Erika Maldonado-Rosado

15 April 2020

BINF Lab 10

```{r}

#1

setwd("~/Desktop")

myT<-read.table("nc101\_scaff\_dataCounts.txt",sep="\t",header=TRUE,row.names=1)

# remove rare genes

myT <- myT[ apply( myT,1, median)> 5,]

myTNorm <- myT

for ( i in 1:ncol(myT))

{

colSum = sum(myT[,i])

myTNorm[,i] =myTNorm[,i]/colSum

}

```

```{r}

#(A) For each row in the spreadsheet, perform a one-way ANOVA with categories “day 2”, “week 12” and “week 18”. Plot out the histogram of all p-values. How many genes are significant at a BH FDR-corrected 0.05 threshold. (see mini-lecture 16B).

PVals <- vector()

for( i in 1:nrow(myTNorm))

{

vals1 <- as.numeric( myTNorm[i,1:3])

vals2 <- as.numeric( myTNorm[i,4:6])

vals3 <- as.numeric(myTNorm[i,7:11])

xvals <-c(vals1,vals2,vals3)

genevals <- factor(c(rep("vals1",length(vals1)),rep("vals2",length(vals2)),rep("vals3",length(vals3))))

geneanova <- lm(xvals~genevals, x = TRUE)

PVals[i] = anova(geneanova)$"Pr(>F)"

}

hist(PVals, main="PValues", breaks = 50, col = "pink")

```

A close up of a logo

Description automatically generated

```{r}

AdjustedPVals <- p.adjust(PVals, method = "BH")

sum(AdjustedPVals <= .05)

```

[1] 612

```{r}

#(B) Next make an ANOVA as a linear regression as a function of time (so 2 days, 86 days and 128 days). Plot out the histogram of all p-values. How many genes are significant at a BH FDR-corrected 0.05 threshold. (see lecture 15)

#Anu and I worked together on the homework

#Jon & Dr. Fodor helped me during tutoring/email by looking at my code. Thank you.

DaysPvals <- vector()

for( i in 1:nrow(myTNorm))

{

Day2 <- as.numeric( myTNorm[i,1:3])

Day86 <- as.numeric( myTNorm[i,4:6])

Day128 <- as.numeric(myTNorm[i,7:11])

Days <-c(Day2,Day86,Day128)

DayLength <- c(rep(2,length(Day2)),rep(86,length(Day86)),rep(128,length(Day128)))

DayTotal <- lm(Days ~ DayLength, x = TRUE)

DaysPvals[i] = anova(DayTotal)$"Pr(>F)"[1]

}

hist(DaysPvals, main="P-Values For Different Days", col = "navy", breaks = 50)

AdjustedPValsDays <- p.adjust(DaysPvals, method = "BH")

sum(AdjustedPValsDays <= .05)

summary(DayTotal)

```

A screenshot of a cell phone

Description automatically generated

#(C) Finally, for each row in the spreadsheet perform an ANVOA comparing the three-parameter model from (A) and the two parameter model from (B). (see mini-lecture 16C). Plot out the histogram of all p-values. For how many genes is there a significant difference between these two models at a BH FDR-corrected threshold.

#Anu and I worked together

PValsOneWay <- vector()

RegPVals <- vector()

ModelPvals <- vector()

Index<- vector()

for( i in 1:nrow(myTNorm))

{

Index[i] <- i

Day2 <- as.numeric( myTNorm[i,1:3])

Week12 <- as.numeric( myTNorm[i,4:6])

Week20 <- as.numeric(myTNorm[i,7:11])

df\_Days <- c(Day2,Week12,Week20)

Types <- factor(c(rep("Day2",length(Day2)),rep("Week12",length(Week12)),rep("Week20",length(Week20))))

Length <- c(rep(3,length(Day2)),rep(86,length(Week12)),rep(128,length(Week20)))

OneWay <- lm(df\_Days ~ Types)

Reg <- lm(df\_Days ~ Length)

PValsOneWay[i] <- anova(OneWay)$"Pr(>F)"

RegPVals[i] <- anova(Reg)$"Pr(>F)"[1]

#Anu helped me with this section

oneway\_res <- sum(residuals(OneWay)^2)

Reg\_res <- sum(residuals(Reg)^2)

difference <- ((Reg\_res - oneway\_res)/2)/(oneway\_res/8)

ModelPvals[i] <- pf(difference, 2, 8)

}

```

```{r}

hist(ModelPvals, main="P-Values of Three and Two Parameter Models", col = "maroon", xlab = "P-Values For The Different Models", breaks = 50)

```

A picture containing pencil

Description automatically generated

```{r}

ForPvals2 <- p.adjust(ModelPvals, method = "BH")

sum(ForPvals2 <= .05)

```

[1] 130

```{r}

# (D)Make three graphs showing the relative abundance of the most significant gene under each of the three ANOVA models. For (A) and (C), the x-axis will the category (day 3, week 12 and week 18) and the y-axis will be the relative abundance. Be sure to properly label and title all graphs and axes. For (B) the x-axis will be time (in days) and the y-axis will be the relative abundance. For the graph of the top hit from (B), include the regression line for the plot from (B).

pValuesOneWayAnova <- vector()

pValuesRegression <- vector()

pValueModelDiff <- vector()

index <- vector()

cats <- factor( c( rep("day2",3),rep("week12",3),rep("week20",5) ))

for( i in 1:nrow(myTNorm))

{

index[i] <- i

#Columns that we need to take

Day2 <- as.numeric( myTNorm[i,1:3])

Week12 <- as.numeric( myTNorm[i,4:6])

Week20 <- as.numeric(myTNorm[i,7:11])

df\_Days <- c(Day2,Week12,Week20)

Length <- c(rep(2,length(Day2)),rep(86,length(Day86)),rep(128,length(Day128)))

#LM of one way and regression

OneWay\_2 <- lm(df\_Days ~ cats)

Reg\_2 <- lm(df\_Days ~ Length, x = TRUE)

pValuesOneWayAnova[i] <- anova(OneWay\_2)$"Pr(>F)"[1]

pValuesRegression[i] <- anova(Reg\_2)$"Pr(>F)"[1]

#Start of Comparison of Models

OneWay\_Res\_2 <- sum(residuals(OneWay\_2)^2)

Reg\_res\_2 <- sum(residuals(Reg\_2)^2)

difference <- ((Reg\_res\_2 - OneWay\_Res\_2)/2)/(OneWay\_Res\_2/8)

pValueModelDiff[i] <- pf(difference, 2, 8,lower.tail = F)

}

```

```{r}

myFrame <- data.frame( index, pValuesOneWayAnova,pValuesRegression,pValueModelDiff)

myFrame <- myFrame[ order(myFrame$pValuesOneWayAnova), ]

boxplot( as.numeric( myTNorm[ myFrame$index[1],]) ~ cats, xlab = "Days and Weeks", ylab = "Abundance", main = "One-Way Anova Model")

```

A screenshot of a cell phone

Description automatically generated

```{r}

myFrame <- data.frame( index, pValuesOneWayAnova,pValuesRegression,pValueModelDiff)

myFrame <- myFrame[ order(myFrame$pValuesRegression), ]

boxplot( as.numeric( myTNorm[ myFrame$index[1],]) ~ Length,xlab = "Days", ylab = "Relative Abundance", main = "Linear Regression Model" )

abline(Reg\_2, col = "red")

```A screenshot of a cell phone

Description automatically generated

```{r}

myFrame <- data.frame( index, pValuesOneWayAnova,pValuesRegression,pValueModelDiff)

myFrame <- myFrame[ order(myFrame$pValueModelDiff), ]

boxplot( as.numeric( myTNorm[ myFrame$index[1],]) ~ cats,xlab = "Day and Weeks", ylab = "Abundance", main = "Model Differences" )

```

A screenshot of a cell phone

Description automatically generated

(E) Overall, do you think the three-parameter model in (A) or the two-parameter model in (B) is more appropriate for these data? Justify your answer.

The three parameter is the best fit. Most of the time when you look at the difference between the model you do not really see any difference (Lecture 16C). However, for a small part of the genes in this dataset the three parameter is a better model. Also, when we looked at the p-value that was corrected with the "BH," at a 95% interval in the two-parameter model only 448 of the genes were a significant hit. While in the three-parameter model there was 612.