Exploratory data of Chagas disease microbiota

Tyrel Long

3/3/2021

Overall, the data looks like it is swayed due to a large amount of zero values. This is most evident in the histograms and the boxplots. There are some differences in the phylum of the different diseases as shown by the boxplots but the qq plots do not look very good. Linear regression models were made to look at the three diseases and the phylum and the model with the Highest AIC was the cardiac model. Finally, the pair plots don’t show anything significant between the three diseases.

rm(list=ls())  
library(readxl)  
chagas=read\_excel("R statistics/final project/chagas table 1.xlsx")  
library(ggplot2)  
names(chagas)

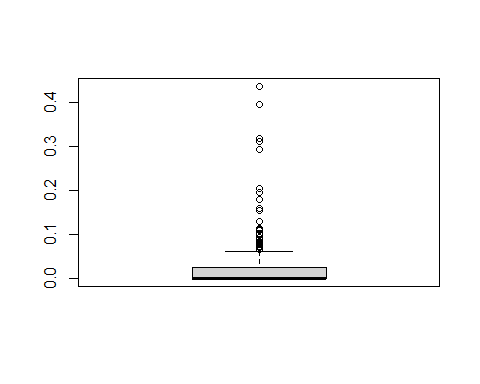
## [1] "Phylum" "Genus"   
## [3] "Control\*" "Indeterminate\*"   
## [5] "Cardiac\*" "Megacolon\*"   
## [7] "FDR\_P" "p-value"   
## [9] "P-value Bonferroni correction"

Square root

chagas$Megacolon.sq=sqrt(chagas$`Megacolon\*`)  
chagas$indeterminate.sq=sqrt(chagas$`Indeterminate\*`)  
chagas$Cardiac.sq=sqrt(chagas$`Cardiac\*`)

Boxplot of Megacolon.sq

boxplot(chagas$Megacolon.sq)



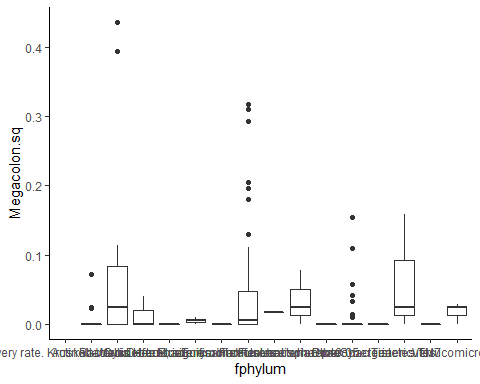
Factors of megacolon, phylum and genus

chagas$fMegacolon=factor(chagas$`Megacolon\*`)  
  
chagas$fphylum=factor(chagas$Phylum)  
chagas$fgenus=factor(chagas$Genus)

Boxplots

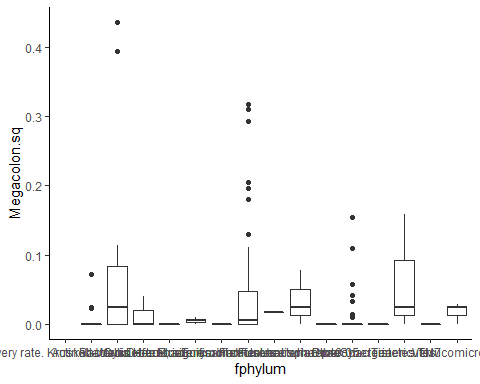
ggplot(chagas ,aes(y=Megacolon.sq,x=fphylum))+geom\_boxplot()+theme\_classic()

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



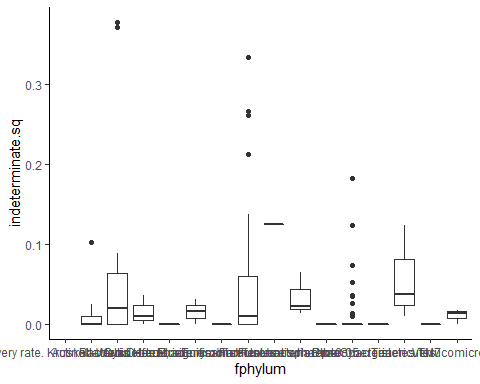
ggplot(chagas ,aes(y=Megacolon.sq,x=fphylum,fgenus))+geom\_boxplot()+theme\_classic()

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



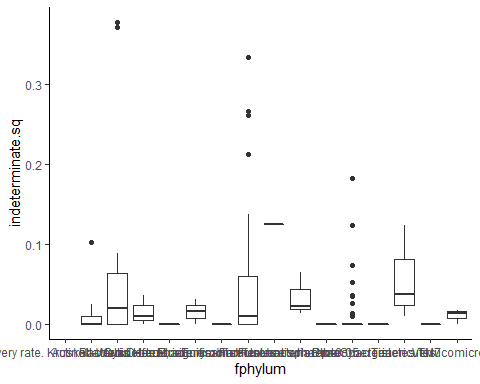
ggplot(chagas ,aes(y=indeterminate.sq,x=fphylum))+geom\_boxplot()+theme\_classic()

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



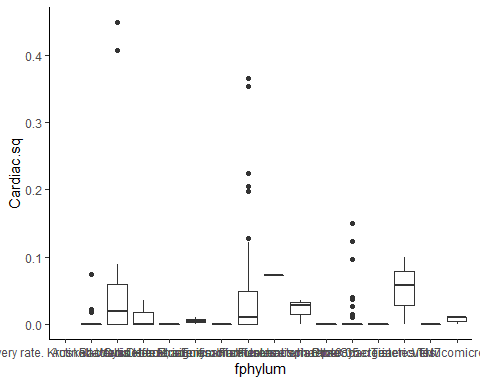
ggplot(chagas ,aes(y=indeterminate.sq,x=fphylum,fgenus))+geom\_boxplot()+theme\_classic()

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



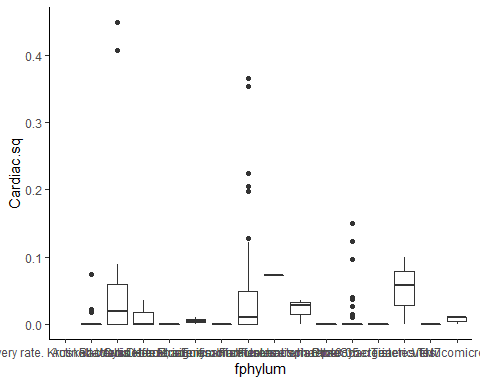
ggplot(chagas ,aes(y=Cardiac.sq,x=fphylum))+geom\_boxplot()+theme\_classic()

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).

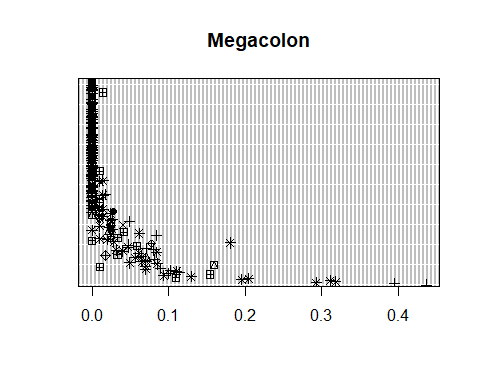


ggplot(chagas ,aes(y=Cardiac.sq,x=fphylum,fgenus))+geom\_boxplot()+theme\_classic()

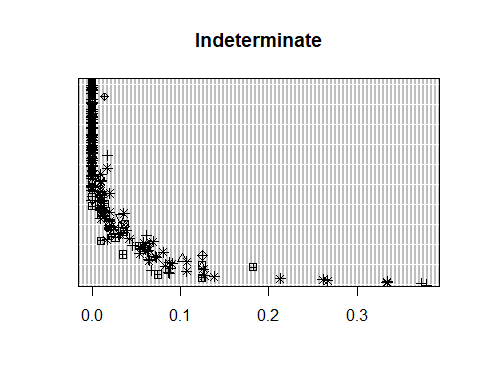
## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).

 Dotcharts

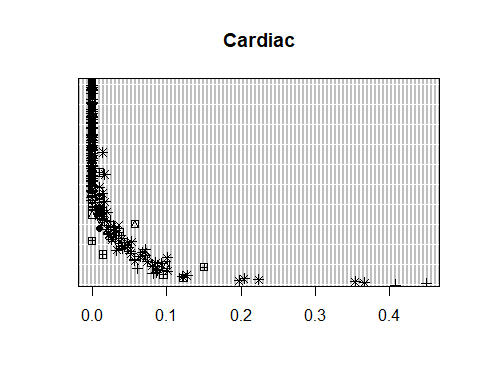
dotchart(chagas$Megacolon.sq,pch = as.numeric(chagas$fphylum),main = "Megacolon")



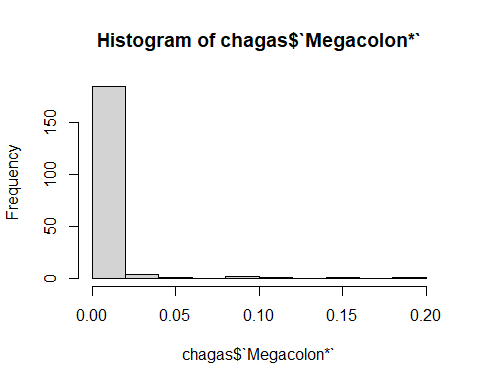
dotchart(chagas$indeterminate.sq,pch = as.numeric(chagas$fphylum),main = "Indeterminate")



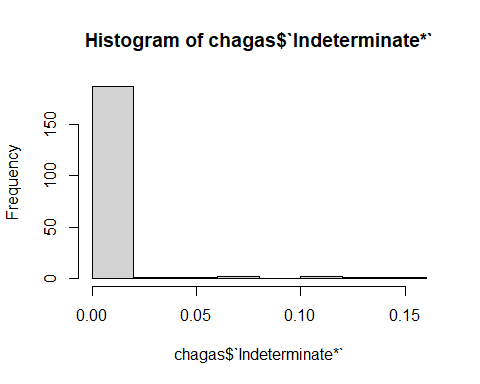
dotchart(chagas$Cardiac.sq,pch = as.numeric(chagas$fphylum),main = "Cardiac")

 Histograms

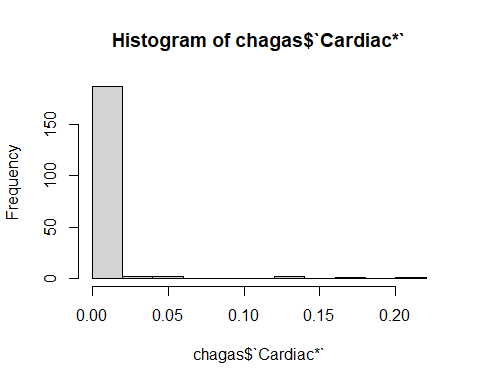
hist(chagas$`Megacolon\*`)



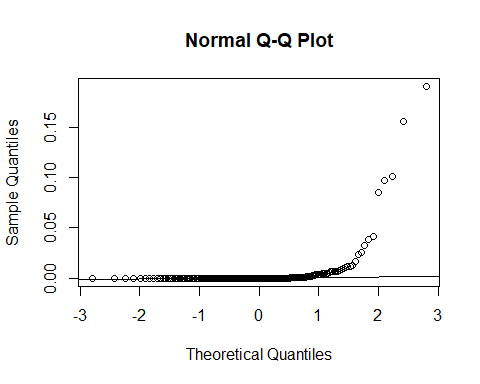
hist(chagas$`Indeterminate\*`)



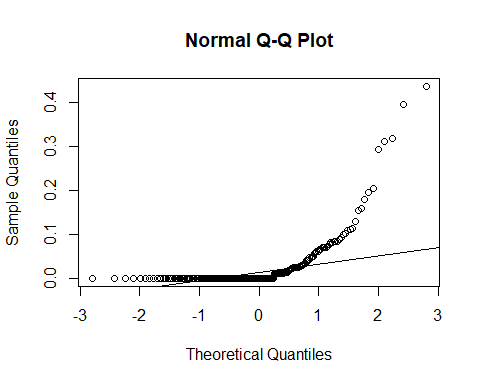
hist(chagas$`Cardiac\*`)

 QQ-plots

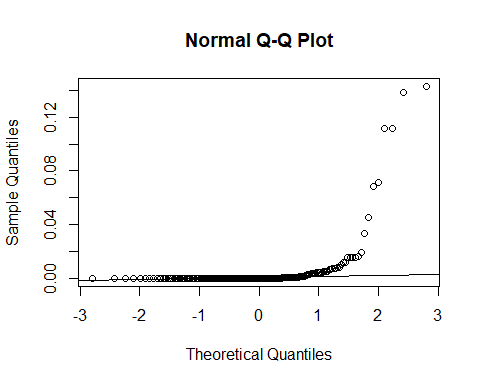
qqnorm(chagas$`Megacolon\*`)  
qqline(chagas$`Megacolon\*`)



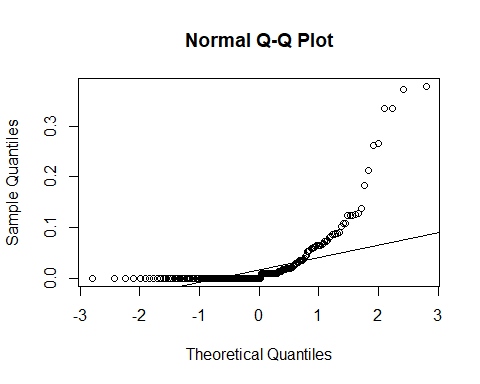
qqnorm(chagas$Megacolon.sq)  
qqline(chagas$Megacolon.sq)



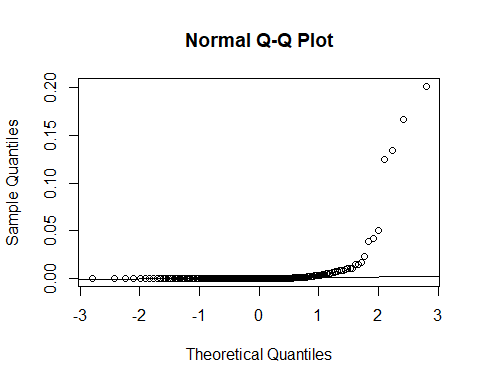
qqnorm(chagas$`Indeterminate\*`)  
qqline(chagas$`Indeterminate\*`)



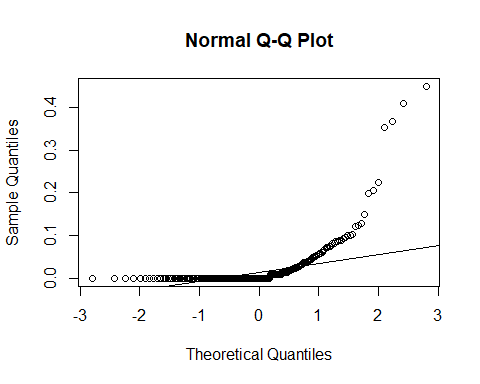
qqnorm(chagas$indeterminate.sq)  
qqline(chagas$indeterminate.sq)



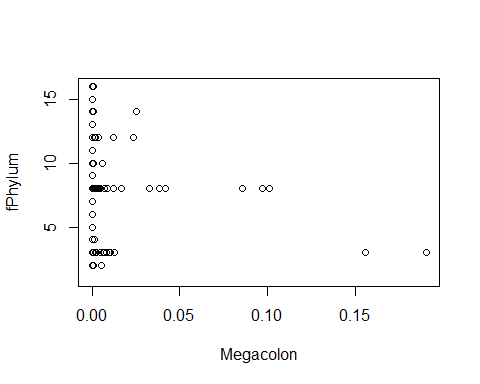
qqnorm(chagas$`Cardiac\*`)  
qqline(chagas$`Cardiac\*`)



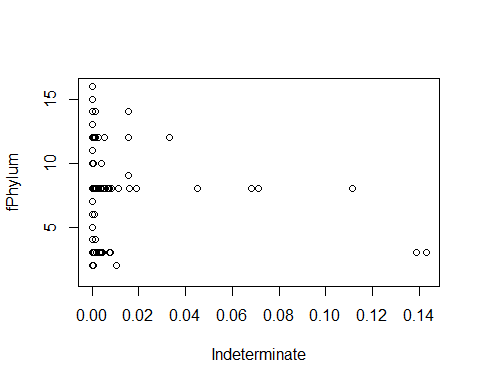
qqnorm(chagas$Cardiac.sq)  
qqline(chagas$Cardiac.sq)

 Scaterplots

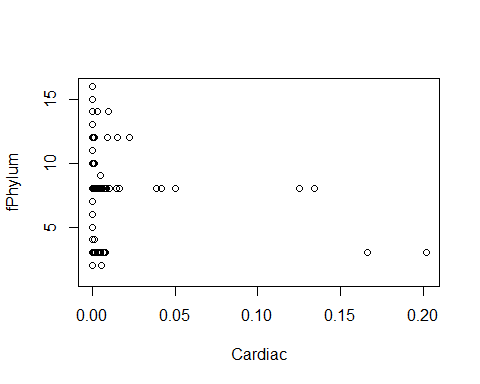
plot(chagas$`Megacolon\*`,chagas$fphylum, xlab="Megacolon",ylab = "fPhylum")



plot(chagas$`Indeterminate\*`,chagas$fphylum, xlab="Indeterminate",ylab = "fPhylum")



plot(chagas$`Cardiac\*`,chagas$fphylum, xlab="Cardiac",ylab = "fPhylum")

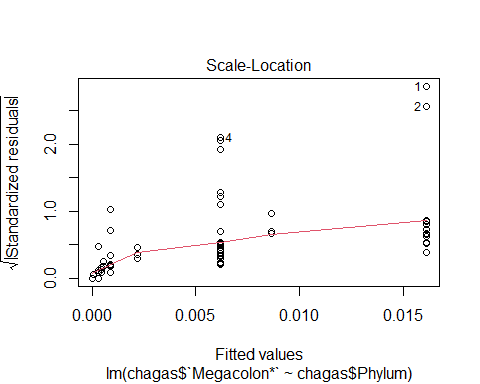
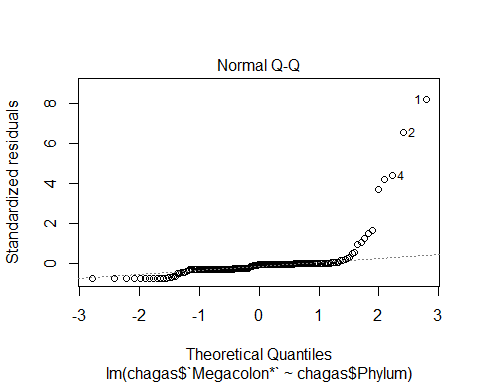
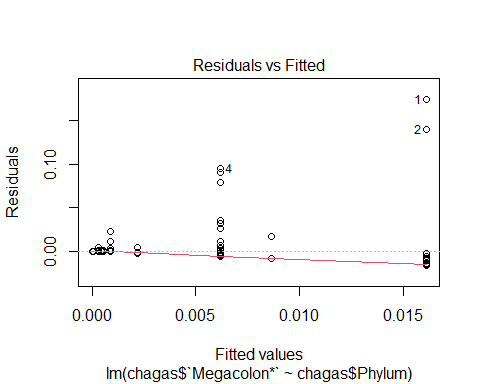
 Linear regresion models

model1=lm(chagas$`Megacolon\*`~chagas$Phylum, data=chagas)  
summary(model1)

##   
## Call:  
## lm(formula = chagas$`Megacolon\*` ~ chagas$Phylum, data = chagas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.016104 -0.006184 -0.000910 -0.000326 0.174296   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.263e-04 4.971e-03 0.066 0.9477   
## chagas$PhylumBacteroidetes 1.578e-02 6.539e-03 2.413 0.0168 \*  
## chagas$PhylumCyanobacteria 2.070e-04 1.346e-02 0.015 0.9877   
## chagas$PhylumDeferribacteres -3.263e-04 2.223e-02 -0.015 0.9883   
## chagas$PhylumElusimicrobia -2.763e-04 1.611e-02 -0.017 0.9863   
## chagas$PhylumEuryarchaeota -3.263e-04 1.611e-02 -0.020 0.9839   
## chagas$PhylumFirmicutes 5.857e-03 5.530e-03 1.059 0.2909   
## chagas$PhylumFusobacteria -2.632e-05 2.223e-02 -0.001 0.9991   
## chagas$PhylumLentisphaerae 1.874e-03 1.346e-02 0.139 0.8895   
## chagas$PhylumOther -3.263e-04 2.223e-02 -0.015 0.9883   
## chagas$PhylumProteobacteria 5.841e-04 5.873e-03 0.099 0.9209   
## chagas$PhylumSynergistetes -3.263e-04 1.611e-02 -0.020 0.9839   
## chagas$PhylumTenericutes 8.340e-03 1.346e-02 0.620 0.5363   
## chagas$PhylumTM7 -3.263e-04 2.223e-02 -0.015 0.9883   
## chagas$PhylumVerrucomicrobia 1.403e-04 1.346e-02 0.010 0.9917   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02167 on 180 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.05549, Adjusted R-squared: -0.01798   
## F-statistic: 0.7553 on 14 and 180 DF, p-value: 0.7161

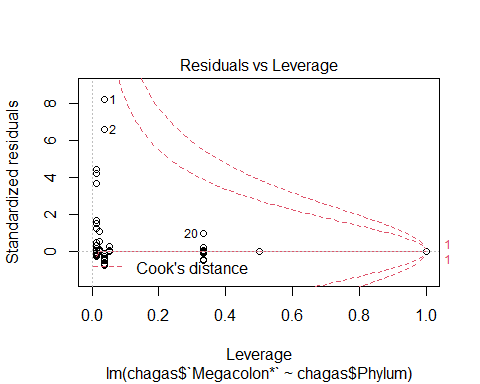
plot(model1)

## Warning: not plotting observations with leverage one:  
## 113, 170, 173



## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced

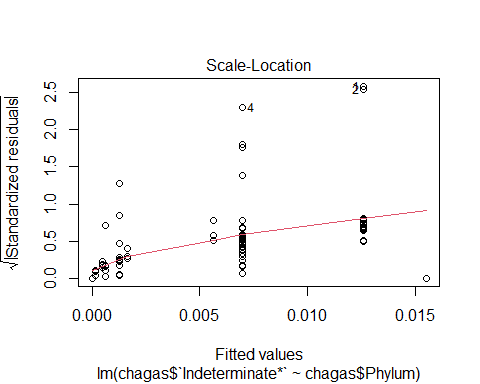
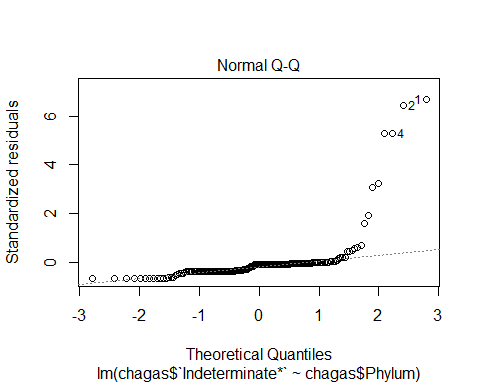
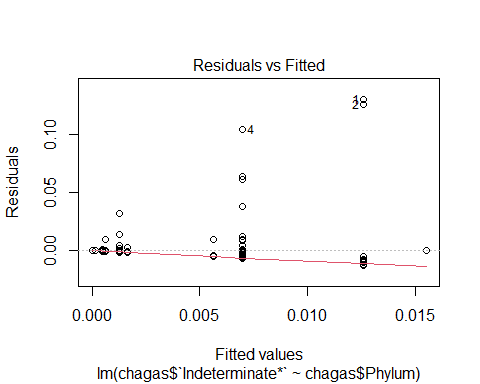


AIC(model1)

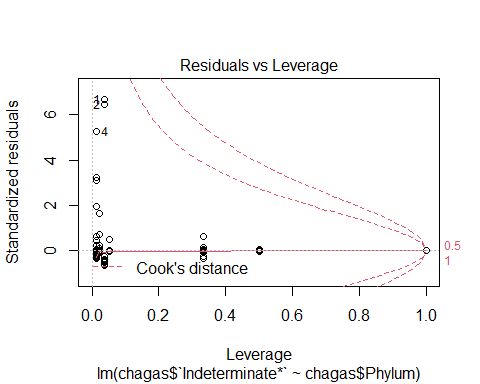
## [1] -924.6928

model2=lm(chagas$`Indeterminate\*`~chagas$Phylum,data = chagas)  
plot(model2)

## Warning: not plotting observations with leverage one:  
## 113, 170, 173



## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced  
  
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced

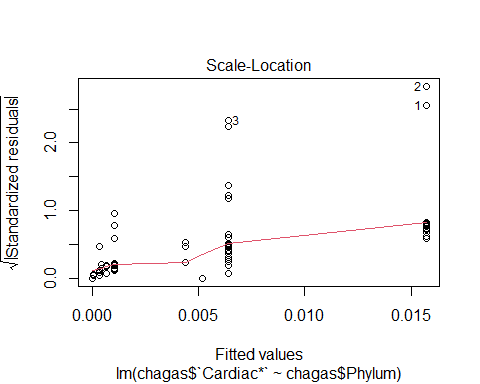
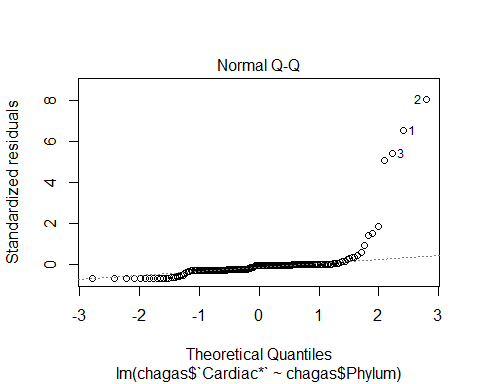
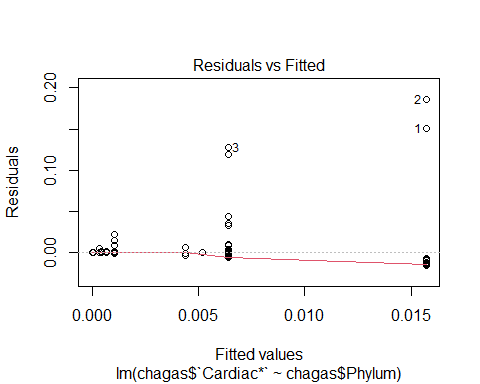


AIC(model2)

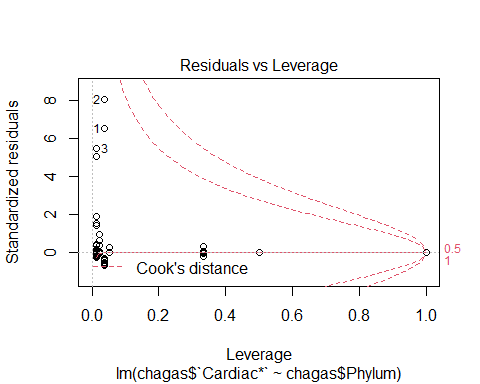
## [1] -957.0602

model3=lm(chagas$`Cardiac\*`~chagas$Phylum,data=chagas)  
plot(model3)

## Warning: not plotting observations with leverage one:  
## 113, 170, 173



## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced  
  
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced



AIC(model3)

## [1] -891.9392

Pair plots

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

ggpairs(chagas [,3:6])

## Warning: Removed 1 rows containing non-finite values (stat\_density).

## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning: Removed 1 rows containing non-finite values (stat\_density).

## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing missing values (geom\_point).  
  
## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning: Removed 1 rows containing non-finite values (stat\_density).

## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing missing values (geom\_point).  
  
## Warning: Removed 1 rows containing missing values (geom\_point).  
  
## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning: Removed 1 rows containing non-finite values (stat\_density).

