Homework #3 **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[,sample_id_filter]
```

Homework #3-1

1. Procedure & R code

이론적 배경은 다음과 같다.

주어진 두 random variable (X,Y)의 n 쌍의 sample data에 대해, Pearson correlation coefficient(추후 r로 서술)은 다음과 같이 정의된다.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Test statistic T = $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ **는 degree of freedom 이 n-2(n : 표본의 개수)인 t분포를 따른다. 본 레포트에서는 1%의 유의수준에서 H0를 기각할 것인지 결정하였다.**
(if p-value > 0.01 : Accept H0, else : reject H0)

H0 : r = 0 (두 random variable은 상관관계가 없다.)
H1 : r != 0 (두 random variable은 양의 상관관계가 있다.)

(1) 상관계수와 pvalue를 저장할 vector를 선언하고, for문을 이용하여 각각의 gene마다 pearson의 correlation test를 진행하고, 그 상관계수 값과 pvalue 값을 vector에 저장한다.
R code :

```
cor.coef<-cor.pvalue<-c()
for(i in 1:dim(expression_filter)[1]){
  survival.time<-clinical_filter$survival_time
```

```
  expression<-expression_filter[i,]
  cor.result<-cor.test(survival.time,expression, method="pearson")
  cor.coef<-c(cor.coef, cor.result$estimate)
  cor.pvalue<-c(cor.pvalue, cor.result$p.value)
}
```

(2) 상관계수가 큰 순서대로 sort한다. 그 후 p-value 값이 0.01보다 작은(H0 reject) 데이터들만 남긴다. 그 후 correlation coefficient가 가장 큰 top 10 gene을 출력한다.

R code :

```
gene.correlation_sort<-row.names(expression_filter)[order(cor.coef,decreasing=TRUE)]
gene.correlation_sort_sort<-gene.correlation_sort[which(cor.pvalue<0.01)]
head(gene.correlation_sort_sort, 10)
```

2. Results

상관계수가 큰 순서대로, top 10 genes는 다음과 같다.

Table 1. Top 10 genes

Top 10 genes
1. KRT5
2. CDH3
3. LRIG2
4. NTSR2
5. STEAP3
6. MALT1
7. RIMS3
8. RBP5
9. C4orf7
10. FLJ40298

Homework #3-2

1. Procedure & R code

이론적 배경은 다음과 같다.

다중선형회귀모형에서, 가장 적합한 regression model을 찾는 것은 k개의 독립변수 중 y를 모형화하기 좋은 가장 적답한 부분집합을 찾는 과정으로서, p-value가 0.1보다 작은 독립변수들로 구성되어 있다.

먼저 모든 k개의 독립변수(본 레포트에서는 k=10)을 이용하여 모형을 적합시킨다. 하나 혹은 여러 개의 독립변수가 0,1 보다 큰 p-값을 가지면, 이에 해당되는 변수를 제거한다.(3-2에서는

기준을 0.1로 설정하였다.) 이를, 모든 p-value가 0.1보다 작게 될 때까지 계속한다. 이를 backward elimination process라고 부른다. 이와 반대로 변수 추가가 필요없을 때 까지 유의한 독립변수를 하나씩 추가해 나가는 과정을 forward selection process라고 말한다. 위 두 방법을 조합하여 사용한다.

본 레포트에서는 편의상, backward elimination process만을 사용하였고, p-value가 가장 큰 gene을 지워나가는 방식으로 process를 적용하였다.

(1) 3-1의 top 10 gene 들을 colnames.data라는 vector에 저장한다. 그 후, multiple linear regression을 진행한 후, 각 gene의 p-value를 관찰한다.

R code :

```
colnames.data<-gene.correlation_sort_sort[c(1:10)]
data<-as.data.frame(t(expression_filter[colnames.data,]))
colnames(data)<-colnames.data
data$survival<-clinical_filter$survival_time
res.simple<-lm(survival~.,data=data)
summary(res.simple)
```

(2) 가장 높은 p-value를 가지는 gene을 확인하고, 모든 gene의 p-value가 0.1보다 작을 때까지 그 gene을 colnames.data에서 없앤 후 (1)의 과정을 반복한다. (첨부한 r script 참조)

2. Results

첫 번째 모델은 10개의 독립변수 모두를 사용한 모델이었으며(Table 1 참조), 그 중 가장 큰 p-value를 가진 gene은 c4orf7이었고, 0.78174의 p-value를 가졌다.

두 번째 모델은 c4orf7을 제외한 9개의 독립변수를 사용한 모델이었으며, 그 중 가장 큰 p-value를 가진 gene은 RIMS3이었고, 0.76697의 p-value를 가졌다.

세 번째 모델은 c4orf7, RIMS3을 제외한 8개의 독립변수를 사용한 모델이었으며, 그 중 가장 큰 p-value를 가진 gene은 STEAP3이었고, 0.618799의 p-value를 가졌다.

네 번째 모델은 c4orf7, RIMS3, STEPA3을 제외한 7개의 독립변수를 사용한 모델이었으며, 그 중 가장 큰 p-value를 가진 gene은 MALT1이었고, 0.409867의 p-value를 가졌다.

다섯 번째 모델은 c4orf7, RIMS3, STEPA3, MALT1을 제외한 6개의 독립변수를 사용한 모델이었으며, 그 중 가장 큰 p-value를 가진 gene은 KRT5이었고, 0.264202의 p-value를 가졌다.

여섯 번째 모델은 c4orf7, RIMS3, STEPA3, MALT1, KRT5을 제외한 5개의 독립변수를 사용한 모델이었으며, 그 중 가장 큰 p-value를 가진 gene은 RBP5이었고, 0.11280의 p-value를 가졌다.

최종 모델은 CDH3, LRIG2, NTSR2, FLJ40298 4개의 독립변수를 사용한 모델이었고, 그 결과값은 다음과 같다.

```
Call:
lm(formula = survival ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2143.6  -827.6  -294.5   569.9  6545.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1195.21     152.46   7.840 3.86e-14 ***
CDH3          158.25      37.67   4.201 3.26e-05 ***
LRIG2         367.31     142.37   2.580 0.010223 *
NTSR2         330.44      84.86   3.894 0.000115 ***
FLJ40298      231.27      65.31   3.541 0.000444 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1302 on 414 degrees of freedom
Multiple R-squared:  0.1123,    Adjusted R-squared:  0.1037
F-statistic: 13.09 on 4 and 414 DF,  p-value: 4.767e-10
```