Lab #7:

Dates: Last day of class. **May 3rd.** This lab (#7) is due.

Final exam. **May 12th;** Thursday May 12th from 11:00 am to 1:30 pm . In the 1st floor classroom. Bioinformatics 104.

Final project also due **May 12th.**

You will have the final study guide before lab on April 28th.

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

Please e-mail code, graphs and answers to questions to [afodor@uncc.edu](mailto:afodor@uncc.edu). Please put lab #7 in the subject line.

This week’s dataset is here:

<http://afodor.github.io/classes/stats2015/prePostPhylum.txt>

(This dataset is described, albeit from a different analysis pipeline, in these papers:

<http://www.sciencemag.org/content/sci/338/6103/120.full.html>

and

<http://www.nature.com/ncomms/2014/140903/ncomms5724/full/ncomms5724.html>

Note that WT and IL10-/- animals are in different cages. So “Cage1\_WT” is a different cage from “Cage1\_10-/-“.

1. Download the dataset. Perform PCA ordination.

(For example:

rm(list=ls())

setwd("C:\\Users\\afodor\\git\\afodor.github.io\\classes\\stats2015\\")

inFileName <- paste("prePostPhylum.txt", sep ="")

myT <-read.table(inFileName,header=TRUE,sep="\t")

numCols <- ncol(myT)

myColClasses <- c(rep("character",4), rep("numeric", numCols-4))

myT <-read.table(inFileName,header=TRUE,sep="\t",colClasses=myColClasses)

myTData<-myT[,5:10]

myPCOA <- princomp(myTData)

1. Graph PCA1 vs. PCA2. Make three versions of the graph. One colored by genotype,

one colored by cage and one colored by timepoint (pre-vs-post)

(SEE NEXT PAGE)

1. Fill in the following table for p-values testing the null hypothesis for PCA 1 and 2. For cage, use a way one-ANOVA. For genotype and timepoint (“pre” vs “post”) use a t-test

|  |  |  |
| --- | --- | --- |
|  | PCA1 | PCA2 |
| Cage |  |  |
| Genotype |  |  |
| Time (pre vs. post) |  |  |

Which variable seems to be most associated with the first PCA axis? Which variable is most associated with the second PCA axis? Does cage seem to be having an effect on these data?

(4)For the POST timepoints only:

1. For each phyla, graph the relative abundance of that phyla vs. cage. Does there appear to be a cage effect across different phyla?
2. For each phyla build a mixed linear model with genotype as the fixed variable and cage as a random variable. Report the intraclass correlation coefficient for each phyla. Are there any phyla that are significantly different for genotype in the mixed model at a 10% false discovery rate?

Hints:

1. If you use par(mfrow=c(3,2)) you can fit all 6 plots for phyla vs. cage on one graph. You can put the p-values and intraclass correlation coefficient in the “main” text above each graph to make a nice summary figure.
2. It can be useful to make a dataframe with just the data you want before building your model. So if you are looping through columns in a “myT” that you’ve read with read.table and i is your column index..

myT <- myT[myT$time == "POST",]

bug <- myT[,i]

cage <- myT$cage

genotype <- myT$genotype

myFrame <- data.frame(bug, cage, genotype)

(and then build your models with data=myFrame…)

1. Getting a p-value out of the mixed linear model could be done with something like:

unclass(summary(M.mixed))$tTable[2,5]

Getting the rho(intraclass correlation coefficient) out of a GLS model can be done with:

coef(M.gls$modelStruct[1]$corStruct,unconstrained=FALSE)[[1]]

(4) You can have both points and boxplots on a scatter graph with something like:

boxplot(myFrame$bug ~myFrame$cage)

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