PSTAT 175 LAB D

Yanjie Qi

11/20/2019

To Set Up

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------ tidyve
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts --------------------------------------------------------------------------- tidyverse_co
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(survival)
library(ggplot2)
```

1.

The data set that we used in lecture hern.txt contains information about treatment participation for heroin addicts in two clinics. In this problem, I want you to use likelihood ratio tests.

```
hern <- read.csv("hern.txt", sep="")
```

(a).

One of the covariates in the data, Prison, is a indicator of whether the subject had served prison time. Fit a Cox PH model to test whether prison has a significant effect on the time spent in the clinic.

```
fit = coxph(Surv(Time,Status)~Prison,data=hern)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(Time, Status) ~ Prison, data = hern)
##
##   n= 238, number of events= 150
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## Prison 0.1838    1.2018   0.1642 1.119    0.263
##
##        exp(coef) exp(-coef) lower .95 upper .95
## Prison     1.202     0.8321    0.8711     1.658
##
## Concordance= 0.536  (se = 0.023 )
## Likelihood ratio test= 1.25  on 1 df,   p=0.3
## Wald test            = 1.25  on 1 df,   p=0.3
## Score (logrank) test = 1.26  on 1 df,   p=0.3
```

According to the table above, we have p-value = 0.3 which is larger than 0.05, meaning that Prison is not statisitcally siginificant.

(b).

The column Clinic tells us which clinic the subject was at. Repeat your test regarding the effect of prison time, but control for the possible confounding effect of the clinic.

```
fit2 = coxph(Surv(Time,Status)~Prison+Clinic,data=hern)
fit3 = coxph(Surv(Time,Status)~Clinic,data=hern)
lrt = 2*(fit2$loglik[2]-fit3$loglik[2])
```

```
## calculate the df
df = length(coef(fit2)) - length(coef(fit3))
df ##output the df = 1
```

```
## [1] 1
```

```
## calculate the p-value
pchisq(lrt,df=1,lower.tail = FALSE)
```

```
## [1] 0.09493244
```

Since p-value = 0.0949 > .05, we could conclude that prison has no significant effect on the time.

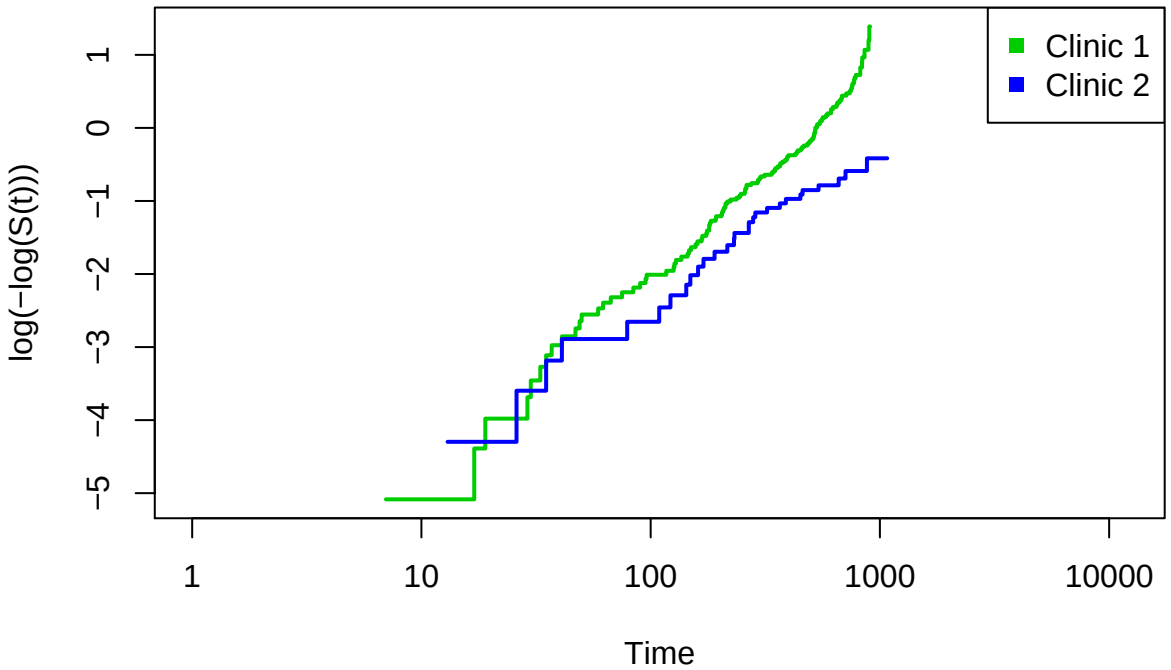
(c).

Use a complementary log-log plot to visualize whether the Cox PH assumption is appropriate for modeling the effect of the clinics. Describe you conclusions from looking at the plot.

```
hern.fit = survfit(Surv(Time,Status)~Clinic,data=hern)
plot(hern.fit,fun="cloglog",xlab = "Time", ylab = "log(-log(S(t)))",col = 3:4,lwd=2,xlim=c(1,12000))
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 1 x value <= 0 omitted from
## logarithmic plot
```

```
legend("topright",legend = c("Clinic 1","Clinic 2"),pch = rep(15,4),col = 3:4)
```



From the plot, we could see that two curves are not parallel and tend to diverge in the end, so we could conclude that the Cox PH assumption is not appropriate for modeling the effect of the clinics.

(d).

Model the effect of prison when you treat the clinics as stratified confounding variables. What do you conclude? Describe in your own words the difference between this model and the model in part (b).

```
fitSC <- coxph(Surv(Time,Status)~Prison+strata(Clinic),data=hern)
summary(fitSC)
```

```
## Call:
## coxph(formula = Surv(Time, Status) ~ Prison + strata(Clinic),
##     data = hern)
##
##   n= 238, number of events= 150
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## Prison 0.3359    1.3992   0.1675 2.005    0.045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## Prison     1.399     0.7147     1.008     1.943
##
## Concordance= 0.544  (se = 0.023 )
## Likelihood ratio test= 3.98  on 1 df,   p=0.05
## Wald test            = 4.02  on 1 df,   p=0.04
## Score (logrank) test = 4.05  on 1 df,   p=0.04
```

From the table above, we have p-value for prison $= 0.045 < 0.05$ so that we could conclude that prison is significant with the effect of clinics as stratified confounding variables.

(e).

Perform a test to see if there is a significant interaction between the prison variable and the clinic variable. We still want to use a stratified model.What do you conclude? Explain what the interaction term means.

First, fit a model with both Prison and Clinic, then fit a second model with only Clinic. Second, preform the LRT such that: H0: the reduced model is preferred (model without Prison) Ha: the full model is preferred (model with Prison).

```
fitSC.int = coxph(Surv(Time,Status)~Prison*strata(Clinic),data=hern)
summary(fitSC.int)
```

```
## Call:
## coxph(formula = Surv(Time, Status) ~ Prison * strata(Clinic),
##     data = hern)
##
##   n= 238, number of events= 150
##
##                                   coef exp(coef) se(coef)      z Pr(>|z|)
## Prison                          0.4055    1.5000   0.1862  2.177   0.0294 *
## Prison:strata(Clinic)Clinic=2 -0.3585    0.6987   0.4236 -0.846   0.3973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                               exp(coef) exp(-coef) lower .95 upper .95
## Prison                           1.5000     0.6667    1.0413     2.161
## Prison:strata(Clinic)Clinic=2    0.6987     1.4312    0.3046     1.603
```

```
##
## Concordance= 0.544  (se = 0.023 )
## Likelihood ratio test= 4.7  on 2 df,   p=0.1
## Wald test            = 4.76  on 2 df,   p=0.09
## Score (logrank) test = 4.82  on 2 df,   p=0.09
```

Then, perform a likelihood ratio test to compare between SC model with and without the interaction term. H0: The reduced model is preferred (without interaction term) Ha: The full model is preferred (with interaction term).

```
#Compute GLRT statistic:
lrt.int = 2*(fitSC.int$loglik[2]-fitSC$loglik[2])
lrt.int ## 0.7183184
```

```
## [1] 0.7183184
```

```
#Compute the df:
df.int =length(coef(fitSC.int)) - length(coef(fitSC))
df.int ## 1
```

```
## [1] 1
```

```
#p-value:
pchisq(lrt.int,df=1,lower.tail = FALSE)
```

```
## [1] 0.3966961
```

Since p-value is 0.3966961 which is larger than 0.05, we fail to reject the null hypothesis and the interaction is not significant. Therefore, we could conclude that effect of prison on time does not depend on Clinic.

2.

The data set retire has information on the life expectancy of individuals living in a senior care facility. We begin by modeling time column which is the survival time in months spent at the facility. The indicator column death will be used as our status variable. We would like to model the difference between men and women so there is a column gender which is 1 for men and 2 for women.

```
## Load the Data Set
retire <- read.table("retire.txt",header=TRUE,skip=2)
```

(a).

Use a Cox Proportional Hazards model to test whether there is a significant difference between men and women. Report the likelihood ratio statistic and the appropriate P value.

```
retire.fit <- coxph(Surv(time,death)~gender,data=retire)
summary(retire.fit)
```

```
## Call:
## coxph(formula = Surv(time, death) ~ gender, data = retire)
##
##   n= 462, number of events= 176
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## gender -0.4280    0.6518   0.1718 -2.492   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
```

```
## gender    0.6518      1.534    0.4655    0.9127
##
## Concordance= 0.54  (se = 0.017 )
## Likelihood ratio test= 5.78  on 1 df,   p=0.02
## Wald test            = 6.21  on 1 df,   p=0.01
## Score (logrank) test = 6.3  on 1 df,   p=0.01
```

From above, we have that Likelihood ratio test statistic = 5.78 on df =1 and the p-value for two gender is 0.02 which is < 0.05 so that we could conclude that there is a statistically significant difference between men and women.

(b).

Fit another model that adjusts for the confounding variable ageentry which gives the age in months of the subject when they entered the facility. Use the anova function to calculate the appropriate likelihood ratio test. Do you come to the same conclusion as in part (a)? How do you explain any difference?

```
retire.fit1 <- coxph(Surv(time,death)~gender+ageentry,data=retire)
retire.fit2 <- coxph(Surv(time,death)~ageentry,data=retire)
anova(retire.fit1)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, death)
## Terms added sequentially (first to last)
##
##           loglik   Chisq Df Pr(>|Chi|)
## NULL     -974.89
## gender   -972.00  5.7786  1    0.01622 *
## ageentry -950.15 43.6985  1  3.831e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above, we notice that p-value for ageentry = 3.81e-11 which means that the confounding variable ageentry has significant impact in the model.

```
anova(retire.fit2,retire.fit1)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, death)
##  Model 1: ~ ageentry
##  Model 2: ~ gender + ageentry
##    loglik  Chisq Df P(>|Chi|)
## 1 -952.39
## 2 -950.15 4.4917  1   0.03406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova table above, we could have the p-value = 0.03406 < 0.05, which means that gender has a significant effect on the time. This is as same as what we concluded in part a. The difference existed is because that we did not consider the ageentry while we do consider it in this part.

(c).

Fit a model with an interaction between age and gender. What do you conclude?

```
retire.fitint <- coxph(Surv(time,death)~age*gender,data = retire)
summary(retire.fitint)
```

```
## Call:
## coxph(formula = Surv(time, death) ~ age * gender, data = retire)
##
##   n= 462, number of events= 176
##
##                 coef exp(coef)  se(coef)      z Pr(>|z|)
## age        -0.011789  0.988280  0.004783 -2.465   0.0137 *
## gender     -5.426193  0.004400  2.619183 -2.072   0.0383 *
## age:gender  0.005012  1.005024  0.002644  1.895   0.0580 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## age           0.9883      1.012 9.791e-01    0.9976
## gender        0.0044    227.282 2.594e-05    0.7463
## age:gender    1.0050      0.995 9.998e-01    1.0102
##
## Concordance= 0.575  (se = 0.022 )
## Likelihood ratio test= 16.03  on 3 df,   p=0.001
## Wald test            = 19.08  on 3 df,   p=3e-04
## Score (logrank) test = 19.62  on 3 df,   p=2e-04
```

From above, we have p-value for the interaction of age and gender $= 0.001 < 0.05$ so we could conclude that the interaction is statistically siginificant.

Then we have the hypothesis test: H0: The model without interation is preferred vs. HA: The model with interation is preferred.

```
retire.fitnlint <- coxph(Surv(time,death)~age+gender,data = retire)
summary(retire.fitnlint)
```

```
## Call:
## coxph(formula = Surv(time, death) ~ age + gender, data = retire)
##
##   n= 462, number of events= 176
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## age    -0.003001  0.997004  0.001171 -2.562   0.0104 *
## gender -0.447285  0.639362  0.171973 -2.601   0.0093 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age       0.9970      1.003    0.9947    0.9993
## gender    0.6394      1.564    0.4564    0.8956
##
## Concordance= 0.583  (se = 0.022 )
## Likelihood ratio test= 12.47  on 2 df,   p=0.002
## Wald test            = 12.72  on 2 df,   p=0.002
## Score (logrank) test = 12.79  on 2 df,   p=0.002
```

```
lrtSC = 2*(retire.fitint$loglik[2]-retire.fitnlint$loglik[2])
lrtSC
```

```
## [1] 3.565097
```

Calculate the df:

```
dfSC = length(coef(retire.fitint)) - length(coef(retire.fitnlint))
dfSC # = 1
```

```
## [1] 1
```

Calculate the p-value:

```
pchisq(lrtSC,df=1,lower.tail = FALSE)
```

```
## [1] 0.0590063
```

From above, we know that p-value = 0.0590063 > 0.05, so we accept the null hypothesis that the model without interaction is prefered.

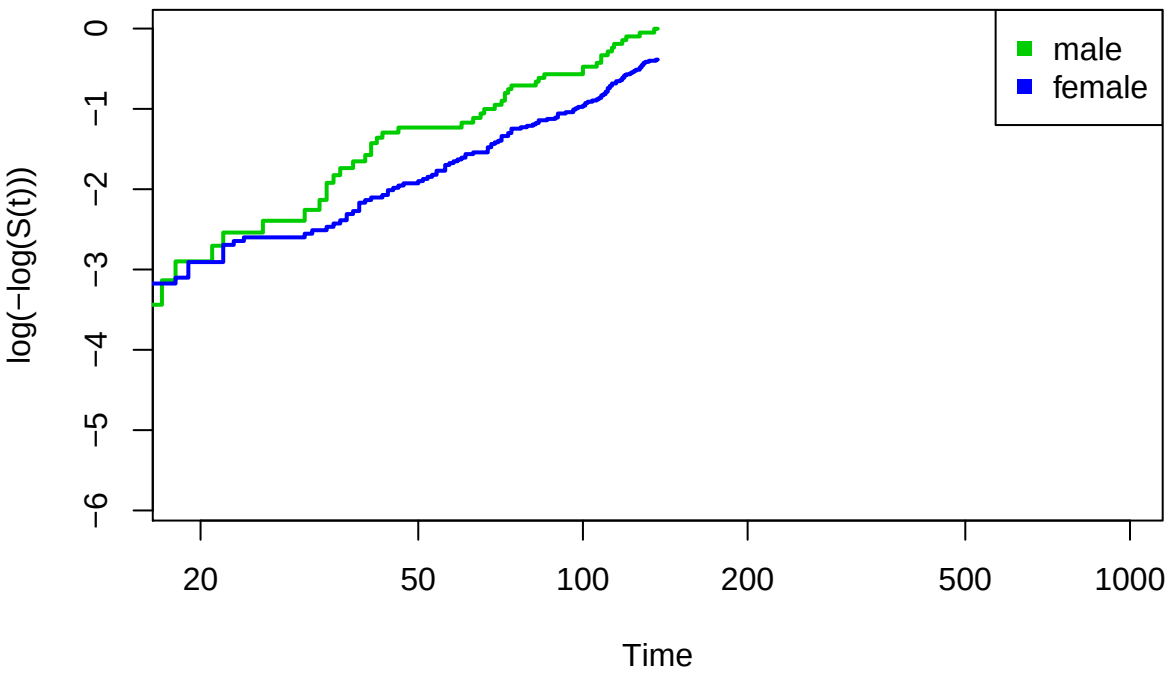
(d).

Plot complementary log-log plot comparing the effect of gender on the survival time. Do you think the proportional hazards assumption is reasonable for this model?

```
retire.fit5 <- survfit(Surv(time,death)~gender,data=retire)
plot(retire.fit5,fun="cloglog",xlab = "Time", ylab = "log(-log(S(t)))",col = 3:4,lwd = 2)
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 1 x value <= 0 omitted from
## logarithmic plot
```

```
legend("topright",legend = c("male","female"),pch = rep(15,4),col = 3:4)
```



From the plot we have above, there is no interaction between two diagrams so that we could conclude that the propotional hazards assumption is reasonable because we could see that hazards are propootional over time.

(e).

Explain clearly why we chose to use ageentry as our covariate and not age which is the age of the subject when the event occurred.

We use ageentry because that age only tell us the events time after entering the facility and we have no idea the age when they entered the facility. This ageentry will help us to figure out that would survival probability

be affected by the age when entering the facility.

3.

Using the same retire data set, I want to fit a Generalized Cox Model where the effect of gender is different before and after 48 months.

(a).

Use the survSplit function to construct a new data frame with additional rows that split the time variable into before and after 48 months.

```
## change 0s to .01, because otherwise survSplit won't run
z<-retire$time==0
for(i in 1:length(z)){
 if (z[i]==TRUE){
 retire$time[i]=.01
 }
 }
```

```
## construct a new data frame that split the time variable into before and after 48 months
retire.Splt <- survSplit(Surv(time,death)~., retire, cut = 48,
                         episode = "timegroup")
## show the first 5 rows of the new data set
head(retire.Splt,5)
```

```
##   obs ageentry age gender tstart time death timegroup
## 1 272      733 870      2      0   48     0         1
## 2 272      733 870      2     48  137     0         2
## 3  67      746 804      2      0   48     0         1
## 4  67      746 804      2     48   58     0         2
## 5  50      748 804      2      0   48     0         1
```

(b).

Use coxph to model the effect of gender including a change of parameter at 48 months. Please include the age at entry if that is still appropriate. Use our likelihood ratio test to determine if gender is still significant in this model.

```
retire.fit3 <- coxph(Surv(tstart,time,death)~gender*strata(timegroup)+ageentry,
                     data = retire.Splt)
summary(retire.fit3)
```

```
## Call:
## coxph(formula = Surv(tstart, time, death) ~ gender * strata(timegroup) +
##     ageentry, data = retire.Splt)
##
##   n= 767, number of events= 176
##
##                                       coef exp(coef)  se(coef)      z
## gender                           -0.574733  0.562855  0.266615 -2.156
## ageentry                          0.007100  1.007126  0.001049  6.770
## gender:strata(timegroup)timegroup=2  0.332986  1.395127  0.350236  0.951
##                                  Pr(>|z|)
## gender                             0.0311 *
## ageentry                         1.28e-11 ***
```

```
## gender:strata(timegroup)timegroup=2   0.3417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                     exp(coef) exp(-coef) lower .95 upper .95
## gender                                 0.5629     1.7767    0.3338    0.9492
## ageentry                               1.0071     0.9929    1.0051    1.0092
## gender:strata(timegroup)timegroup=2    1.3951     0.7168    0.7023    2.7716
##
## Concordance= 0.648  (se = 0.023 )
## Likelihood ratio test= 50.38  on 3 df,   p=7e-11
## Wald test            = 52.64  on 3 df,   p=2e-11
## Score (logrank) test = 53.84  on 3 df,   p=1e-11
```

```
retire.fit4 <- coxph(Surv(tstart,time,death)~ageentry,
                     data = retire.Splt)
summary(retire.fit4)
```

```
## Call:
## coxph(formula = Surv(tstart, time, death) ~ ageentry, data = retire.Splt)
##
##   n= 767, number of events= 176
##
##              coef exp(coef) se(coef)     z Pr(>|z|)
## ageentry 0.007166  1.007192 0.001039 6.897 5.32e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## ageentry     1.007     0.9929     1.005     1.009
##
## Concordance= 0.642  (se = 0.023 )
## Likelihood ratio test= 44.99  on 1 df,   p=2e-11
## Wald test            = 47.56  on 1 df,   p=5e-12
## Score (logrank) test = 48.22  on 1 df,   p=4e-12
```

```
anova(retire.fit3,retire.fit4)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(tstart, time, death)
##  Model 1: ~ gender * strata(timegroup) + ageentry
##  Model 2: ~ ageentry
##    loglik  Chisq Df P(>|Chi|)
## 1 -949.70
## 2 -952.39 5.3922  2   0.06747 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above, we have p-value $= 0.06747 > 0.05$, so conclude that gender is not significant in this model after split time into before and after 48 months.

(c).

Give 95% confidence intervals for the hazard ratio for men before and after the 48 month cut off.

```
retire.male <-coxph(Surv(tstart,time,death)~ageentry+(gender==1)*strata(timegroup==1),data=retire.Splt)
exp(confint(retire.male,level = 0.95))
```

```
##                                                                    2.5 %   97.5 %
## ageentry                                                        1.0050576 1.009198
## gender == 1TRUE                                                 0.8159375 1.987571
## gender == 1TRUE:strata(timegroup == 1)timegroup == 1=TRUE 0.7022516 2.771629
```

```
retire.male2 <- coxph(Surv(tstart,time,death)~ageentry+(gender==1)*strata(timegroup==2),data=retire.Spl
exp(confint(retire.male2,level = 0.95))
```

```
##                                                                    2.5 %   97.5 %
## ageentry                                                        1.0050576 1.009198
## gender == 1TRUE                                                 1.0535639 2.996031
## gender == 1TRUE:strata(timegroup == 2)timegroup == 2=TRUE 0.3607987 1.423991
```

So 95% confidence intervals for the hazard ratio for men before and after the 48 month cut off is (0.7022516 2.771629) and (0.3607987 1.423991).

(d).

Would you conclude that it is important to consider a change in the effect of gender before and after 4 years in the retirement facility?

```
retire.fit6 <- coxph(Surv(tstart,time,death)~gender:strata(timegroup) + ageentry,data=retire.Splt)
summary(retire.fit6)
```

```
## Call:
## coxph(formula = Surv(tstart, time, death) ~ gender:strata(timegroup) +
##     ageentry, data = retire.Splt)
##
##   n= 767, number of events= 176
##
##                                        coef exp(coef)  se(coef)      z
## ageentry                           0.007100  1.007126  0.001049  6.770
## gender:strata(timegroup)timegroup=1 -0.574733  0.562855  0.266615 -2.156
## gender:strata(timegroup)timegroup=2 -0.241748  0.785254  0.227129 -1.064
##                                    Pr(>|z|)
## ageentry                           1.28e-11 ***
## gender:strata(timegroup)timegroup=1   0.0311 *
## gender:strata(timegroup)timegroup=2   0.2872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                    exp(coef) exp(-coef) lower .95 upper .95
## ageentry                              1.0071     0.9929    1.0051    1.0092
## gender:strata(timegroup)timegroup=1    0.5629     1.7767    0.3338    0.9492
## gender:strata(timegroup)timegroup=2    0.7853     1.2735    0.5031    1.2256
##
## Concordance= 0.648  (se = 0.023 )
## Likelihood ratio test= 50.38  on 3 df,   p=7e-11
## Wald test            = 52.64  on 3 df,   p=2e-11
## Score (logrank) test = 53.84  on 3 df,   p=1e-11
```

Based on the summary we had before, we could see that when in the timegroup=1(before 4 years), gender is significant (p-value=0.0311) while it is not significant in the timegroup=2(after 4 years) (p-value=0.2872).Therefore, it is important to consider a change in the effect of gender before and after 4 years in the retirement facility.