Exercises

Section 10.2 Fitting the Simple Linear Regression Model

10.4 The time between eruptions ...

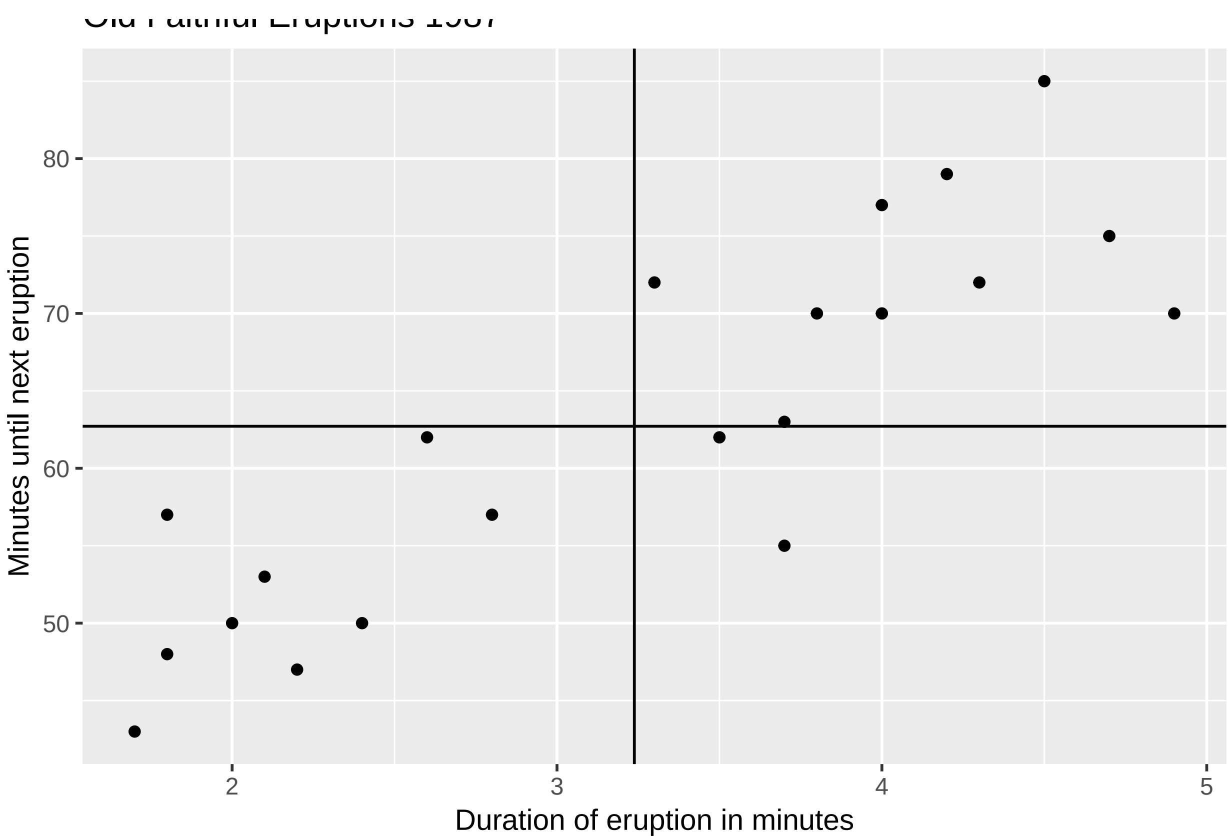
```
library(ggplot2)
```

```
oldfaithful <- read.csv(file='data/1987oldfaithful.csv',header=TRUE,sep=',')
head(oldfaithful)
```

```
##   Observation.Number Duration.of.Eruption Time.Between.Eruptions
## 1                  1                  2.0                     50
## 2                  2                  1.8                     57
## 3                  3                  3.7                     55
## 4                  4                  2.2                     47
## 5                  5                  2.1                     53
## 6                  6                  2.4                     50
```

Assume the time between eruptions is linearly dependent on the duration of the last eruption.

```
x <- oldfaithful$'Duration.of.Eruption'
y <- oldfaithful$'Time.Between.Eruptions'
ggplot(data=oldfaithful, aes(x,y)) +
  geom_point() +
  geom_vline(xintercept = mean(x)) +
  geom_hline(yintercept = mean(y)) +
  labs(
    title = 'Old Faithful Eruptions 1987',
    x = 'Duration of eruption in minutes',
    y = 'Minutes until next eruption'
  )
```



Let

- $n = 21$
- $(x_i)_{i \in [n]}$ be the values of the independent variable.
- $(Y_i)_{i \in [n]}$ be random variables for which $(y_i)_{i \in [n]}$ are the obserations.

Assume for each $i \in [n]$ that Y_i is linearly dependent on x_i. In particular that there exists two real values β_0 and β_1 and a random error $\epsilon_i \sim N(0, \sigma^2)$ such that

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

We can estimate β_0 and β_1 by the least squares method:

$$\min_{(\beta_0, \beta_1) \in \mathbb{R}^2} Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 \cdot x_i))^2$$

From calculus we know a function Q has a critical point at $(\hat{\beta}_0, \hat{\beta}_1)$ if $\nabla Q(\hat{\beta}_0, \hat{\beta}_1) = 0$.

$$\frac{\partial Q}{\partial \beta_0} = -2 \cdot \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 \cdot x_i))$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \cdot \sum_{i=1}^{n} x_i \cdot (y_i - (\beta_0 + \beta_1 \cdot x_i))$$

Therefore

$$n \cdot \hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i \right) \cdot \hat{\beta}_1 = \left(\sum_{i=1}^{n} y_i \right)$$

$$\left(\sum_{i=1}^{n} x_i\right) \cdot \hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2\right) \cdot \hat{\beta}_1 = \left(\sum_{i=1}^{n} x_i \cdot y_i\right)$$

By defining $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, one can show

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

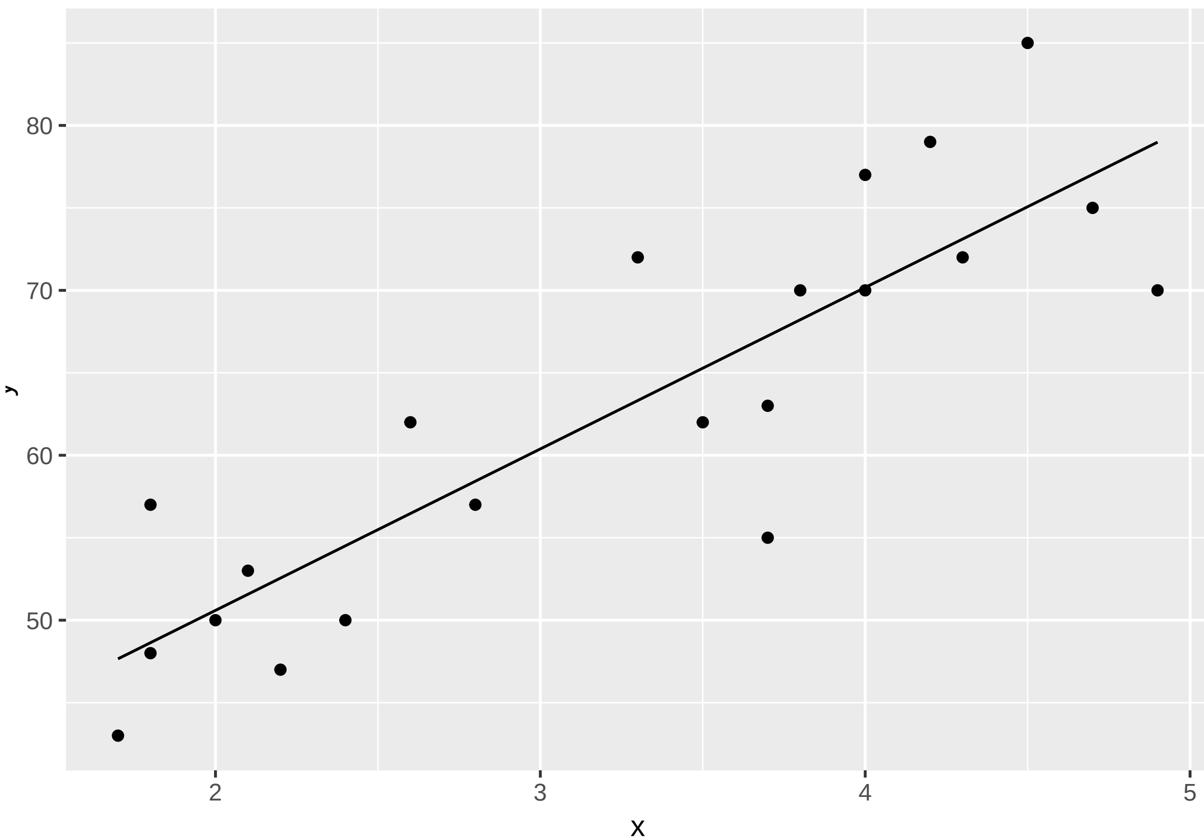
```
S <- function(x,y) {
  sum( (x-mean(x))*(y-mean(y)) )
}
```

```
betahat_1 <- S(x,y)/S(x,x) # slope
betahat_0 <- mean(y) - betahat_1*mean(x) # y-intercept
```

This gives us a linear function to estimate the time between eruptions.

$$\hat{y}(x) = \hat{\beta}_1 \cdot x + \hat{\beta}_0$$

```
yhat <- function(x) return(betahat_1*x + betahat_0)
ggplot(oldfaithful, aes(x,y)) +
  geom_point() +
  stat_function(fun = yhat)
```



So if an eruption of Old Faithful lasted 3 minutes we could estimate the next eruption to occur in

```
yhat(3)
```

```
## [1] 60.38332
```

minutes.

The residue plot to check linearity

```
y - yhat(x) # residuals
```

```
##  [1]  -0.5932479   8.3647659 -12.2363652  -5.5512617   1.4277452
##  [6]  -4.5092755   5.5327107  -1.4253031   8.6796624  -3.2783514
## [11]  -4.2363652   1.7846279   9.9315796  -2.0264342   6.8266141
## [16]  -0.1733859  -4.6562272  -0.6352341  -8.9844480   6.8686003
## [21]  -1.1104066
```

```
ggplot(oldfaithful, aes(x,y-yhat(x))) + geom_point()
```