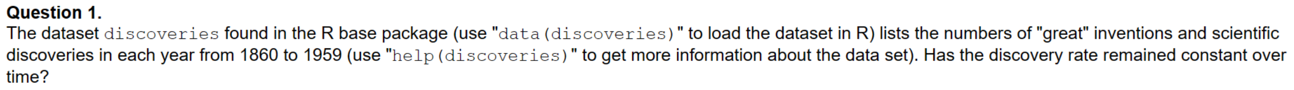
Discrete analysis_HW4

ID : 111024517 Name : 鄭家豪

Problem 1

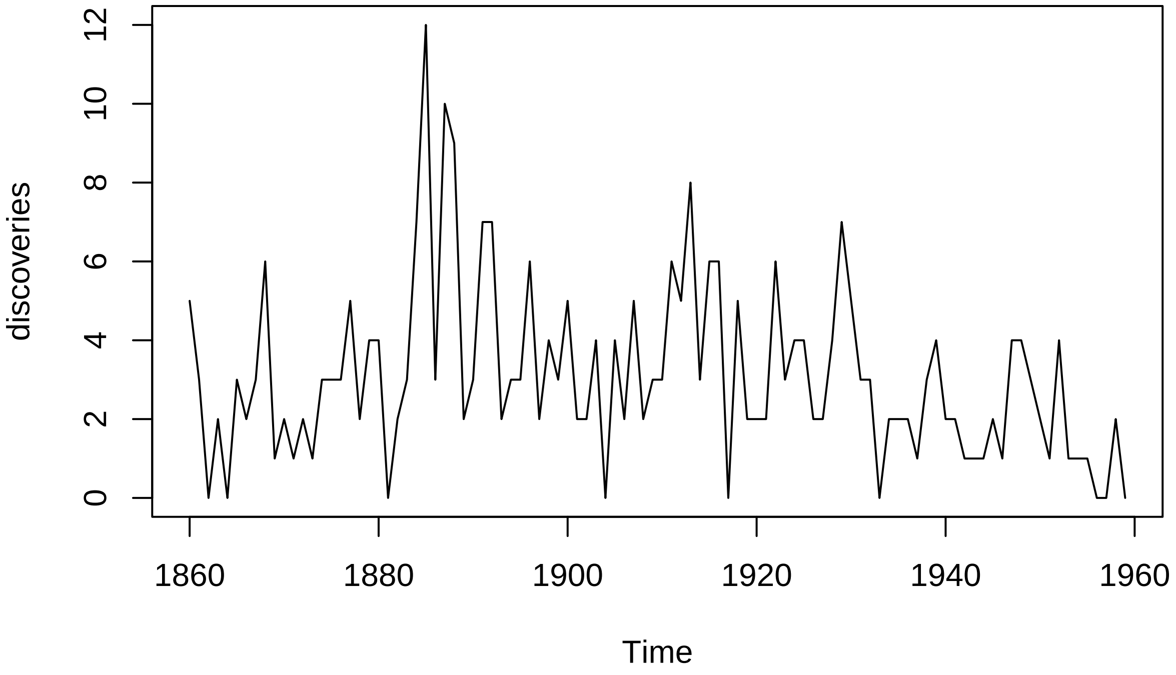


Question 1.
The dataset `discoveries` found in the R base package (use "`data(discoveries)`" to load the dataset in R) lists the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959 (use "`help(discoveries)`" to get more information about the data set). Has the discovery rate remained constant over time?

Sol.

First, we observe the time series plot of 'discoveries' data by 'ts.plot' funciton in R:

```
data(discoveries)
ts.plot(discoveries)
```



Averagelly, there seems to be a tendency for discovery counts (y_t)to decrease as time increases. Then, with a significant level 0.05, do the hypothesis test to compare two models:

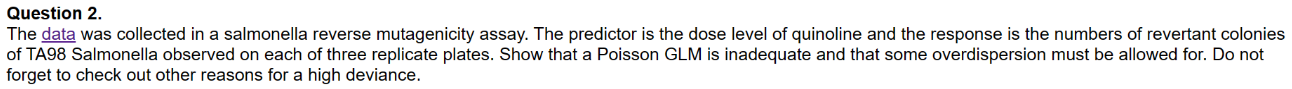
$$H_0 : \text{null model(only intercept)} v.s. H_1 : log(y_t) = \beta_0 + \beta_1 \times time$$

```
time = 1860:1959
model1 = glm( discoveries ~ 1+ time, data= discoveries, family ="poisson")
summary(model1)
```

```
Call:
glm(formula = discoveries ~ 1 + time, family = "poisson", data = discoveries)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.354807   3.775677   3.007  0.00264 **
time        -0.005360   0.001982  -2.705  0.00683 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 164.68  on 99  degrees of freedom
Residual deviance: 157.32  on 98  degrees of freedom
AIC: 430.32

Number of Fisher Scoring iterations: 5
```

The test statistic, $D_{H_0} - D_{H_1} = 164.68 - 157.32 = 7.36 \geq \chi^2_{0.05,1} = 3.841459$, so we reject H_0 at level 0.05. That is, it is confident to reject 'discoveries rate is not constant over time' at level 0.05.

Problem 2



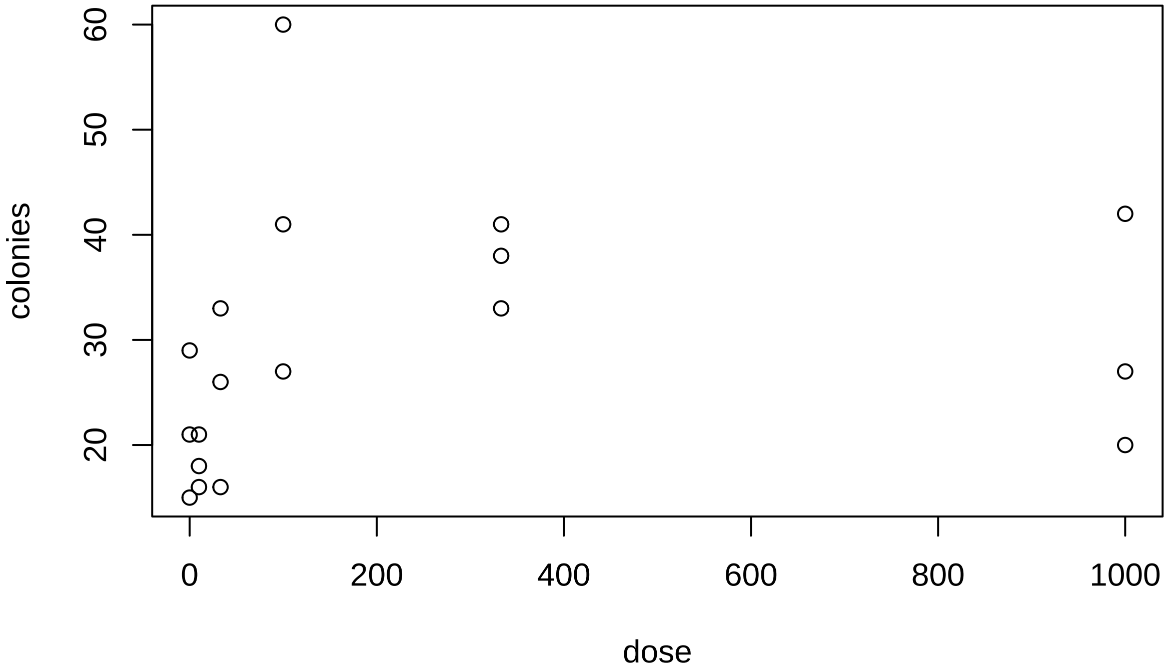
Question 2.
The data was collected in a salmonella reverse mutagenicity assay. The predictor is the dose level of quinoline and the response is the numbers of revertant colonies of TA98 Salmonella observed on each of three replicate plates. Show that a Poisson GLM is inadequate and that some overdispersion must be allowed for. Do not forget to check out other reasons for a high deviance.

Sol.

```
data = read.table("salmonella.txt")
dim(data) ; unique(data$dose)
```

```
[1] 18  2

[1]    0   10   33  100  333 1000
```

```
plot(x = data$dose,y = data$colonies, xlab= "dose", ylab = "colonies")
```



First, We should convert dose variable into an ordered factor variable:

```
data$dose = as.ordered(factor(data$dose))
unique(data$dose )
```

```
[1] 0    10   33   100  333  1000
Levels: 0 < 10 < 33 < 100 < 333 < 1000
```

Then, fit the Poisson model with significant level 0.05:

```
model = glm(formula = colonies ~ ., data = data, family = "poisson")
summary(model)
```

```

Call:
glm(formula = colonies ~ ., family = "poisson", data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.32778    0.04561  72.968  < 2e-16 ***
dose.L       0.50669    0.11530   4.394 1.11e-05 ***
dose.Q      -0.22791    0.11078  -2.057 0.039668 *  
dose.C      -0.41331    0.11383  -3.631 0.000282 ***
dose^4       0.15583    0.11263   1.384 0.166501    
dose^5       0.13253    0.10577   1.253 0.210187    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 78.358  on 17  degrees of freedom
Residual deviance: 33.496  on 12  degrees of freedom
AIC: 138.03

Number of Fisher Scoring iterations: 4
```

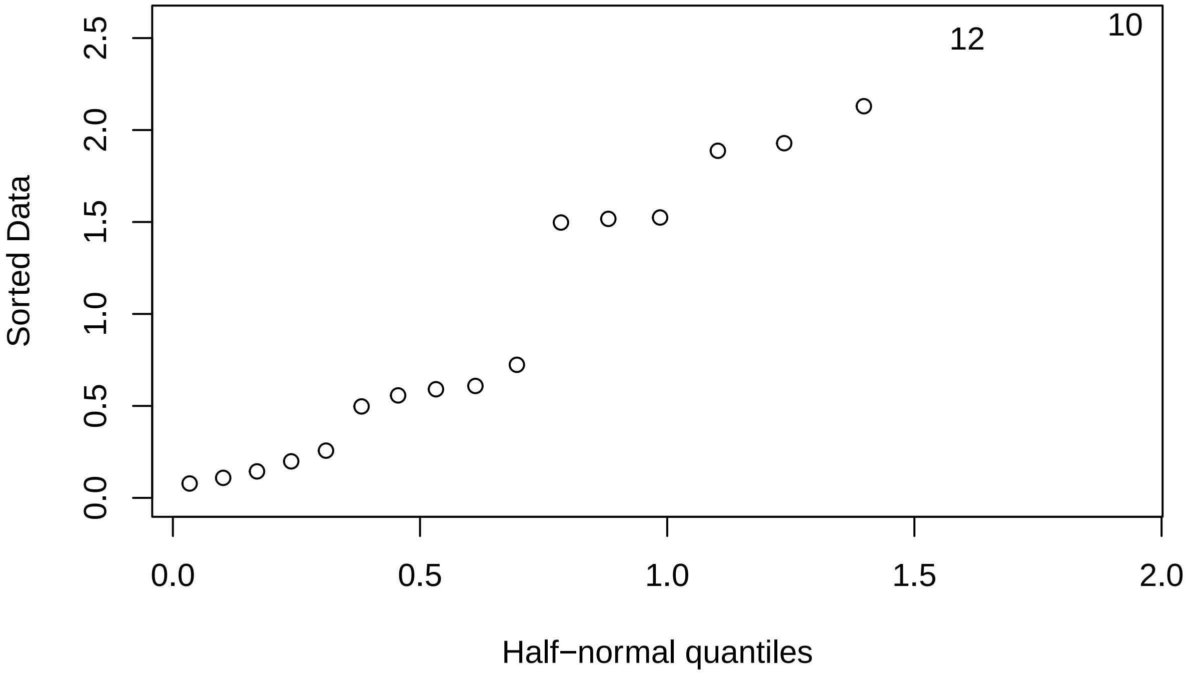
```
model$deviance > qchisq(p = 0.95,df = 12)
```

```
[1] TRUE
```

The deviance is too larger, then we should check whether this data has outliers or may be overdispersed:

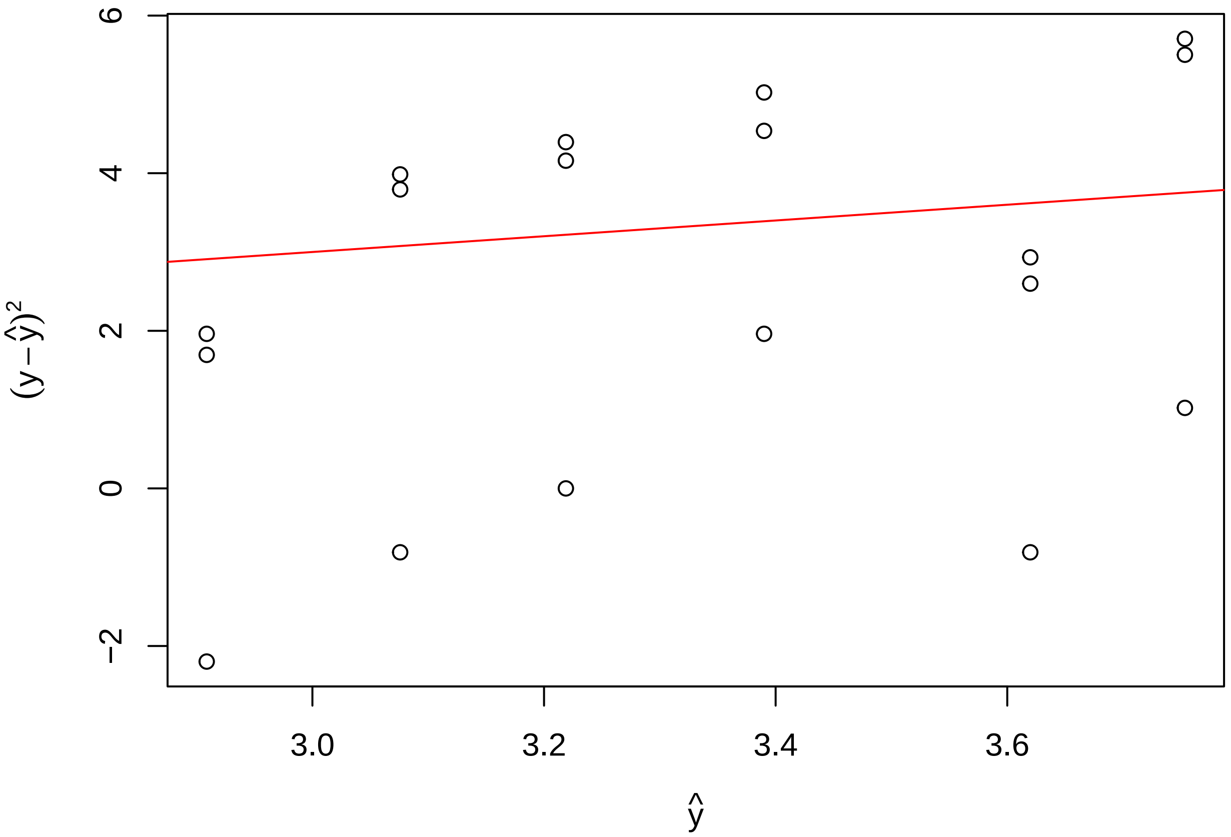
- For detecting outlier (using half-normal plot):

```
library(faraway)
halfnorm( residuals(model))
```



- For detecting overdispersion (comparing mean and variance of fitted model):

```
pois.fit.mean = log( fitted(model) )
pois.fit.var = log( (data$colonies - exp(pois.fit.mean))^2 )
par(mar=c(4.5,4.5,0.5,0.5))
plot( x= pois.fit.mean, y= pois.fit.var,
      xlab = expression( hat(y)),
      ylab = expression( (y- hat(y))^2) )
abline(0,1, cex=1.5, col="red")
```



Obviously, there are some outliers and there is a clear difference between mean and variance. Therefore, we refit this model after adding dispersion parameter:

```
sigma.hat = sum( residuals( model, type = "pearson")^2) / model$df.residual
summary(model, dispersion = sigma.hat)
```

```

Call:
glm(formula = colonies ~ ., family = "poisson", data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.32778    0.07601  43.779  < 2e-16 ***
dose.L       0.50669    0.19218   2.637  0.00838 ** 
dose.Q      -0.22791    0.18465  -1.234  0.21710    
dose.C      -0.41331    0.18972  -2.178  0.02937 *  
dose^4       0.15583    0.18772   0.830  0.40649    
dose^5       0.13253    0.17628   0.752  0.45217    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 2.777999)

    Null deviance: 78.358  on 17  degrees of freedom
Residual deviance: 33.496  on 12  degrees of freedom
AIC: 138.03

Number of Fisher Scoring iterations: 4
```

Finally, we can find that the $s.e.(\hat{\beta})$'s are corrected, and the significant effects are different compared with the model without adding dispersion parameter.

Problem 3

Question 3.
The data was collected from subjects in Denmark. Build a Poisson model to describe how marital status (i.e., single, married, or divorced) changes with age, taking care to give an understandable interpretation of the important parameter estimates. Under your model, predict the probability that a Dane, aged 55, is divorced.

Sol.

```
data = read.table("maritaldane.txt",header = T)
data$Age = as.ordered(data$Age)
head(data)
```

```
    Age Single Married Divorced Total
1 17-21     17       1        0    18
2 21-25     16       8        0    24
3 25-30      8      17        1    26
4 30-40      6      22        4    32
5 40-50      5      21        6    32
6 50-60      3      17        8    28
```

For Poisson model analysis, convert this data into long format:

```
library(tidyverse)
long.data = data %>%
  gather(status, count, -Total, -Age)
long.data$status = factor(long.data$status,levels = c("Single", "Married","Divorced"))
head(long.data)
```

```
    Age Total status count
1 17-21    18 Single    17
2 21-25    24 Single    16
3 25-30    26 Single     8
4 30-40    32 Single     6
5 40-50    32 Single     5
6 50-60    28 Single     3
```

It is observed that this data is collected under the different value of size variable (Total), so use Poisson rate model to analyse it:

```
model = glm(count~offset(log(Total))+(.-Total)^2, data = long.data,family = "poisson")
summary(model)
```

```
Call:
glm(formula = count ~ offset(log(Total)) + (. - Total)^2, family = "poisson",
    data = long.data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.460e+00  1.955e-01  -7.469 8.09e-14 ***
Age.L                -2.060e+00  6.475e-01  -3.181 0.001467 **
Age.Q                 7.424e-01  6.255e-01   1.187 0.235298
Age.C                 8.130e-02  5.894e-01   0.138 0.890287
Age^4                -2.020e-01  5.417e-01  -0.373 0.709281
Age^5                 6.103e-02  5.084e-01   0.120 0.904437
Age^6                -1.778e-01  4.823e-01  -0.369 0.712415
Age^7                -1.057e-01  4.451e-01  -0.238 0.812215
statusMarried         5.227e-01  2.574e-01   2.031 0.042304 *
statusDivorced       -6.353e+00  1.231e+04  -0.001 0.999588
Age.L:statusMarried   3.163e+00  9.232e-01   3.426 0.000612 ***
Age.Q:statusMarried  -2.514e+00  8.954e-01  -2.808 0.004991 **
Age.C:statusMarried   6.064e-01  8.022e-01   0.756 0.449716
Age^4:statusMarried  -1.615e-01  6.943e-01  -0.233 0.816040
Age^5:statusMarried  -3.409e-02  6.116e-01  -0.056 0.955546
Age^6:statusMarried   2.168e-01  5.534e-01   0.392 0.695190
Age^7:statusMarried   8.911e-02  5.004e-01   0.178 0.858671
Age.L:statusDivorced  2.622e+01  4.623e+04   0.001 0.999547
Age.Q:statusDivorced -1.483e+01  3.800e+04   0.000 0.999689
Age.C:statusDivorced  2.164e+00  3.688e+04   0.000 0.999953
Age^4:statusDivorced  6.217e+00  4.144e+04   0.000 0.999880
Age^5:statusDivorced -7.916e+00  3.583e+04   0.000 0.999824
Age^6:statusDivorced  5.637e+00  2.186e+04   0.000 0.999794
Age^7:statusDivorced -2.028e+00  8.407e+03   0.000 0.999808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.2038e+02  on 23  degrees of freedom
Residual deviance: 3.0331e-10  on  0  degrees of freedom
AIC: 128.04

Number of Fisher Scoring iterations: 21
```

Since most of the effects were not significant, model selection was performed:

```
while (prod(coef(summary(model))[,"Pr(>|z|)"]<=0.05)!=1) {
  model_mat = model.matrix(model)[,-1]
  minaic = 1000
  for (i in 1:dim(model_mat)[2]) {
    t_dat = data.frame(count = long.data$count, model_mat[,-i],
                       Total = long.data$Total)
    fit = glm(count~offset(log(Total))+(.-Total), data = t_dat, family = "poisson")
    if (fit$aic<minaic) {
      minaic = fit$aic
      selectX = i
    }
  }
t_dat = data.frame(count = long.data$count, model_mat[,-selectX], Total = data$Total)
model = glm(count~offset(log(Total))+(.-Total), data = t_dat, family = "poisson")
}
summary(model)
```

```
Call:
glm(formula = count ~ offset(log(Total)) + (. - Total), family = "poisson", 
    data = t_dat)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -1.8441     0.1486 -12.410  < 2e-16 ***
Age.L                 -3.2406     0.4711  -6.879 6.05e-12 ***
statusMarried          0.9727     0.2025   4.804 1.56e-06 ***
Age.L.statusMarried    3.7454     0.6327   5.920 3.22e-09 ***
Age.Q.statusMarried   -1.3857     0.4275  -3.242  0.00119 ** 
Age.L.statusDivorced   5.9468     0.8390   7.088 1.36e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 120.384  on 23  degrees of freedom
Residual deviance:  12.371  on 18  degrees of freedom
AIC: 104.41

Number of Fisher Scoring iterations: 5
```

As seen from the above summary, the interaction effects between Age and marital status are significant, which means that the number of people in different marital statuses will change with different age groups. Finally, predict the probability of (Age=55 ∈ (50,60), status = Divorced):

```r
which1 = which(long.data$Age == "50-60" & long.data$status =="Divorced")
which2 = which(long.data$Age == "50-60")
cat("P(Divoced, age=55) = ", fitted(model)[which1]/ (sum(fitted(model)[which2])))
```

```
P(Divoced, age=55) =  0.2904925
```