Homework #4 **20180047 김경수**

*rds 파일들은 다음과 같이 implement 하였고, subtype이 NA인 데이터들을 다음과 같이 배제시켰다.

```
clinical<-readRDS("clinical.rds") # Clinical annotation
expression<-readRDS("expression.rds") # Gene expression
mutation<-readRDS("mutation.rds") # Mutation

sample_id_filter<-intersect(colnames(expression),
intersect(unique(mutation$sample_id),
clinical$sample_id[which(!is.na(clinical$subtype))]))
clinical_filter<-dplyr::filter(clinical,sample_id %in% sample_id_filter)
mutation_filter<-dplyr::filter(mutation, sample_id %in% sample_id_filter)
expression_filter<-expression[,as.character(clinical_filter$sample_id)]
```

Homework #4-1

Describe the purpose of Kaplan-Meier log-rank test and Cox regression and their difference.

(1) Kaplan-Meier log-rank test의 목적은 다음과 같다. 주어진 유의수준 하에서 두 가지 이상의 그룹들 사이의 survival curve의 차이가 있음을 확인하기 위함이다.
O_i : group i에서 관찰된 event의 수, E_i : group I에서의 expected event의 수라고 할 때, $\sum \frac{(O_i - E_i)^2}{E_i}$가 자유도가 (group의 개수 - 1)인 카이제곱 분포를 따름을 이용해서 이를 판단한다. difference가 존재하는지 여부를 확인할 수 있으나, 여러 explanatory variable까지 동시에 고려할 수 없다.

(2) Cox regression은, log-rank test와는 달리 survival curve에 여러 explanatory variable이 미치는 영향을 파악할 수 있고(다변수 처리 가능), binary, continuous한 변수들에도 적용될 수 있다.
Cox regression의 목적 또한 log-rank test처럼 survival curve의 차이가 있음을 확인하기 위함이다. 앞서 말한 바와 같이, log-rank test와는 달리 사건의 발생에 영향을 줄 수 있는 다른 변수들의 분석도 가능하다. 각 사건이 발생한 시점에서 사건 발생에 영향을 줄 수 있는 변수에 대해, 사건이 발생한 경우와 그렇지 않은 경우의 위험의 수준을 비교하여 결과를 도출한다.

Homework #4-2

Perform Kaplan-Meier test and Cox regression using stage and subtype data. Interpret those results with survival plot or forest plot, and corresponding p-values.
normal-like subtype도 포함하여 진행했습니다.

1. Procedure & R code

이론적 배경은 4-1에서 설명하였다.

(1) 먼저 stage에 대해 Kaplan-Meier test와 Cox regression을 진행하기 위해, 다음과 같이 vector들을 설정한다. "survival" package를 이용한다.
R code :

```
clinical_filter$stage <- factor(clinical_filter$stage)
vital_status <- 1*(clinical_filter$vital_status == 0)
survival_time <- as.numeric(clinical_filter$survival_time)
su <- survival::Surv(survival_time, vital_status)
```

(2) survfit 함수를 이용해서 Kaplan-Meier test를 진행하고, 그 결과를 ggsurvplot 함수를 이용하여 plot을 그려 관찰한다.
R code :

```
fit.survival<-survfit(su~stage, data=clinical_filter)
fit.survival
ggsurvplot(fit.survival, data=clinical_filter, pval=TRUE)
```

(3) coxph 함수를 이용해서 cox regression을 진행하고, 그 결과를 ggforest 함수를 이용하여 plot을 그려 관찰한다.
R code :

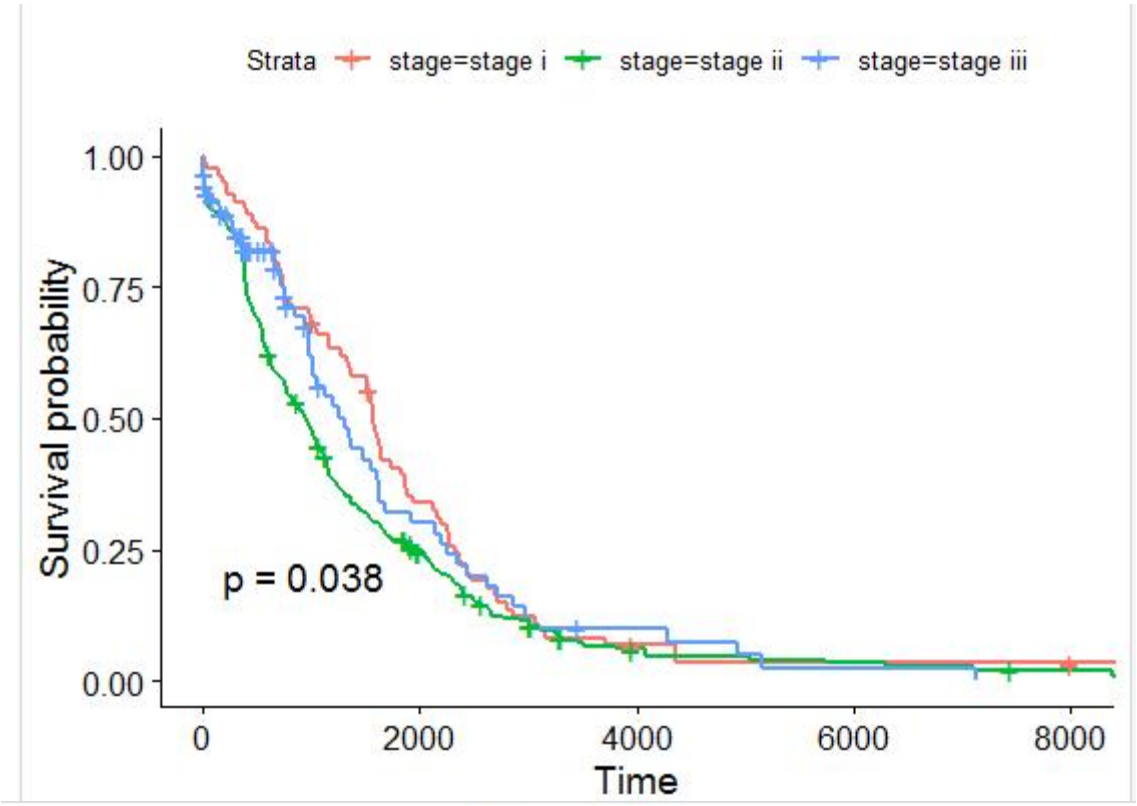
```
fit.cox<-coxph(su~stage, data=clinical_filter)
fit.cox
ggforest(fit.cox, data=clinical_filter)
```

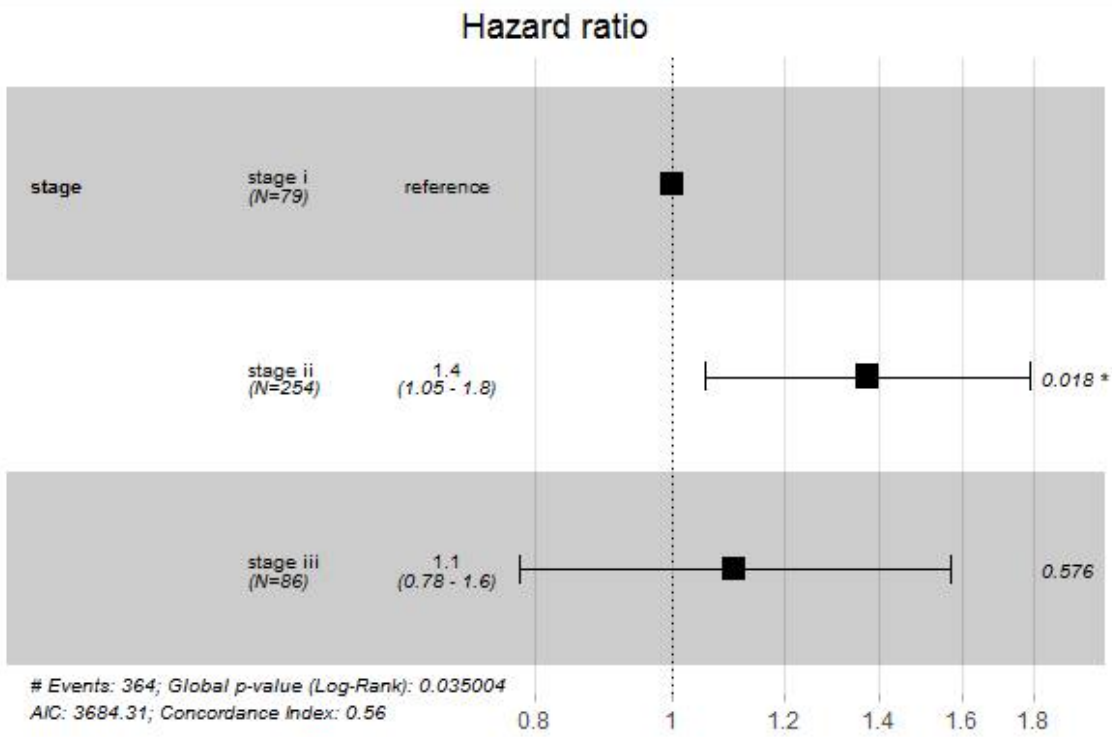
(4) subtype에 대해서도 (1)~(3) 과정을 반복한다.
R code :

```
clinical_filter$subtype <- factor(clinical_filter$subtype)
fit.survival<-survfit(su~subtype, data=clinical_filter)
fit.survival
ggsurvplot(fit.survival, data=clinical_filter, pval=TRUE)
fit.cox<-coxph(su~subtype, data=clinical_filter)
fit.cox
ggforest(fit.cox, data=clinical_filter)
```

2. Results

(1) Stage data에 관한 Kapaln-Meier test와 cox regression의 plot은 다음과 같다.

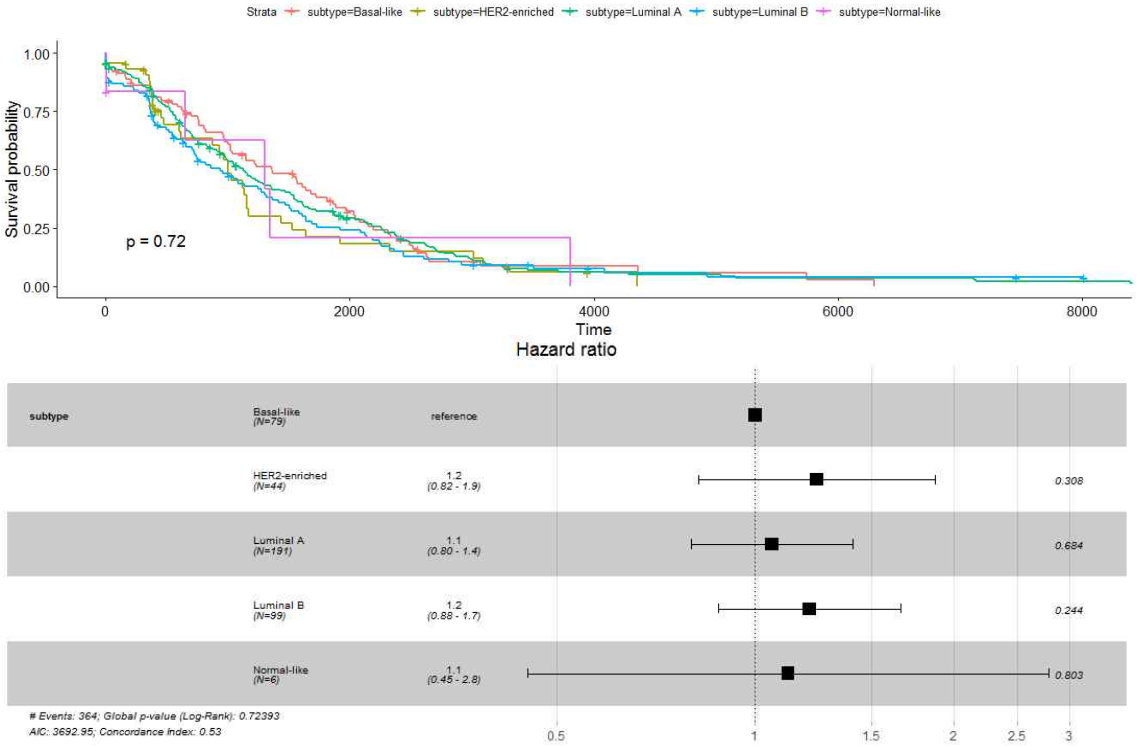




Kapaln-Meier test의 경우 p-value가 0.038이 나왔다. 따라서 0.05 유의수준에서 survival curve의 difference가 존재한다고 할 수 있다. Cox regression의 경우 stage1을 reference로 할 시, hazard ratio의 confidence interval이 stage2의 경우 1.05~1.8(p-value : 0.018) /

stage3의 경우 0.78~1.6(p-value : 0.576) 으로 측정되었다. 따라서 0.05유의수준에서 stage1과 stage2의 survival curve에는 difference가 있다고 판단할 수 있다.

(2) Subtype data에 관한 Kapaln-Meier test와 cox regression의 plot은 다음과 같다.



Kapaln-Meier test의 경우 p-value가 0.72이 나왔다. 따라서 0.05 유의수준에서 survival curve의 difference가 존재하지 않다고 할 수 있다. Cox regression의 경우 Basal-like를 reference로 할 시, hazard ratio의 confidence interval이 HER2-enriched의 경우 0.82~1.9(p-value : 0.308) / Luminal A의 경우 0.80~1.4(p-value : 0.684) / Luminal B의 경우 0.88~1.7(p-value : 0.244) / Normal-like의 경우 0.45-~2.8(p-value : 0.803) 으로 측정되었다. 따라서 0.05유의수준에서 subtype에 따른 survival curve의 difference는 존재하지 않는다고 결론내릴 수 있다.