Lab #6

Due April 14th (2 week lab); send what you have to [afodor@uncc.edu](mailto:afodor@uncc.edu) with “Lab #6” in the subject line. As usual, show all of your code.

1. We again return to our RNA seq dataset of E. Coli genes from mice.

The URL is here:

http://afodor.github.io/classes/stats2015/longitdunalRNASeqData.zip

As before, read and normalize the counts table ( “nc101\_scaff\_dataCounts.txt “ into R). For example:

setwd("C:\\somewhere")

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

}

(The first 3 columns are “day 2”, the next 3 columns are “week 12” and the last 5 are “week 18” (even though they say w20)).

1. 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.
2. 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.
3. Finally, for each row in the spreadsheet perform an ANVOA comparing the three-parameter model from (A) and the two parameter model from (B). 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.
4. 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).
5. 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.

HINTS:

In your for loop, get each row of data and cast it type of numeric…

-----------------

for( i in 1:nrow(myTNorm))

{

myData <- as.numeric( myTNorm[i,] )

## build your linear models with myData as the y-variable

}

Don’t forget that if myLm is a linear model, you can get the p-value with

anova(myLm)$ "Pr(>F)"[1]

(but for question (C) you will need to calculate the p-value with pf )

To make a box-plot for the most significant hits, you can keep track of the row-index to go along with each p-value

So if you set up your for loop like this…

pValuesOneWayAnova <- vector()

pValuesRegression <- vector()

pValueModelDiff <- vector()

index <- vector()

cats <- factor( c( rep("day3",3),rep("week12",3),rep("week20",5) ))

for( i in 1:nrow(myTNorm))

{

index[i] <- i

# populate your p-values

}

Then once you have your p-values, you can make a data-frame, order it so the smallest p-value is on top and generate your box-plots like for example…

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

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

boxplot( as.numeric( myTNorm[ myFrame$index[1],]) ~ cats )

That will generate the boxplot for the most significant hit under the one-way ANOVA model (and you can follow a similar logic for the other two models).