STAT350 ASSIGNMENT1

Yuhan Zhang   301345627

Q1 and Q2

1. n indept pairs of (x_i, y_i). Show $\sum_{i=1}^{n} x_i e_i = 0$.

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{dS}{d\hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)$$

$$= -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i$$

$$= \sum_{i=1}^{n} e_i \cdot x_i = 0.$$

2. Show $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

$$\frac{dS}{d\hat{\beta}_0} = \sum 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)'$$

$$= -2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$= \sum y_i - \hat{y}_i = \boxed{\sum e_i = 0}$$

$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i$$

$$= \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i$$

$\because \sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i x_i = 0$ (from Q_1)

$\therefore \sum_{i=1}^{n} \hat{y}_i e_i = 0.$

Q3

3. $y = \beta X + \epsilon$. $\epsilon \sim N(0, \sigma^2)$.

$\because \epsilon \sim N(0, \sigma^2)$ $\therefore \epsilon$ is normal distubution.

$\therefore f(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\beta X)^2/2\sigma^2}$

$L = f(\epsilon_1) \cdot f(\epsilon_2) \cdots f(\epsilon_i)$

$= \prod_{i=1}^{n} f(\epsilon_i)$

$= (2\pi\sigma^2)^{-n/2} e^{-\sum(y_i-\beta x_i)^2/2\sigma^2}$

$l = \ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum(y_i-\beta x_i)^2}{2\sigma^2}$

$0 = \frac{dl}{d\beta} = -\frac{1}{2\sigma^2}\sum(y_i-\beta x_i)^2$

$\therefore \frac{d(\sum(y_i-\beta x_i)^2)}{d\beta} = 0$

$\Rightarrow$ same as OLE

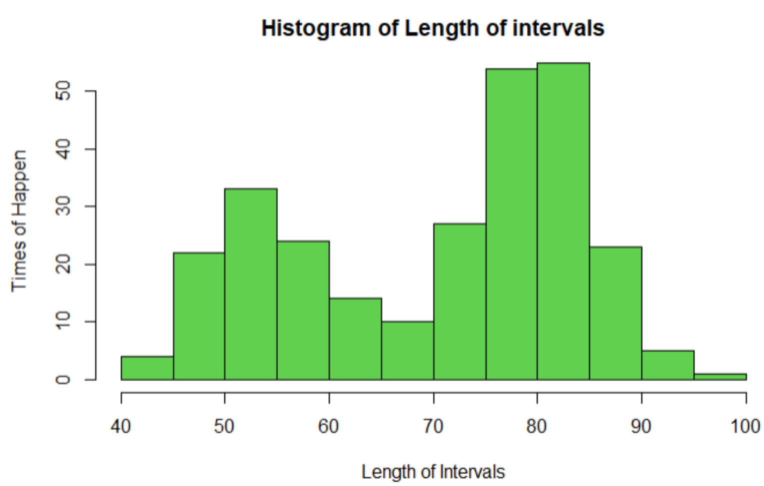
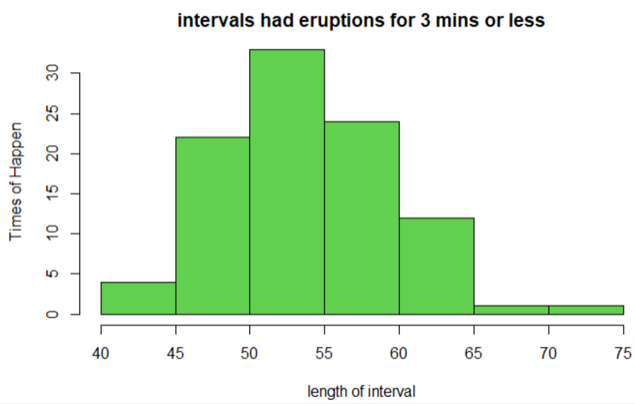
$0 = \frac{dl}{d\sigma^2} = -\frac{n}{2}\left(\frac{2\pi}{2\pi\sigma^2}\right) + \frac{\sum(y_i-\beta x_i)^2}{2\sigma^4}$

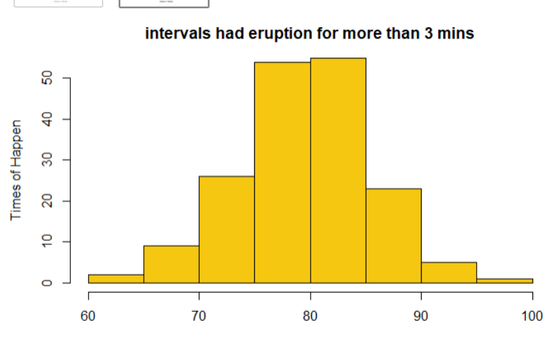
$\therefore \frac{n}{2}\frac{1}{\sigma^2} = \frac{\sum(y_i-\beta x_i)^2}{2\sigma^4} = \frac{\sum e_i^2}{2\sigma^4}$

$\sigma n^2/2 = \sum e_i^2/2$

$\hat{\sigma}^2 = \sum e_i^2/n$. ~ MLE which is biased. as the unbiased estimator is $\hat{\sigma}^2 = \sum e_i^2/n-2$.

Q4

```
#####Stat350 Homework1
```{r}
mydata <- read.csv("geyser.csv",header=TRUE)
mydata
```
```

X <int>	eruptions <dbl>	waiting <int>
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51
10	4.350	85

```
###a)
```{r}
hist(mydata$waiting, main="Histogram of Length of intervals", xlab="Length of Intervals", ylab= "Times of Happen",
col=3)
```
```



As there are two peaks, the data is not a normal distribution.
Therefore, it is not faithful and is not reliable.

````
###b)
```{r}
mean(mydata$waiting)
sd(mydata$waiting)
```
````

```
[1] 70.89706
[1] 13.59497
```

No, I would not use mean and standard deviation to describe the data as the data is not a normal distribution.

````
###c)
```{r}
less_than_3 = df[mydata$eruptions <= 3,]
more_than_3 = df[mydata$eruptions > 3,]
hist(less_than_3$waiting, xlab="length of interval", ylab= "Times of Happen", main="intervals had eruptions for 3 mins or less", col=3)
hist(more_than_3$waiting, xlab="length of interval", ylab= "Times of Happen", main="intervals had eruption for more than 3 mins", col=7)
```
````




````
###b)
```{r}
mean(mydata$waiting)
sd(mydata$waiting)
```
````

```
[1] 70.89706
[1] 13.59497
```

No, I would not use mean and standard deviation to describe the data as the data is not a normal distribution.

````
###c)
```{r}
less_than_3 = df[mydata$eruptions <= 3,]
more_than_3 = df[mydata$eruptions > 3,]
hist(less_than_3$waiting, xlab="length of interval", ylab= "Times of Happen", main="intervals had eruptions for 3 mins or less", col=3)
hist(more_than_3$waiting, xlab="length of interval", ylab= "Times of Happen", main="intervals had eruption for more than 3 mins", col=7)
```
````

```{r}
mean(less_than_3$waiting, )
sd(less_than_3$waiting, )
mean(less_than_3$waiting) +c(-1,1)*sd(less_than_3$waiting)        ##68%
mean(less_than_3$waiting) +2*c(-1,1)*sd(less_than_3$waiting)      ##95%
mean(less_than_3$waiting) +3*c(-1,1)*sd(less_than_3$waiting)      ##99.7%
```

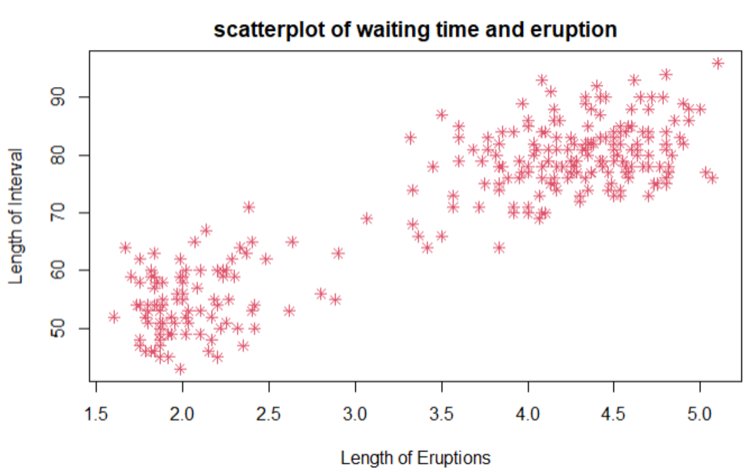
```
[1] 54.49485
[1] 5.840098
[1] 48.65475 60.33494
[1] 42.81465 66.17504
[1] 36.97455 72.01514
```

```{r}
mean(more_than_3$waiting, )
sd(more_than_3$waiting, )
mean(more_than_3$waiting) +c(-1,1)*sd(more_than_3$waiting)
mean(more_than_3$waiting) +2*c(-1,1)*sd(more_than_3$waiting)
mean(more_than_3$waiting) +3*c(-1,1)*sd(more_than_3$waiting)
```

```
[1] 79.98857
[1] 5.994239
[1] 73.99433 85.98281
[1] 68.00009 91.97705
[1] 62.00585 97.97129
```

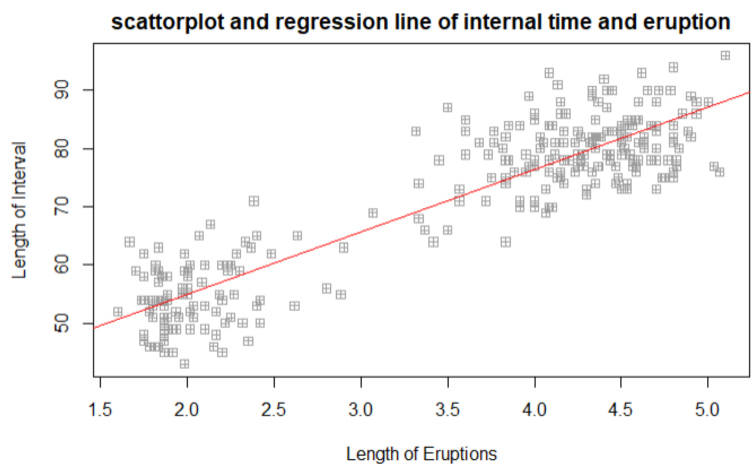
###d)

```{r}
plot(mydata$waiting~df$eruptions, main="scatterplot of waiting time and eruption", xlab="Length of
Eruptions",ylab="Length of Interval",col=2, pch=8)
```



The length of internal between two eruption increases with the length of the precious eruption.
Yes, there is a positive linear relation between length of eruption and length of waiting interval.

~~~~
###e)
```{r}
plot(mydata$waiting~mydata$eruptions, main="scattorplot and regression line of internal time and eruption",
xlab="Length of Eruptions",ylab="Length of Interval", col=8, pch=12)
abline(regre_plot,col="red")
```
~~~~




~~~~
```{r}
data.2 = data.frame(eruptions=2)
predict(regre_plot,data.2)
```
~~~~

~~~~
       1
54.93368
~~~~