is 55.0 years, it's estimated directly from the survival curve. If the curve doesn't contain 0.5(by the end of the study less than half women develop the menopause), we will not have the estimated median survival time.

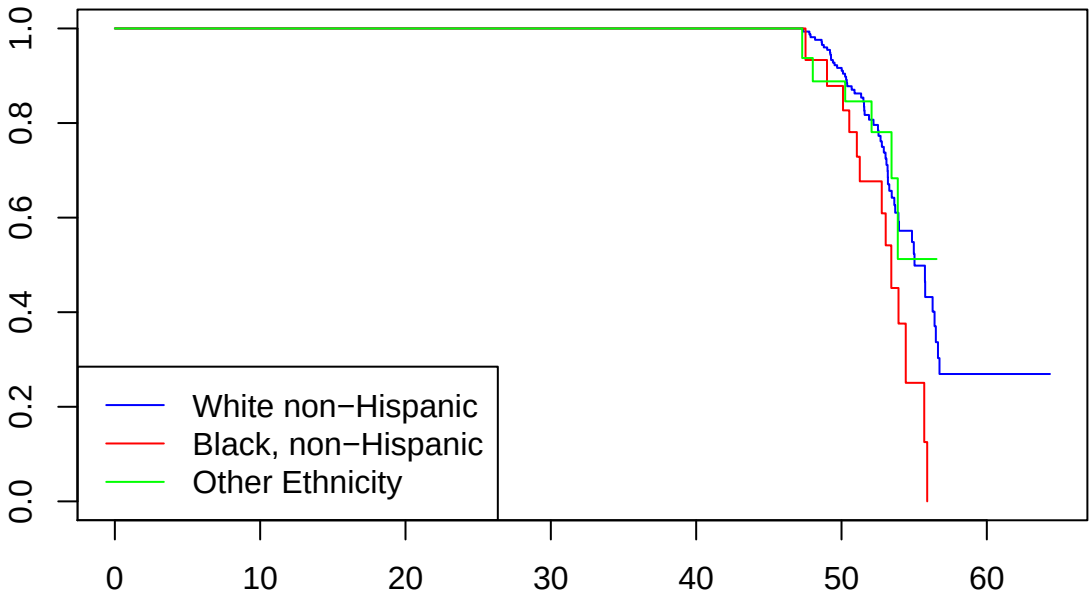
The estimate based on the exponential distribution is 12.2. For the parametric estimate, we first fit the data to a exponential model, and find the 1/2 quantile of the cdf as the median, so even if the curve doesn't contain 0.5, we will still get the estimated median survival time.

Because of the memorylessness of exponential distribution, $\Pr(T > s + t \mid T > s) = \Pr(T > t), \quad \forall s, t \geq 0$, it doesn't matter what the at intake age they entered the study, it is estimated that half of these women will develop menopause in the following 12.2 years.

IV. How race relates to menopause_age.

```
my.fit4 <- survfit(my.surv.object ~ Race, data = data)
b = summary(my.fit4)

plot(my.fit4, col = c("blue","red", "green") )
legend("bottomleft", legend = c("White non-Hispanic", "Black, non-Hispanic", "Other Ethnicity"), lty = 1
```




```
# logrank test to assess the effect of race
# (surv_diff <- survdiff(my.surv.object ~ Race, data = data)) # Error: Right censored data only
cox.fit4 <- coxph(my.surv.object ~ Race, data = data)
summary(cox.fit4)
```

```
## Call:
## coxph(formula = my.surv.object ~ Race, data = data)
##
##   n= 380, number of events= 75
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
```

```
## RaceBlack  0.7636    2.1460   0.3118  2.449   0.0143 *
## RaceOther -0.1026    0.9025   0.4325 -0.237   0.8125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## RaceBlack    2.1460      0.466    1.1647     3.954
## RaceOther    0.9025      1.108    0.3866     2.107
##
## Concordance= 0.537  (se = 0.029 )
## Likelihood ratio test= 5.4  on 2 df,   p=0.07
## Wald test            = 6.35  on 2 df,   p=0.04
## Score (logrank) test = 6.67  on 2 df,   p=0.04
```

P-value for logrank test is 0.04, so the survival distributions for the three race groups are different.

V. Whether race provides additional information about menopause_age beyond that provided by education.

Va. Is race a significant predictor of menopause_age after adjusting for education?

```
cox.fit5 <- coxph(my.surv.object ~ Education + Race, data = data)
summary(cox.fit5)
```

```
## Call:
## coxph(formula = my.surv.object ~ Education + Race, data = data)
##
##   n= 380, number of events= 75
##
##                            coef exp(coef)  se(coef)      z Pr(>|z|)
## EducationSome College  0.665261  1.944998  0.434052  1.533  0.12536
## EducationCollege Grad  0.005844  1.005861  0.438044  0.013  0.98936
## EducationPost-Grad     0.662156  1.938968  0.407687  1.624  0.10434
## RaceBlack              0.917290  2.502501  0.334301  2.744  0.00607 **
## RaceOther             -0.053662  0.947753  0.433777 -0.124  0.90155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## EducationSome College     1.9450     0.5141    0.8307     4.554
## EducationCollege Grad     1.0059     0.9942    0.4263     2.374
## EducationPost-Grad        1.9390     0.5157    0.8721     4.311
## RaceBlack                 2.5025     0.3996    1.2996     4.819
## RaceOther                 0.9478     1.0551    0.4050     2.218
##
## Concordance= 0.585  (se = 0.038 )
## Likelihood ratio test= 11.98  on 5 df,   p=0.03
## Wald test            = 12.63  on 5 df,   p=0.03
## Score (logrank) test = 13.05  on 5 df,   p=0.02
```

Race is still a significant predictor of menopause_age after adjusting for education, as the p-value for hazard ratio for black compared to white is <0.05.

Vb. Provide point and 95% confidence interval estimates for the relative risk of menopause_age for a Black Patient with an Other Ethnicity patient controlling for education. Interpret.

```
data0 = read.table("Menopause.dat")
colnames(data0) = c("Patient_ID","Intake_Age","Menopause_Age",
                    "Menopause","Race","Education")
data5 = data0 %>%
  mutate(Race = ifelse(Race == 0,"White",ifelse(Race == 1,"Black","Other")),
         Education = ifelse(Education == 0,"Post-Grad",
                            ifelse(Education == 1,"College Grad",
                                   ifelse(Education == 2,"Some College","HS or less"))),
         Race = factor(Race),
         Race = relevel(Race, ref = "Other"),
         Education = factor(Education, ordered = FALSE, levels = c("HS or less","Some College", "Colleg

my.surv.object2 <- Surv(data5$Intake_Age, data5$Menopause_Age, data5$Menopause)
cox.fit5.2 <- coxph(my.surv.object2 ~ Education + Race, data = data5)
summary(cox.fit5.2)
```

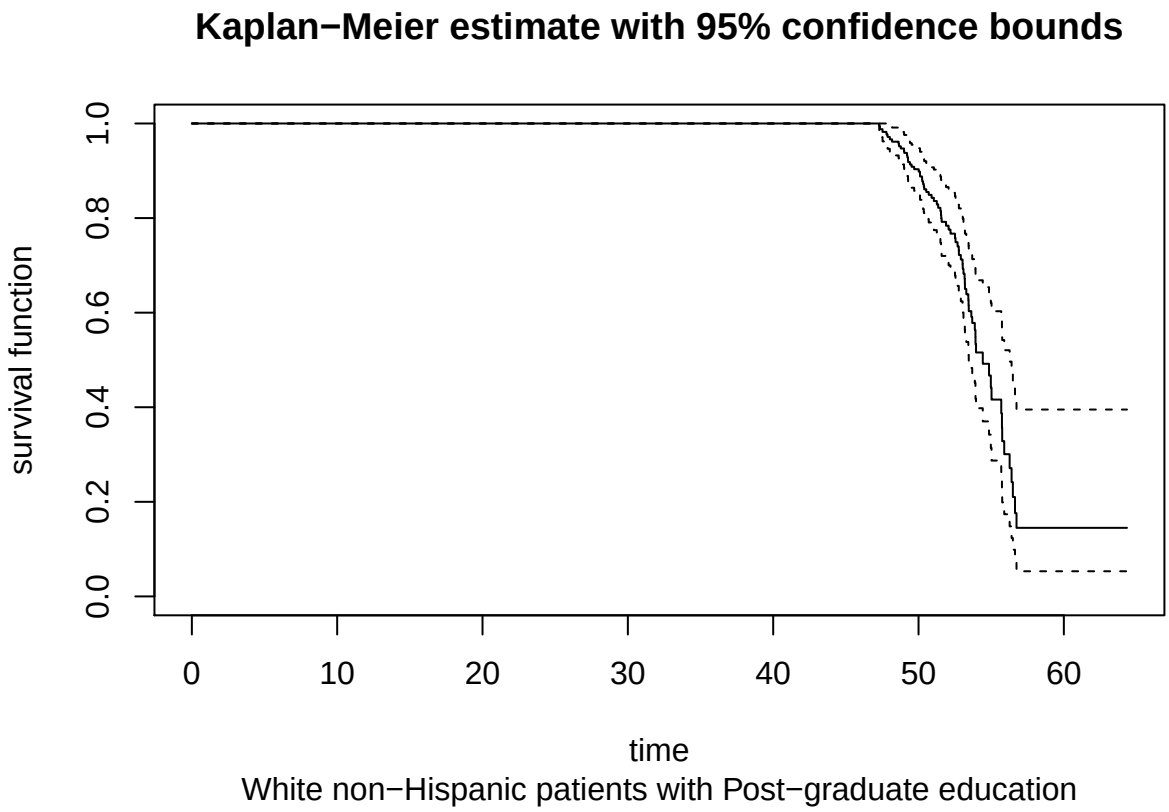
```
## Call:
## coxph(formula = my.surv.object2 ~ Education + Race, data = data5)
##
##   n= 380, number of events= 75
##
##                            coef exp(coef) se(coef)     z Pr(>|z|)
## EducationSome College 0.665261  1.944998 0.434052 1.533   0.1254
## EducationCollege Grad 0.005844  1.005861 0.438044 0.013   0.9894
## EducationPost-Grad    0.662156  1.938968 0.407687 1.624   0.1043
## RaceBlack             0.970952  2.640458 0.504297 1.925   0.0542 .
## RaceWhite             0.053662  1.055128 0.433777 0.124   0.9015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## EducationSome College     1.945     0.5141    0.8307     4.554
## EducationCollege Grad     1.006     0.9942    0.4263     2.374
## EducationPost-Grad        1.939     0.5157    0.8721     4.311
## RaceBlack                 2.640     0.3787    0.9827     7.095
## RaceWhite                 1.055     0.9478    0.4509     2.469
##
## Concordance= 0.585  (se = 0.038 )
## Likelihood ratio test= 11.98  on 5 df,   p=0.03
## Wald test            = 12.63  on 5 df,   p=0.03
## Score (logrank) test = 13.05  on 5 df,   p=0.02
```

The relative risk of menopause_age for a Black Patient with an Other Ethnicity patient controlling for education is 2.64. The 95% CI is (0.9827, 7.095).

When t is fixed and adjusting for education, the hazard ratio of a Black Patient with an Other Ethnicity patient is 2.64, the 95% CI of the hazard ratio includes 1 so there's no significant difference in hazard ratio between a Black Patient with an Other Ethnicity patient.

Vc. Based on the regression model for menopause_age as a function of race and education, produce an estimate of the baseline survival function for White non-Hispanic patients with Post-graduate education.

```
my.fit5c = survfit(cox.fit5, newdata = data.frame(Race = "White", Education = "Post-Grad"))
plot(my.fit5c, main = "Kaplan-Meier estimate with 95% confidence bounds",  xlab = "time", ylab = "survi
```



Vd. Check the proportional hazards assumption.

The proportional hazards (PH) assumption can be checked using statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals. In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption. The function cox.zph() provides a convenient solution to test the proportional hazards assumption for each covariate included in a Cox refression model fit. For each covariate, the function cox.zph() correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole. The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refuted by a significant relationship.

```
# To test for the proportional-hazards (PH) assumption
(test.ph <- cox.zph(cox.fit5))
```

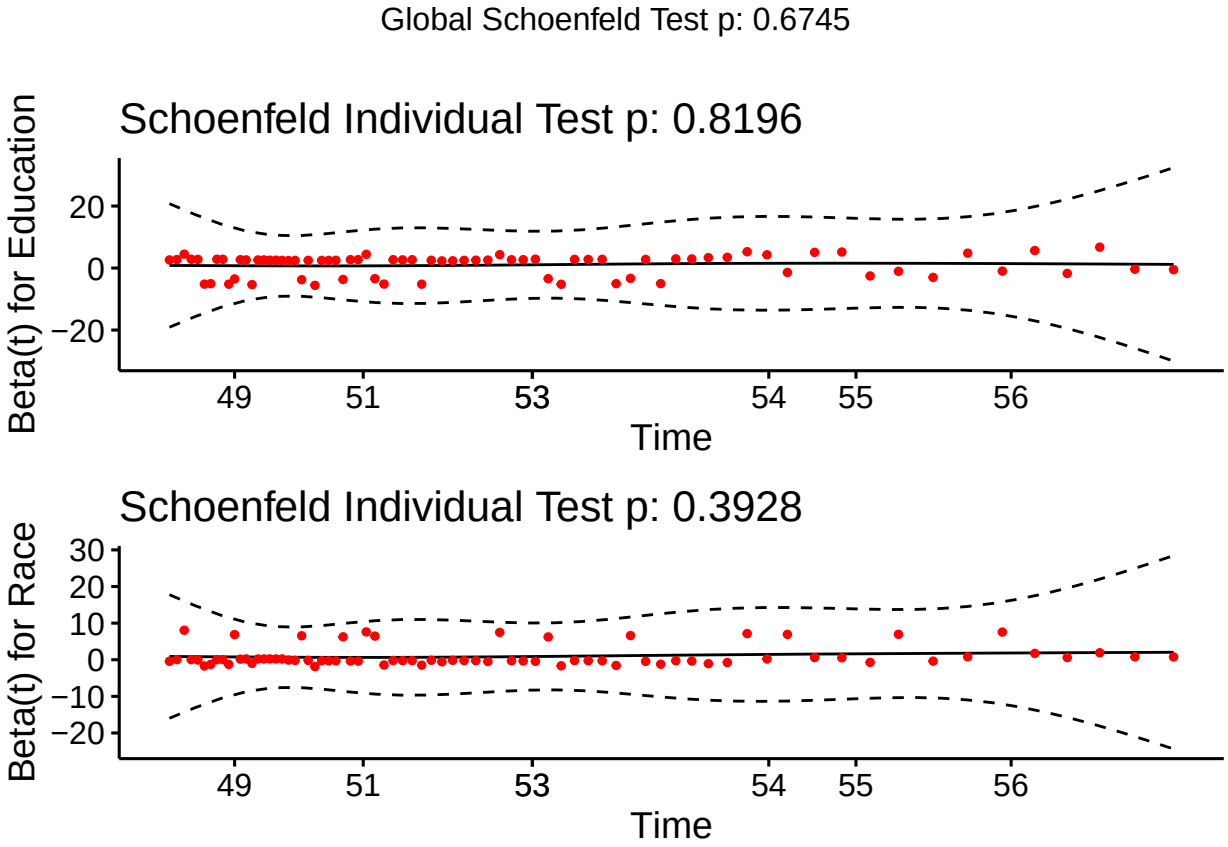
```
##           chisq df    p
## Education 0.924  3 0.82
## Race      1.869  2 0.39
## GLOBAL    3.165  5 0.67
```

From the output above, the test is not statistically significant for each of the covariates, and the global test is also not statistically significant. Therefore, we can assume the proportional hazards.

It's possible to do a graphical diagnostic using the function ggcoxzph(), which produces, for each covariate, graphs of the scaled Schoenfeld residuals against the transformed time.

```
ggcoxzph(test.ph)
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



In the figure above, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a +/- 2-standard-error band around the fit.

Note that, systematic departures from a horizontal line are indicative of non-proportional hazards, since proportional hazards assumes that estimates do not vary much over time.

From the graphical inspection, there is no pattern with time. The assumption of proportional hazards appears to be supported for the covariates Race and Education.