# Assignment 6, EDDA 2017

Martin de la Riva(11403799) and Kieran O'Driscoll(11426438), Group 23

18 May 2017

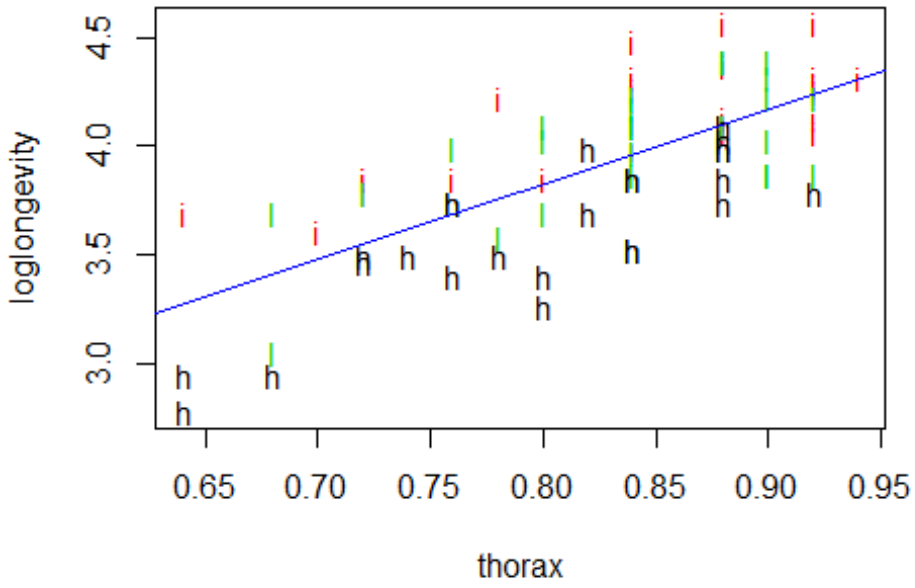
## Assignment 6

### Exercise 1

```
fruitflies=read.table('fruitflies.txt', header=TRUE)
fruitflies$loglongevity <- log(fruitflies$longevity)
```

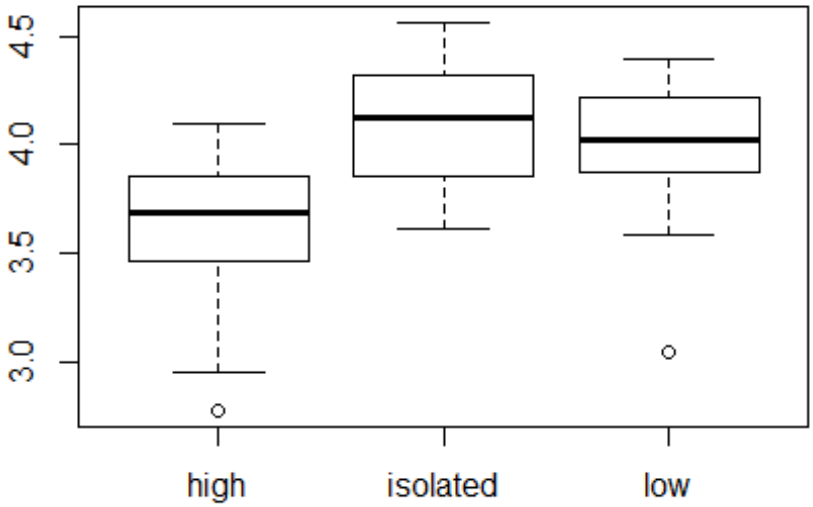
#### 2.

```
plot(loglongevity~thorax, pch=as.character(activity),
col=as.numeric(activity), data=fruitflies)
abline(lm(loglongevity~thorax, data=fruitflies), col="blue")
```



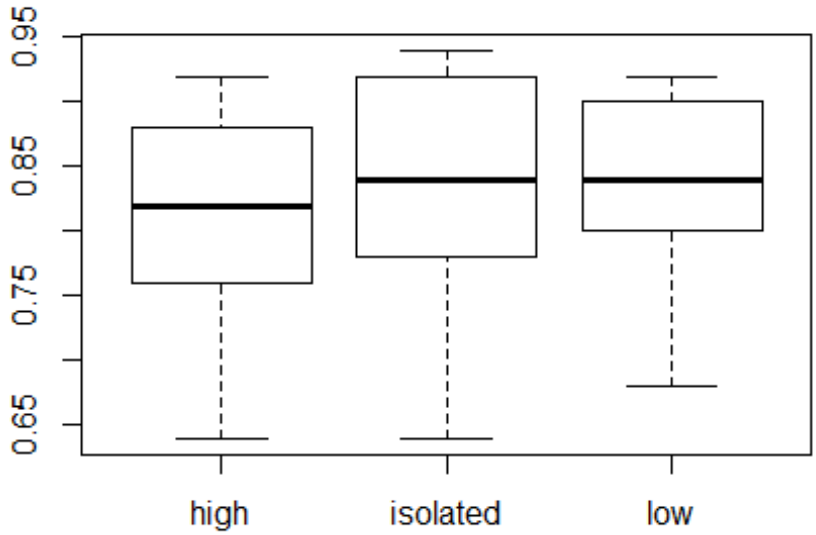

In the plot above, longevity seems to increase with torax.

```
# per activity:
boxplot(fruitflies$loglongevity~fruitflies$activity)
```




In these boxplots, it can be seen how high activity has a significantly low longevity than low and isolated activity. Those two last activities remain in similar longevity values.

```
# thorax per activity (is there a significant difference between them?)
boxplot(fruitflies$thorax~fruitflies$activity)
```




Lastly, we plot activity with respect to thorax, to see if thorax lengths are equally distributed along the activity groups. This is important, as if one group was full of bigger thorax fruitflies, this could affect the result of the experiment. The activity groups are seen to actually have similar thorax values.

### 3.

```
flieslm = lm(loglongevity~activity, data=fruitflies)
anova(flieslm)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value    Pr(>F)
## activity   2 3.6665  1.8333  19.421 1.798e-07 ***
## Residuals 72 6.7966  0.0944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sexual activity seems to significantly influence longevity, with a p-value of 1.798e-07.

### 4.

```
confint(flieslm)
```

```
##                       2.5 %    97.5 %
## (Intercept)      3.4796296 3.7246190
## activityisolated 0.3439909 0.6904582
## activitylow      0.2244780 0.5709453
```

The more sexual activity the less longetivity. In the Confidence Intervals for the different activities, this can be seen: Confidence Intervals for $\mu_{high}$ activity: [3.4796296, 3.7246190] Confidence Intervals for $\mu_{isolated}$ - $\mu_{high}$: [0.3439909, 0.6904582] Confidence Intervals for $\mu_{low}$ - $\mu_{high}$: [0.2244780, 0.5709453] Therefore, high has the lowest longevity, and isolated the highest (isolated > low > high).

## 5.

```
flieslm2 = lm(loglongevity~thorax+activity, data=fruitflies)
anova(flieslm2)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax     1 5.4322  5.4322 132.175 <2e-16 ***
## activity   2 2.1129  1.0565  25.705  4e-09 ***
## Residuals 71 2.9180  0.0411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(flieslm2,test="F")
```

```
## Single term deletions
##
## Model:
## loglongevity ~ thorax + activity
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                2.9180 -235.50
## thorax    1    3.8786 6.7966 -174.08  94.374 1.139e-14 ***
## activity  2    2.1129 5.0309 -198.64  25.705 4.000e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After testing it (both with anova and drop1, we get very similar results on both) we can conclude that both factors (thorax and activity) have a significant influence on the longevity.

## 6.

```
summary(flieslm2)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = fruitflies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4858 -0.1612  0.0104  0.1510  0.3574
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       1.21893    0.24865   4.902 5.79e-06 ***
## thorax            2.97899    0.30665   9.715 1.14e-14 ***
## activityisolated  0.40998    0.05839   7.021 1.07e-09 ***
## activitylow       0.28570    0.05849   4.885 6.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2027 on 71 degrees of freedom
## Multiple R-squared:  0.7211, Adjusted R-squared:  0.7093
## F-statistic:  61.2 on 3 and 71 DF,  p-value: < 2.2e-16
```

As said before, the more sexual activity, the lowest longevity. Therefore, sexual activity influences negatively longevity (Isolated > Low > High). final model: 1.21893 + 2.97899*thorax + 0.40998 (isolated) + 0.28570 (low) - 0.69568 (high)

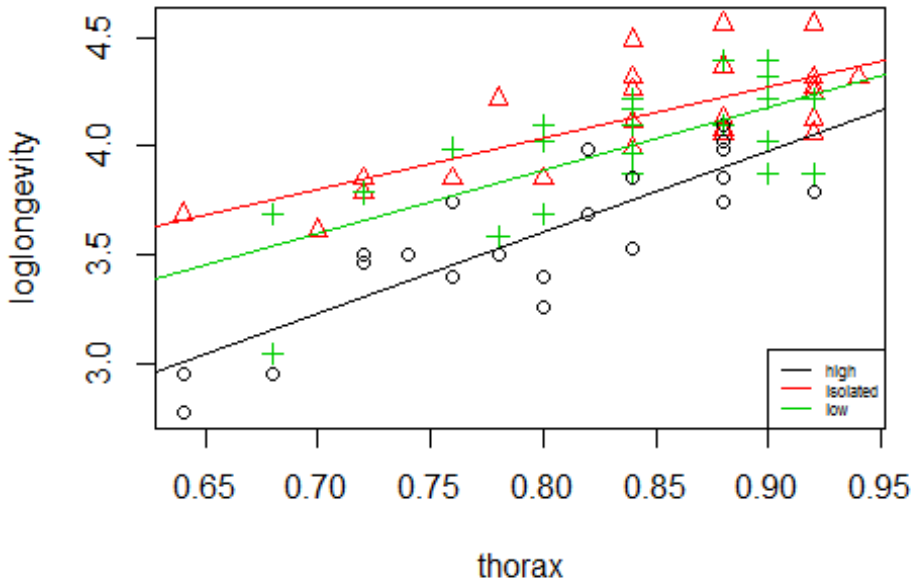
For a fly with an average thorax:

```
mean(fruitflies$thorax)

## [1] 0.8245333

###    loglongevity_isolated= 1.21893 + 2.97899*0.8245333 + 0.40998 =
4.085186
1.21893 + 2.97899*mean(fruitflies$thorax) + 0.40998

## [1] 4.085187

###    loglongevity_low     = 1.21893 + 2.97899*0.8245333 + 0.28570 =
3.960906
1.21893 + 2.97899*mean(fruitflies$thorax) + 0.28570

## [1] 3.960907

###    loglongevity_high    = 1.21893 + 2.97899*0.8245333 - 0.69568 =
2.979526
1.21893 + 2.97899*mean(fruitflies$thorax) - 0.69568

## [1] 2.979527

min(fruitflies$thorax)

## [1] 0.64

###    loglongevity_isolated= 1.21893 + 2.97899*0.64 + 0.40998 = 3.535464
1.21893 + 2.97899*min(fruitflies$thorax) + 0.40998

## [1] 3.535464

###    loglongevity_low     = 1.21893 + 2.97899*0.64 + 0.28570 = 3.411184
1.21893 + 2.97899*min(fruitflies$thorax) + 0.28570

## [1] 3.411184
```

```
###    loglongevity_high    = 1.21893 + 2.97899*0.64 - 0.69568 = 2.429804
1.21893 + 2.97899*min(fruitflies$thorax) - 0.69568

## [1] 2.429804
```

## 7.

```
plot(loglongevity~thorax, pch=unclass(activity), col=as.numeric(activity),
data=fruitflies)
for (i in 1:3) abline(lm(loglongevity~thorax,
data=fruitflies[as.numeric(fruitflies$activity)==i,]), col=i)
legend("bottomright", legend=levels(fruitflies$activity), col=c(1,2,3),
lty=1, cex=0.5)
```



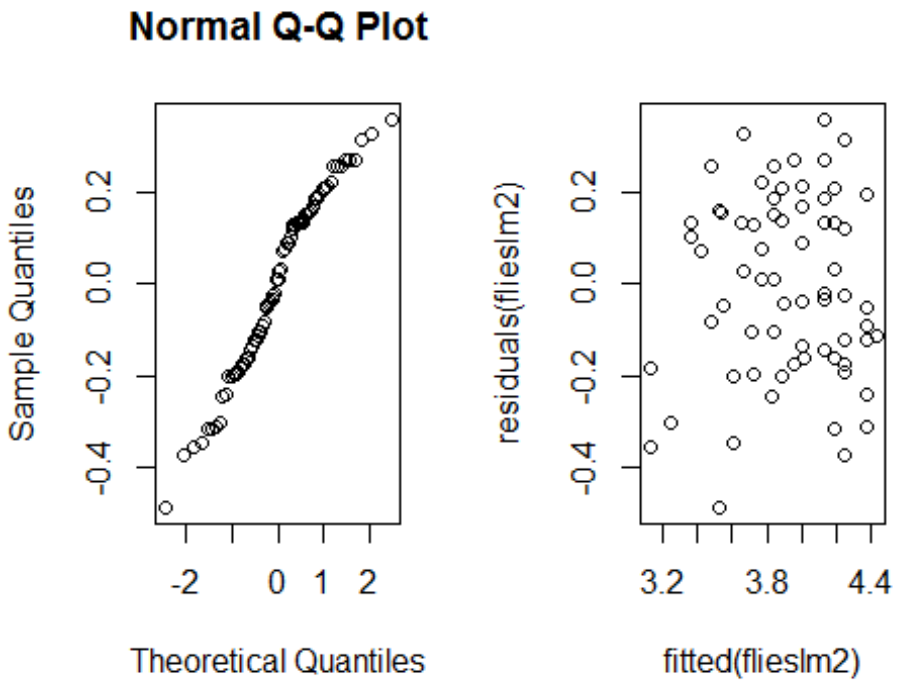

Lines have similar slope, which denotes that thorax influences similarly between activities.

## 8.

We prefer the analysis with thorax, as it is more complete. Also, it is possible to detect if there could be a difference in longevity between groups because of flies with significantly different thorax, instead of sexual activity. The analysis are fine, both factors do influence longevity.

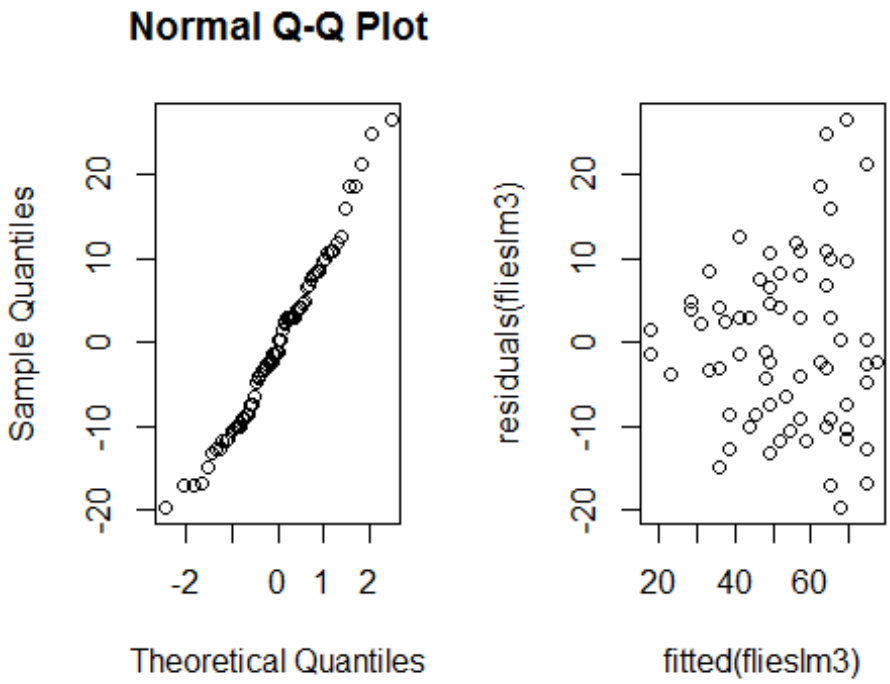
## 9.

```
par(mfrow=c(1,2))
qqnorm(residuals(flieslm2))
plot(fitted(flieslm2),residuals(flieslm2))
```




Normality is clearly met by the residuals, although heteroscedasticity is not that clear.

```
par(mfrow=c(1,2))
flieslm3 = lm(longevity~thorax+activity, data=fruitflies)
qqnorm(residuals(flieslm3))
plot(fitted(flieslm3),residuals(flieslm3))
```



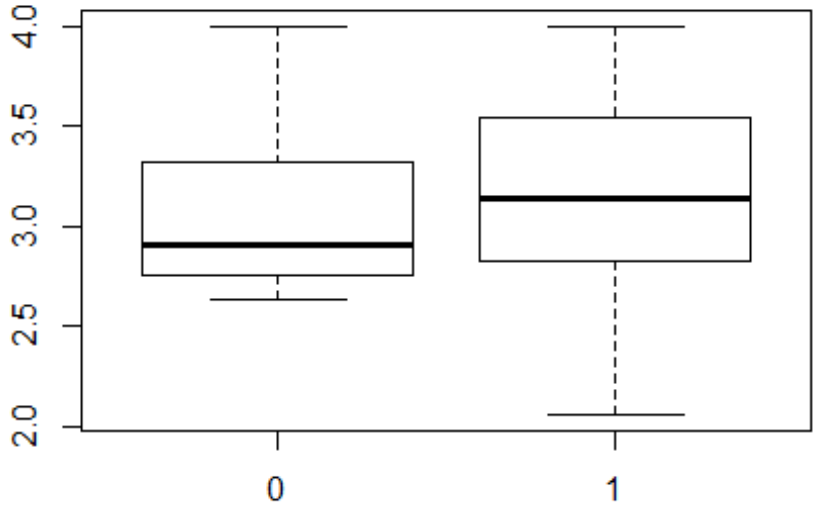

In this case heteroscedasticity seems more clear, which may mean that taking the logarithm of longevity was not a good decition.

## Exercise 2

```
psi=read.table('psi.txt', header=TRUE)
```

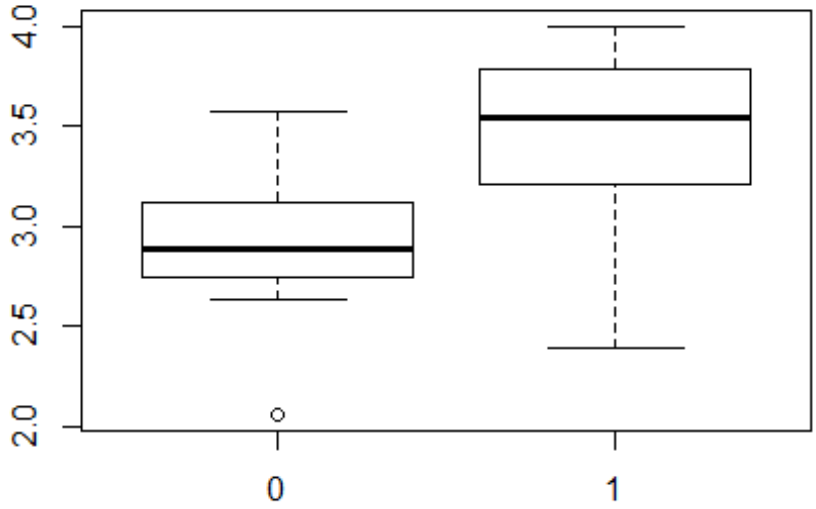
### 1.

```
boxplot(psi$gpa~psi$psi)
```




In this boxplot of gpa which is based on whether the student was instructed using psi or not, it can be seen that the ones with psi have a slightly better gpa, although the plot also shows the existence of some outliers.

```
boxplot(psi$gpa~psi$passed)
```




In this plot, we took the same approach as before but having whether the student passed or not. As expected, the students that passed have a significantly higher gpa.

```
xtabs(~passed+psi,data=psi)
```

```
##       psi
## passed  0  1
##      0 15  6
##      1  3  8
```

In this table we can see the relation between students that passed/not passed and students that had psi or not. It can be clearly seen how there are relatively more people that had psi and passed, than those studying traditionally that passed.

## 2.

```
psi$psi=factor(psi$psi)
psi$passed=factor(psi$passed)
psilm = glm(passed~psi+gpa, data=psi, family=binomial)
```

## 3.

```
drop1(psilm,test="Chisq")
```

```
## Single term deletions
##
## Model:
## passed ~ psi + gpa
##        Df Deviance    AIC    LRT Pr(>Chi)
```

```
## <none>      26.253 32.253
## psi     1   32.418 36.418 6.1647 0.013033 *
## gpa     1   35.342 39.342 9.0885 0.002572 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(psilm)
```

```
##
## Call:
## glm(formula = passed ~ psi + gpa, family = binomial, data = psi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8396  -0.6282  -0.3045   0.5629   2.0378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -11.602      4.213  -2.754  0.00589 **
## psi1           2.338      1.041   2.246  0.02470 *
## gpa            3.063      1.223   2.505  0.01224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 26.253  on 29  degrees of freedom
## AIC: 32.253
##
## Number of Fisher Scoring iterations: 5
```

According to the tests above, psi has a significant influence in whether the student passes or not, and the estimation on when psi is 1 is a positive number. Therefore, it that influence is positive, which would mean that psi does improve the learning of students.

### 4.

```
# Student with psi with a gpa is 3
1/(1+exp(-(-11.602 + 2.338*1 + 3.063*3)))
```

```
## [1] 0.4812588
```

```
# Student without psi with a gpa is 3
1/(1+exp(-(-11.602 + 2.338*0 + 3.063*3)))
```

```
## [1] 0.08218674
```

### 5.

```
exp(2.338)
```

```
## [1] 10.36049
```

The odds of passing the assignment rendered by students instructed with psi is 10.36049. This means that having studied using the psi method increases the chance of passing the assingment by 10.36049%. This value is independent from gpa.

### 6.

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

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.0265
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02016297 0.95505763
## sample estimates:
## odds ratio
##  0.1605805
```

15 is the number of students who did not receive psi and didn't showed improvement. 6 is the number of students who received psi and didn't show impovement. The hypothesis tested by Fisher's exact test is if there is independence between the 2 factors (having psi and improving). This is rejected (p-value < 0.05) which proves that there is a dependence between them, and concluding that psi is more helpful to improve than the previous teaching method.

### 7.

It does not take into account the amount of improvement found per student, but besides that it is a correct experiment.

### 8.

1st: advantage: It takes more information into account, and can estimate probability of passing having the data of the student. disadvantage: It is a more complex method and may be sensitive to outliers.

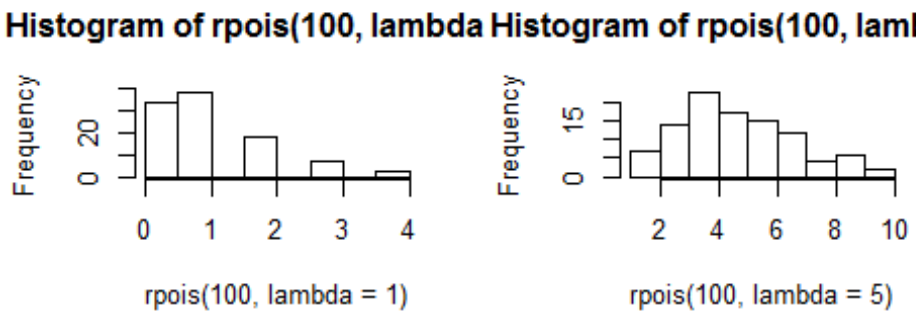
2nd: advantage: It is a simple method to find significance on improvement. disadvantage: It does not take into account the amount of improvement found per student

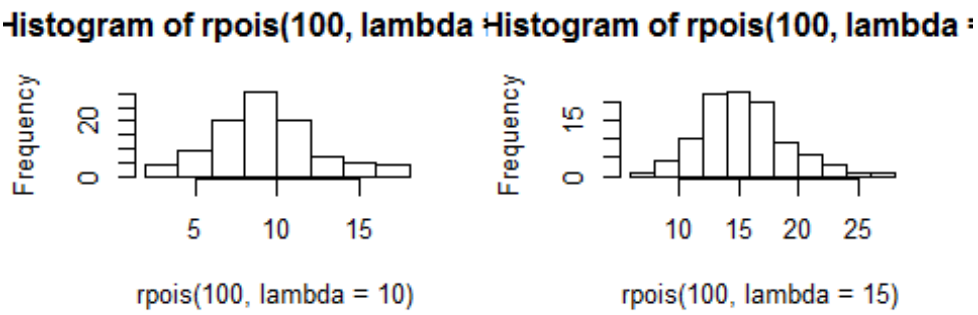
## Exercise 3

### 1.

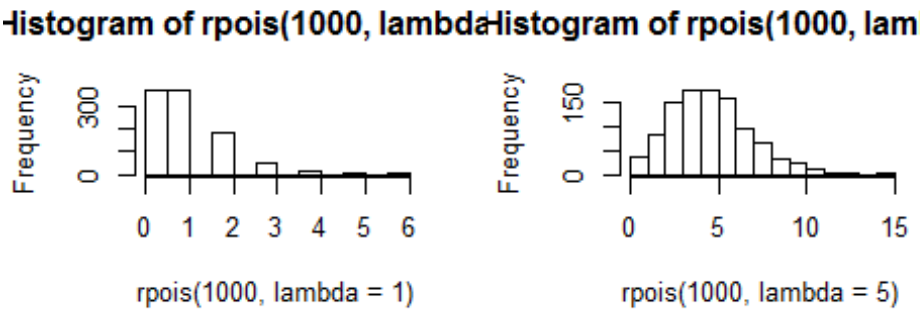
```
africa_data=read.table("africa.txt", header=TRUE)
par(mfrow=c(2,2))
hist(rpois(100, lambda = 1))
hist(rpois(100, lambda = 5))
```

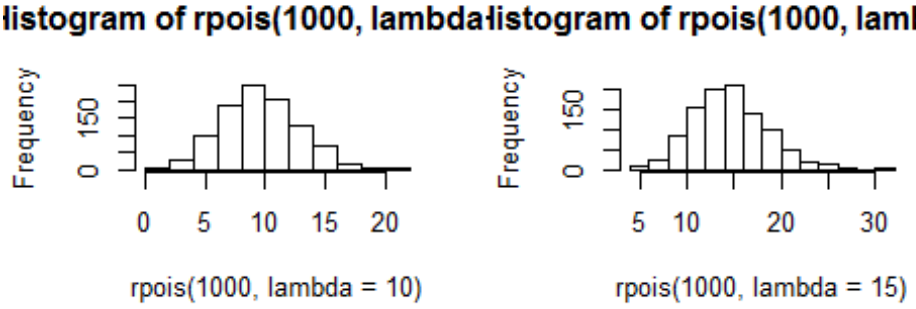
```r
hist(rpois(100, lambda = 10))
hist(rpois(100, lambda = 15))
```




```r
par(mfrow=c(2,2))
hist(rpois(1000, lambda = 1))
hist(rpois(1000, lambda = 5))
hist(rpois(1000, lambda = 10))
hist(rpois(1000, lambda = 15))
```

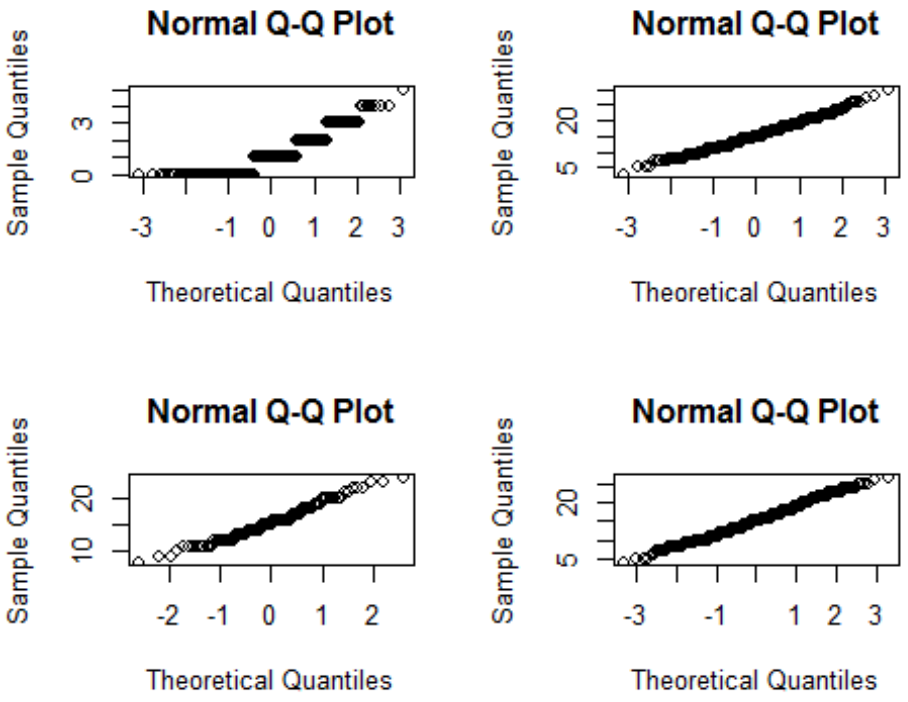



Increasing lambda and r pois seem to show a normal distribution.

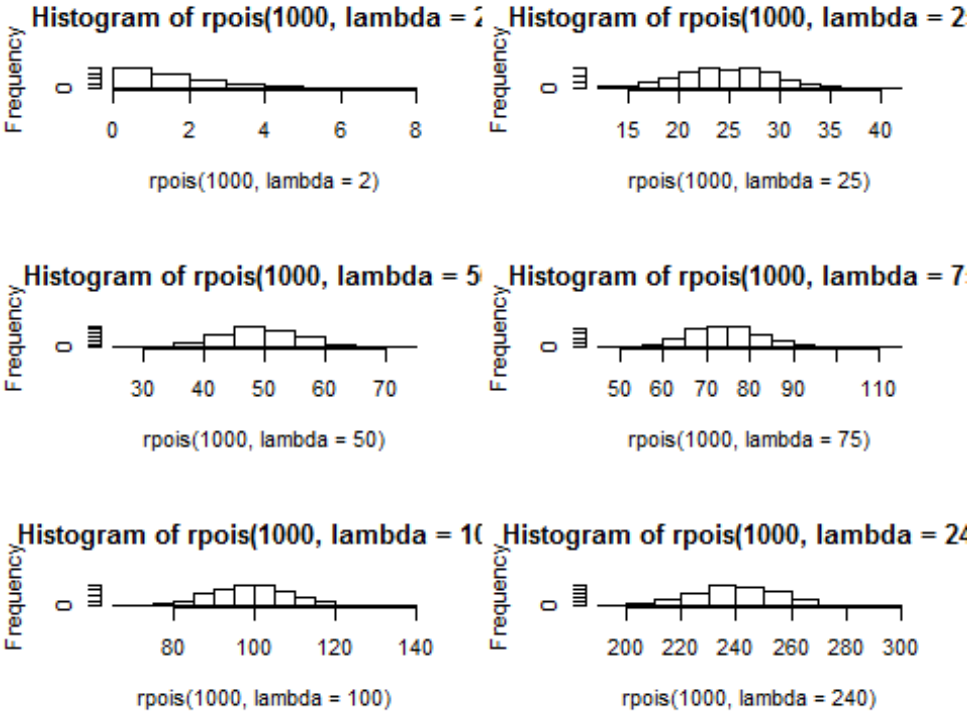
## 2.

```
par(mfrow=c(2,2))
qqnorm(rpois(500, lambda = 1))
qqnorm(rpois(500, lambda = 15))
qqnorm(rpois(100, lambda = 15))
qqnorm(rpois(1000, lambda = 15))
```




What can be seen from question one is that as lambda and n increases, the poisson distribution better approximates a normal distribution. This is seen in the histograms in question 1 and is further backed up by the QQ plots. As the poisson distribution gets closer to approximating the normal distribution, the poisson distributions would be in the same location-scale family as they would have properties similar to a normal distribution.

```
africa_data=read.table("africa.txt", header=TRUE)
par(mfrow=c(3,2))
hist(rpois(1000, lambda = 2))
hist(rpois(1000, lambda = 25))
hist(rpois(1000, lambda = 50))
hist(rpois(1000, lambda = 75))
hist(rpois(1000, lambda = 100))
hist(rpois(1000, lambda = 240))
```




## 3

```
#Analysisof variance
galaglm_full=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec
+numregim,family=poisson,data=africa_data)
error1 = summary(galaglm_full)$r.squared
summary(galaglm_full)

##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data =
africa_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3443  -0.9542  -0.2587   0.3905   1.6953
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5102693  0.9053301  -0.564  0.57301
## oligarchy    0.0730814  0.0345958   2.112  0.03465 *
## pollib      -0.7129779  0.2725635  -2.616  0.00890 **
## parties      0.0307739  0.0111873   2.751  0.00595 **
## pctvote      0.0138722  0.0097526   1.422  0.15491
## popn         0.0093429  0.0065950   1.417  0.15658
## size        -0.0001900  0.0002485  -0.765  0.44447
```

```
## numelec     -0.0160783  0.0654842  -0.246  0.80605
## numregim     0.1917349  0.2292890   0.836  0.40303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.668  on 27  degrees of freedom
## AIC: 111.48
##
## Number of Fisher Scoring iterations: 6
```

### 4.

Removing numelec has it had the highest p-value.

```
#Analysisof variance
galaglm=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size
+numregim,family=poisson,data=africa_data)
summary(galaglm)

##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numregim, family = poisson, data = africa_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3997  -0.9381  -0.2666   0.4220   1.6998
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.6078028  0.8239267  -0.738  0.46070
## oligarchy    0.0781368  0.0277656   2.814  0.00489 **
## pollib      -0.6773897  0.2290130  -2.958  0.00310 **
## parties      0.0296786  0.0102888   2.885  0.00392 **
## pctvote      0.0131290  0.0092895   1.413  0.15756
## popn         0.0089313  0.0063746   1.401  0.16120
## size        -0.0002021  0.0002436  -0.830  0.40682
## numregim     0.1758198  0.2210498   0.795  0.42639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.728  on 28  degrees of freedom
## AIC: 109.54
```

```
##
## Number of Fisher Scoring iterations: 5
```

Removing numregim.

```
galaglm=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size,family=poisso
n,data=africa_data)
summary(galaglm)

##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size, family = poisson, data = africa_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3522  -0.9651  -0.1945   0.4833   1.6179
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.1126871  0.5163030  -0.218 0.827228
## oligarchy    0.0859620  0.0259100   3.318 0.000908 ***
## pollib      -0.6894029  0.2278572  -3.026 0.002481 **
## parties      0.0291944  0.0101954   2.863 0.004190 **
## pctvote      0.0141588  0.0091980   1.539 0.123723
## popn         0.0062736  0.0053994   1.162 0.245272
## size        -0.0001950  0.0002425  -0.804 0.421378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 29.363  on 29  degrees of freedom
## AIC: 108.17
##
## Number of Fisher Scoring iterations: 5
```

Removing size.

```
galaglm=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn,family=poisson,dat
a=africa_data)
summary(galaglm)

##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn, family = poisson, data = africa_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.4109  -0.9943  -0.1399   0.5516   1.6125
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.244466   0.495708  -0.493  0.62190
## oligarchy    0.083168   0.025437   3.270  0.00108 **
## pollib      -0.652830   0.221234  -2.951  0.00317 **
## parties      0.029800   0.010294   2.895  0.00379 **
## pctvote      0.013842   0.009282   1.491  0.13591
## popn         0.005587   0.005378   1.039  0.29883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 30.044  on 30  degrees of freedom
## AIC: 106.85
##
## Number of Fisher Scoring iterations: 5
```

Removing popn.

```
galaglm=glm(miltcoup~oligarchy+pollib+parties+pctvote,family=poisson,data=afr
ica_data)
summary(galaglm)

##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote,
##     family = poisson, data = africa_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5456  -0.9841  -0.1881   0.5948   1.6705
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.093657   0.463279  -0.202  0.83979
## oligarchy    0.095358   0.022421   4.253 2.11e-05 ***
## pollib      -0.666615   0.217564  -3.064  0.00218 **
## parties      0.025630   0.009502   2.697  0.00699 **
## pctvote      0.012134   0.009056   1.340  0.18031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 31.081  on 31  degrees of freedom
```

```
## AIC: 105.89
##
## Number of Fisher Scoring iterations: 5
```

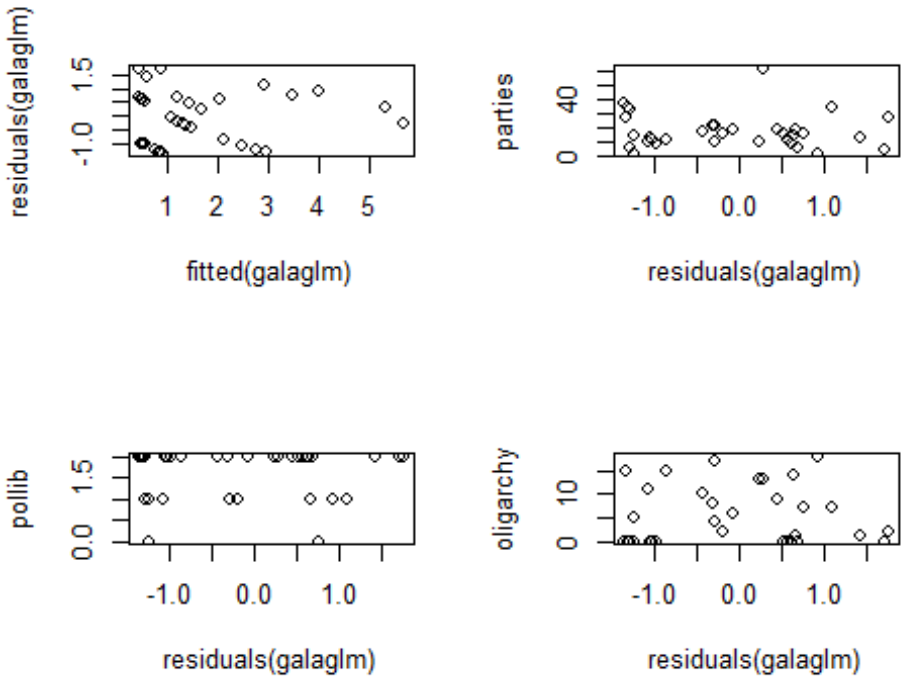
And finally removing pctvote.

```
galaglm=glm(miltcoup~oligarchy+pollib+parties,family=poisson,data=africa_data
)
summary(galaglm)
```

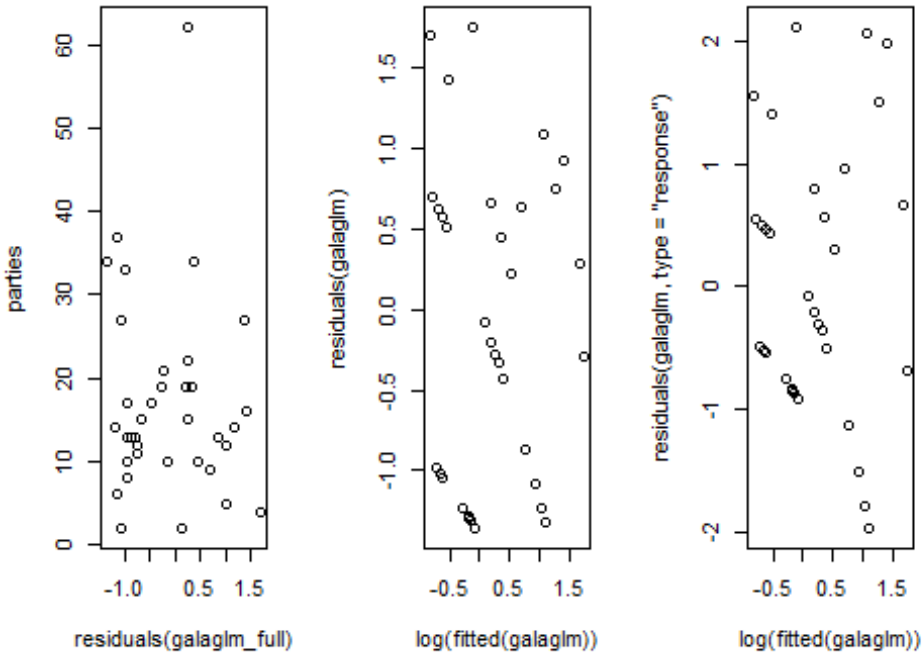
```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
##     data = africa_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3583  -1.0424  -0.2863   0.6278   1.7517
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.251377   0.372689   0.674  0.50000
## oligarchy    0.092622   0.021779   4.253 2.11e-05 ***
## pollib      -0.574103   0.204383  -2.809  0.00497 **
## parties      0.022059   0.008955   2.463  0.01377 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 32.856  on 32  degrees of freedom
## AIC: 105.66
##
## Number of Fisher Scoring iterations: 5
```

## 5.

```
par(mfrow=c(2,2))
attach(africa_data)
plot(fitted(galaglm),residuals(galaglm))
plot(residuals(galaglm),parties)
plot(residuals(galaglm),pollib)
plot(residuals(galaglm),oligarchy)
```



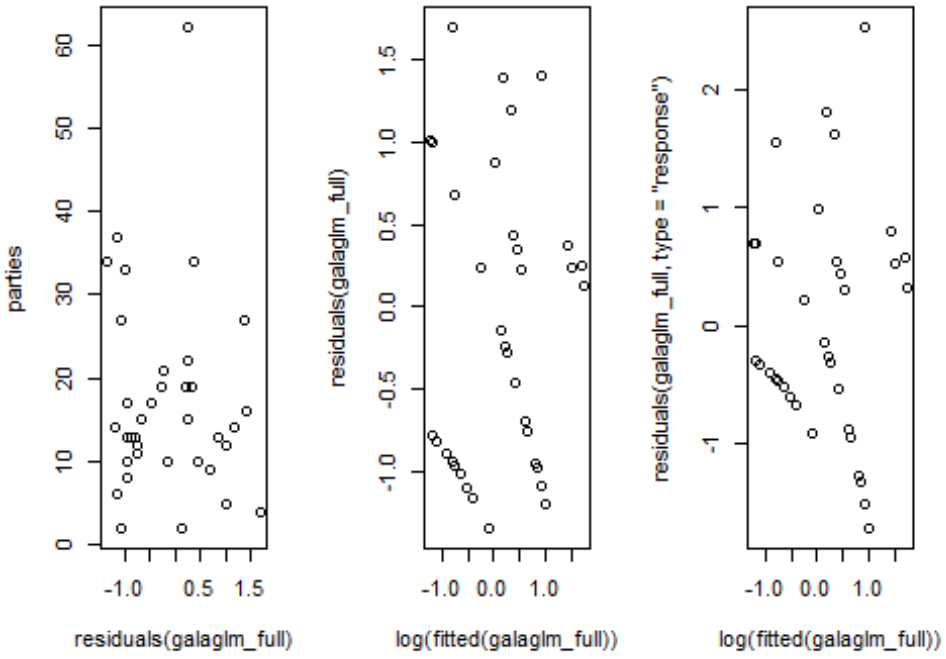

```
par(mfrow=c(1,3))
plot(residuals(galaglm_full),parties)
plot(log(fitted(galaglm)),residuals(galaglm))
plot(log(fitted(galaglm)),residuals(galaglm,type="response"))
```




The response residuals clearly increase with the (logarithm) of the fitted values, as expected under a Poisson model.

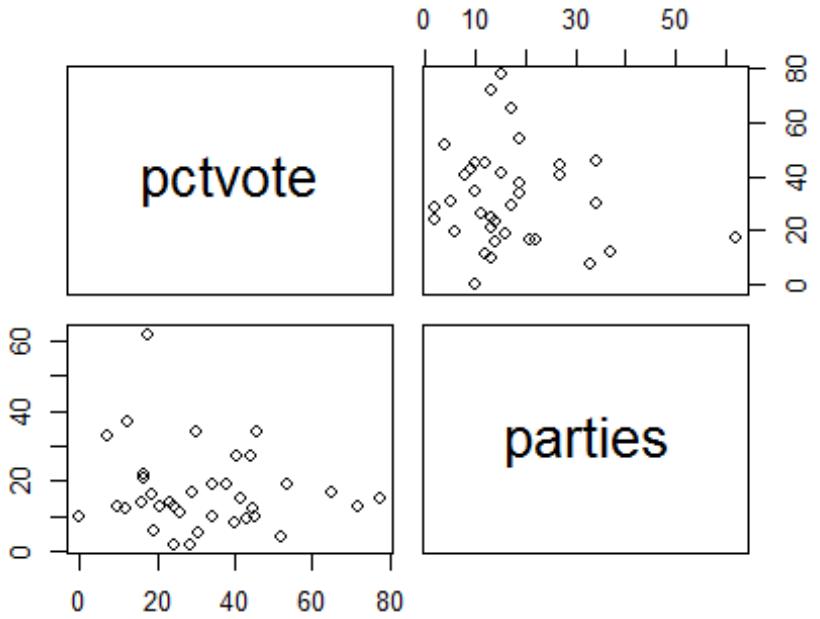
Next, the model from Q3 will be investigated to see if it follows the same pattern.

```
par(mfrow=c(1,3))
plot(residuals(galaglm_full),parties)
plot(log(fitted(galaglm_full)),residuals(galaglm_full))
plot(log(fitted(galaglm_full)),residuals(galaglm_full,type="response"))
```




While the response residuals clearly increase with the logarithm of the fitted values, the step down approach looks to contain less clear patterns in it's data. In the full model with all explanatory variables, there looks to be much clearer linear structures within the data compared to the step down model. This could be due to the removed variables. During the process of the step down approach in Q4 what can be seen is that some of the variables become less significant as variables are removed. For example, in the last step the variable pctvote is removed. This reduces the p-value of the variable parties. By removing pctvote, the relationship between pctvote and parties is removed which could affect the model if the is a dependency between these two variables.

```
pairs(pctvote~parties)
```




However, from the pairs graph there does not seem to be a strong correlation.