

Module 8 Homework

Yanhe Wen

3/12/2017

Question 1

(a)

```
library(ALL)
data(ALL)
library(lmtest)
ALL_BB1234 <- ALL[,ALL$BT %in% c("B", "B1", "B2", "B3", "B4")]
y_109at <- exprs(ALL_BB1234)["109_at",]
anova(lm(y_109at~ALL_BB1234$BT))

## Analysis of Variance Table
##
## Response: y_109at
##          Df Sum Sq Mean Sq F value Pr(>F)
## ALL_BB1234$BT  4  2.1053  0.52632   3.4829 0.01082 *
## Residuals    90 13.6006  0.15112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

cat("p-value is less than 0.05, we reject null hypothesis, there is difference
    among different stages of patients, so the stages do affect the mean")

## p-value is less than 0.05, we reject null hypothesis, there is difference
##    among different stages of patients, so the stages do affect the mean
```

(b)

```
cat("B3 genen mean expression value is:", mean(exprs(ALL[,ALL$BT %in% c("B3")])["109_at",]))

## B3 genen mean expression value is: 6.685333
```

(c)

```
pairwise.t.test(y_109at, ALL_BB1234$BT)

##
## Pairwise comparisons using t tests with pooled SD
##
## data: y_109at and ALL_BB1234$BT
##
##      B      B1      B2      B3
## B1 1.00 -      -      -
## B2 0.52 1.00 -      -
## B3 1.00 1.00 0.37 -
```

```
## B4 1.00 0.20 0.01 0.61
##
## P value adjustment method: holm
cat("From the pairwise test, there is no difference between B and other groups!")

## From the pairwise test, there is no difference between B and other groups!
```

(d)

```
pairwise.t.test(y_109at, ALL_BB1234$BT, p.adjust.method = "fdr")

##
## Pairwise comparisons using t tests with pooled SD
##
## data: y_109at and ALL_BB1234$BT
##
##      B      B1      B2      B3
## B1 0.40 -      -      -
## B2 0.19 0.48 -      -
## B3 0.57 0.48 0.15 -
## B4 0.62 0.11 0.01 0.20
##
## P value adjustment method: fdr
cat("From the pairwise test with fdr, we still can't find difference
    between B and other groups!")

## From the pairwise test with fdr, we still can't find difference
##      between B and other groups!
```

(e)

```
shapiro.test(residuals(lm(y_109at~ALL_BB1234$BT)))

##
## Shapiro-Wilk normality test
##
## data: residuals(lm(y_109at ~ ALL_BB1234$BT))
## W = 0.97839, p-value = 0.1177
bptest(lm(y_109at~ALL_BB1234$BT))

##
## studentized Breusch-Pagan test
##
## data: lm(y_109at ~ ALL_BB1234$BT)
## BP = 1.4516, df = 4, p-value = 0.8352
cat("Both diagnostics tests show the p-values are greater than 0.05,
    so we don't need to apply robust ANOVA tests")

## Both diagnostics tests show the p-values are greater than 0.05,
##      so we don't need to apply robust ANOVA tests
```

Question 2

(a)

```
y_2 <- exprs(ALL_BB1234)
p_values <- apply(y_2, 1, function(x) kruskal.test(x~ALL_BB1234$BT)$p.value)
p_fdrs <- p.adjust(p_values, method = "fdr")
cat("There are",sum(p_fdrs<0.05),"genes expressed differently.")

## There are 423 genes expressed differently.
```

(b)

```
for(i in 1:5){
cat(rownames(y_2)[order(p_fdrs, decreasing = FALSE)][i],"\n")
}

## 1389_at
## 38555_at
## 40268_at
## 1866_g_at
## 40155_at
```

Question 3

(a)

```
ALLBg <- ALL[,which(ALL$BT %in% c("B1","B2","B3","B4") & ALL$sex %in% c("M","F"))]
y_3 <- exprs(ALLBg)["38555_at",]
Bcell<-ALLBg$BT
gender <- ALLBg$sex
anova(lm(y_3~Bcell*gender))

## Analysis of Variance Table
##
## Response: y_3
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Bcell       3 24.436   8.1453 19.1179 1.818e-09 ***
## gender       1  0.032   0.0319  0.0748  0.7851
## Bcell:gender  3  0.230   0.0768  0.1803  0.9095
## Residuals   81 34.511   0.4261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

cat("B stages do affect the expression values, but gender doesn't affect,
    B stages and genders combined factors doesn't affect expression values either")

## B stages do affect the expression values, but gender doesn't affect,
##      B stages and genders combined factors doesn't affect expression values either
anova(lm(y_3~Bcell+gender))

## Analysis of Variance Table
```

```
##
## Response: y_3
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Bcell      3 24.436   8.1453 19.6945 9.286e-10 ***
## gender     1  0.032   0.0319  0.0771    0.782
## Residuals 84 34.741   0.4136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

cat("From the test result, we conclude the same as above.")

## From the test result, we conclude the same as above.
summary(lm(y_3~Bcell+gender))

##
## Call:
## lm(formula = y_3 ~ Bcell + gender)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.0392 -0.4896 -0.0535  0.4686  1.6918
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.94064     0.16954  40.937 < 2e-16 ***
## BcellB2      -0.79014     0.18361  -4.303 4.52e-05 ***
## BcellB3      -1.42256     0.20071  -7.088 3.92e-10 ***
## BcellB4      -1.34133     0.23714  -5.656 2.09e-07 ***
## genderM       -0.04005     0.14428  -0.278    0.782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6431 on 84 degrees of freedom
## Multiple R-squared:  0.4132, Adjusted R-squared:  0.3853
## F-statistic: 14.79 on 4 and 84 DF,  p-value: 3.459e-09

cat("Here we can see that all the B stages are have extremely small p-values,
    but genders p-value is greater than 0.05. Therefore, we can agree the conclusion above.")

## Here we can see that all the B stages are have extremely small p-values,
##      but genders p-value is greater than 0.05. Therefore, we can agree the conclusion above.
```

(b)

```
shapiro.test(residuals(lm(y_3~Bcell*gender)))

##
## Shapiro-Wilk normality test
##
## data:  residuals(lm(y_3 ~ Bcell * gender))
## W = 0.96926, p-value = 0.03291

bptest(lm(y_3~Bcell*gender))

##
```

```
## studentized Breusch-Pagan test
##
## data:  lm(y_3 ~ Bcell * gender)
## BP = 8.1338, df = 7, p-value = 0.3209

cat("Shapiro test and bptest show the p-values are identical,
    so we agree with the assumption from part a")

## Shapiro test and bptest show the p-values are identical,
##    so we agree with the assumption from part a
```

Question 4

```
ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")]
y_4 <- exprs(ALLB123)["1242_at",]
group <- ALLB123$BT[,drop=T]
n <- length(y_4)
#T.obs <- anova(lm(y_4~group))$F[1]
T.obs <- max(by(y_4,group,mean)) - min(by(y_4,group,mean))
n.perm <- 2000
T.perm <- rep(NA, n.perm)
for(i in 1:n.perm) {
  y_4.perm = sample(y_4, n, replace=F)
  #T.perm[i] = anova(lm(y_4.perm~group))$F[1]
  T.perm[i] = max(by(y_4.perm,group,mean))-min(by(y_4.perm,group,mean))
}
mean(T.perm>=T.obs)

## [1] 0.4905

cat("p-value > 0.05, so there is no significant difference among 1242_at")

## p-value > 0.05, so there is no significant difference among 1242_at
```