1. Use the given link Data Set.

Answer the below questions:

a. What are the assumptions of ANOVA, test it out?

Ans:

To use the ANOVA test we made the following assumptions:

>Each group sample is drawn from a normally distributed population

>All populations have a common variance

>All samples are drawn independently of each other

>Within each sample, the observations are sampled randomly and independently of each other

>Factor effects are additive

>data<-read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/yeast/yeast.data",

col.names=c("Sequence Name","mcg","gvh","alm","mit","erl","pox","vac","nuc","Class"))

>head(data)

>str(data)

>df<-data.frame(data$nuc,data$Class)

>head(df)

>data$Class<-as.numeric(data$Class)

>data$mcg<-as.factor(data$mcg)

>data$erl<-as.factor(data$erl)

>data$mit<-as.factor(data$mit)

>data$alm<-as.factor(data$alm)

>data$pox<-as.factor(data$pox)

>data$vac<-as.factor(data$vac)

>data$nuc<-as.factor(data$nuc)

>data$gvh<-as.factor(data$gvh)

>my\_modle<-lm(mcg~mit+alm,data)

>my\_modle1<-lm(Class~vac+nuc+pox+erl+mit+alm+gvh+mcg ,data)

>anova(my\_modle1)

#TESTING ASSUMPTIONS

#Generate residual and predicted values

>data$resids <- residuals(my\_modle1)

>data$preds <- predict(my\_modle1)

>data$sq\_preds <- data$preds^2

>head(data)

#Look at a plot of residual vs. predicted values

>plot(resids ~ preds, data = data,

xlab = "Predicted Values",

ylab = "Residuals")

#Perform a Shapiro-Wilk test for normality of residuals

>shapiro.test(data$resids)

#Perform Levene's Test for homogenity of variances

#install.packages("car")

>library(car)

>leveneTest(Class ~ mcg, data = data)

>leveneTest(Class ~ erl, data = data) #Unnecessary, because not comparing Blocks

#Perform a Tukey 1-df Test for Non-additivity

>my\_1df\_mod<-lm(Class~vac+nuc+pox+erl+mit+alm+gvh+mcg+ sq\_preds, data)

>anova(my\_1df\_mod)

>boxplot(Class~vac, data = data,

main = "amino acid content of vacuolar and extracellular proteins",

xlab = "Score of amino acid content of vacuolar and extracellular proteins",

ylab = "class")

b. Why ANOVA test? Is there any other way to answer the above question?

Ans:

For the one-way ANOVA, the most common nonparametric alternative tests

are the Kruskal-Wallis test and the median test.

> kruskal.test(Class~vac ,data)

result: Kruskal-Wallis rank sum test

data: Class by vac

Kruskal-Wallis chi-squared = 66.205, df = 47, p-value = 0.03376