Lab 10 | Jon Lee

Code

Lab 10 | BINF 6310 | Spring 2020 | Jon Lee

By the beginning of the next lab (April 9th), send what you have to afodor@uncc.edu (mailto:afodor@uncc.edu) with “Lab #10” 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 (http://afodor.github.io/classes/stats2015/longitdunalRNASeqData.zip)

As before, read and normalize the counts table (“nc101_scaff_dataCounts.txt “ into R). For example:

myT<-read.table(“nc101_scaff_dataCounts.txt”,sep=",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”).

Hide

```
countData <- read.table("nc101_scaff_dataCounts.txt", sep = "\t", header = TRUE, row.nam
es = 1)

#remove rare genes
countData <- countData[apply(countData, 1, median)> 5,]

countNorm <- countData
for (i in 1:ncol(countData))
{
  colSum = sum(countData[,i])
  countNorm[,i] = countNorm[,i]/colSum
}
```

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).

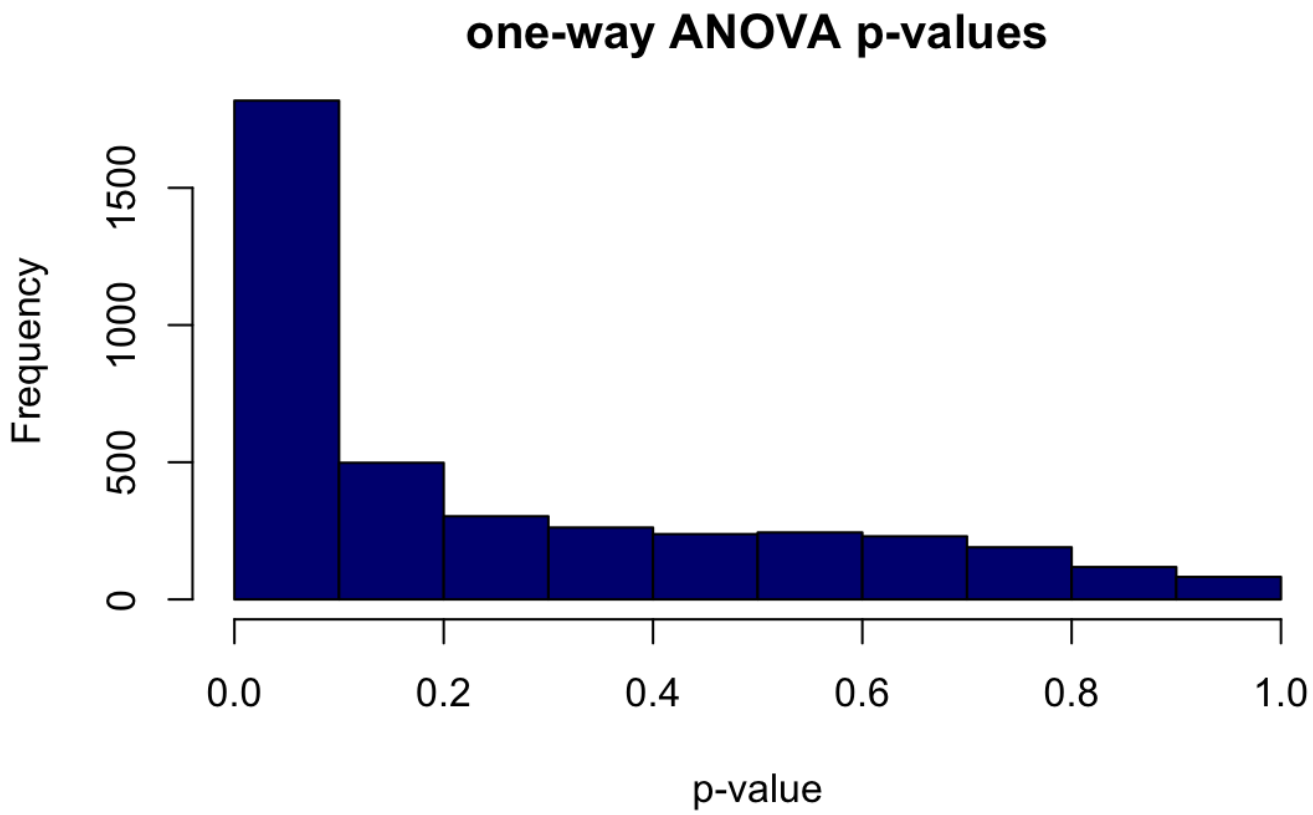
Hide

```
#one-way ANOVA
pValA <- vector()
for(i in 1:nrow(countNorm))
{
  countRow <- as.numeric(countNorm[i,])
  timePoint <- c(rep("day2", 3), rep("week12", 3), rep("week18", 5))
  countLM <- lm(countRow ~ timePoint, x = TRUE)
  pValA[i] <- anova(countLM)$"Pr(>F)"[1]
}

#number of significant genes at a BH FDR-corrected 0.05 threshold
pValA_BH <- p.adjust(pValA, method = "BH")
print(sum(pValA_BH < 0.05))
```

```
[1] 612
```

Hide

```
#histogram of one-way ANOVA p-values
hist(pValA, xlab = "p-value", main = "one-way ANOVA p-values", col = "navy")
```



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)

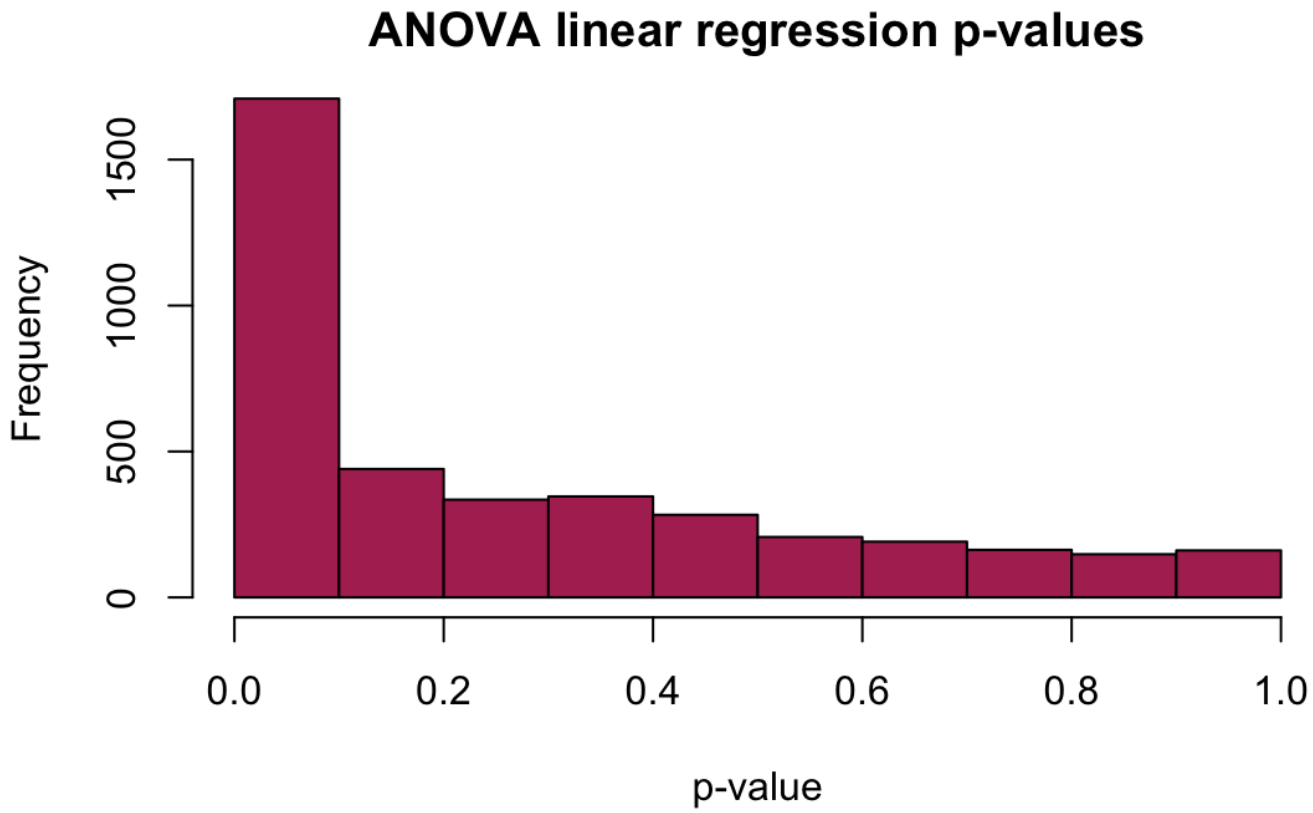
Hide

```
#ANOVA linear regression as a function of time
pValB <- vector()
for(i in 1:nrow(countNorm))
{
  countRow <- as.numeric(countNorm[i,])
  timePoint <- c(rep(2, 3), rep(86, 3), rep(128, 5))
  countLM <- lm(countRow ~ timePoint, x = TRUE)
  pValB[i] <- anova(countLM)$"Pr(>F)"[1]
}

#number of significant genes at a BH FDR-corrected 0.05 threshold
pValB_BH <- p.adjust(pValB, method = "BH")
print(sum(pValB_BH < 0.05))
```

```
[1] 448
```

Hide

```
#histogram of ANOVA linear regression p-values
hist(pValB, xlab = "p-value", main = "ANOVA linear regression p-values", col = "maroon")
```



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.

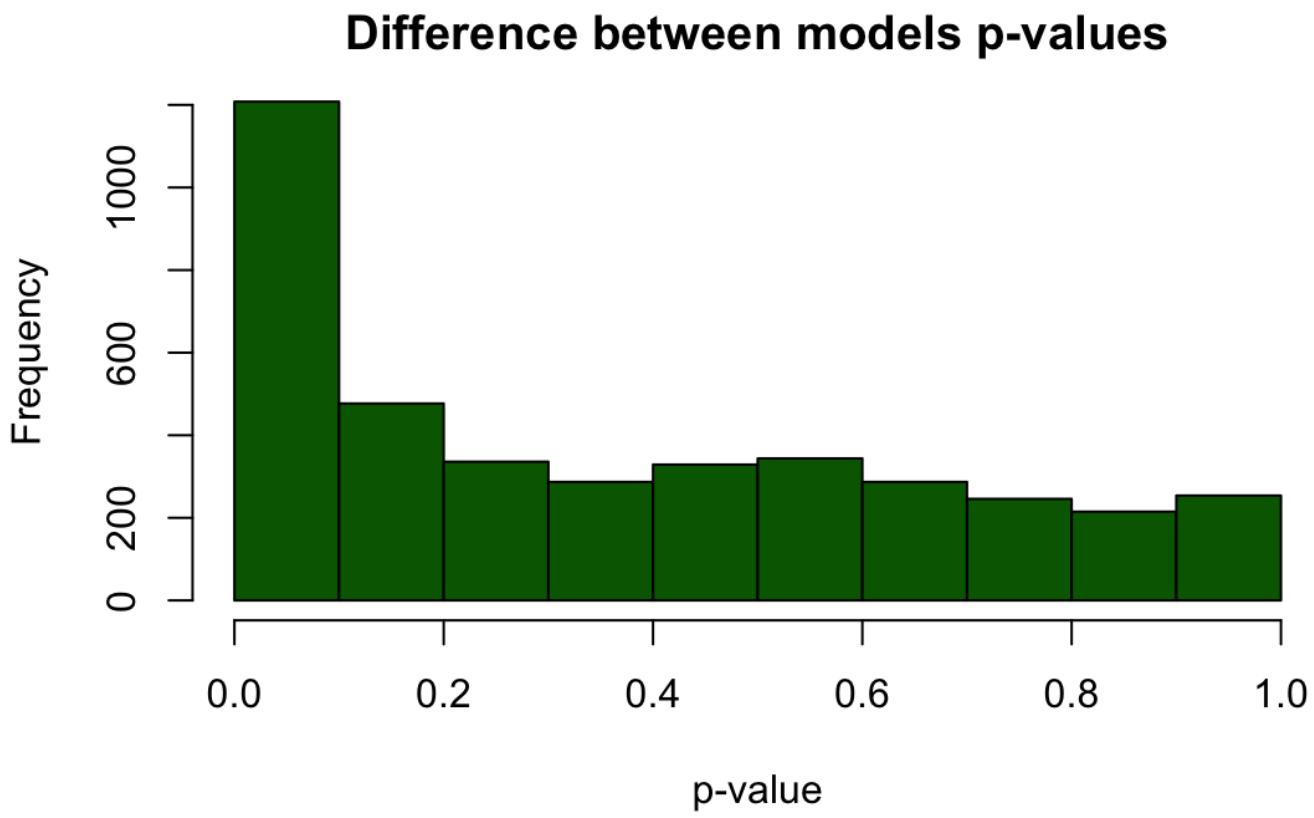
Hide

```
#ANOVA comparing 3 and 2 parameter models
pValANOVA <- vector()
pValLR <- vector()
pValDiff <- vector()
index <- vector()

for(i in 1:nrow(countNorm))
{
    index[i] <- i
    countRow <- as.numeric(countNorm[i,])

    #populate p-values one-way ANOVA
  timeANOVA <- c(rep("day2", 3), rep("week12", 3), rep("week18", 5))
  modelANOVA <- lm(countRow ~ timeANOVA, x = TRUE)
  pValANOVA[i] <- anova(modelANOVA)$"Pr(>F)"[1]

  #populate p-values ANOVA LR
  timeLR <- c(rep(2, 3), rep(86, 3), rep(128, 5))
  modelLR <- lm(countRow ~ timeLR, x = TRUE)
  pValLR[i] <- anova(modelLR)$"Pr(>F)"[1]

  #populate modle difference p-value with 9 full degrees of fredom
  fullErr <- sum(residuals(modelANOVA)^2)
  reducedErr <-sum(residuals(modelLR)^2)
  Fstat <- ((reducedErr - fullErr)/1)/(fullErr/9)
  pValDiff[i] <- pf(Fstat, 1, 9, lower.tail = FALSE)
}

#number of significant difference between genes at a BH FDR-corrected 0.05 threshold
pValDiff_BH <- p.adjust(pValDiff, method = "BH")
print(sum(pValDiff_BH < 0.05))
```

```
[1] 136
```

Hide

```
#histogram of ANOVA linear regression p-values
hist(pValDiff, xlab = "p-value", main = "Difference between models p-values", col = "dar
k green")
```



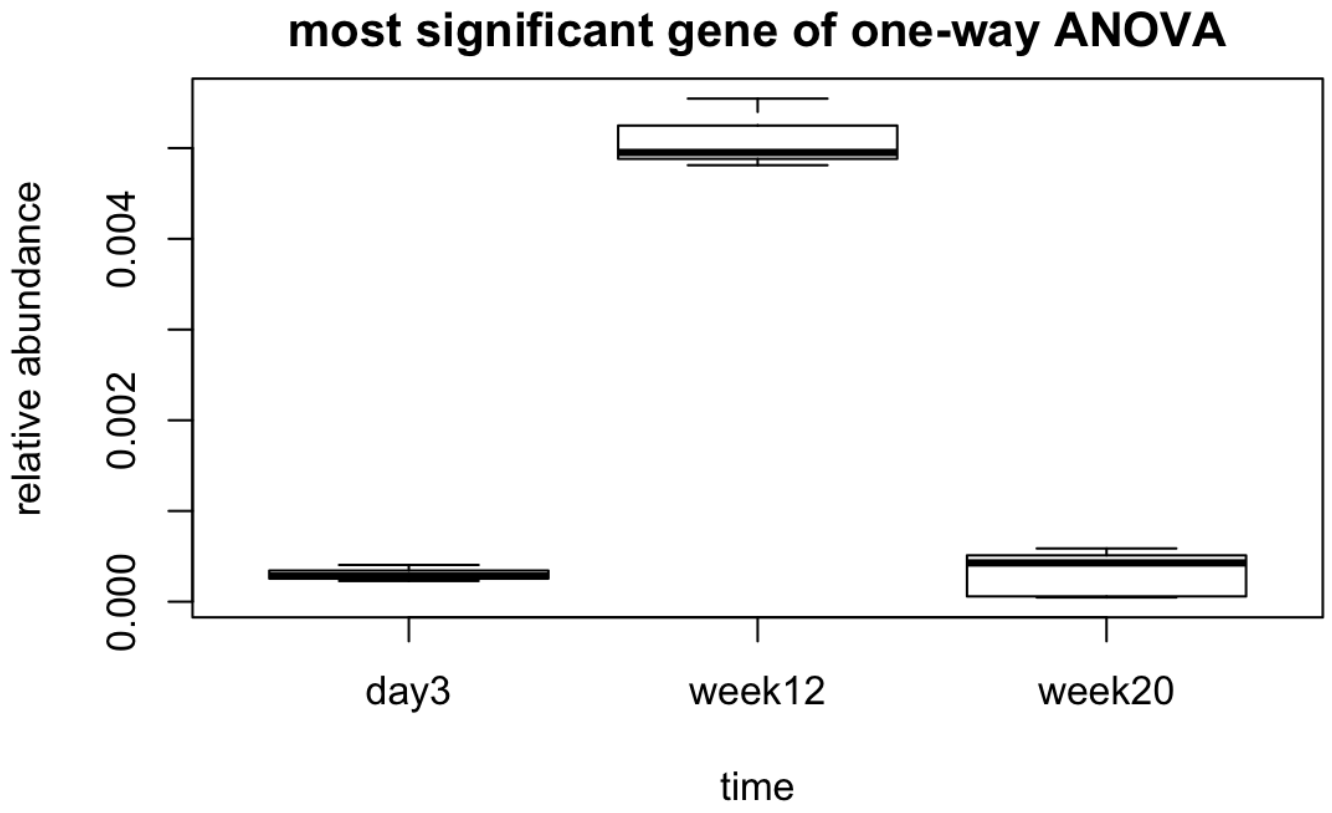
D. Make three graphs showing the relative abundance of the most significant gene under each of the three ANOVA models. For (A) and (C), the x-axis will the category (day 3, week 12 and week 18) and the y-axis will be the relative abundance. Be sure to properly label and title all graphs and axes. For (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).

Hide

```
#x-axis labels for abundacne graphs (boxplots)
categoriesAC <- factor(c(rep("day3", 3), rep("week12", 3), rep("week20", 5)))
categoriesB <- factor(c(rep(2, 3), rep(86, 3), rep(128, 5)))

#abundance data frame
abundFrame <- data.frame(index, pValANOVA, pValLR, pValDiff)
```

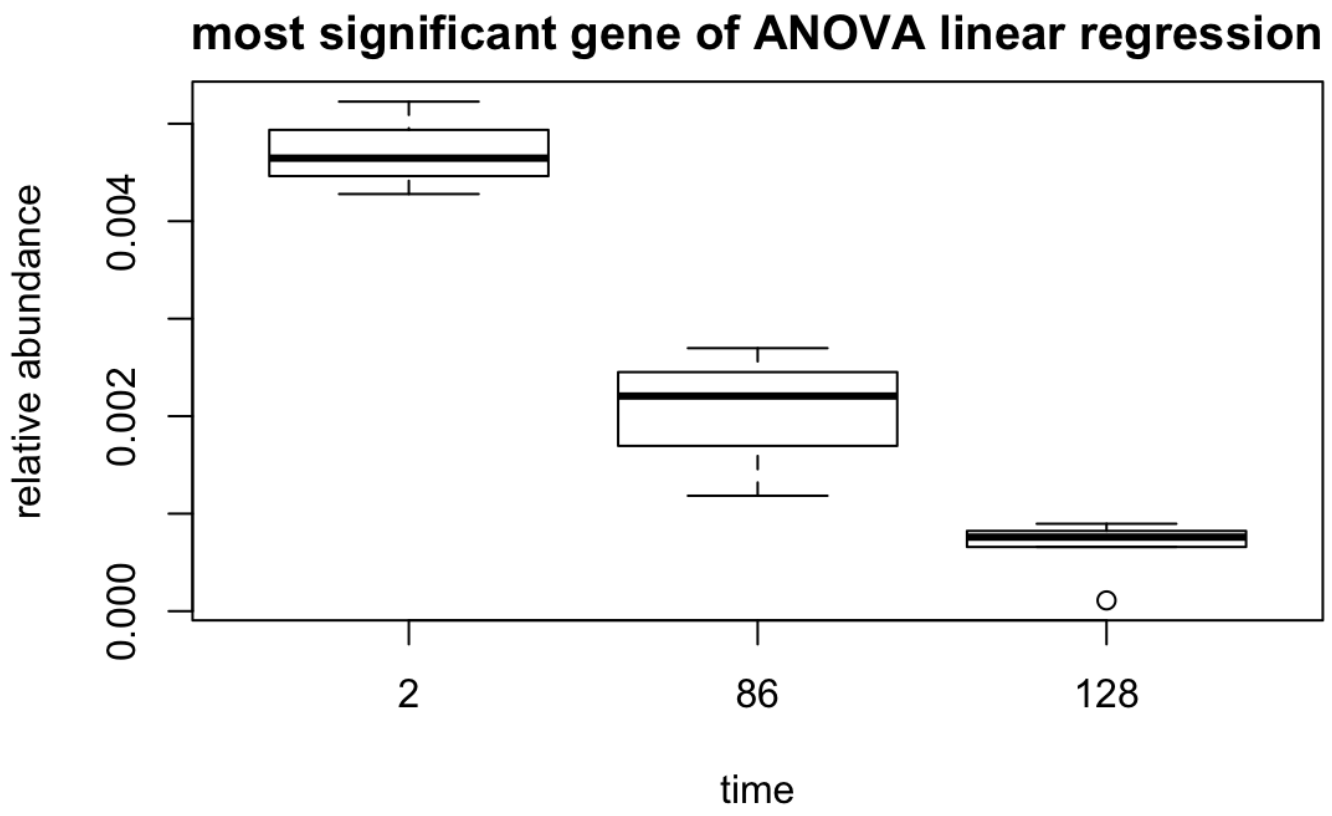
Hide

```
#relative abundance graph of most significant gene for one-way ANOVA
abundFrame <- abundFrame[order(abundFrame$pValANOVA),]
boxplot(as.numeric(countNorm[abundFrame$index[1],]) ~ categoriesAC, xlab = "time", ylab
 = "relative abundance", main = "most significant gene of one-way ANOVA")
```



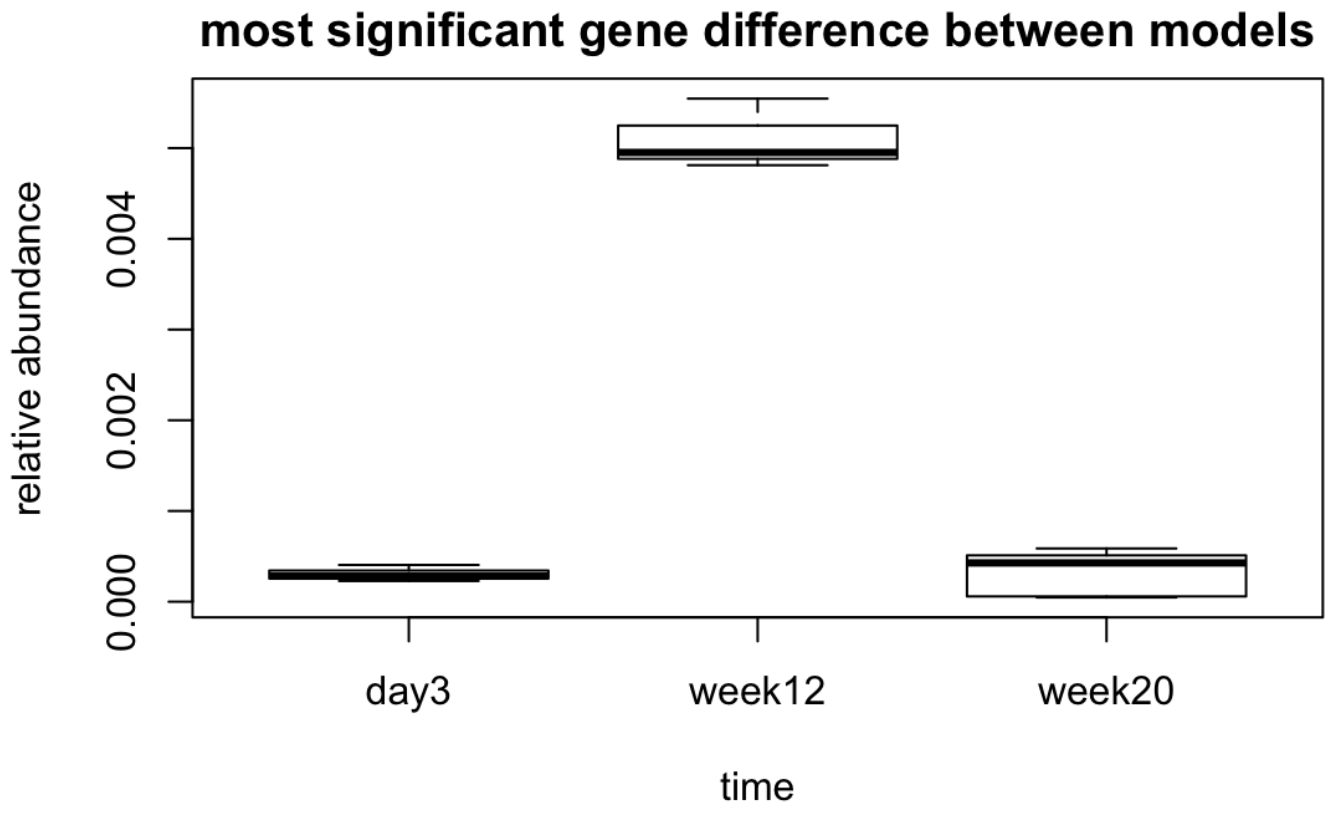
Hide

```
#relative abundance graph of most significant gene for ANOVA linear regression
abundFrame <- abundFrame[order(abundFrame$pValLR),]
boxplot(as.numeric(countNorm[abundFrame$index[1],]) ~ categoriesB, xlab = "time", ylab =
"relative abundance", main = "most significant gene of ANOVA linear regression")
```



Hide

```
#relative abundance graph of most significatn gene difference between models
abundFrame <- abundFrame[order(abundFrame$pValDiff),]
boxplot(as.numeric(countNorm[abundFrame$index[1],]) ~ categoriesAC, xlab = "time", ylab
 = "relative abundance", main = "most significant gene difference between models")
```



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.

ANSWER:

When looking at the difference between the models, we see that the three parameter model in (A) is only more appropriate for 136 of out of the 3983 total genes. This is only for a small margin of genes out of the total, so we can conclude that for majority of the time there is no difference between the two models, but for some genes the three parameter is better than the two parameter modele.

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 – see mini-lecture 16C

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).