STAT350 Assignment 1 Solution

Question 1

Given that $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - x_i\hat{\beta}_1$, where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ and $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_iy_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$. Then

$$
\begin{aligned}
\sum_{i=1}^n x_ie_i &= \sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1) \\
&= \sum_{i=1}^n x_iy_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
&= \sum_{i=1}^n x_iy_i - (\sum_{i=1}^n y_i/n - \hat{\beta}_1 \sum_{i=1}^n x_i/n) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
&= \sum_{i=1}^n x_iy_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} + \hat{\beta}_1 \left(\frac{(\sum_{i=1}^n x_i)^2}{n} - \sum_{i=1}^n x_i^2 \right) \\
&= \sum_{i=1}^n x_iy_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} + \frac{\sum_{i=1}^n x_iy_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} \left(\frac{(\sum_{i=1}^n x_i)^2}{n} - \sum_{i=1}^n x_i^2 \right) \\
&= \sum_{i=1}^n x_iy_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} - \sum_{i=1}^n x_iy_i + \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} \\
&= 0.
\end{aligned}
$$

Question 2

Given that $\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}_1$, then

$$\begin{aligned}
\sum_{i=1}^{n} \hat{y}_i e_i &= \sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} (\hat{\beta}_0 + x_i\hat{\beta}_1)(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1) \\
&= \hat{\beta}_0 \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - n\hat{\beta}_0^2 - 2\hat{\beta}_0\hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 \\
&= \frac{(\sum_{i=1}^{n} y_i)^2}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} y_i)^2}{n} + 2\hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1^2 \frac{(\sum_{i=1}^{n} x_i)^2}{n} \\
&\quad - 2\hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + 2\hat{\beta}_1^2 \frac{(\sum_{i=1}^{n} x_i)^2}{n} - \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 \\
&= -\hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + \hat{\beta}_1^2 \frac{(\sum_{i=1}^{n} x_i)^2}{n} - \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 \\
&= -\hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} - \sum_{i=1}^{n} x_i y_i \right) + \hat{\beta}_1^2 \left(\frac{(\sum_{i=1}^{n} x_i)^2}{n} - \sum_{i=1}^{n} x_i^2 \right) \\
&= -\hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} - \sum_{i=1}^{n} x_i y_i \right) + \hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} - \sum_{i=1}^{n} x_i y_i \right) \\
&= 0.
\end{aligned}$$

Question 3

The joint likelihood of independent $(y_1, y_2, ..., y_n)$ is

$$L = \prod_{i=1}^{n} f(y_i|\beta, \sigma) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \right).$$

Then the log-likelihood is

$$l = \log(L) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

To get the maximum likelihood estimator for β, we set the derivative to 0 and solve

$$\frac{\delta l}{\delta \beta} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i) x_i = 0 \Longrightarrow \sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 = 0.$$

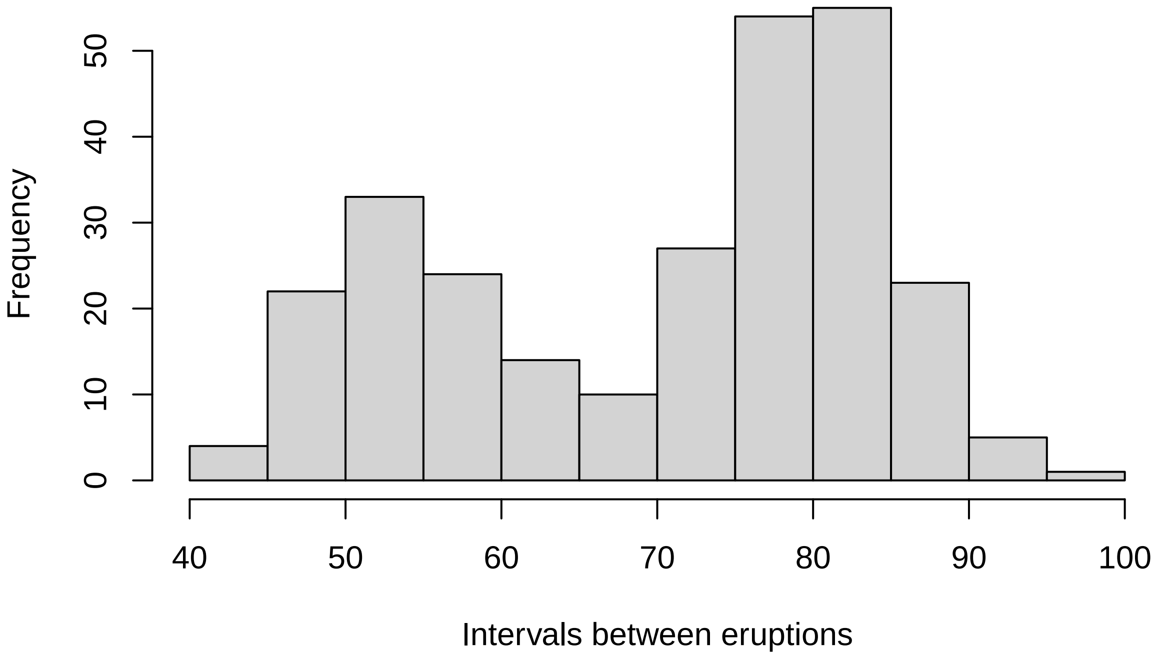
Thus,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Question 4

(a)

```
mydata=read.csv("/Users/dbingham/Desktop/geyser.csv",header=TRUE)
hist(mydata$waiting,main="",xlab="Intervals between eruptions")
```



The interval between eruptions has a bimodal distribution. The Old Faithful is not that reliable, and you may end up waiting a very long time to see the next eruption.

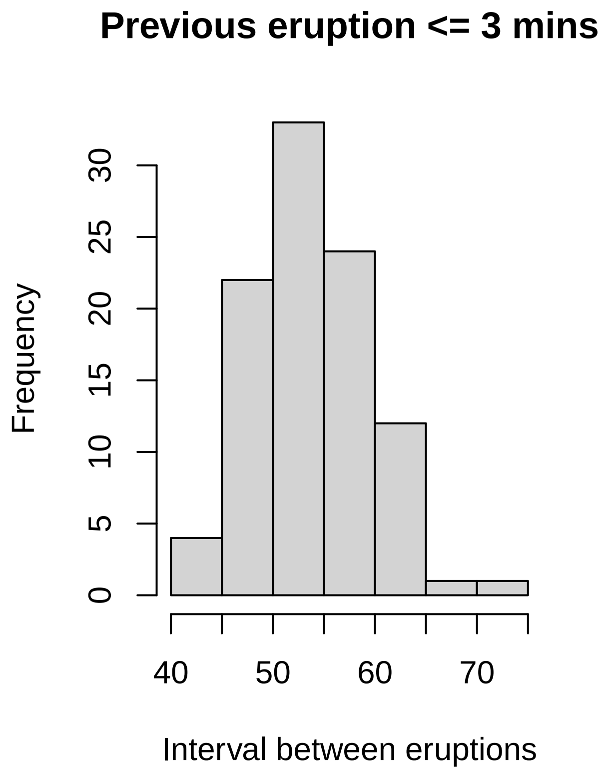
(b)

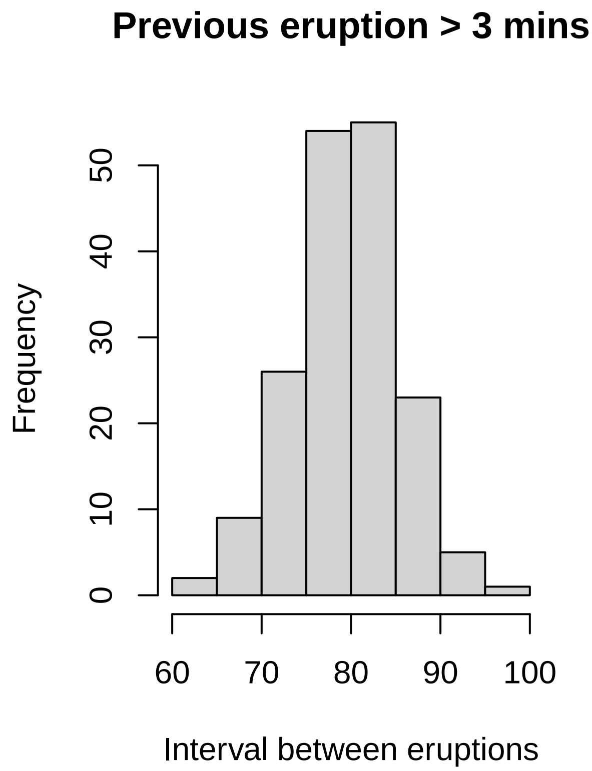
Standard descriptive statistic, such as mean, is not appropriate to describe a bimodal distribution, as well as the standard deviation which is just a function of mean.

(c)

```
interval_1=mydata[which(mydata$eruptions<=3),]$waiting
interval_2=mydata[which(mydata$eruptions>3),]$waiting

par(mfrow=c(1,2))
hist(interval_1,
     main="Previous eruption <= 3 mins",
     xlab="Interval between eruptions")
hist(interval_2,
     main="Previous eruption > 3 mins",
     xlab="Interval between eruptions")
```






```
mean(interval_1)-2*sd(interval_1);mean(interval_1)+2*sd(interval_1)
```

```
## [1] 42.81465
```

```
## [1] 66.17504
```

```
mean(interval_2)-2*sd(interval_2);mean(interval_2)+2*sd(interval_2)
```

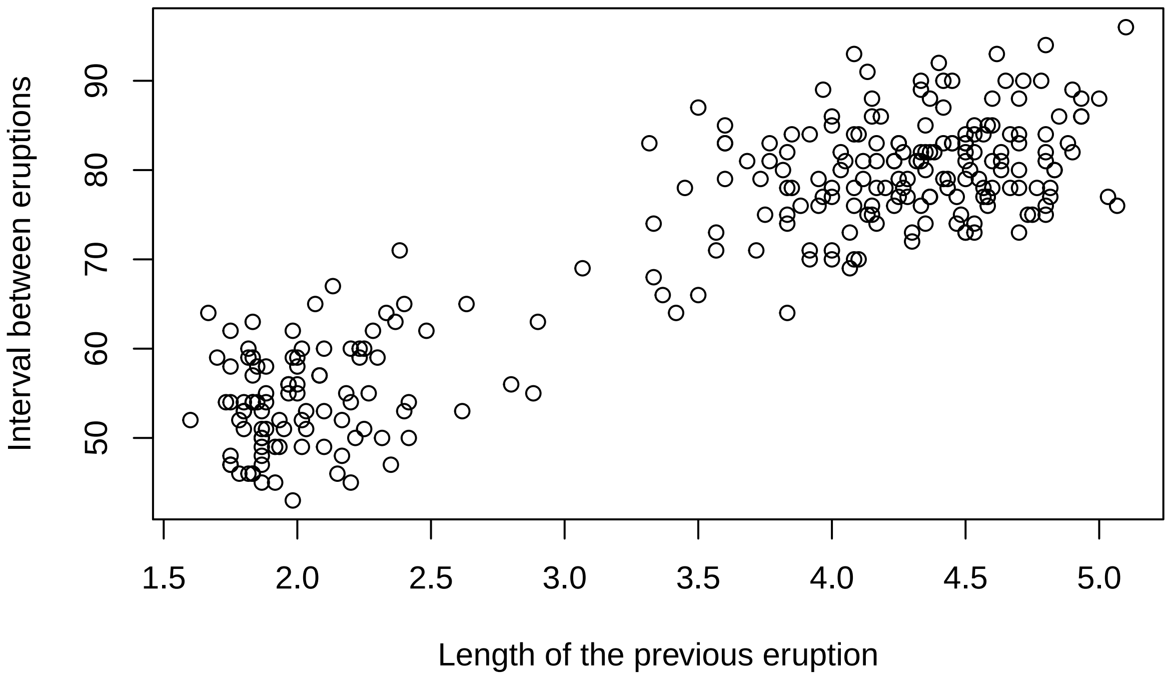
```
## [1] 68.00009
```

```
## [1] 91.97705
```

The histograms of intervals between eruptions for both sets of data are approximately normal with bell shape. Based on 68-95-99.7 empirical rule, if the previous eruption is 3 mins or less, then we can say there is a 95% probability that the interval between eruptions will between 42.815 and 66.175 mins. So we would recommend a person return in 42.815 mins so they have a 97.5% chance of seeing the geyser. On the other hand, if the previous eruption is greather than 3 mins, there is a 95% probability that the interval between eruption is between 68 and 91.977 mins. Therefore, the recommendation would be to return in 68 mins to have a 97.5% probability of seeing the geyser.

(d)

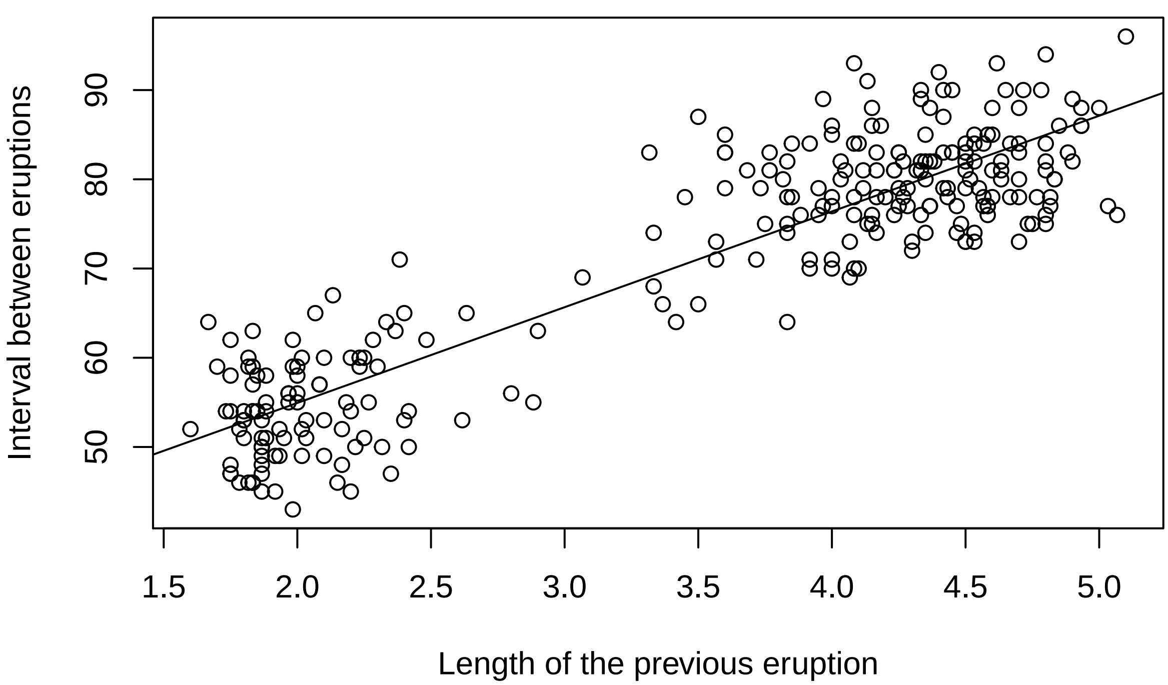
```
plot(mydata$eruptions,mydata$waiting,
     xlab="Length of the previous eruption",
     ylab="Interval between eruptions")
```



There is a linear relationship between the interval between eruptions and the length of the previous eruption, so linear regression could be used here.

(e)

```
lm=lm(waiting~eruptions,data=mydata)
plot(mydata$eruptions,mydata$waiting,
     xlab="Length of the previous eruption",
     ylab="Interval between eruptions")
abline(lm)
```



(f)

```
new.eruption=data.frame(eruptions=2)
predict(lm,newdata=new.eruption,interval="prediction",level=0.95)
```

```
##        fit      lwr      upr
## 1 54.93368 43.23248 66.63488
```

Lots of wayt to look at this. For example, suppose the length of the previous eruption was 2 mins, the expected interval between previous and next eruptions is 54.934 mins with 95% prediction interval (43.232, 66.635) mins. Any of the following conclusions would be considered correct: 1. We expect to wait 54.934 mins until the next eruption. 2. There is a 95% probability that the interval between eruptions will be between 43.232 and 66.635 mins. So we would recommend that a person return in 43.232 mins so they have a 97.5% chance of seeing the next eruption. 3. We would say that there is a 95% chance of seeing the next erutpion if a person return in between 43.232 and 66.635 mins.