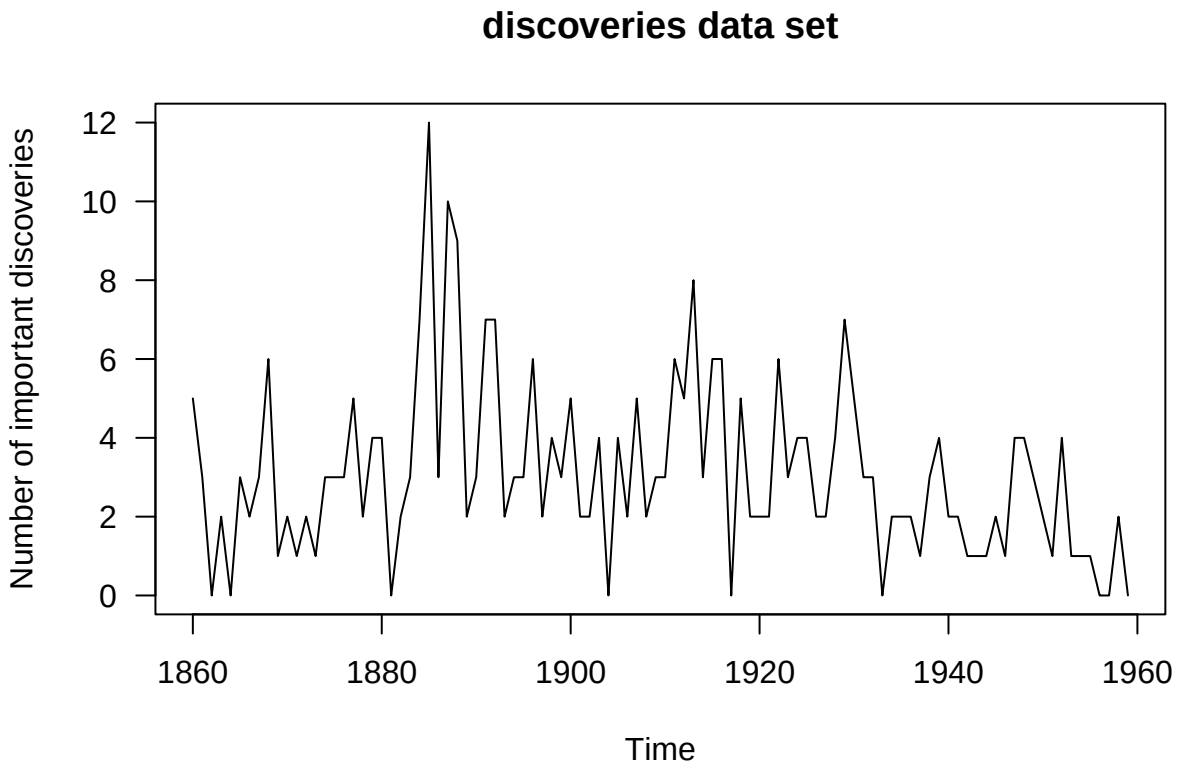
Discrete Analysis Homework 4

110024516 邱繼賢

Problem 1.

```
data(discoveries)
plot(discoveries, ylab = "Number of important discoveries",
     las = 1)
title(main = "discoveries data set")
```



We can see that the **numbers** of great inventions and scientific discoveries in each year from 1860 to 1959 does not look like a constant over time. However, we do not know the **total numbers** of inventions and

scientific discoveries in each year, namely *size variable*. We can not infer that the higher number of great discoveries means the higher discovery rate. We need more information.

If we believe that the **total numbers** of discoveries in each year are almost equal (or regard **one year** as size variable), then we can compare the discovery rate by comparing the great discovery numbers. Let's construct the Poisson GLM.

$$y_i \sim Poi(\mu_i)$$
$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 \text{ Year}$$

```
year = 1860:1959
modp = glm(discoveries ~ year, family = poisson)
summary(modp)
```

```
##
## Call:
## glm(formula = discoveries ~ year, family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8112  -0.9482  -0.3533   0.6637   3.5504
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.354807   3.775677   3.007  0.00264 **
## year        -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
```

Although the model does not fit well enough (look at the Residual deviance), we can still observe that the variable *Year* significantly involves the response. The great discovery numbers (rate) hasn't remained

constant over time.

Problem 2.

First, fit the Poisson GLM

$$y_i \sim Poi(\mu_i)$$
$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 \text{ dose}$$

```
salmon = read.table("salmonella.txt")
mod2.1 = glm(colonies ~ dose, family = poisson, salmon)
summary(mod2.1)
```

```
##
## Call:
## glm(formula = colonies ~ dose, family = poisson, data = salmon)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6482  -1.8225  -0.2993   1.2917   5.1861
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.3219950  0.0540292  61.485   <2e-16 ***
## dose        0.0001901  0.0001172   1.622    0.105
## ---
## 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: 75.806  on 16  degrees of freedom
## AIC: 172.34
##
## Number of Fisher Scoring iterations: 4
```

We can see that the Residual deviance is really large, and the goodness-of-fit test is rejected (p-value = $P(\chi^2_{16} > 75.806) < 0.05$).

The reasons of this situation may be

1. Wrong $X\beta$ structure

2. Outliers

3. Over-dispersion

Let's try the complicated model by adding the quadratic and cubic terms of dose

$$
\begin{aligned}
y_i &\sim Poi(\mu_i) \\
\log(\mu_i) &= \eta_i = \beta_0 + \beta_1 \text{ dose} + \beta_2 \text{ dose}^2 + \beta_3 \text{ dose}^3
\end{aligned}
$$

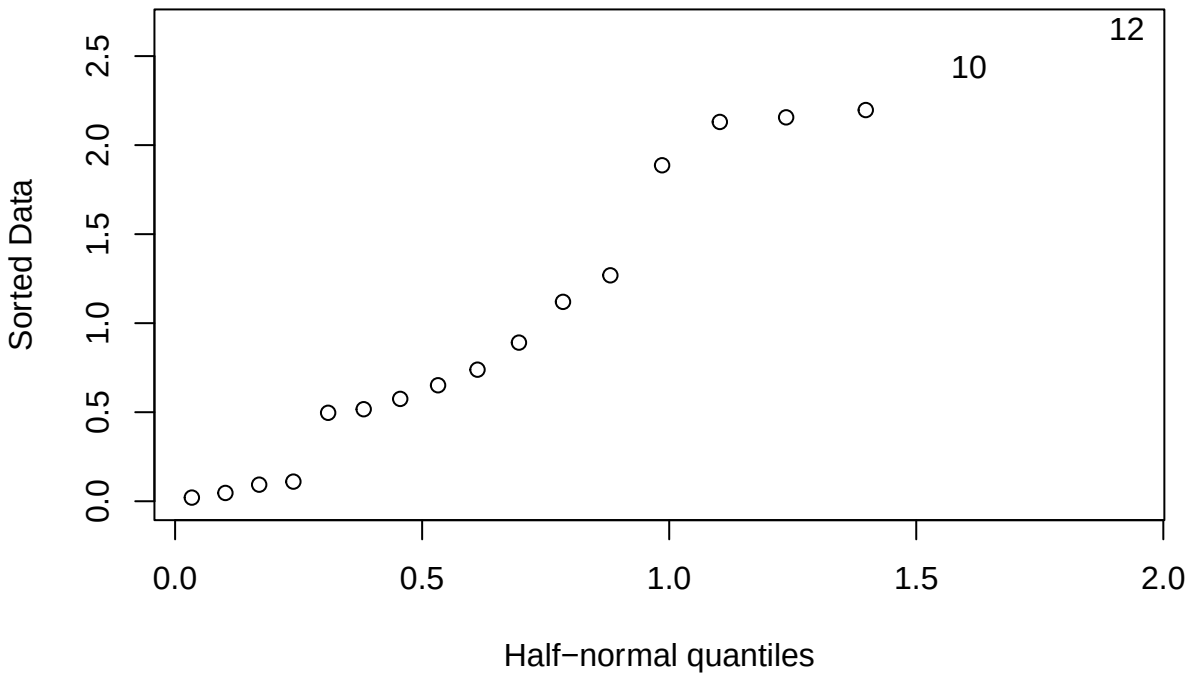
```
mod2.2 = update(mod2.1, .~.+I(dose^2)+I(dose^3))
summary(mod2.2)
```

```
##
## Call:
## glm(formula = colonies ~ dose + I(dose^2) + I(dose^3), family = poisson,
##     data = salmon)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.43608  -0.85295  -0.07833   0.56028   2.65580
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.930e+00  8.988e-02  32.595  < 2e-16 ***
## dose         1.141e-02  2.051e-03   5.563 2.65e-08 ***
## I(dose^2)   -3.653e-05  7.602e-06  -4.805 1.54e-06 ***
## I(dose^3)    2.558e-08  5.668e-09   4.514 6.37e-06 ***
## ---
## 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: 36.055  on 14  degrees of freedom
```

```
## AIC: 136.59
##
## Number of Fisher Scoring iterations: 4
```

The Residual deviance became smaller slightly but the goodness-of-fit test is still rejected (p-value = $P(\chi^2_{14} > 36.055) < 0.05$).
We can add more explanatory terms in the model to reduce the deviance, but the model will become very hard to explain. Let's check whether the large deviance is caused by outliers.

```
"halfnorm" <- function (x, nlab = 2, labs = as.character(1:length(x)), ylab = "Sorted Data") {
    x <- abs(x)
    labord <- order(x)
    x <- sort(x)
    i <- order(x)
    n <- length(x)
    ui <- qnorm((n + 1:n)/(2 * n + 1))
    plot(ui, x[i], xlab = "Half-normal quantiles", ylab = ylab, ylim=c(0,max(x)),
         type = "n")
    if(nlab < n)
        points(ui[1:(n - nlab)], x[i][1:(n - nlab)])
    text(ui[(n - nlab + 1):n], x[i][(n - nlab + 1):n], labs[labord][(n - nlab + 1):n])
}
halfnorm(residuals(mod2.2))
```



We do not see any clear evidence of outlier in the half-normal plot.

The only reason left is the over-dispersion. We can solve this situation by

1. Adding a dispersion parameter σ^2

2. Refitting the model as a Negative Binomial GLM

Problem 3.

Regard *Age* as numerical variable and construct Poisson GLM (rate model)

$$y_i \sim Poi(\mu_i)$$

$$\log(\mu_i) = \eta_i' = \log(Total) + \eta_i \sim \text{offset}(Total) + \text{unclass}(Age) * Status$$

```
marital = read.table("maritaldane.txt")
data = cbind(stack(marital[,c(2,3,4)]),rep(as.factor(marital$Age),3),rep(marital$Total,3))
colnames(data) = c("count","Status","Age","Total")
mod3.1 = glm(count ~ offset(log(Total)) + Status*unclass(Age), family = poisson, data)
summary(mod3.1)
```

```
##
## Call:
## glm(formula = count ~ offset(log(Total)) + Status * unclass(Age),
##     family = poisson, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.85773  -0.73135  -0.06458   0.54341   1.36452
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.31100    0.25647   1.213 0.225273
## StatusMarried                -1.33700    0.36443  -3.669 0.000244 ***
## StatusDivorced               -4.72730    0.72988  -6.477 9.37e-11 ***
## unclass(Age)                 -0.42301    0.07888  -5.363 8.19e-08 ***
## StatusMarried:unclass(Age)    0.50940    0.09456   5.387 7.15e-08 ***
## StatusDivorced:unclass(Age)   0.92728    0.13565   6.836 8.16e-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:  21.828  on 18  degrees of freedom
## AIC: 113.87
##
## Number of Fisher Scoring iterations: 5
```

隨著變數 *Age* 每上升一個階級

1. *Status* = *single* 所佔的比例變為原本的 $e^{-0.42301} = 0.6550721$ 倍
2. *Status* = *married* 所佔的比例變為原本的 $e^{-0.42301+0.50940} = 1.090231$ 倍
3. *Status* = *divorced* 所佔的比例變為原本的 $e^{-0.42301+0.92728} = 1.655776$ 倍

Predict the probability of

$$(Age\ ,\ Status)\ =\ (55\ ,\ \text{divorced})\ \Rightarrow\ (\text{unclass}(Age)\ ,\ Status)\ =\ (6\ ,\ \text{divorced})\ \Rightarrow\ x_0\ =\ (1,0,1,6,0,6)^T$$

$$\log(\hat{p}_{x_0})\ =\ \log\left(\frac{\hat{\mu}_{x_0}}{28}\right)\ =\ \hat{\eta}_{x_0}\ =\ x_0^T\hat{\beta}\ =\ -1.390718$$

$$\Rightarrow \hat{p}_{x_0}\ =\ \exp\left(\hat{\eta}_{x_0}\right)\ =\ 0.2488965$$

```
x0 = c(1,0,1,6,0,6)
eta = sum(mod3.1$coefficients*x0)
mu = exp(eta)
mu
```

```
## [1] 0.2488965
```

The 95% confidence interval of $\hat{p}_{x_0}$

$$\left[\exp\left(\hat{\eta}_{x_0} - Z_{0.975}\ \text{se}\left(\hat{\eta}_{x_0}\right)\right)\ ,\ \exp\left(\hat{\eta}_{x_0} + Z_{0.975}\ \text{se}\left(\hat{\eta}_{x_0}\right)\right)\right]\ =\ [0.1740123\ ,\ 0.3560064]$$

where

$$\text{se}\left(\hat{\eta}_{x_0}\right)\ =\ \sqrt{x_0^T\ \hat{\Sigma}\ x_0}$$

```
mod3.1_sum = summary(mod3.1)
cm = mod3.1_sum$cov.unscaled
se = sqrt(t(x0) %*% cm %*% x0)[1,1]
exp(eta+c(-1,1)*qnorm(0.975)*se)
```

```
## [1] 0.1740123 0.3560064
```