Pstat 175 Project

Yangcheng Lai, Anthony Yang, Anson Lo

Due: 2020/12/19

Description of data

This data is a study of factors associated with death from cardiovascular disease or other cause. The follow up time is days after admission to a hospital. The cause of death was recorded as Coronary Vascular Disease (CVD), Other Cause (OC) or Censored. The primary event we are interested in is CVD. The covariates are age (participants' age in years), gender (participants' gender), and BMI (participants' Body Mass Index in kg/m^2).

Scientific questions

The main question we are interested in from this data is what factors are associated with death from CVD or other causes? Some interesting covariates that might help us explore this problem would be BMI; we can explore whether participants with a higher Body Mass Index tends to have a lower survival rate from CVD or other causes. The confounding variable for this problem might be participants' age, because participants' age could potentially influence the survival rate greatly thus affecting our study. Due to it has 2 tpyes of events (CVD and OC), we can use risk competing model to analysis these two events seperately, e.g. treat OC as censored while we analysis CVD.

```r
library(survival)  ##load the library survival
```

```r
id <- comprisk$V1
age <- comprisk$V2
gender <- comprisk$V3
bmi <- comprisk$V4
time <- comprisk$V5
ev_typ <- comprisk$V6
```

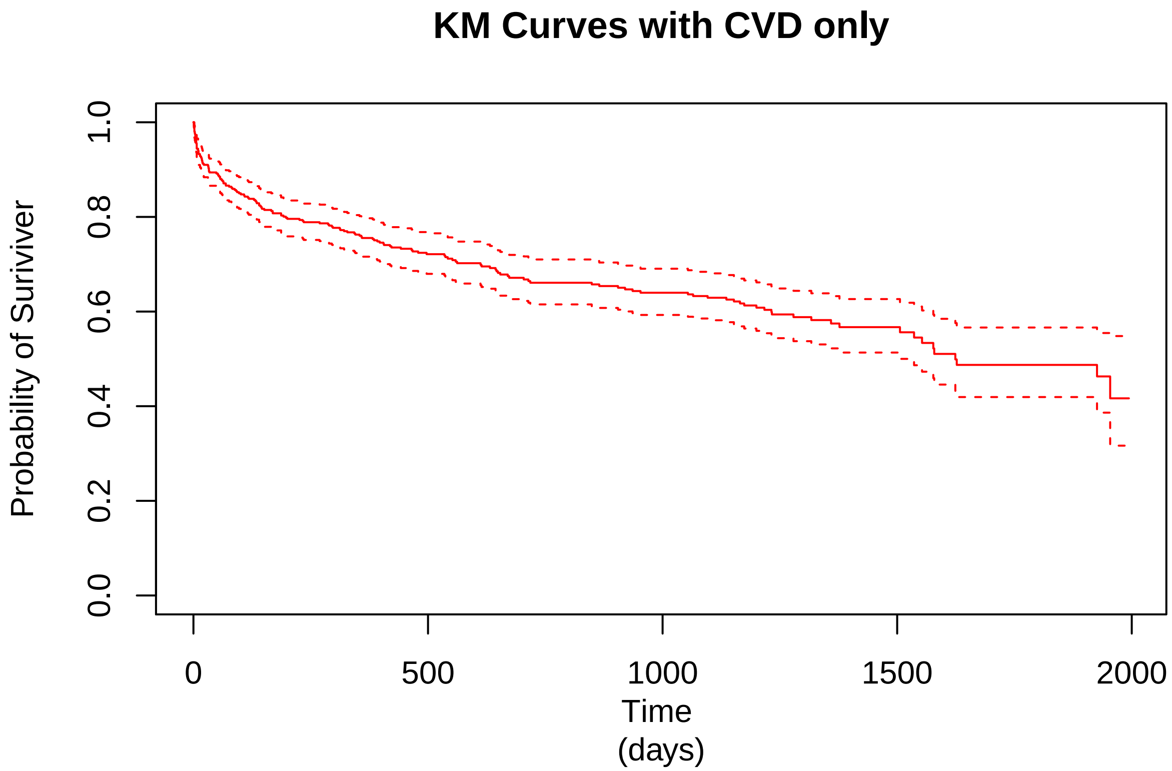
KM graph

case 1: CVD

Here we transform all OC events into censored, then the data only have events of CVD and failure, and then we can focus on one specific cause (CVD cause only)).

```r
comprisk$V7 <- comprisk$V6
comprisk$V7[comprisk$V6 == 2 ] <- 0
```

```
time.sur1 <- Surv(time, comprisk$V7)
time.fit1 = survfit(time.sur1~1)

plot(time.fit1,main="KM Curves with CVD only",
     xlab="Time \n(days)",ylab="Probability of Suriviver",
     col = "red")
```



According to this KM graph, CVD occurs frequently at the first half period of time in the study, and the the second part becomes flatter and survival rate decreases faster. Thus we may think that, as time becomes longer, there are more censored observations and few death due to CVD. And if we only focus on CVD, the survival rate will only decrease till around 0.5

```
cvd.cox <- coxph(time.sur1~age+gender+bmi,data = comprisk)

summary(cvd.cox)
```

```
## Call:
## coxph(formula = time.sur1 ~ age + gender + bmi, data = comprisk)
##
##   n= 453, number of events= 167
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## age     0.085313  1.089058  0.005921 14.408  < 2e-16 ***
## gender -0.146217  0.863971  0.156959 -0.932    0.352
## bmi     0.061992  1.063953  0.014628  4.238 2.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age        1.089     0.9182    1.0765     1.102
## gender     0.864     1.1574    0.6352     1.175
## bmi        1.064     0.9399    1.0339     1.095
##
## Concordance= 0.85  (se = 0.013 )
## Likelihood ratio test= 306.2  on 3 df,   p=<2e-16
## Wald test            = 218.6  on 3 df,   p=<2e-16
## Score (logrank) test = 281.2  on 3 df,   p=<2e-16
```

As we can see, if we only treat CVD as event, and regard OC as censored, we find out that the p values for age, gender, and bmi are 2e-16, 0.352, and 2.26e-05 respectively. Therefore we can conclude that there is a significant difference between the survival rates of different observations' age and bmi. Instead, the gender seems not too significant.

now, we use AIC to select covariates.

```
model1 <- coxph(time.sur1~age, data = comprisk)
model2 <- coxph(time.sur1~gender, data = comprisk)
model3 <- coxph(time.sur1~bmi, data = comprisk)
AIC(model1, model2, model3)
```

```
##        df      AIC
## model1  1 1584.892
## model2  1 1866.102
## model3  1 1851.093
```

Age is our first pick! Then choose second from gender and bmi.

```
model1.2 <- coxph(time.sur1~age+gender, data = comprisk)
model1.3 <- coxph(time.sur1~age+bmi, data = comprisk)
AIC(model1.2, model1.3)
```

```
##          df      AIC
## model1.2  2 1586.086
## model1.3  2 1569.974
```

Therefore our final model will be age + bmi + gender.

```
model.full <- coxph(time.sur1~age+bmi+gender, data = comprisk)
summary(model.full)
```

```
## Call:
## coxph(formula = time.sur1 ~ age + bmi + gender, data = comprisk)
##
##   n= 453, number of events= 167
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## age     0.085313  1.089058  0.005921 14.408  < 2e-16 ***
## bmi     0.061992  1.063953  0.014628  4.238 2.26e-05 ***
## gender -0.146217  0.863971  0.156959 -0.932    0.352
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age        1.089     0.9182    1.0765     1.102
## bmi        1.064     0.9399    1.0339     1.095
## gender     0.864     1.1574    0.6352     1.175
##
## Concordance= 0.85  (se = 0.013 )
## Likelihood ratio test= 306.2  on 3 df,   p=<2e-16
## Wald test            = 218.6  on 3 df,   p=<2e-16
## Score (logrank) test = 281.2  on 3 df,   p=<2e-16
```

By the summary result, we see that age truely has the highest hazard rate for CVD, and gender has the smallest coefficient. In additional, higher the age and bmi are, higher the hazard rates are, thus a lower survival rate.

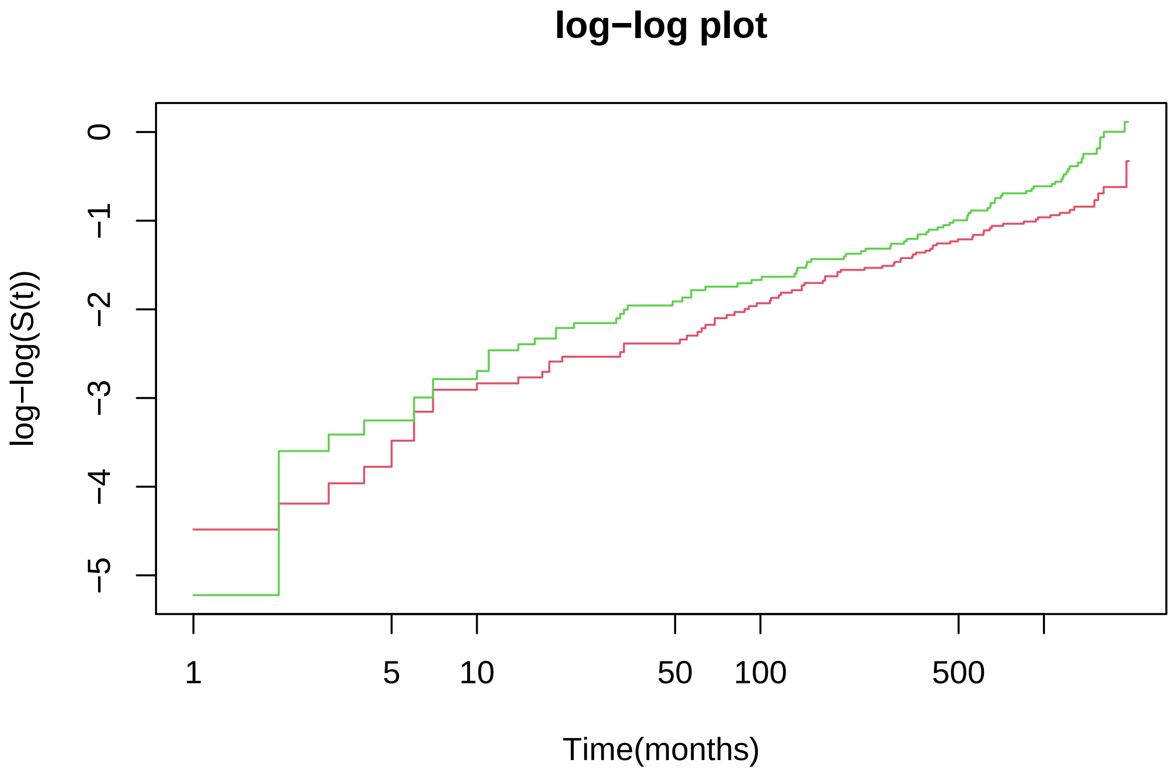
```
# use cox.zph to check our assumption
cox.zph(model.full)
```

```
##          chisq df      p
## age      2.873  1 0.0901
## bmi      7.505  1 0.0062
## gender   0.311  1 0.5772
## GLOBAL 11.258  3 0.0104
```

Due to p value of bmi is smaller than 0.05, the significant level, thus our assumption could be not very appropriate.

Check Assumptions (c-loglog plot)

```
plot(survfit(time.sur1~comprisk$V3),
fun="cloglog",
col = c(2,3),
ylab = "log-log(S(t))",
xlab = "Time(months)",
main = "log-log plot")
```

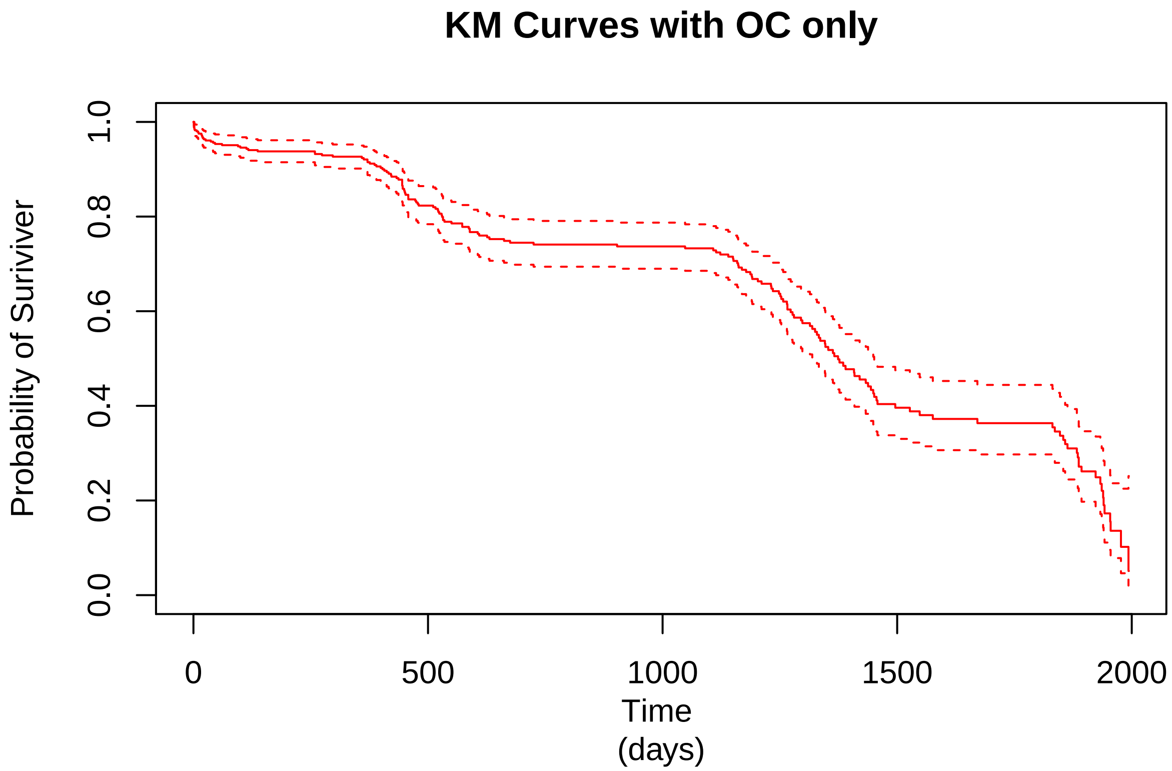


From the log-log plot above, we can see that there are some narrowing of the gap between the curves. But overall the graph is parallel and it suggests that there is not much concerned about our proportional hazards assumption.

case 2: OC

Then we transform all CVD events into censored (OC cause only).

```
comprisk$V8 <- comprisk$V6
comprisk$V8[comprisk$V6 == 1] <- 0
comprisk$V8[comprisk$V6 == 2] <- 1
time.sur2 <- Surv(time, comprisk$V8)
time.fit2 = survfit(time.sur2~1)
plot(time.fit2,main="KM Curves with OC only",
xlab="Time \n(days)",ylab="Probability of Suriviver",
col = "red")
```



According to the KM graph which only take OC as event, we see that the survival rate decreases pretty unpredictable. It sometimes drops faster and sometimes slow.

```
oc.cox <- coxph(time.sur2~age+gender+bmi,data = comprisk)

summary(oc.cox)
```

```
## Call:
## coxph(formula = time.sur2 ~ age + gender + bmi, data = comprisk)
##
##   n= 453, number of events= 170
##
##              coef exp(coef)  se(coef)      z Pr(>|z|)
## age    -0.028371  0.972027  0.004889 -5.803 6.52e-09 ***
## gender  0.069182  1.071632  0.171055  0.404    0.686
## bmi    -0.066547  0.935619  0.015825 -4.205 2.61e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age       0.9720     1.0288    0.9628    0.9814
## gender    1.0716     0.9332    0.7664    1.4985
## bmi       0.9356     1.0688    0.9070    0.9651
##
## Concordance= 0.668  (se = 0.021 )
## Likelihood ratio test= 57.7  on 3 df,   p=2e-12
## Wald test            = 54.33  on 3 df,   p=1e-11
## Score (logrank) test = 55.19  on 3 df,   p=6e-12
```

Compared with case 1, it's similar that both cases show that age and bmi are factors that cause difference, and gender has no significant difference. However, in case 2, the p value for age is a lot smaller than that of case 1. Then we may think that age in case 2 has relatively smaller effect on the survival rate than age in case 1. It makes some sence to me, because cardiocascular happens frequently on older people, and death of "other event" could be not too relative with age.

now, we use AIC to select covariates.

```
model1 <- coxph(time.sur2~age, data = comprisk)
model2 <- coxph(time.sur2~gender, data = comprisk)
model3 <- coxph(time.sur2~bmi, data = comprisk)
AIC(model1, model2, model3)
```

```
##        df      AIC
## model1  1 1641.561
## model2  1 1678.379
## model3  1 1658.492
```

Our first pick is age;

```
model1.2 <- coxph(time.sur2~age+gender, data = comprisk)
model1.3 <- coxph(time.sur2~age+bmi, data = comprisk)
AIC(model1.2, model1.3)
```

```
##          df      AIC
## model1.2  2 1643.097
## model1.3  2 1624.046
```

Second pick is bmi, and third pick is gender. Thus, our last model in this case is also age + bmi + gender.

```
model.full2 <- coxph(time.sur2~age+bmi+gender, data = comprisk)
summary(model.full2)
```

```
## Call:
## coxph(formula = time.sur2 ~ age + bmi + gender, data = comprisk)
##
##   n= 453, number of events= 170
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## age    -0.028371  0.972027  0.004889 -5.803 6.52e-09 ***
## bmi    -0.066547  0.935619  0.015825 -4.205 2.61e-05 ***
## gender  0.069182  1.071632  0.171055  0.404    0.686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age       0.9720     1.0288    0.9628    0.9814
## bmi       0.9356     1.0688    0.9070    0.9651
## gender    1.0716     0.9332    0.7664    1.4985
```

```
##
## Concordance= 0.668  (se = 0.021 )
## Likelihood ratio test= 57.7  on 3 df,   p=2e-12
## Wald test            = 54.33  on 3 df,   p=1e-11
## Score (logrank) test = 55.19  on 3 df,   p=6e-12
```

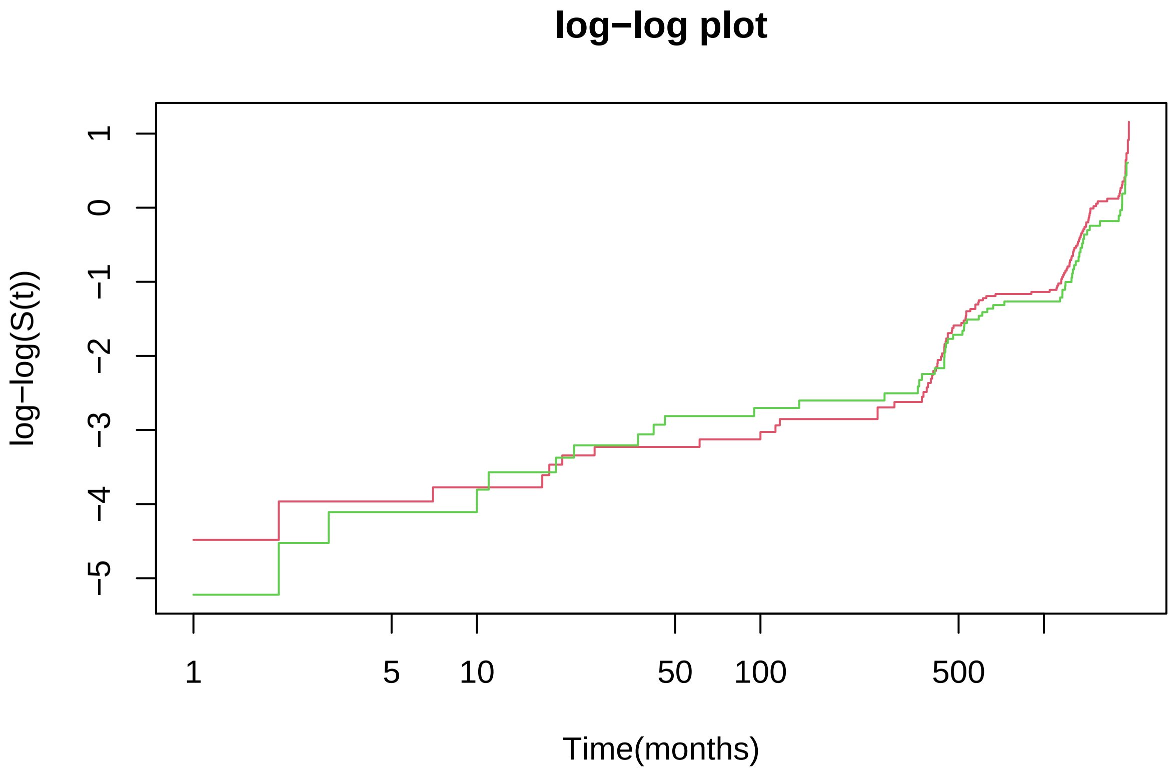
```
# use cox.zph to check our assumption
cox.zph(model.full2)
```

```
##        chisq df     p
## age     3.42  1 0.064
## bmi     3.67  1 0.055
## gender  0.84  1 0.359
## GLOBAL  7.32  3 0.062
```

The p values are all greater than 0.05, the significant level, thus we fail to reject the Null hypothesis and claim there is no significant evidence for us to abandon the PH assumption.

c-loglog plot

```
plot(survfit(time.sur2~comprisk$V3),
fun="cloglog",
col = c(2,3),
ylab = "log-log(S(t))",
xlab = "Time(months)",
main = "log-log plot")
```

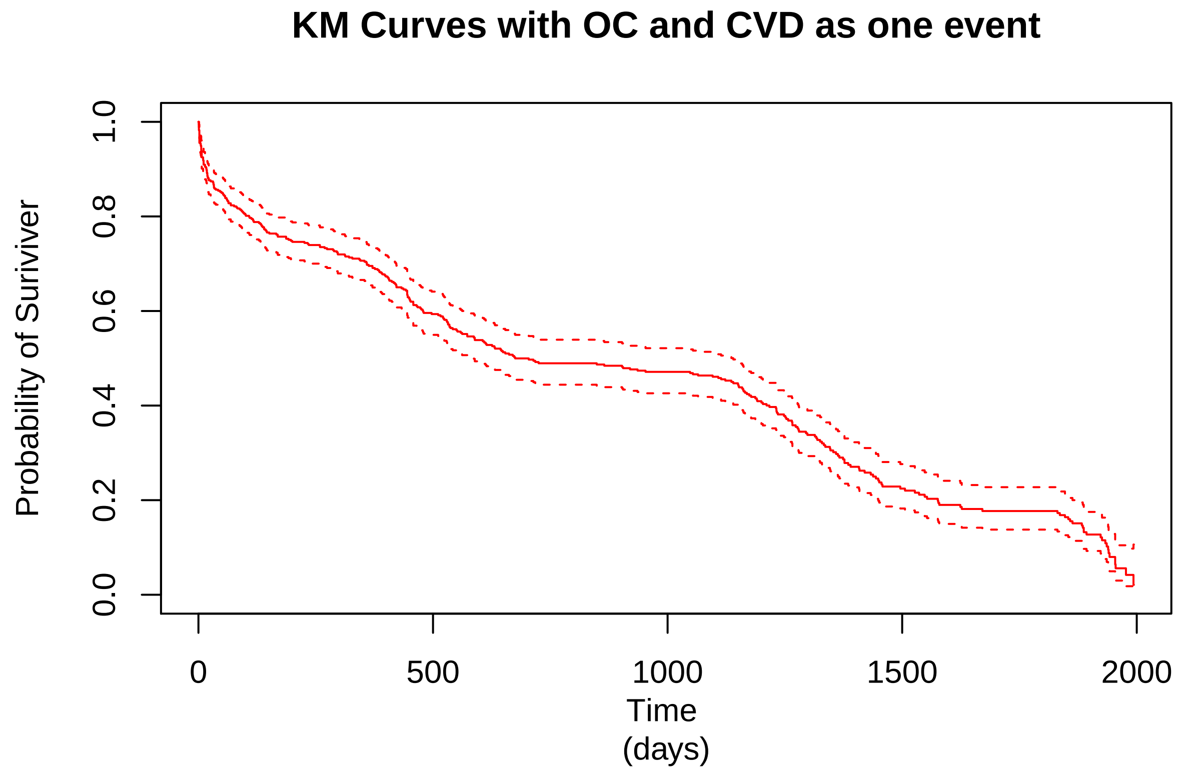


From the above log-log plot, we can see that it crosses after 10 months. This suggests that we should be concerned about our proportional hazards assumption.

Case 3 both CVD and OC

Lastly, we include both CVD and OC as one single event, so the event will only be event or censored (All causes).

```
comprisk$V9 <- comprisk$V6
comprisk$V9[comprisk$V6 == 2] <- 1
time.sur3 <- Surv(time, comprisk$V9)
time.fit3 = survfit(time.sur3~1)
plot(time.fit3,main="KM Curves with OC and CVD as one event",
xlab="Time \n(days)",ylab="Probability of Suriviver",
col = "red")
```



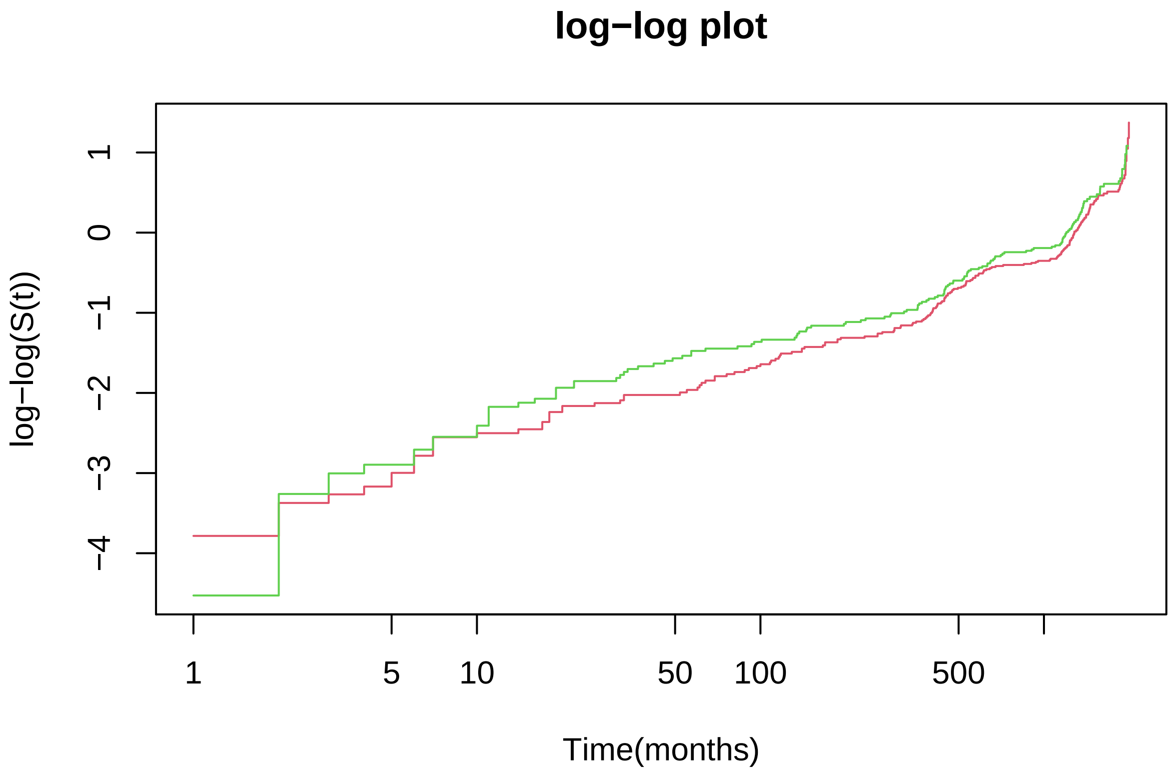
```
both.cox <- coxph(time.sur3~age+gender+bmi,data = comprisk)

summary(both.cox)
```

```
## Call:
## coxph(formula = time.sur3 ~ age + gender + bmi, data = comprisk)
##
##   n= 453, number of events= 337
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## age     0.024083  1.024375  0.003244  7.425 1.13e-13 ***
## gender -0.101214  0.903740  0.113891 -0.889    0.374
## bmi    -0.001643  0.998358  0.010329 -0.159    0.874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age       1.0244     0.9762    1.0179     1.031
## gender    0.9037     1.1065    0.7229     1.130
## bmi       0.9984     1.0016    0.9783     1.019
##
## Concordance= 0.666  (se = 0.014 )
## Likelihood ratio test= 57.48  on 3 df,   p=2e-12
## Wald test            = 56.24  on 3 df,   p=4e-12
## Score (logrank) test = 57.2  on 3 df,   p=2e-12
```

From here, if we treat both CVD and OC as an event, then the age will be the only factor that there is a significant difference.

Check Assumptions (log-log plot)

```
plot(survfit(time.sur3~comprisk$V3),
fun="cloglog",
col = c(2,3),
ylab = "log-log(S(t))",
xlab = "Time(months)",
main = "log-log plot")
```



From the loglog plot above, we can see that overall the graph is parallel although at the end the gap is narrowing so that there is not much concerned about our proportional hazards assumption.

Competing risk

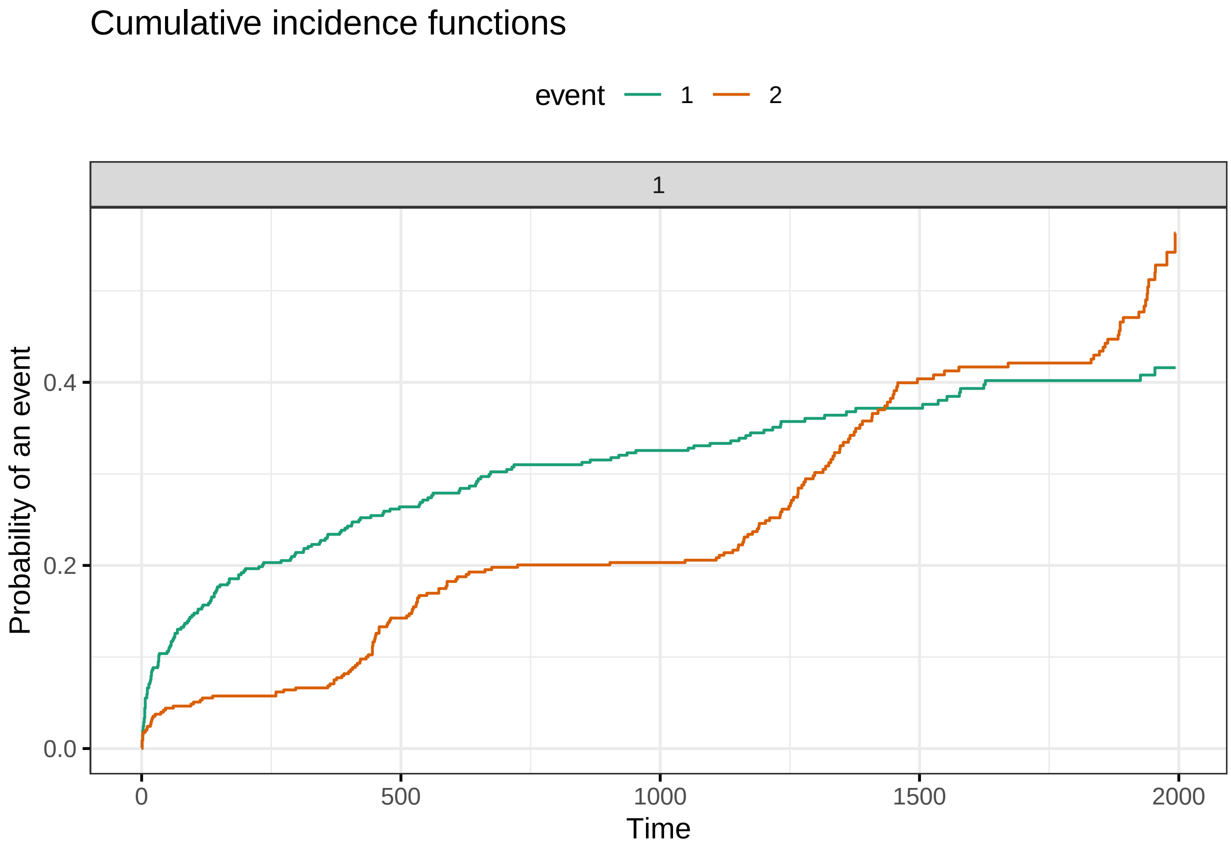
```
library("cmprsk")
```

```
## Warning: package 'cmprsk' was built under R version 4.0.3
```

```
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 4.0.3

## Loading required package: ggplot2

## Loading required package: ggpubr

## Warning: package 'ggpubr' was built under R version 4.0.3
```

```
fit3 <- cuminc(ftime = time, fstatus = ev_typ)
ggcompetingrisks(fit3, palette = "Dark2", legend = "top", ggtheme = theme_bw())
```



```
print(fit3)
```

```
## Estimates and Variances:
## $est
##           500      1000      1500
## 1 1 0.2639860 0.3255391 0.3716929
## 1 2 0.1425074 0.2031471 0.4038211
##
## $var
##              500         1000         1500
## 1 1 0.0004328607 0.0005084718 0.0005758851
## 1 2 0.0002778898 0.0003822172 0.0007174177
```

```
# 1 is CVD, 2 is OC
```

From the graph, we see that the probability of CVD is higher than OC for the most time, and then become lower after around 1400 days.

Conclusion

By analyzing CVD and OC separately, we find out that the most essential factor associated with death from CVD or other causes is the observations' age, and the following is bmi, body mass index, and finally observations' gender. At first, we thought that bmi could be the most important covariate that causes CVD. After comparing AIC and coefficiant, instead, we then find out that age is a more important covariate for CVD. By using risk competing model, we also conclude that the best models for these 2 event types are the same, which is age + bmi + gender; both of the probability of CVD and OC increase if observations' age increase, but in contrast, age has more impact on CVD than that of OC, and from the cumulative graph, after approximately 1400 days of observation, the probability of OC will exceed probability of CVD.

Citation

Our data is from Chapter 9 of Hosmer, D.W. and Lemeshow, S. and May, S. (2008) Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition, John Wiley and Sons Inc., New York, NY. (it's also on GauchoSpace)