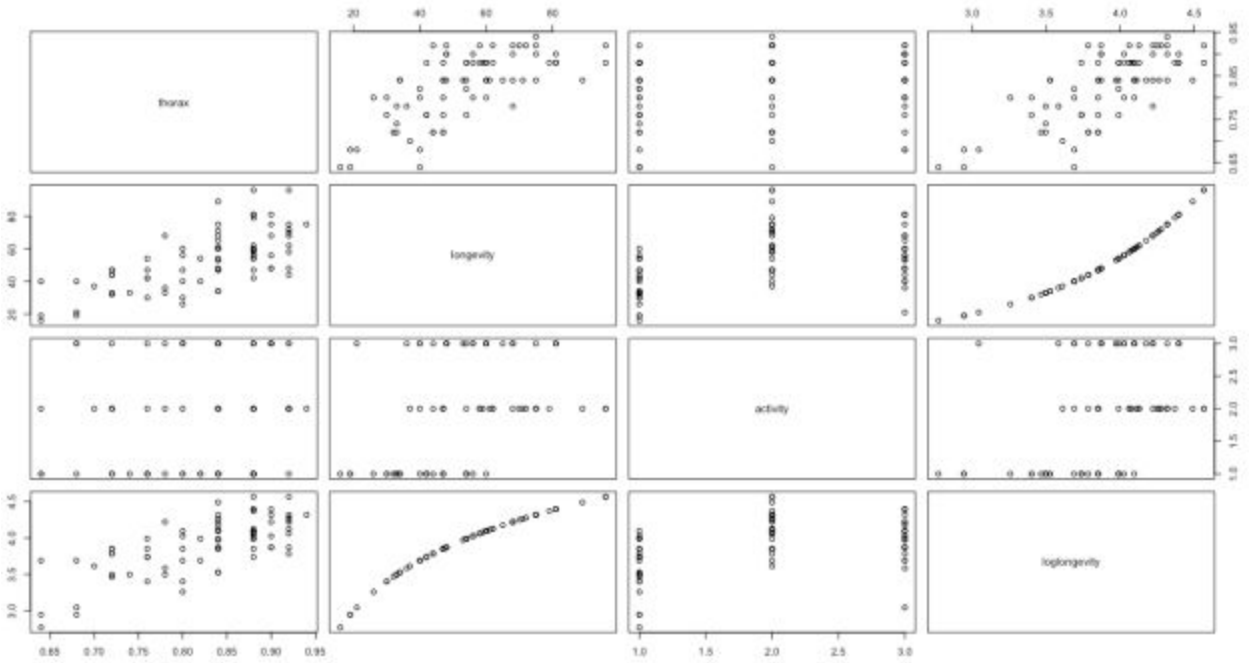
March 19,2018

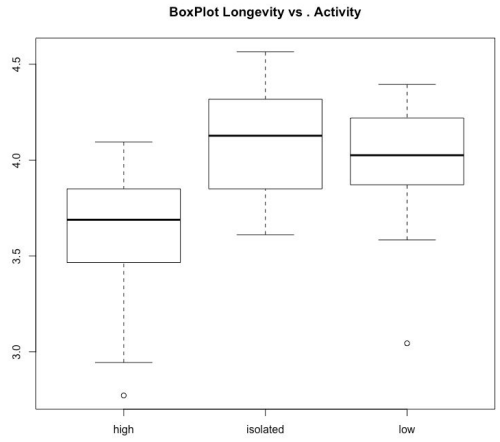
Assignment 4

Chao Zhang & Ibrahim Kanj
Group 13

Exercise 1)

2)





3) Since the loglongevity is not normally distributed, we used kruskal-wallis to check whether there is a significant difference between the activity and longevity. Kruskal-wallis shows that there is a significant difference where the p-value () is less than 0.05.

4) using longevity rather than loglongevity we can say that estimation of days for isolated is 24.840 days more than high activity and estimation of days for low activity is 18.040 more than the high activity. So sexual activity decreases longevity.

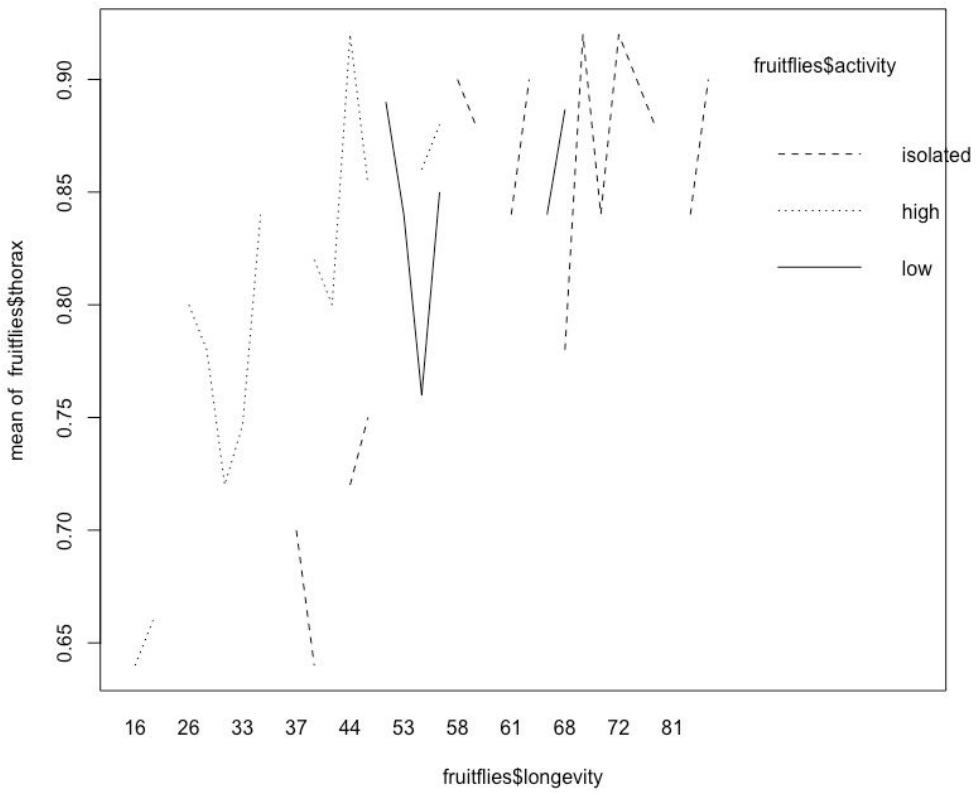
5) By applying anova on thorax*activity we see that the thorax has an effect on the longevity but both thorax as well as activity have no significant effect on longevity.

6) By looking at the anova of thorax + activity we can see that the sexual activity has an effect on longevity where the p-value = 4e-09.

The linear regression model in this case is 1.21893 + 2.97899* thorax + error + (0.40998 if isolated) or (0.28570 if low) or (-0.69568 if high)

so a fly with average thorax (0.84) could have a longevity of 4.13 (isolated) , 4 (low) and 3.02 (high)
and a fly with minimum thorax (0.64) could have a longevity of 3.54 (isolated) , 3.41 (low) and 2.43 (high)

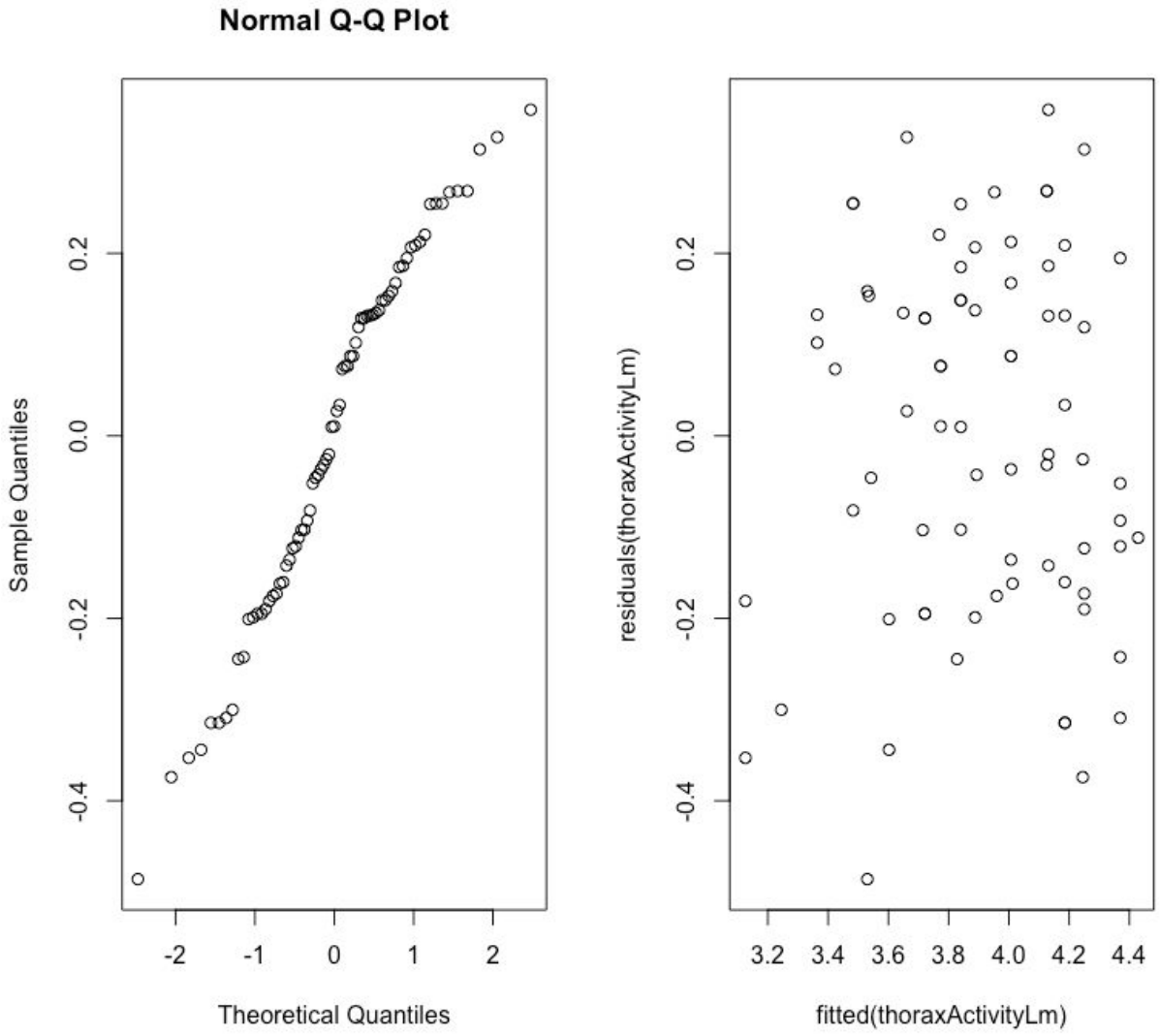
7)



From the interaction plot we can see a similar effect for the length of the thorax on the longevity under the different activities.

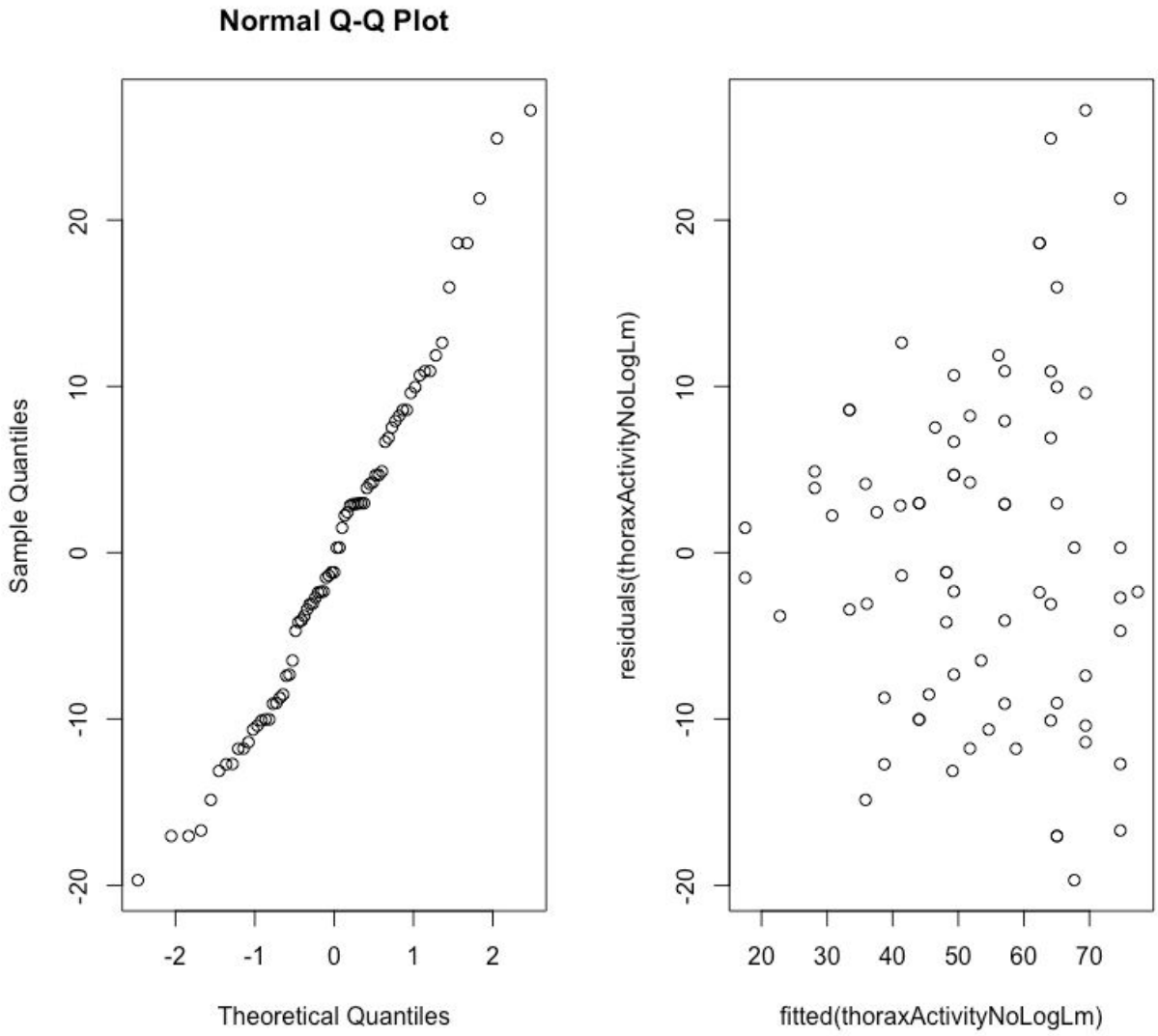
8)

9)



From the qq-plot as well as the histogram it seems that the residuals are normal. Also the fitted graph vs residuals is more or less centered at 0 with a range of [-0.5,0.4]

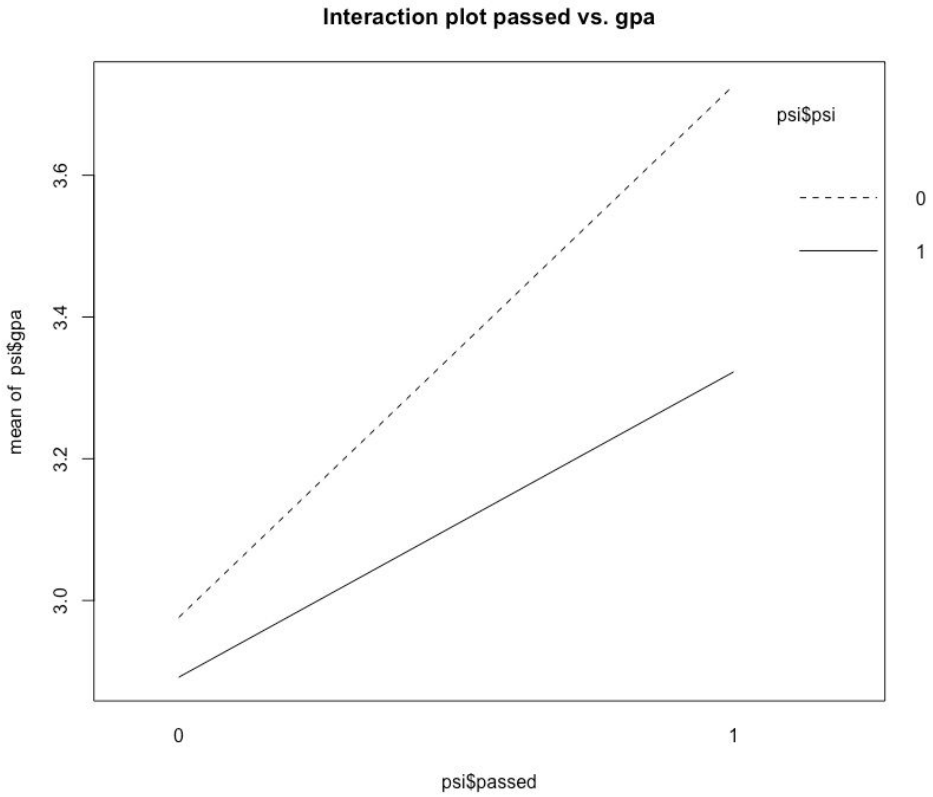
10)



Actually the qq-plot is now more normal than in #10 and also the fitted graph vs residuals is more centered around 0 with a range of [-21,26] thus the use of a logarithm function did not really help in this exercise as the data was already normal.

Exercise 2)

1)



From the data and the above plot we can see that students who took psi although they had a low gpa, some were able to pass the assignment. Another conclusion could be that more students passed the assignment while taking the psi compared to others that didn't take it.

2) The logistic regression model is 1/(1+e^-x) where x is = -11.602 + 3.063*gpa + 2.338 (if psi)

3) According to drop1, the p-value for psi is less than 0.05 therefore psi has a significance different on whether the test has been passed or not. And from the summary of the model we can see that if psi is taken the log odds of passing the test increases by 2.338

4) gpa 3 with psi = 0.48 ~ around 50% chance of passing the assignment
gpa 3 without psi = 0.082 ~ high chance of failure

5) As we said in ex 3, if psi is taken the log odds of passing the test increases by 2.338 and this number is not dependent on gpa.

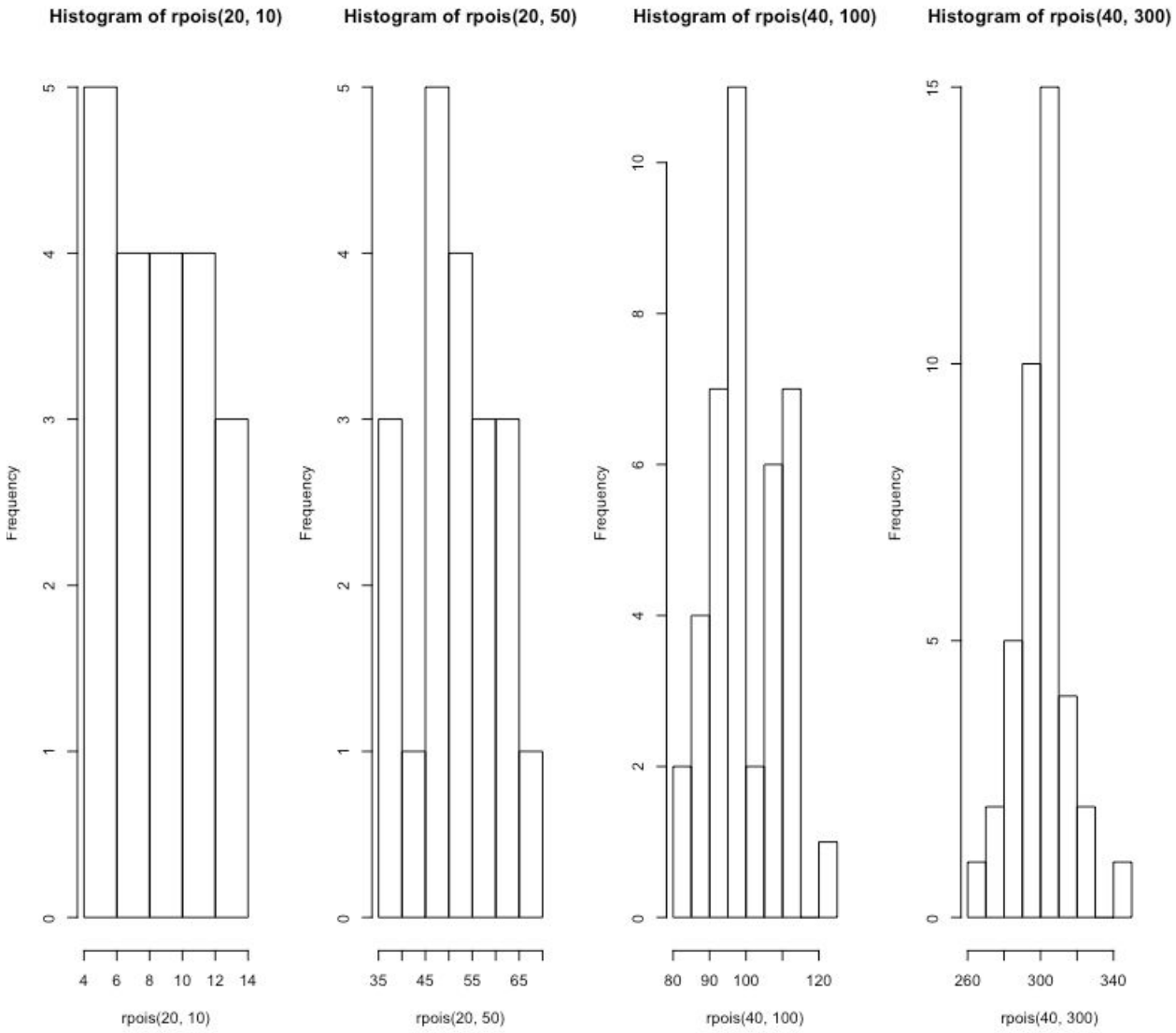
6) The 15 are the students who didn't undergo psi and didn't show improvement in the assignment, the 6 is students who took psi and didn't show improvement. The conclusion that can be made is that 3/18 didn't take psi yet showed some improvement while taking psi 8/14 student showed improvement => psi could improve a students score

7) Since in the summary of the logistic regression model we see the effect of gpa on passing an exam, we can not just ignore the fact that students differ in gpa. Meaning both gpa and psi plays a role in a student passing the assignment or not.

8) The advantage of the first model is that it allows you to predict whether a student would pass an assignment based on his gpa and whether he took psi or not. The advantage of the second model is to have an initial judgment whether psi plays a role or have a significant effect on a student passing the assignment.

Exercise 3)

1)

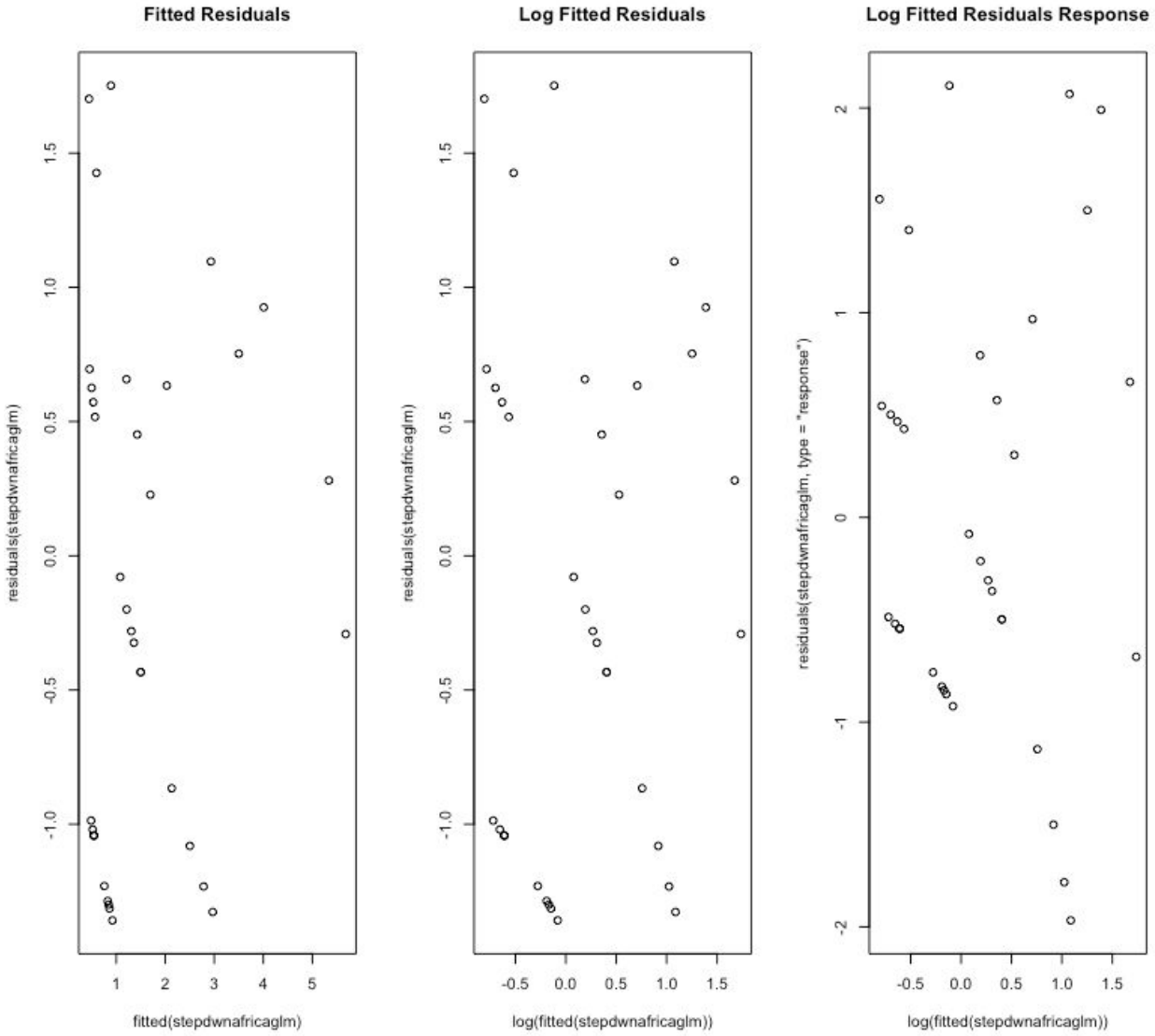


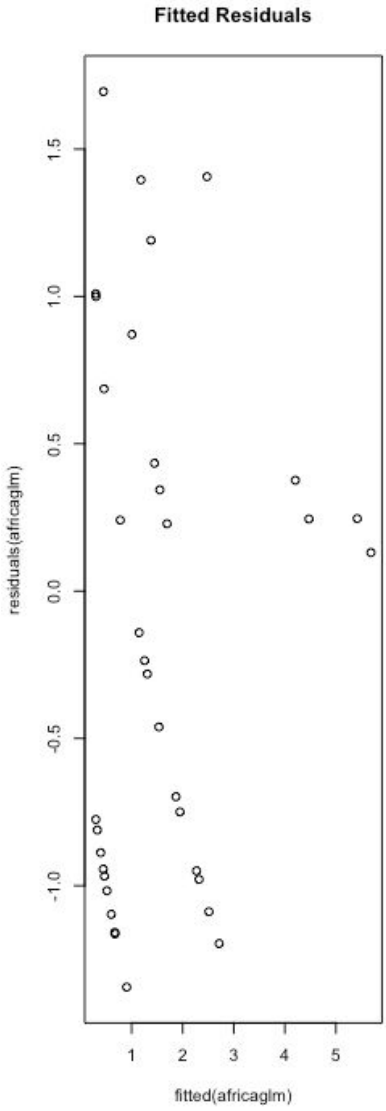
From the above poisson distribution we see that as n and lambda are increasing the more normal the distribution becomes.

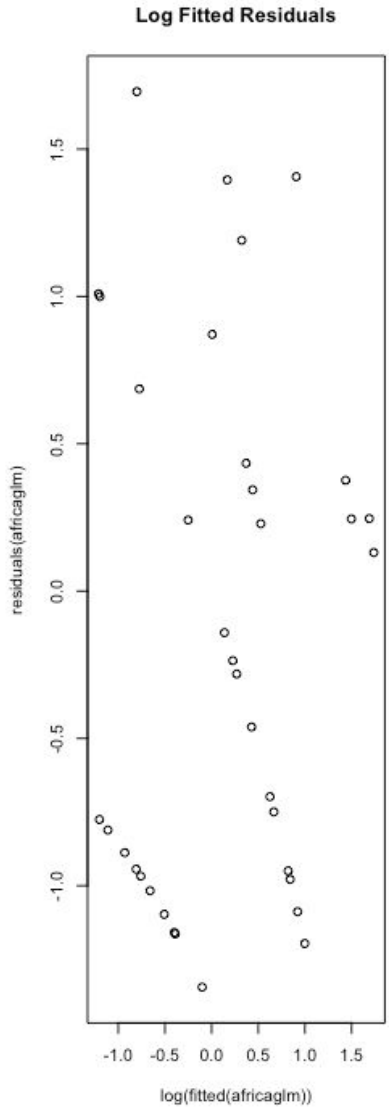
3) code in R

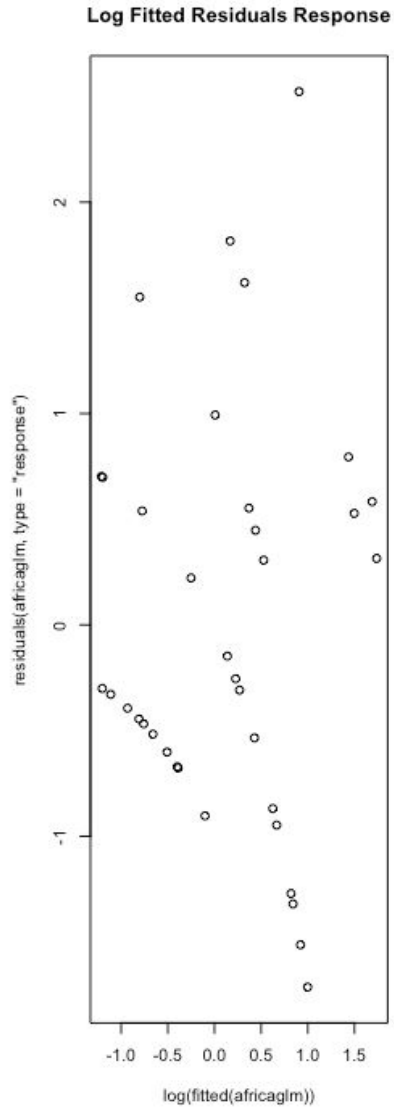
4) Using the step down method we first drop **numelec** then **numeregim** then **size** then **popn** then **pctvote**

5)




Fitted Residuals
residuals(africaglm)
fitted(africaglm)
1.5
1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5


Log Fitted Residuals
residuals(africaglm)
log(fitted(africaglm))
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
1.5


Log Fitted Residuals Response
residuals(africaglm, type = "response")
log(fitted(africaglm))
2
1
0
-1
-1.0
-0.5
0.0
0.5
1.0
1.5

Appendix

Full R Code

```
remove(list = ls())
par(mfrow=c(1,1))

#Exercise 1)
fruitflies = read.table("fruitflies.txt",header=TRUE)

#1)
fruitflies$loglongevity = log(fruitflies$longevity)

#2)
pairs(fruitflies)
boxplot(fruitflies$loglongevity~fruitflies$activity) + title("BoxPlot Longevity vs . Activity")

#3)
qqnorm(fruitflies$loglongevity)
qqline(fruitflies$loglongevity)
# since it is not normally distributed, we use kruskal-wallis
kruskal.test(fruitflies$activity,fruitflies$loglongevity)

#4)
summary(lm(fruitflies$longevity~fruitflies$activity))

#5)
anova(lm(fruitflies$longevity~fruitflies$activity*fruitflies$thorax))

#6)
summary(lm(fruitflies$loglongevity~fruitflies$thorax+fruitflies$activity))
anova(lm(fruitflies$loglongevity~fruitflies$thorax+fruitflies$activity))

#7)
interaction.plot(fruitflies$longevity,fruitflies$activity,fruitflies$thorax)

#8)

#9)
par(mfrow=c(1,2))
thoraxActivityLm = lm(fruitflies$loglongevity~fruitflies$activity+fruitflies$thorax)
qqnorm(residuals(thoraxActivityLm))
qqline(residuals(thoraxActivityLm))
```

```
plot(fitted(thoraxActivityLm), residuals(thoraxActivityLm))

#10)
thoraxActivityNoLogLm = lm(fruitflies$longevity~fruitflies$thorax+fruitflies$activity)
qqnorm(residuals(thoraxActivityNoLogLm))
qqline(residuals(thoraxActivityNoLogLm))
plot(fitted(thoraxActivityNoLogLm), residuals(thoraxActivityNoLogLm))


#Exercise 2)
remove(list = ls())
par(mfrow=c(1,1))

psi = read.table("psi.txt",header=TRUE)

#1)

boxplot(psi$gpa~psi$psi)
boxplot(psi$gpa~psi$passed)
interaction.plot(psi$passed,psi$psi,psi$gpa) + title("Interaction plot passed vs. gpa")

#2)

newpassed = psi$passed[1:32]
newpsi = psi$psi[1:32]
newgpa = psi$gpa[1:32]
dataFramePsi = data.frame(newpassed,newpsi,newgpa)

#barplot(xtabs(dataFramePsi$newpassed~factor(dataFramePsi$newgpa),data =
dataFramePsi))

psiGlm=glm(formula =
factor(dataFramePsi$newpassed)~dataFramePsi$newgpa+factor(dataFramePsi$newpsi),data=
dataFramePsi,family=binomial)
summary(psiGlm)

drop1(psiGlm,test="Chisq")
plot(fitted(psiGlm),residuals(psiGlm),main = "Residual vs. fitted")

#4)
newpassed = NULL
newgpa = NULL
newpsi = NULL
```

```
newgpa = 0
newpsi = c("0")
newData = data.frame(newgpa = 3, newpsi = "0")
predict(psiGlm,newData,type = "response")
```

#5)

#6)

```
x=matrix(c(3,15,8,6),2,2)
fisher.test(x)
```

#Exericse 3)

```
remove(list = ls())
par(mfrow=c(1,1))

africa = read.table("africa.txt",header=TRUE)
```

#1)

```
par(mfrow=c(1,4))
hist(rpois(20,10))
hist(rpois(20,50))
hist(rpois(40,100))
hist(rpois(40,300))
```

#2)

#3)

```
africaglm=glm(africa$miltcoup~africa$oligarchy+africa$pollib+africa$parties+africa$pctvote+afri
ca$popn+africa$size+africa$numelec+africa$numregim,family=poisson,data=africa)
summary(africaglm)
```

#4)

```
stepdwnafricaglm=glm(africa$miltcoup~africa$oligarchy+africa$pollib+africa$parties,family=pois
son,data=africa)
summary(stepdwnafricaglm)
```

#5)

```
par(mfrow=c(1,3))
plot(fitted(stepdwnafricaglm),residuals(stepdwnafricaglm),main = "Fitted Residuals")
plot(log(fitted(stepdwnafricaglm)),residuals(stepdwnafricaglm),main = "Log Fitted Residuals")
plot(log(fitted(stepdwnafricaglm)),residuals(stepdwnafricaglm,type="response"),main = "Log
Fitted Residuals Response")
```

```
par(mfrow=c(1,3))
plot(fitted(africaglm),residuals(africaglm),main = "Fitted Residuals")
plot(log(fitted(africaglm)),residuals(africaglm),main = "Log Fitted Residuals")
plot(log(fitted(africaglm)),residuals(africaglm,type="response"),main = "Log Fitted Residuals
Response")
```