**Module 8 Home Work**

**Problem 1: (40 points)**

On the ALL data set, consider the ANOVA on the gene with the probe “109\_at ”expression values on B-cell patients in 5 groups: B, B1, B2, B3 and B4.

(a)Conduct the one-way ANOVA. Do the disease stages affect the mean gene expression value?

(b)From the linear model fits, find the mean gene expression value among B3 patients.

(c) Which group’s mean gene expression value is different from that of group B?

(d)Use the pairwise comparisons at FDR=0.05 to find which group means are different. What is your conclusion?

(e) Check the ANOVA model assumptions with diagnostic tests? Do we need to apply robust ANOVA tests here? If yes, apply the appropriate tests and state your conclusion.

Answer the question in each part directly. Relevant R outputs should be displayed to support your conclusion.

**Answer:**

(a)

ALLB1234 <- ALL[,which(ALL$BT %in% c("B","B1","B2","B3","B4"))]

a1 <- exprs(ALLB1234)["109\_at",]

anova(lm(a1 ~ ALLB1234$BT))

**Output:**

p = 0.01082 .Here as the p value is less than 0.05 we can reject the null hypothesis. Therefore we can say that the disease stages affect the mean gene expression value.

(b)

ALLB1234 <- ALL[,which(ALL$BT %in% c("B","B1","B2","B3","B4"))]

a1 <- exprs(ALLB1234)["109\_at",]

summary(lm(a1 ~ ALLB1234$BT-1))

**Output:**

The mean gene expression value among B3 patients is 6.68533

(c)

**Output:**

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

Above output includes different p-values of all the groups. From the output it is evident that the values are greater than 0.05 hence we cannot reject null hypothesis. Therefore none of the groups(B1,B2,B3,B4) are different from the mean expression of group B.

(d)

ALLB1234 <- ALL[,ALL$BT%in%c("B","B1", "B2", "B3","B4")]

a1 <- exprs(ALLB1234)["109\_at",]

pairwise.t.test(a1,ALLB1234$BT,p.adjust.method='fdr')

Output:

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

None of the means of B1,B2,B3,B4 are different from B after fdr Adjustment.

Conclusion:

Mean expression of B2,B4 (p=0.01) are different as p<0.01.

(e)

ALLB1234 <- ALL[,ALL$BT%in%c("B","B1", "B2", "B3","B4")]

a1 <- exprs(ALLB1234)["109\_at",]

shapiro.test(residuals(lm(a1 ~ ALLB1234$BT)))

**Output:**

W = 0.97839, p-value = 0.1177

Here p=0.1177 which is >0.05 and we can accept the null hypothesis and that the data follows Normal distribution.

library(lmtest)

bptest(lm(a1 ~ ALLB1234$BT), studentize = FALSE)

**Output:**

BP = 1.1702, df = 4, p-value = 0.883

Here p=0.883 which is >0.05 and we can accept the null hypothesis of equal variances.

Both the tests are positive and above tests are appropriate. Hence we don’t need other robust tests.

**Problem 2: (20 points)**

Apply the nonparametric Kruskal-Wallis tests for every gene on the B-cell ALL patients in stage B, B1, B2, B3, B4 from the ALL data. (Hint: use the apply() function.)

(a) Use FDR adjustments at 0.05 level. How many genes are expressed differently in some of the groups?

(b) Find the probe names for the top five genes with smallest p-values.

Please submit your R commands together with your answers to each part of the question.

**Answer:**

(a)

ALLB1234 <- ALL[,which(ALL$BT %in% c("B","B1","B2","B3","B4"))]

ALLa2 <- exprs(ALLB1234)[,]

b1Data <- apply(ALLa2,1,function(x) kruskal.test(x ~ ALLB1234$BT)$p.value)

b1Data.fdr <- p.adjust(p=b1Data,method="fdr")

sum(b1Data.fdr<0.05)

**Output:**

By the output we can tell that 423 genes are expressed differently in some of the groups.

(b)

Allnames(sort(p.values)[1:5]

**Output:**

The Top five genes with small p-­‐values

[1] "1389\_at" "40268\_at" "38555\_at" "1866\_g\_at" "40155\_at"

**Problem 3: (20 points)**

On the ALL data sex, we consider the ANOVA on the gene with the probe “38555\_at” expression values on two factors. The first factor is the disease stages: B1, B2, B3 and B4 (we only take patients from those four stages). The second factor is the gender of the patient (stored in the variable ALL$sex).

(a)Conduct the appropriate ANOVA analysis. Does any of the two factors affects the gene expression values? Are there interaction between the two factors?

(b)Check the ANOVA model assumption with diagnostic tests? Are any of the assumptions violated?

Please submit your R commands together with your answers to each part of the question. Relevant R outputs should be displayed to support your conclusion.

**Answer:**

ALLd <- ALL[,which(ALL$BT %in% c("B1","B2","B3","B4") & ALL$sex %in% c("M", "F"))]

ALLD <- exprs(ALLdatasex)["38555\_at",]

Bcell<-ALLdatasex$BT

sexg<-ALLdatasex$sex

anova(lm(ALLdataSex~ Bcell\*sexg))

**Output:**

Response: ALLdataSex

Df Sum Sq Mean Sq F value Pr(>F)

Bcell 3 24.436 8.1453 19.1179 1.818e-09 \*\*\*

sexg 1 0.032 0.0319 0.0748 0.7851

Bcell:sexg 3 0.230 0.0768 0.1803 0.9095

Residuals 81 34.511 0.4261

For Bcell, p=1.818e-­‐09 which is <0.05 and can say that B group affects gene expression

For sexg ,p=0.7851 which is >0.05 and can say that it does not affect gene expression.

Since both their ratio p=0.9095 which is >0.05 and can say there is no significant statistical interaction.

(b)

shapiro.test(residuals(lm(ALLdataSex ~ Bcell\*sexg)))

**Output:**

Shapiro-Wilk normality test

data: residuals(lm(ALLdataSex ~ Bcell \* sexg))

W = 0.96926, p-value = 0.03291

Here p=0.03291 which is <0.05, we can reject null hypothesis and normal distribution assumption is violated.

**Breusch-­‐Pagan test:**

bptest(lm(ALL ~ Bcell+sexg), studentize = FALSE)

**Output:**

Breusch-Pagan test

data: lm(ALLdataSex ~ Bcell + sexg)

BP = 4.5761, df = 4, p-value = 0.3336

Here p=0.3336 which is >0.05, We can accept null hypothesis of equal variances and the normal distribution assumption is violated.

**Problem 4: (20 points):**

We wish to conduct a permutation test for ANOVA on (y1….,yn), with the group identifiers stored in the vector ‘group’. We wish to use !!!! ( !!!! j –𝜇)2 as the test statistic. Here 𝜇j is the j-th group sample mean, and 𝜇 =!! 𝜇 !!!! j.

(a)Program this permutation test in R.

(b)Run this permutation test on the Ets2 repressor gene 1242\_at on the patients in stage B1, B2, and B3 from the ALL data set.

Submit your R script for (a) and the answer and R outputs for (b).

Hint: the sample group means can be found by R command by(y,group,mean).

Answer:

(a)

stg <- #Example: c("B1","B2")

gen <- #Example: "109\_at"

Stats <- function(stages,gene){

ALLB <- ALL[,which(ALL$BT %in% stages)]

data <- exprs(ALLB)[gene,]

group <- ALLB$BT[,drop=T]

g <- length(stages)

Means <- summary(lm(data ~ group-1))[["coefficients"]][1:g]

TotalMean <- (1/g)\*sum(Means)

Uj\_U <- NULL

for (i in 1:g){

Uj\_U[i] <- (Means[i]-TotalMean)^2

}

T.obs <- (1/(g-1))\*sum(Uj\_U) #Observed statistic

n <- length(data)

n.perm = 2000

T.perm = NULL

for(i in 1:n.perm) {

data.perm = sample(data, n, replace=F)

MeansPerm <- summary(lm(data.perm ~ ALLB$BT-1))[["coefficients"]][1:g]

TotalMeanPerm <- (1/g)\*sum(MeansPerm)

Uj\_U1 <- NULL

for (k in 1:g){

Uj\_U1[k] <- (MeansPerm[k]-TotalMeanPerm)^2

}

T.perm[i] = (1/(g-1))\*sum(Uj\_U1) #Permuted statistic

}

mean(T.perm>=T.obs) #p-value

}

(b)

stg <- c("B1","B2","B3")

gen <- "1242\_at"

Stats(stg,gen)

**Output:**

[1] 0.526

Here p=0.526 which is >0.05, so we fail to reject the null hypothesis Hence the (nonparametric) permutation ANOVA test shows no such evidence for B1,B2 and B3 patients to have different mean gene 1242\_at expression values.