# question 1

# question 1

library(ISLR)

sapply(Caravan,table)

table(Caravan$PVRAAUT)

table(Caravan$AVRAAUT)

Caravan

nrow(Caravan)

colnames(Caravan)

d0 = subset(Caravan,select = c(-PVRAAUT,-AVRAAUT))

ncol(Caravan)

ncol(d0)

#a)

str(d0)

levels(d0$Purchase) = c("0","1")

d0$Purchase=as.numeric(d0$Purchase)-1

d0$Purchase

dtrain = d0[1:1000,]

nrow(d0)

dtest = d0[1001:5822,]

head(d0$Purchase)

d0$Purchase

#b)

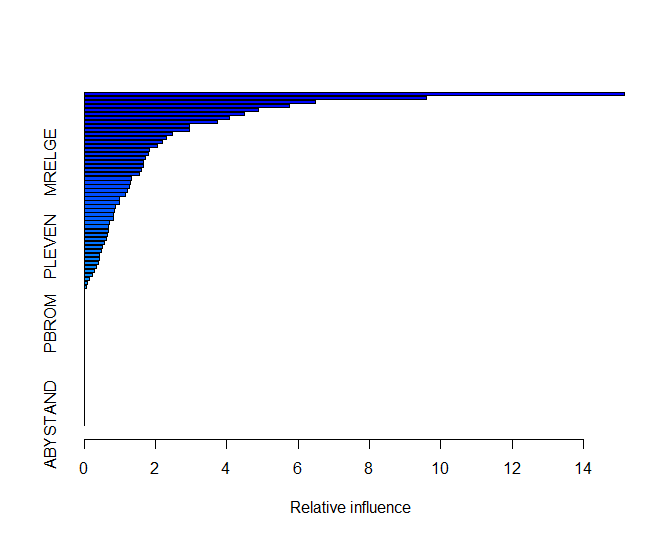
library(gbm)

set.seed(1)

boost1 = gbm(Purchase~.,data = dtrain,distribution="bernoulli",n.trees = 1000,shrinkage = 0.01)

summary(boost1)

# PPERSAUT seems to be the most important



#c)

yhat.boost = predict(boost1,newdata = dtest,n.trees=1000,type = "response")

yhat.boost

yhat = ifelse(yhat.boost>0.2,"1","0")

tb = table(yhat,dtest$Purchase)

tb

# yhat 0 1

# 0 4410 256

# 1 123 33

aux = prop.table(tb)

aux

# yhat 0 1

# 0 0.914558275 0.053090004

# 1 0.025508088 0.006843633

# 0.025508088 of people predicted to make a purchase do in fact make one

1-sum(diag(aux))

# test error rateis 0.07859809

# question 2

#a)

library(cluster)

d0 = utilities

d0$X6 = as.numeric(d0$X6)

d1=d0[,-1]

set.seed(1)

seg.k = kmeans(d1,centers = 4)

seg.k$cluster

summary(seg.k)

table(seg.k$cluster)

# [1] 4 3 4 3 3 4 3 1 4 3 2 3 3 4 3 2 3 4 1 3 3 4

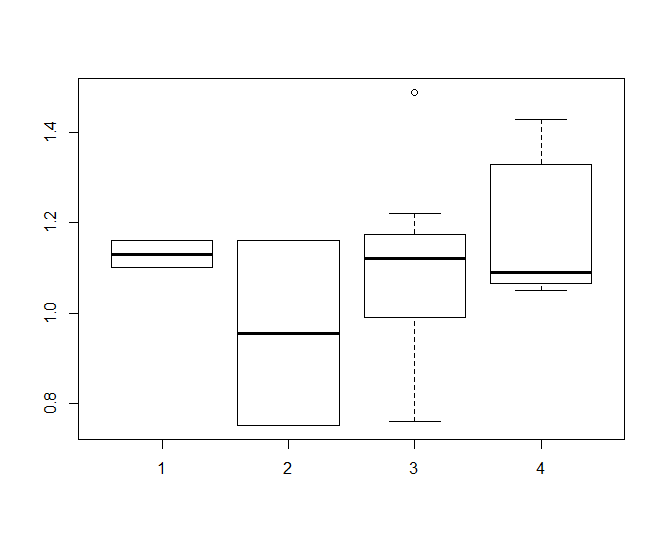
# 1 2 3 4

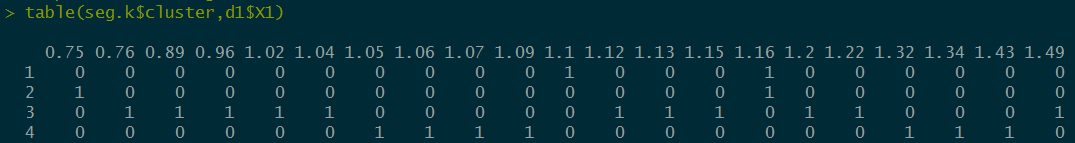
# 2 2 11 7

#b)

table(seg.k$cluster,d1$X1)

boxplot(d1$X1 ~ seg.k$cluster)

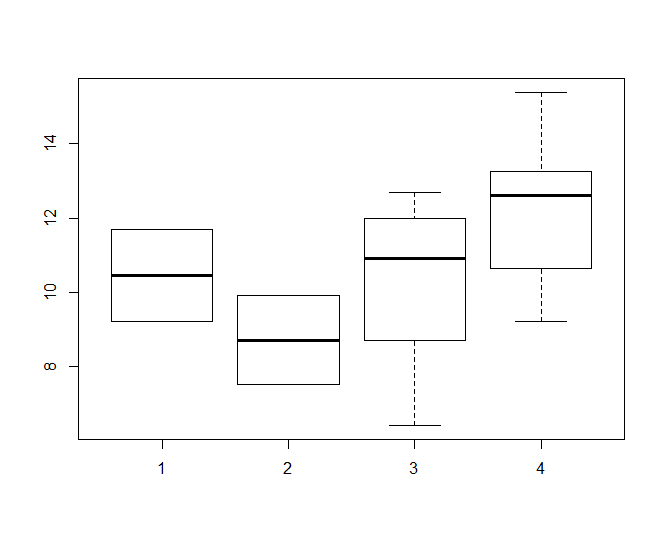
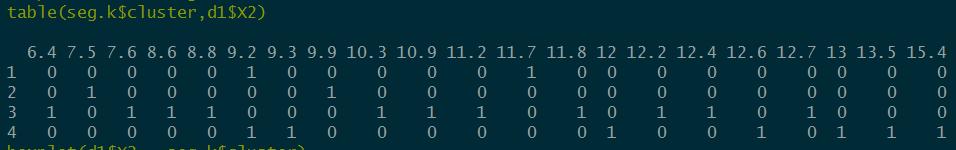




# group 2 is smaller than others in X1

table(seg.k$cluster,d1$X2)

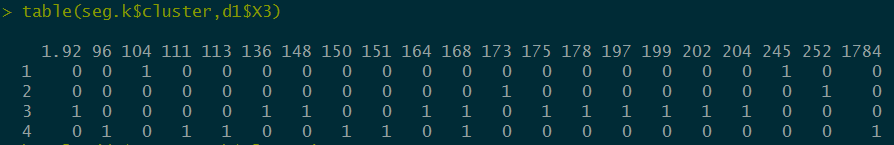
boxplot(d1$X2 ~ seg.k$cluster)

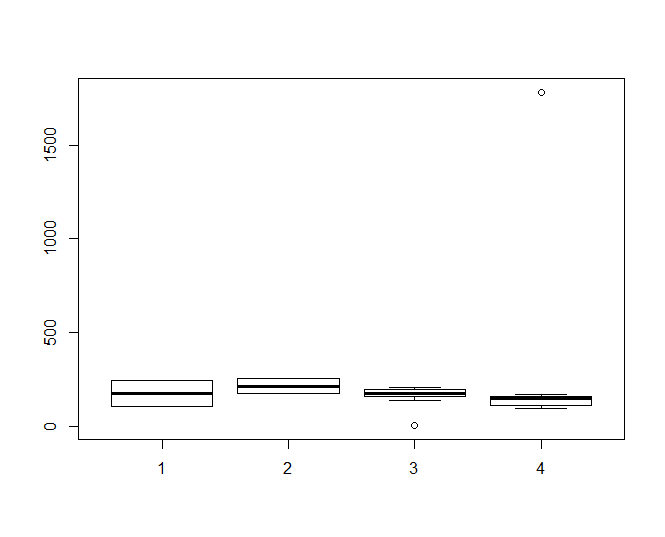


# no big diff

table(seg.k$cluster,d1$X3)

boxplot(d1$X3 ~ seg.k$cluster)

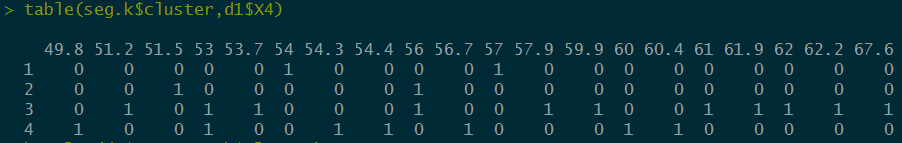


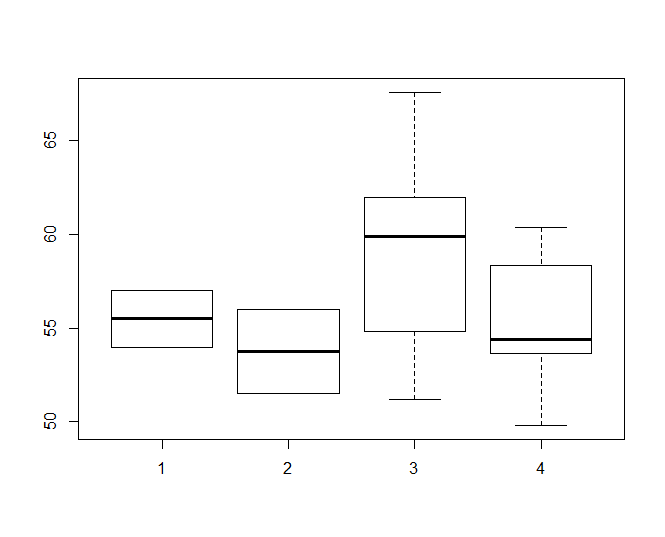


# no big diff

table(seg.k$cluster,d1$X4)

boxplot(d1$X4 ~ seg.k$cluster)

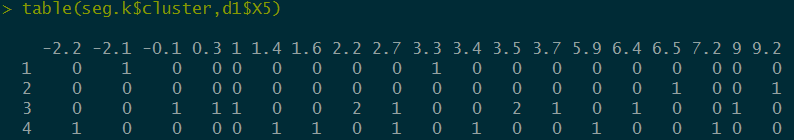


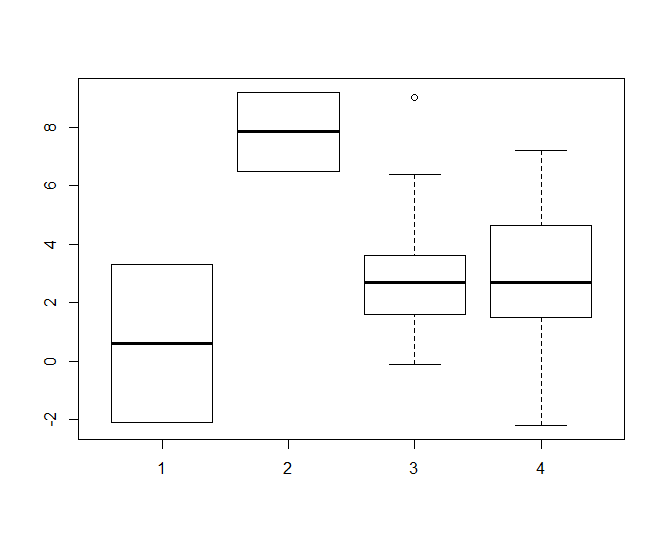


# group 3 is slightly higher than 1,2,4 in X4

table(seg.k$cluster,d1$X5)

boxplot(d1$X5 ~ seg.k$cluster)

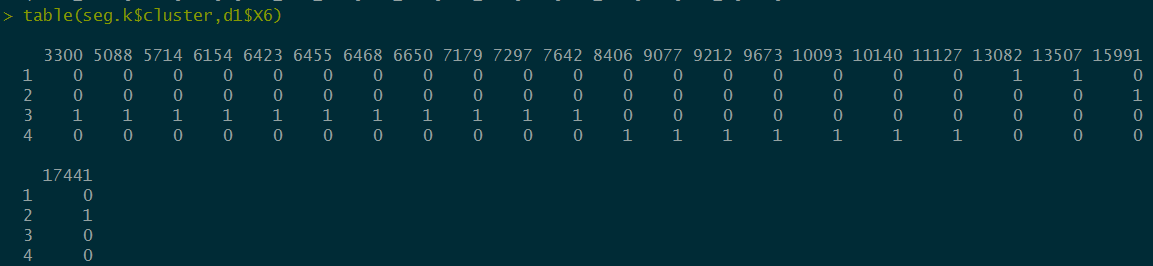


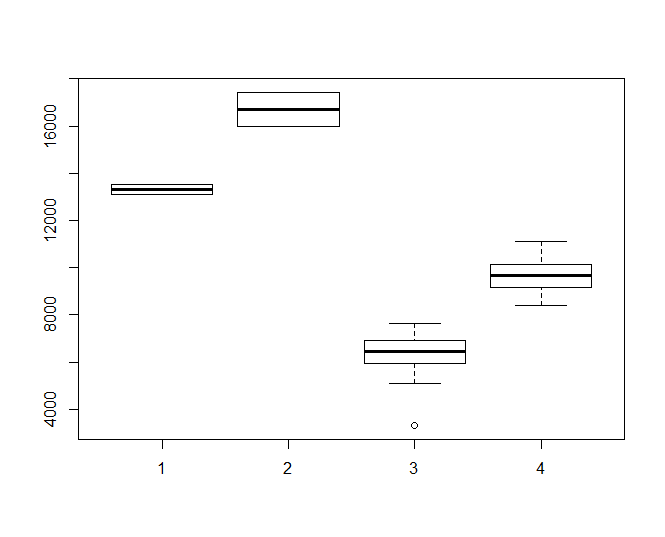


# group 2 is higher than group 1,3,4 in X5

table(seg.k$cluster,d1$X6)

boxplot(d1$X6 ~ seg.k$cluster)

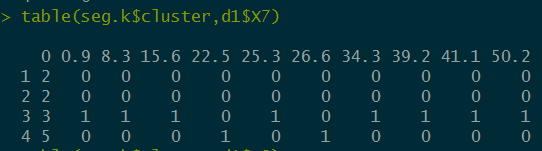


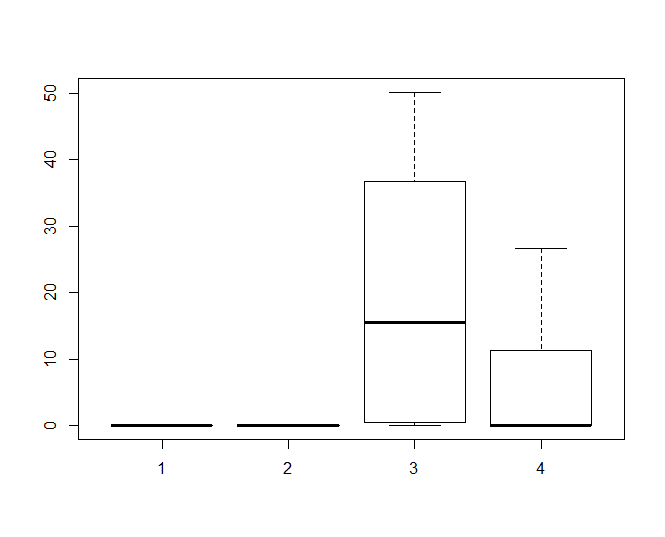


# group 1,2,3 and 4 are all different in X6, group 2 is highest, followed by group 1,4,and 3.

table(seg.k$cluster,d1$X7)

boxplot(d1$X7 ~ seg.k$cluster)

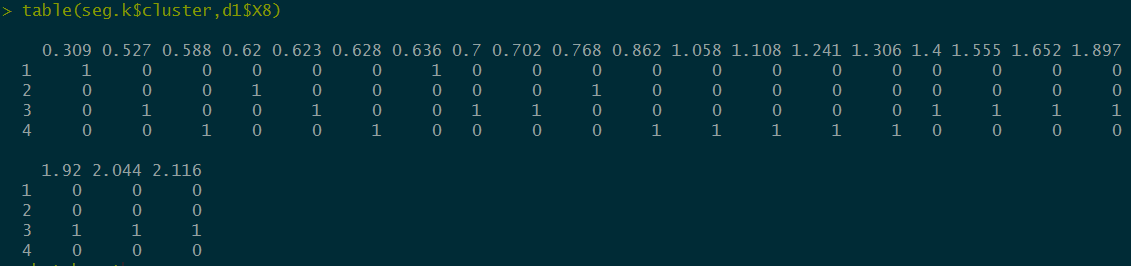


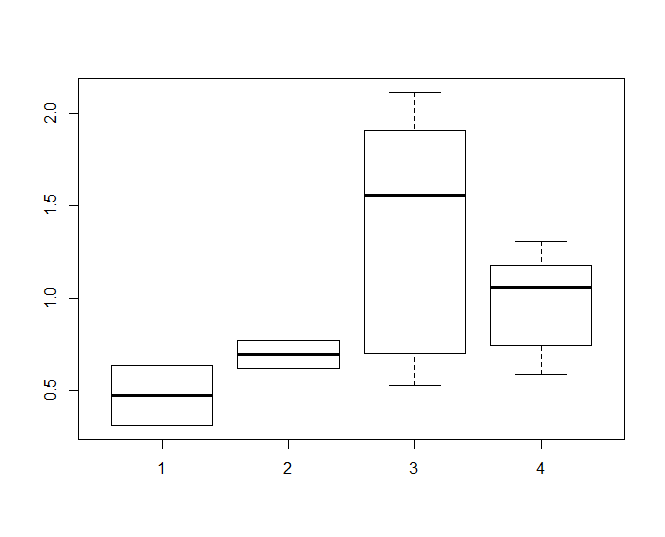


# group 3 is higher than others in X7

table(seg.k$cluster,d1$X8)

boxplot(d1$X8 ~ seg.k$cluster)





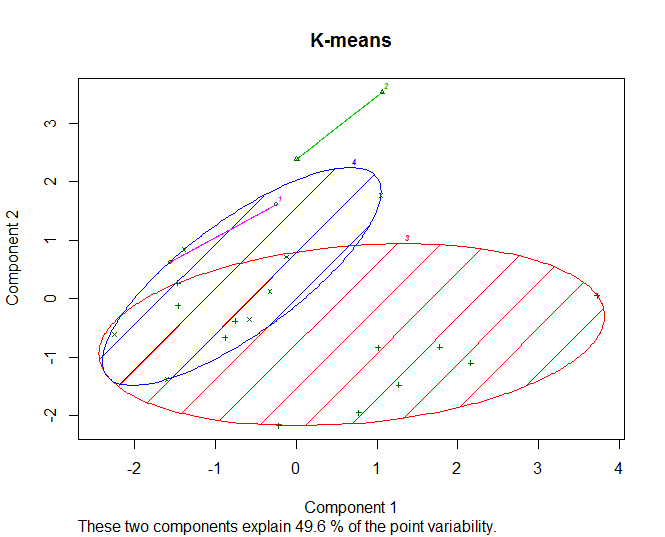
# group 3 is higher than and group 1 is lower than others in X8

#c)

pic=clusplot(d1,seg.k$cluster,color=T,shade=T,labels = 1,lines=0,main="K-means",cex=0.5)

# the green ellipse, which is group 2 is different from others, because 1,3,4 are overlapping each other

# and group 1 is different from group 3 and 2. Group 3 and 4 is similar



#d)

pr = prcomp(d1,scale=T)

rot = pr$rotation

rot

pr$x

as.matrix(scale(d1))%\*%rot

PC1.scaled = rot[,1]

PC1.scaled

# X1 X2 X3 X4 X5 X6 X7 X8

# 0.4451538 0.5647231 -0.1333447 -0.3561666 -0.2684790 0.1404794 0.1970019 -0.4557031

# is the PC1 eigonvector for scaled X

# the defining equation of scaled data is:

# PC1 = 0.4451538\*C1+0.5647231\*C2-0.1333447\*C3-0.3561666\*C4-0.2684790\*C5+0.1404794\*C6+0.1970019\*C7-0.4557031\*C8

# and C[i] = (X[i]-meanX[i])/std(X[i])

# question 3

#a)

getwd()

d0 = read.table("species.txt")

d1 = d0

colnames(d1) = c("pH","Biomass","Species")

d1$Biomass = as.numeric(d1$Biomass)

d1$Species = as.numeric(d1$Species)

aux = as.numeric(d1$pH)

aux

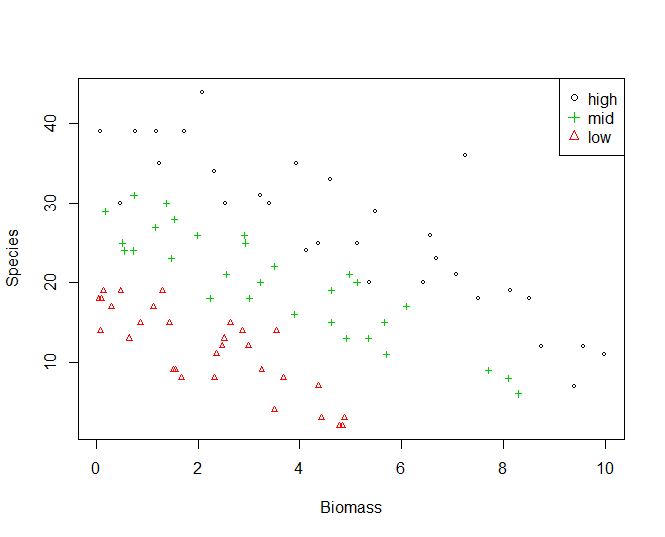
d1$pH

levels(d1$pH)

plot(Species~Biomass,d1,col = aux,pch = aux,cex = 0.6)

legend("topright",c("high","mid","low"),pch = c(1,3,2),col = c(1,3,2) )

# according to the plot, we can not find evidence saying that the slope for each PH level is different



#b)

m1 = glm(Species~Biomass\*pH,d1,family = poisson)

#c）

m2 = glm(Species~.,d1,family = poisson)

anova(m2,m1, test = "Chisq")

# Analysis of Deviance Table

# Model 1: Species ~ pH + Biomass

# Model 2: Species ~ Biomass \* pH

# Resid. Df Resid. Dev Df Deviance Pr(>Chi)

# 1 86 99.242

# 2 84 83.201 2 16.04 0.0003288 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# P-value is small, we can not reject m1(Model 2: Species ~ Biomass \* pH) is better than m2(Model 1: Species ~ pH + Biomass)

#d)

llow = rep("low",101)

lmid = rep("mid",101)

lhigh = rep("high",101)

lbio = seq(0,10,0.1)

dlow = data.frame(pH = llow,Biomass = lbio)

dmid = data.frame(pH = lmid,Biomass = lbio)

dhigh = data.frame(pH = lhigh,Biomass = lbio)

ylow = predict(m1,dlow,type = "response")

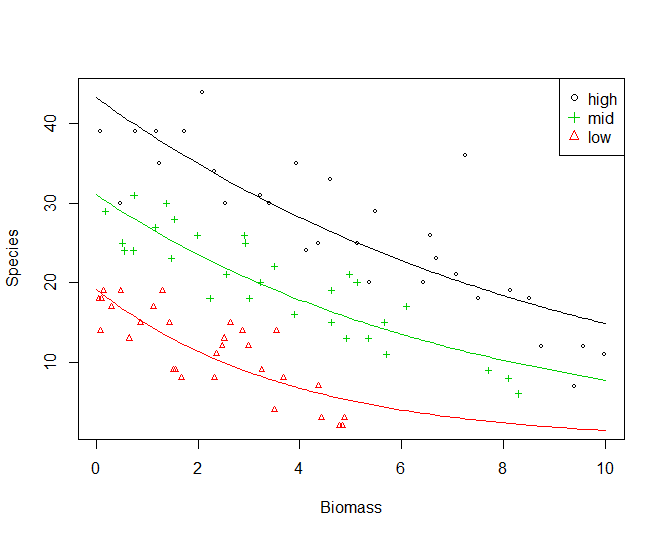
ymid = predict(m1,dmid,type = "response")

yhigh = predict(m1,dhigh,type = "response")

lines(ylow~lbio,col= 2,pch=".",cex = 0.6,xlim = c(0,10),ylim = c(0,40))

lines(ymid~lbio,col= 3,pch=".",cex = 0.6,xlim = c(0,10),ylim = c(0,40))

lines(yhigh~lbio,col= 1,pch=".",cex = 0.6,xlim = c(0,10),ylim = c(0,40))



#e)

newval = data.frame(pH = "low",Biomass = 2.0)

alpha = 0.05

yhat = predict(m1, newval, se.fit=T, type="link")

lower95 = exp(yhat$fit - qnorm(1-alpha/2)\*yhat$se.fit)

lower95 = round(lower95,3)

upper95 = exp(yhat$fit + qnorm(1-alpha/2)\*yhat$se.fit)

upper95 = round(upper95,3)

cat("(",lower95,",",upper95,")\n")

# ( 10.173 , 12.639 )

#f)

yhat = predict(m1,d1,type = "response")

res = yhat-d1$Species

which.max(abs(res))

d1[18,]

# pH Biomass Species

# 18 high 7.242062 36

# the number of species is 36, which associated with largest residual