Lab # 11:

Please e-mail code, graphs and answers to questions to [afodor@uncc.edu](mailto:afodor@uncc.edu) with the subject Lab #11.

Please have lab submitted (whatever you have) before class on Wed, April 19.

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

This week's dataset is the case/control colorectal adenomas data set, but this time starting with raw counts at the family level:

http://afodor.github.io/classes/stats2015/familyPivotedTaxaAsColumnsNotNormalized.txt

(1) Normalize the dataset so that the number of counts is (approximately) the same in each sample.

For example:

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

rowSums <- apply( myT, 1, sum)

avgPerSample <- mean(rowSums)

for( i in 1:nrow(myT))

{

rowSum = sum(myT[i,])

myT[i,] = avgPerSample \* myT[i,] / rowSum

}

myT <- round(myT)

(2) Consider only the bugs (columns) that have at least one zero count:

(A) Plot the log(p-values) for each bug from a null hypothesis that case/control is from the same distribution. On one axis of your plot put p-vales from a simple linear model (in which you have log transformed the data with for example log(myT+1)) and on the other axis put the p-values from the zero inflated negative binomial (ZINB) model in which both the binomial and negative binomial parts have a term for case/control.(You can get the p-values for the ZINB by "lrtest" in package "lmtest”)

Do the p-values from the two models seem broadly similar? At a 10% false discovery rate, which model produces more significantly different hits?

(B) Plot the AIC values for the two models built for each of the bugs. For most of the bugs, which model has the lowest AIC?

(C) For each microbe significant by either model, make a boxplot comparing case and control. You can collect these in a PDF (no need to send me the PDF!). There is no single “right” answer to this question, but do you have a sense looking at the data as to which model (linear vs. mixed) might be more reliable? Or do both models seem like they are working equally well?

(see next page for graph hints)

You can make boxplots for each taxa (including plotting all the datapoints) with code something like this…

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

someCol <- 33

caseControl <- ifelse( grepl("case", row.names(myT)), "case", "control")

bug <- myT[,someCol]

boxplot( bug ~ caseControl,main=paste(names(myT)[someCol],"p-values or other info here"))

myFrame <- data.frame(bug, caseControl)

stripchart(bug~ caseControl, data = myFrame,vertical = TRUE, pch = 21, add=TRUE )

You can open up a pdf like this…

pdf("someFile.pdf")

Putting the above code in a for loop across every column that you want to graph will make a nice file that will visualize your whole dataset.

You do need to tell R when you are done writing to your PDF, which you can do with:

dev.off()

(If you call par(mfrow=c(2,2)) you can make your boxplots 4 to a page..)