Assignment 4_Group11

Name: Jiamian Liu **VU student number:** 2632301
Name: Xiaoyu Yang **VU student number:** 2640948
Name: Fangzheng Lyu **VU student number:** 2644757

Exercise 1

1. The codes adding loglongevity column are shown below as figure 1.

```
> setwd("/Users/flora/Downloads/EDDA/assignment4/data")
> flies = read.table("fruitflies.txt",header = TRUE)
> loglongevity = log10(flies[,2])
> flies <- within(flies,{loglongevity <- loglongevity})
```

Figure 1: Codes

2. The plot is shown as figure 2, using the code

```
> plot(loglongevity~thorax,pch=as.character(activity))
```

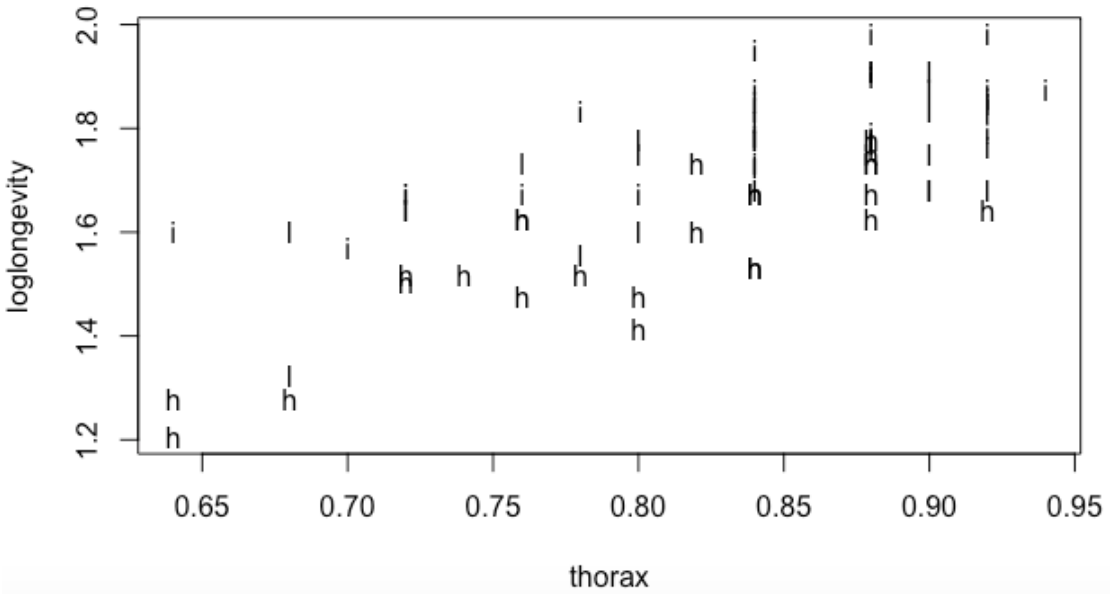

Figure 2: Informative plot

3. According to the question we can assume H0: Sexual activity do not influence longevity. And from the Anova we can see that p-values is smaller than 0.05, so H0 is rejected, which means sexual activity influences longevity.

```
> aovpen = lm(loglongevity~activity,data = flies)
> anova(aovpen)
Analysis of Variance Table

Response: loglongevity
          Df  Sum Sq Mean Sq F value    Pr(>F)
activity   2 0.69154 0.34577  19.421 1.798e-07 ***
Residuals 72 1.28191 0.01780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Anova

4. From the summary we can see that the sexual activity decrease longevity. And the estimated loglongevities for the high, low and isolated conditions are 1.56, 1.74 and 1.79. So the longevities for these three conditions are 36.31, 54.95 and 61.66.

```
> summary(lm(loglongevity~activity,data = flies))

Call:
lm(formula = loglongevity ~ activity, data = flies)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41489 -0.05792  0.01108  0.09073  0.21377

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.56438    0.02669  58.621  < 2e-16 ***
activityisolated  0.22463    0.03774   5.952 8.82e-08 ***
activitylow       0.17272    0.03774   4.577 1.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1334 on 72 degrees of freedom
Multiple R-squared:  0.3504,    Adjusted R-squared:  0.3324
F-statistic: 19.42 on 2 and 72 DF,  p-value: 1.798e-07
```

Figure 4: Summary

5. According to the question we can assume H0: Sexual activity do not influence longevity. And from the Anova we can see that p-values is smaller than 0.05, so H0 is rejected, which means sexual activity influences longevity (including thorax length as an explanatory variable).

```
> aovpen2 = lm(loglongevity~activity+thorax,data = flies)
> anova(aovpen2)
Analysis of Variance Table

Response: loglongevity
          Df  Sum Sq Mean Sq F value    Pr(>F)
activity   2 0.69154 0.34577  44.606 2.838e-13 ***
thorax     1 0.73155 0.73155  94.374 1.139e-14 ***
Residuals 71 0.55037 0.00775
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> drop1(aovpen2,test='F')
Single term deletions

Model:
loglongevity ~ activity + thorax
         Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>                 0.55037 -360.60
activity  2   0.39852 0.94888 -323.75  25.705 4.000e-09 ***
thorax    1   0.73155 1.28191 -299.19  94.374 1.139e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: Anova

6. From the summary we can see that the sexual activity decrease longevity. From the summary we could use the formula:

$$longevity = 10^{loglongevity} = 10^{0.53+0.18*\text{activityisolated}+0.12*\text{activitylow}+1.29*\text{thorax}}$$

And for a fly with average thorax length(0.82) the estimated longevities for the isolated, low and high conditions are 58.59, 51.03 and 38.71.
For a typical fly as small as the smallest in the data set(0.64), the estimated longevities for the high, low and isolated conditions are 34.32, 29.90 and 22.68.

```
> summary(lm(loglongevity~activity+thorax,data = flies))

Call:
lm(formula = loglongevity ~ activity + thorax, data = flies)

Residuals:
      Min        1Q    Median        3Q       Max
-0.210991 -0.069993  0.004518  0.065568  0.155204

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.52938    0.10799   4.902 5.79e-06 ***
activityisolated  0.17805    0.02536   7.021 1.07e-09 ***
activitylow       0.12408    0.02540   4.885 6.18e-06 ***
thorax            1.29376    0.13318   9.715 1.14e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08804 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
```

Figure 6: Summary

7. From the Figure 7 we can see longevity generally increases with increasing thorax, but this dependence is not the same under three conditions of sexual activities.

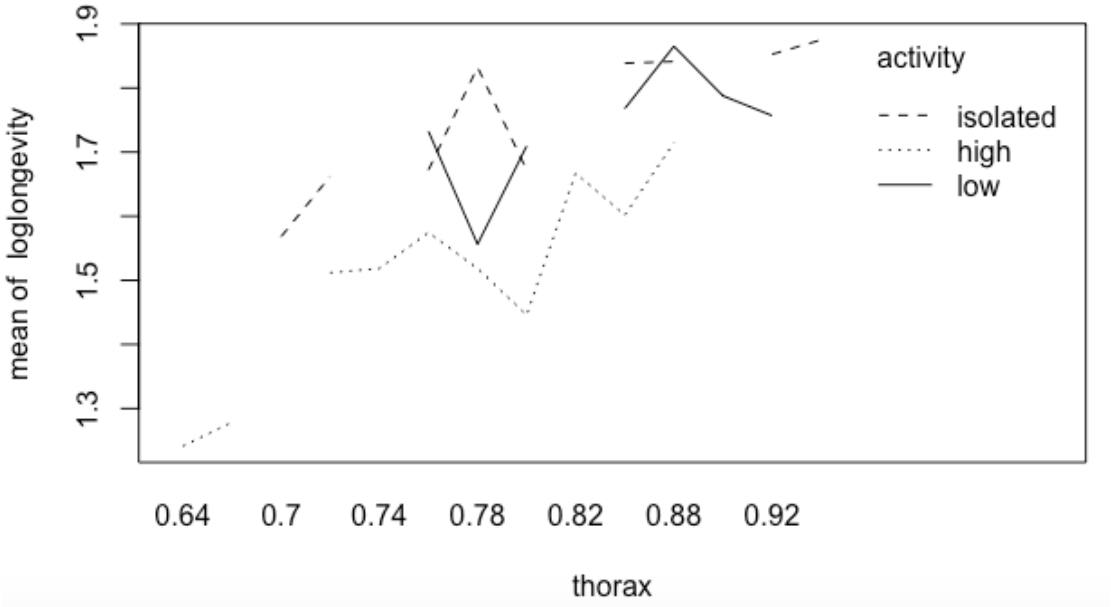

Figure 7: Interaction plot

8. I prefer the analyses including thorax length as an explanatory variable. Because for this test the sexual activity can be fixed at three levels and the thorax of flies can not be controlled, dependence of longevity on the numerical variable thorax is a-priori evident, and the variable is included only to increase the precision of the analysis. And we can also consider the uncontrollable factors using ANCOVA.
9. The QQ-plot of residuals and fitted plot are shown as Figure 8 and 9. From the QQ-plot we can see it is approximating a straight line and from the Shapiro.test we can see the p-value is greater than 0.05. And also we can see the p-value of studentized Breusch-Pagan test is also greater than 0.05. So the sample can be considered as normal and have no heteroscedasticity.

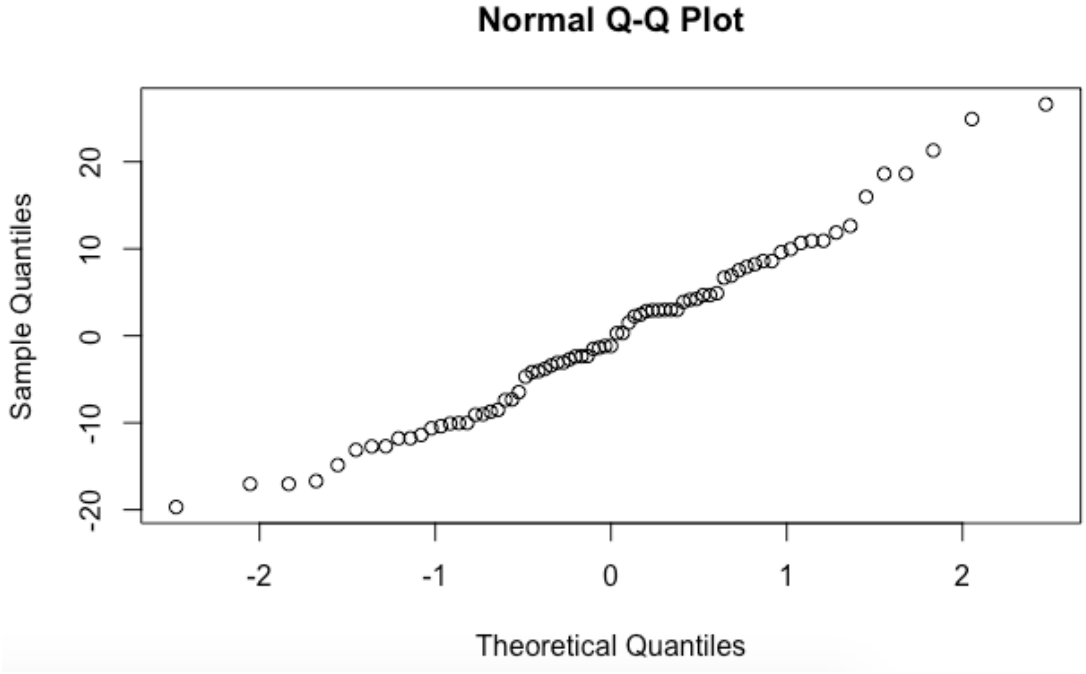

Figure 8: QQ-plot

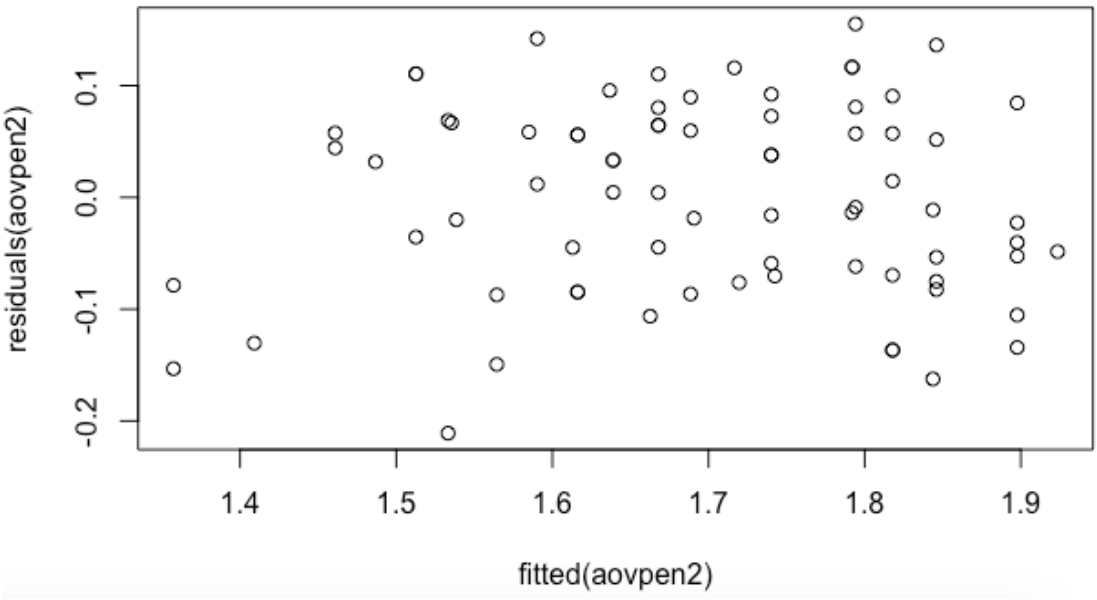

Figure 9: Fitted plot

```
> shapiro.test(residuals(aovpen2))

        Shapiro-Wilk normality test

data:  residuals(aovpen2)
W = 0.96838, p-value = 0.05748
```

Figure 10: Shapiro.test

```
> bptest(aovpen2)

        studentized Breusch-Pagan test

data:  aovpen2
BP = 2.5333, df = 3, p-value = 0.4693
```

Figure 11: bptest

10. We can use the parameter generated from ANCOVA in Figure 11 to build the following formula:

    $longevity = -67.375 + 20.066 * activityisolated + 13.054 * activitylow + 132.62 * thorax$.

    And as p-value of shapiro-test is greater than 0.05 and p-value of studentized Breusch-Pagan test is smaller than 0.05. So the residuals of longevity is normal but have heteroscedasticity, thus it's better to use loglongevity rather than use longevity.

```
> aovpen3 = lm(longevity~activity+thorax,data = flies)
> summary(aovpen3)

Call:
lm(formula = longevity ~ activity + thorax, data = flies)

Residuals:
    Min      1Q  Median      3Q     Max
-19.688  -8.622  -1.176   6.790  26.605

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -67.375     12.750  -5.284 1.33e-06 ***
activityisolated  20.066      2.994   6.701 4.13e-09 ***
activitylow       13.054      2.999   4.352 4.43e-05 ***
thorax           132.618     15.725   8.434 2.62e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.4 on 71 degrees of freedom
Multiple R-squared:  0.6749,    Adjusted R-squared:  0.6611
F-statistic: 49.12 on 3 and 71 DF,  p-value: < 2.2e-16
```

Figure 12: Summary

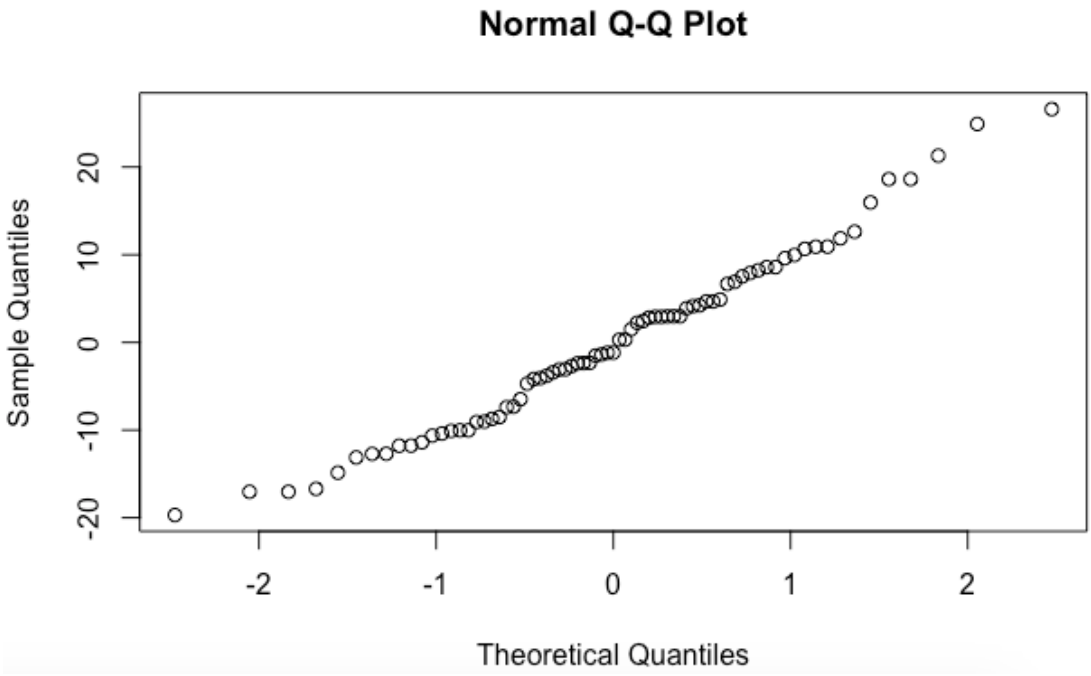

Figure 13: QQ-plot

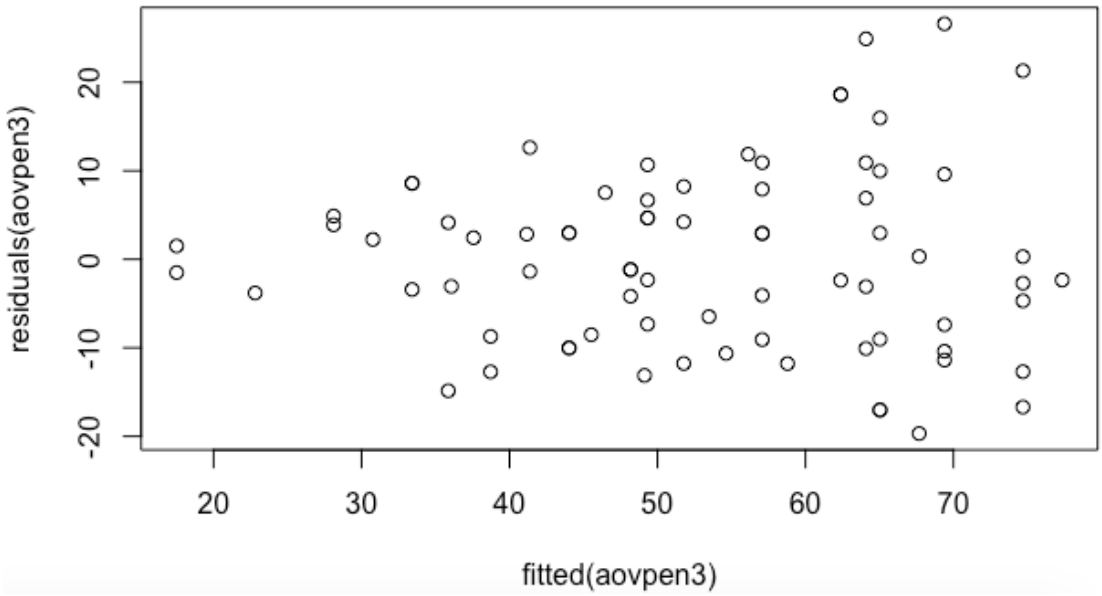

Figure 14: Fitted plot

```
> shapiro.test(residuals(aovpen3))

        Shapiro-Wilk normality test

data:  residuals(aovpen3)
W = 0.98091, p-value = 0.3176
```

Figure 15: Shapiro.test

```
> bptest(aovpen3)

        studentized Breusch-Pagan test

data:  aovpen3
BP = 10.516, df = 3, p-value = 0.01465
```

Figure 16: bptest

Exercise 2

1. After loading data, we first try to summary the percentage of individuals with cancer for every combination of psi. The code and output is shown as follows:

```
> teach=read.table("psi.txt",header=TRUE)
> tot=xtabs(~psi,data=teach)
> tot
psi
 0  1
18 14
> round(xtabs(passed~psi,data=teach)/tot,2)
psi
   0    1
0.17 0.57
```

Figure 16: Code and output of passed~psi

From Fig.16 we could know that if students receive psi, they will get a much higher pass rate.

We also show the gpa distribution; the code is: hist(teach$gpa)

The histogram is shown in Fig.17.

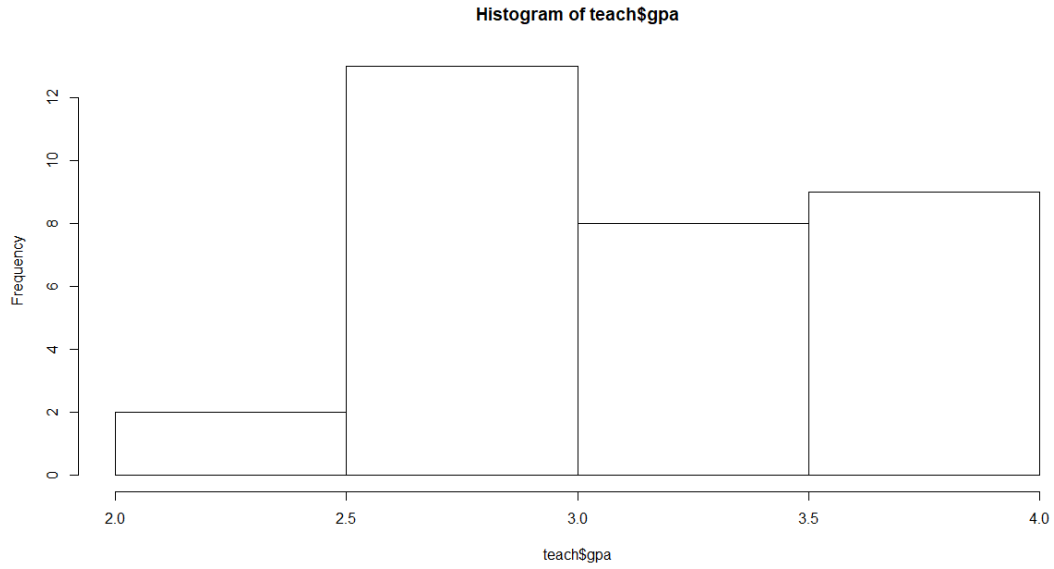

Figure 17: gpa distribution

However, the gpa varies with each other. By classifying gpa we could show the pass per gpa group. We add a new column for the origin data, named “grade”. We define gpa in (3.5,4] is grade “A”, gpa in (3,3.5] is grade “B”, gpa in (2.5,3] is grade “C” and gpa in [2,2.5] is grade D. We then build a new table. The code and the first 5 row of the new table are shown in Fig.18.

```
> grade <- c()
> for (count in 1:32){
+   if(teach[count,3]>3.5){grade[count]<-"A"}
+   else if(teach[count,3]>3){grade[count]<-"B"}
+   else if(teach[count,3]>2.5){grade[count]<-"C"}
+   else {grade[count]<-"D"}
+ }
> newteach<-data.frame(teach,grade)
> newteach[1:5,]
  passed psi  gpa grade
1      0   0 2.66     C
2      0   0 2.89     C
3      0   0 3.28     B
4      0   0 2.92     C
5      1   0 4.00     A
```

Figure 18: New table with column "grade"

Now we could summary the percentage of individuals with cancer for every combination of grade. The code and output is shown as follows:

```
> totgrade=xtabs(~grade,data=newteach)
> round(xtabs(passed~grade,data=newteach)/totgrade,2)
grade
   A    B    C    D
0.67 0.38 0.08 0.50
```

Figure 19: Code and output of passed~grade

We could also summary the percentage per grade-group. The code and the barplot for are shown in Fig.20 and Fig.21.

```
> barplot(xtabs(psi~grade,data=teach)/totgrade)
```

Figure 20: Code for barplot

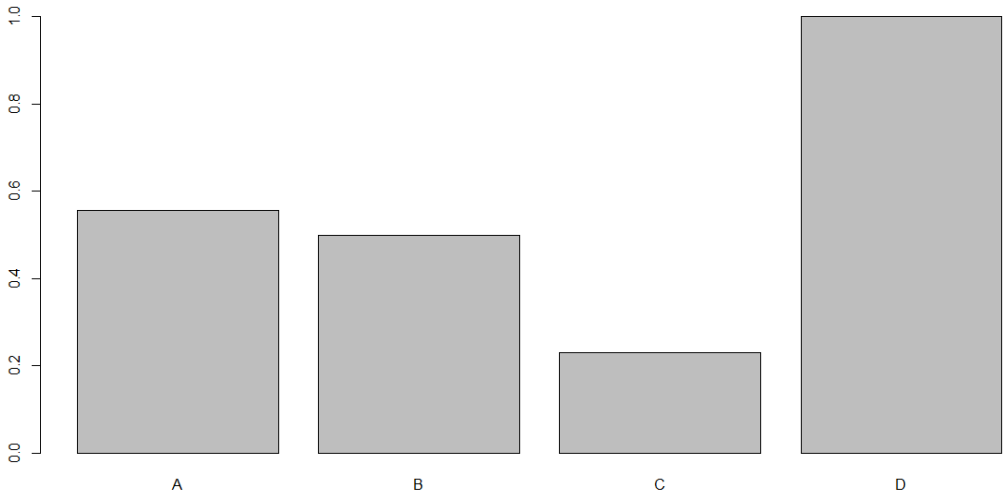

Figure 21: Barplot for grade group

2. We use function glm to generate the logistic model, the code and output is as shown in Fig.22.

```
> resultglm=glm(passed~psi+gpa,data=teach,family=binomial)
> summary(resultglm)

Call:
glm(formula = passed ~ psi + gpa, family = binomial, data = teach)

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 **
psi            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
```

Figure 22: Code and output for logistic model

From the output we could know that the result is:

Pr(passed=1) = ψ (-11.602 + 2.338*psi + gpa*3.063 + error)

Where ψ (x) is logistic function and is $\psi(x)=1/(1+e^{-x})$

3. According to the glm function, the positive signs of the parameter estimates mean that higher values of these variables give higher probability of "passed". Because "psi" value is 2.338>0 and is a necessary sign, the psi thus works. Besides, the p-value of "psi is 0.02470, also shows "psi" has important influence on "passed".

4. According to the formula in 2, when the gpa of a student equal to 3 who receives psi, means gpa=3 and psi=1, the calculation and results are shown in Fig.23.

```
> c=-11.602 + 2.338*1 + 3*3.063
> Pr=1/(1 + exp(1)^-c)
> Pr
[1] 0.4812588
>
```

Figure 23: Probability for student gpa=3, receives psi

We could know in that situation; the probability is about 0.481.

According to the formula in 2, when the gpa of a student equal to 3 who does not receive psi, means gpa=3 and psi=0, the calculation and results are shown in Fig.24.

```
> c=-11.602 + 2.338*0 + 3*3.063
> Pr=1/(1 + exp(1)^-c)
> Pr
[1] 0.08218674
```

Figure 24: Probability for student gpa=3, does not receive psi

We could know in that situation; the probability is about 0.082.

5. When instructing an arbitrary student with psi rather than the standard method means the psi from 0 changes to 1, means the linear predictor by 2.338 and increases the odds of cancer by a factor e^2.33 8= 10.36049.
   The interpretation of this number is the increase of odds, which is an outcome that has probability p of arising, which are defined as: o=p/(1-p). The odds mean (1) the betting game in which you gain 1 unit if the outcome occurs (which has probability p) and loose o units otherwise (an "1 − o against bet") is fair if o is the odds, which is increased via "psi" changes; (2) if the odds is k, then the probability of winning is k times as big as the probability of loosing, which is increased via "psi" changes.)
   The increase is independent on gpa, because gpa is not changed so it does contribute to the increase of odds according to the representation of formula.

6. 15 means the number of students who did not receive psi and did not show improvement. 6 means the number of students who received psi but did not show improvement. The result is shown in Fig.25.

```
> x=matrix(c(3,15,8,6),2,2)
> x
     [,1] [,2]
[1,]    3    8
[2,]   15    6
> 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
```

Figure 25: Results of fisher test

The conclusion is that the p-value is 0.0265, which means the variables are not independent, we reject the null hypothesis H0: p1 = p2, where The experiments in the first sequence have success probability p1, those in the second p2.

7. The second approach is not wrong, because properties and quantities of the data meets the requirements of establishing the fisher test. It shows the dependency between two group of data and the test result instructs us to reject or not reject the H0 hypothesis.

8. First approach: advantage: logistic regression works well for predicting categorical outcomes and predict multinomial outcomes.

First approach: disadvantage: Logistic regression attempts to predict outcomes based on a set of independent variables, but logit models are vulnerable to overconfidence. That is, the models can appear to have more predictive power than they actually do as a result of sampling bias.
Second approach advantage: Fisher's test is a statistical significance test used in the analysis of contingency tables. The significance of the deviation from a null hypothesis can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity.
Second approach disadvantage: it is conservative. its actual rejection rate is below the nominal significance level. It is not good to use of fixed significance levels when dealing with discrete problems.

EXERCISE 3

3.1

Firstly, we change the n. The number of times an event occurs in an interval increase and the distribution is more like the theoretical shape with the n increasing.

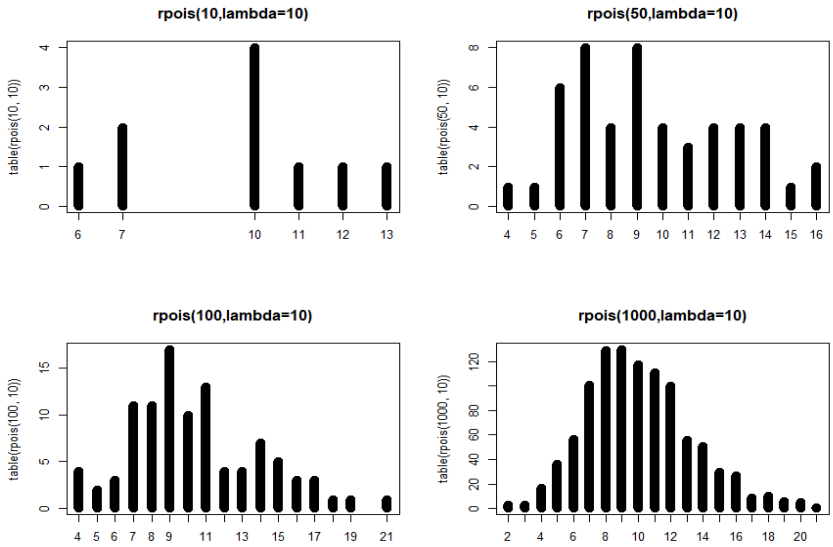

Fig21. change the n 10, 50,100, 1000

Secondly, we change the lambda from 2 to 20. We can find that with the lambda increasing, the distribution is more like normal distribution, which means for very large λ the Poisson(λ)-distribution is approximately equal to a normal distribution with mean $\mu = \lambda$ and variance $\sigma 2 = \lambda$.

Then, the mean of distribution increase with the larger λ value. The mean and variance of the Poisson-distribution both equal λ. Hence, the larger the parameter, the larger the values of Y on average and the larger the spread in the values of Y.

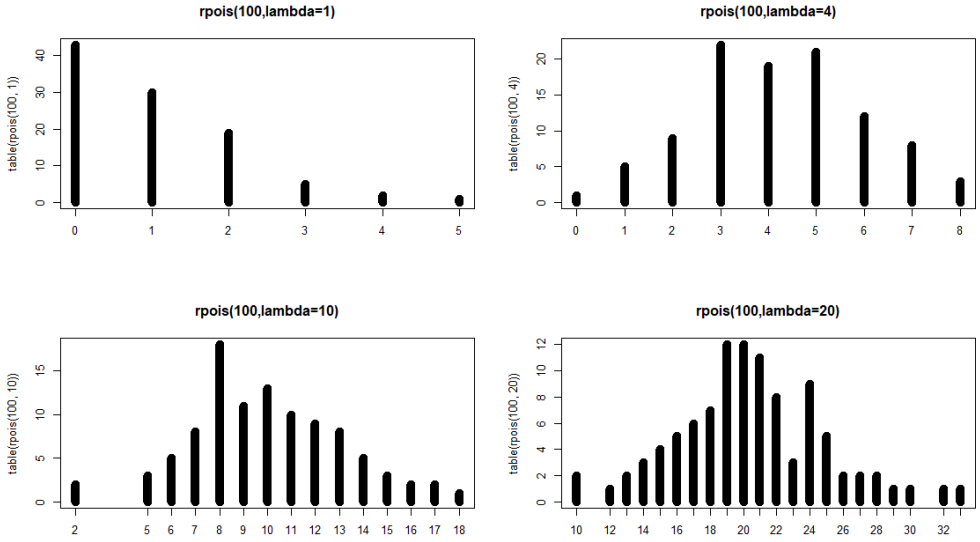

Fig.22 change the lambda 2, 4,10, 20

Code of 3.1:

```
par(mfrow=c(2,2))
plot(table(rpois(100,1)), type = "h", col = "black", lwd=10,main="rpois(100,lambda=1)")
plot(table(rpois(100,4)), type = "h", col = "black", lwd=10,main="rpois(100,lambda=4)")
plot(table(rpois(100,10)), type = "h", col = "black", lwd=10,main="rpois(100,lambda=10)")
plot(table(rpois(100,20)), type = "h", col = "black", lwd=10,main="rpois(100,lambda=20)")

par(mfrow=c(2,2))
plot(table(rpois(10,10)), type = "h", col = "black", lwd=10,main="rpois(10,lambda=10)")
plot(table(rpois(50,10)), type = "h", col = "black", lwd=10,main="rpois(50,lambda=10)")
plot(table(rpois(100,10)), type = "h", col = "black", lwd=10,main="rpois(100,lambda=10)")
plot(table(rpois(1000,10)), type = "h", col = "black", lwd=10,main="rpois(1000,lambda=10)")
```

3.2

The different Poission distributions cannot be in the same location-scale family. The reason is that the change of λ will lead a very different shape of distribution. For example, when we use λ=1, most of the values are in the left. However, when λ increases to 20, most of the values are in the middle (similar to normal distribution).

The random variable X and distribution function of Y = a + b X also belongs to the location scale family. In Poisson-regression, the parameter λ is modelled as: log λ = β0 + β1X1 + . . . βpXp. where the expression on the right indicates the combination of explanatory variables.

For each observation Y the parameter λ is modelled differently, since the corresponding values of X1, . . ., Xp will differ in general. Hence, the variances in different observations are different as well. This means that residuals Yn − Y ˆn do not come from one fixed distribution. Therefore, a normal QQ-plot of these response residuals is not useful.)

Instead, the deviance residuals are useful for diagnostic plots. Deviance is a measure of the discrepancy between the full model and the model under consideration. Deviance residuals are response residuals scaled by the deviance of that observation.

3.3

First, we study the collinearity. We can see some collinearity between the miltcoup and other explanatory variables.

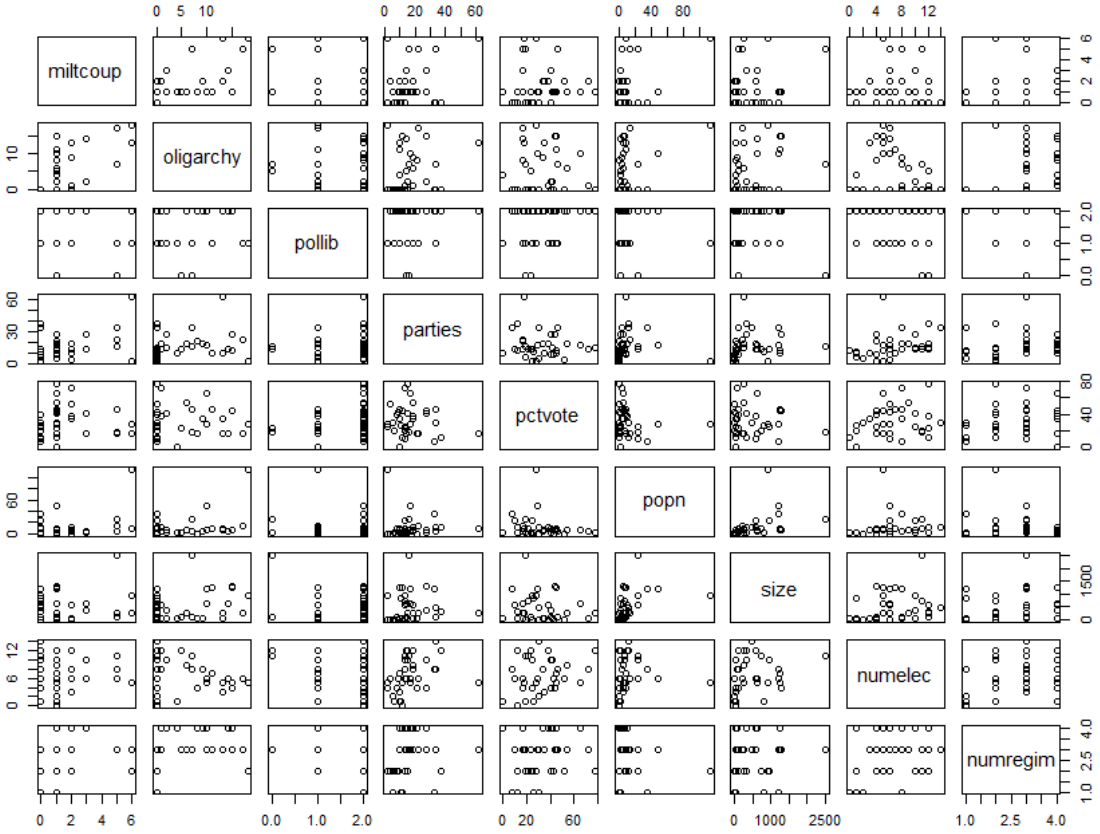

Fig.23 Collinearity of Africa

Then, from summary we can find that oligarchy, pollib, parties have significant effet on number years country ruled by military oligarchy (p-value <0.05, reject H0). The positive signs of the parameter (oligarchy, parties) estimates mean that higher values of these variables give higher numbers of muiltcoup. Oppositely, negative signs of the parameter (pollib) estimates mean that lower values of these variables give lower numbers of muiltcoup.

```
> summary(africaglm)

Call:
glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
    popn + size + numelec + numregim, family = poisson, data = africa)

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
```

The output shows 95 % confidence intervals.

```
> confint(africaglm)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept) -2.4335049109  1.148089620
oligarchy    0.0045915288  0.141483576
pollib      -1.2570629668 -0.182012570
parties      0.0080568606  0.052321186
pctvote     -0.0054171503  0.032940743
popn        -0.0038404317  0.022244262
size        -0.0007146351  0.000272539
numelec     -0.1438197483  0.114689702
numregim    -0.2632334399  0.643070807
```

The coefficients table shows that oligarchy, parties and numregim have positive relevant with number years country ruled by military oligarchy. Pollib has negative relevant with number years country ruled by military oligarchy.

```
> coef(africaglm)
  (Intercept)     oligarchy        pollib       parties       pctvote
-0.5102692854  0.0730813725 -0.7129778804  0.0307739289  0.0138722128
         popn          size       numelec      numregim
 0.0093429334 -0.0001899975 -0.0160783349  0.1917349158
```

Code of 3.3:

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

confint(africaglm)
coef(africaglm)
```

3.4

We use step down approach to reduce the step-down approach (delete 'numelec' → delete 'numregim' → delete 'size' → delete 'popn' → delete 'pctvote'). Finally, all explanatory variables (oligarchy, pollib, parties) in the model are significant.

```
> summary(africaglm_step)

Call:
glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
    data = africa)

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
```

Code of 3.4:

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