

Soution:

1a)

Analysis of Variance Table

Response: y

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

ALLB1234$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

As the value of p = 0.01082 is small, we reject the null-hypothesis and conclude that these disease stages affect the mean gene expression value.

1b)

Call:

lm(formula = y ~ ALLB1234$BT - 1)

Residuals:

Min 1Q Median 3Q Max

-1.09026 -0.27845 0.03999 0.26618 0.71532

Coefficients:

Estimate Std. Error t value Pr(>|t|)

ALLB1234$BTB 6.81021 0.17385 39.17 <2e-16 \*\*\*

ALLB1234$BTB1 6.57951 0.08918 73.78 <2e-16 \*\*\*

ALLB1234$BTB2 6.47503 0.06479 99.94 <2e-16 \*\*\*

ALLB1234$BTB3 6.68533 0.08106 82.48 <2e-16 \*\*\*

ALLB1234$BTB4 6.91417 0.11222 61.61 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3887 on 90 degrees of freedom

Multiple R-squared: 0.9967, Adjusted R-squared: 0.9966

F-statistic: 5513 on 5 and 90 DF, p-value: < 2.2e-16

From the table, we can see that the mean gene expression value among B3 patient is 6.68533

1c)

Pairwise comparisons using t tests with pooled SD

data: y and ALLB1234$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

Comparing the p-values from the table above, we can see that all are more than 0.05, which means the expressions are equal, thus we cannot reject the null hypothesis. So we can say that none of B1, B2, B3, B4 is different from mean expression of group B

1d)

Pairwise comparisons using t tests with pooled SD

data: y and ALLB1234$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

Even after using FDR adjustment, we can see that mean expression of B is not different from that of B1, B2, B3, B4, as the p-values are more that 0.05.

Mean expression of B2, B4 are different as p< 0.01

1e)

> shapiro.test(residuals(lm(y ~ ALLB1234$BT)))

Shapiro-Wilk normality test

data: residuals(lm(y ~ ALLB1234$BT))

W = 0.9784, p-value = 0.1177

Using shapiro test, we get p > 0.05, we accept the null hypothesis that the data follows Normal distribution.

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

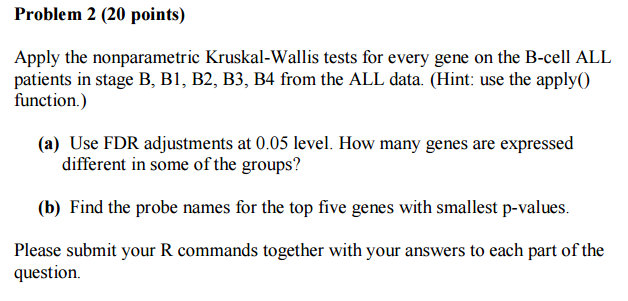
Breusch-Pagan test

data: lm(y ~ ALLB1234$BT)

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

Using Breusch pagan test, we get p>0.05, we accept the nnull hypothesis of equal variances.

As both the test for ANOVA assumptions of mean and variances were positive, we do not need to do the ROBUST TEST



Solutions:

2a)

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

> data.ALL <- exprs(ALLB1234)[,]

> ALLData <- apply(data.ALL,1,function(x) kruskal.test(x ~ ALLB1234$BT)$p.value)

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

> sum(ALLData.fdr<0.05)

[1] 423

At 0.05 level FDR adjustment, 423 genes have their p values less than 0.05. Hence we reject the null hypothesis of equal distribution for them and consider that the expression is different in them.

2b)

> order.ALLfdr <- order(ALLData.fdr, decreasing=FALSE)

> k=1

> names = NULL

> for (i in order.ALLfdr[1:5]){names[k] <- names(ALLData.fdr[i])

+ k=k+1}

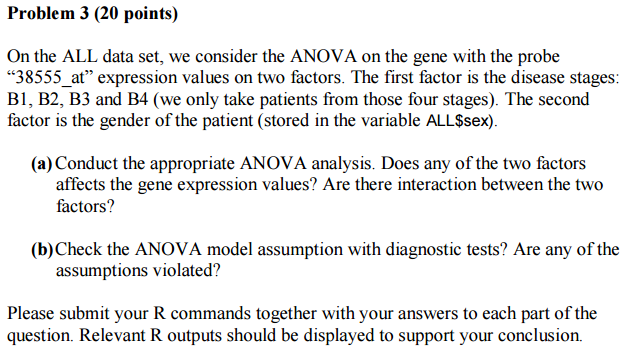
> print('Top five genes with smallest p-values =')

[1] "Top five genes with smallest p-values ="

> print(names)

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

The top five genes with smallest p-values is : 1389\_at , 38555\_at, 40268\_at, 1866\_g\_at, 40155\_at



Solutions:

3a)

Analysis of Variance Table

Response: ALL.dataSex

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

B.cell 3 24.436 8.1453 19.1179 1.818e-09 \*\*\*

ALL.sex 1 0.032 0.0319 0.0748 0.7851

B.cell:ALL.sex 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

As the p-value is less than 0.05 for Bcell, we can say that B group affects gene expression

As the p value for ALL.sex is more that 0.05, we say that gender does not effect gene expression.

For B.cell:ALL.sex has a p-value that is more than 0.05, we conclude that there is no statistically significant interaction

3b)

Shapiro-Wilk normality test

data: residuals(lm(ALL.dataSex ~ B.cell + ALL.sex))

W = 0.971, p-value = 0.04335

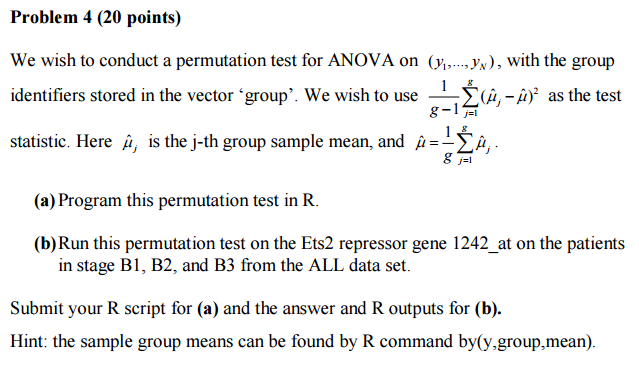
For shapiro-wilk normality test, the p-value is less than 0.05, we can reject the null hypothesis. Thus the normal distribution assumption is violated.

Breusch-Pagan test

data: lm(ALL.dataSex ~ B.cell + ALL.sex)

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

For Breusch-pagan we accept the null hypothesis of equal variances as the p-value is greter than 0.05. Thus the normal distribution assumption is not violated



Solutions:

4a) the R code for the problem is given below

> stages <- #Example: c("B1","B2","B3")

+ gene <- #Example: "109\_at"

+ Statistics <- function(stages,gene){

+ ALL.B <- ALL[,which(ALL$BT %in% stages)]

+ data <- exprs(ALL.B)[gene,]

+ group <- ALL.B$BT[,drop=T]

+ g <- length(stages)

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

+ Total.Mean <- (1/g)\*sum(Means)

+ MUj\_MU <- NULL

+ for (i in 1:g){

+ MUj\_MU[i] <- (Means[i]-Total.Mean)^2

+ }

+ T.obs <- (1/(g-1))\*sum(MUj\_MU) #Observed statistic

+ n <- length(data)

+ n.perm = 2000

+ T.perm = NULL

+ for(i in 1:n.perm) {

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

+ Means.Perm <- summary(lm(data.perm ~ ALL.B$BT-1))[["coefficients"]][1:g]

+ Total.MeanPerm <- (1/g)\*sum(Means.Perm)

+ MUj\_MU1 <- NULL

+ for (k in 1:g){

+ MUj\_MU1[k] <- (Means.Perm[k]-Total.MeanPerm)^2

+ }

+ T.perm[i] = (1/(g-1))\*sum(MUj\_MU1) #Permuted statistic

+ }

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

+ }

4b)

> stages <- c("B1","B2","B3")

> gene <- "1242\_at"

> Statistics(stages,gene)

[1] 0.5425

As the p-value is greater than 0.05, we accept the null hypothesis, of equal distribution of expression values.