STAT3008 Exercise 2 Solutions
(2011 - 2012 2nd Semester)

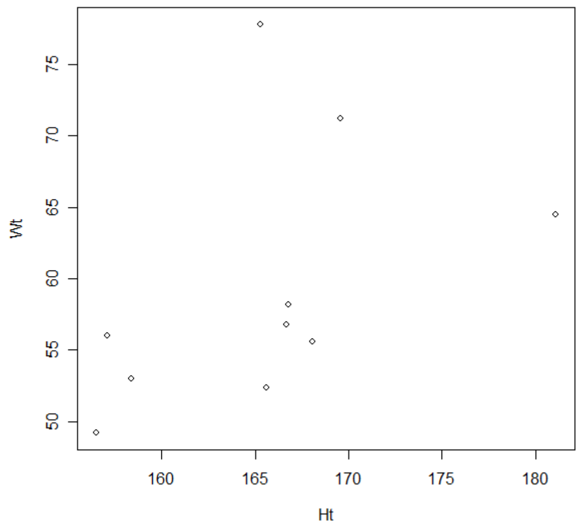
Q1.

Problem 2.1

2.1.1

R Codes:

```
library(alr3)
data(htwt)
plot(htwt)
```



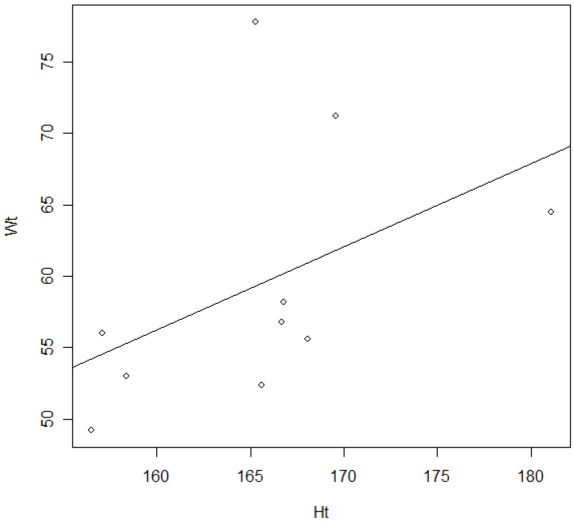
With only 10 points, it is hard to determine whether a simple linear regression is enough or not.

2.1.2

R Codes:

```
library(alr3)
data(htwt)
plot(htwt)
fit=lm(Wt~Ht,data=htwt)      #Fit the regression line
abline(fit)
average=mean(htwt)      #Compute the means
s=(10-1)*cov(htwt)      #cov() computes the sample covariance matrix
s      #Display Sxx,Syy,Sxy
average      #Display the means
```

summary(fit) #Display the summary of the fitted model



Using R, we have $\bar{x} = 165.52$, $\bar{y} = 59.47$, $SXX = 472.076$, $SYY = 731.961$,

$SXY = 274.786$. $\hat{\beta}_0 = -36.8759$ and $\hat{\beta}_1 = 0.5821$.

2.1.3

R Codes:

summary(fit) #Display the summary of the fitted model

vcov(fit) #Display the variance-covariance matrix of the estimates

Using the above commands, we get $se(\hat{\beta}_0) = 64.4728$, $se(\hat{\beta}_1) = 0.3892$ and

$\hat{\sigma}^2 = 8.456^2 = 71.502$. Also, $\widehat{Cov}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 4156.7419 & -25.0700 \\ -25.0700 & 0.1515 \end{pmatrix}$.

2.1.4

R Codes:

anova(fit) #Display the anova table of the fitted model

We can get from the results that $F - statistic = 2.237$ and $p - value = 0.1731$. Therefore, there is no strong evidence that the slope is significant.

Q2.

2.3

2.3.1

α is the value of $E(Y \mid X = \bar{x})$.

2.3.2

$$\begin{aligned} RSS(\alpha, \beta_1) &= \sum (y_i - \alpha - \beta_1(x_i - \bar{x}))^2 \\ &= \sum (y_i - \alpha)^2 - 2\beta_1 \sum (y_i - \alpha)(x_i - \bar{x}) + \beta_1^2 \sum (x_i - \bar{x})^2 \end{aligned}$$

And also, $\sum (y_i - \alpha)(x_i - \bar{x}) = \sum y_i(x_i - \bar{x}) - \alpha \sum (x_i - \bar{x})$
$= SXY - 0 = SXY$

Therefore, $RSS(\alpha, \beta_1) = \sum (y_i - \alpha)^2 - 2\beta_1 SXY + \beta_1^2 SXX$

$$\left.\frac{\partial RSS}{\partial \alpha}\right|_{\alpha=\hat{\alpha}} = -2\sum (y_i - \hat{\alpha}) = 0$$

$$\hat{\alpha} = \bar{y}$$

$$\left.\frac{\partial RSS}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2SXY + 2\hat{\beta}_1 SXX = 0$$

$$\hat{\beta}_1 = \frac{SXY}{SXX}$$

2.3.3

$Var(\hat{\alpha}) = Var(\bar{y}) = \dfrac{\sigma^2}{n}$, $Var(\hat{\beta}_1) = Var(\dfrac{SXY}{SXX}) = \dfrac{\sigma^2}{SXX}$

Q3.

2.4.1

R Codes:

```
library(alr3)
data(heights)
model1=lm(Dheight~Mheight,data=heights)
summary(model1)
anova(model1)
```

From the above commands, we can get $\hat{\sigma}^2 = 2.27^2 = 5.1529$, $\hat{\beta}_0 = 29.917$,

$\hat{\beta}_1 = 0.542$, $se(\hat{\beta}_0) = 1.623$ and $se(\hat{\beta}_1) = 0.026$. $R^2 = 0.241$ means 0.241 of the

variability of the data is explained by the model. The ANOVA table shows that

$F - statistic = 435$ with $p - value =< 2\times10^{-16}$. F-test suggests that β_1 is significant or $\beta_1 \neq 0$.

Q4.

a)

$$\text{Since } y_i = \beta_0 + \beta_1 x_i + e_i \qquad \ldots\ldots(1)$$

$$\sum y_i = n\beta_0 + \beta_1 \sum x_i + \sum e_i$$

$$\text{Then } \bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{e} \qquad \ldots\ldots(2)$$

$$\text{(1)-(2), } y_i - \bar{y} = (\beta_0 + \beta_1 x_i + e_i) - (\beta_0 + \beta_1 \bar{x} + \bar{e})$$

$$\text{As a result, } y_i - \bar{y} = \beta_1 (x_i - \bar{x}) + (e_i - \bar{e})$$

b)

e_i 's are i.i.d. with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. Thus, $E(e_i^2) = \sigma^2$ and

$E(\sum e_i^2) = n\sigma^2$.

Also, $E(\bar{e}) = 0$ and $Var(\bar{e}) = \sigma^2 / n$. Thus, $E(\bar{e}^2) = \sigma^2 / n$.

$$(y_i - \bar{y})^2 = \beta_1^2 (x_i - \bar{x})^2 + (e_i - \bar{e})^2 + 2\beta_1 (e_i - \bar{e})(x_i - \bar{x})$$

$$\sum (y_i - \bar{y})^2 = \beta_1^2 \sum (x_i - \bar{x})^2 + \sum (e_i - \bar{e})^2 + 2\beta_1 \sum (e_i - \bar{e})(x_i - \bar{x})$$

$$E[\sum (y_i - \bar{y})^2] = \beta_1^2 \sum (x_i - \bar{x})^2 + \sum E(e_i - \bar{e})^2 + 2\beta_1 \sum (x_i - \bar{x})[E(e_i - \bar{e})]$$

$$E(e_i - \bar{e}) = 0$$

$$\begin{aligned} E(e_i - \bar{e})^2 &= E(e_i^2 + \bar{e}^2 - 2\bar{e}e_i) \\ &= E(e_i^2) + E(\bar{e}^2) - 2E(\bar{e}e_i) \\ &= \sigma^2 + \frac{\sigma^2}{n} - 2E(\frac{e_i^2}{n}) \\ &= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \end{aligned}$$

$$\text{Therefore, } E[\sum (y_i - \bar{y})^2] = \beta_1^2 \sum (x_i - \bar{x})^2 + (n-1)\sigma^2$$

c)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i$$

$$y_i - \hat{y}_i = y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i$$

$$y_i - \hat{y}_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$$

$$(y_i - \hat{y}_i)^2 = (y_i - \bar{y})^2 + \hat{\beta}_1^{\,2} (x_i - \bar{x})^2 - 2\hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y})$$

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 + \hat{\beta}_1^{\,2} \sum (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 + \hat{\beta}_1^{\,2} \sum (x_i - \bar{x})^2 - 2\hat{\beta}_1 \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \times \sum (x_i - \bar{x})^2$$

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 + \hat{\beta}_1^{\,2} \sum (x_i - \bar{x})^2 - 2\hat{\beta}_1^{\,2} \sum (x_i - \bar{x})^2$$

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 - \hat{\beta}_1^{\,2} \sum (x_i - \bar{x})^2$$

d)

From the results of (c), we get

$$E[\sum (y_i - \hat{y}_i)^2] = E[\sum (y_i - \bar{y})^2] - E(\hat{\beta}_1^{\,2}) \sum (x_i - \bar{x})^2$$

From the results of (b) and $E(\hat{\beta}_1^{\,2}) = Var(\hat{\beta}_1^{\,2}) + [E(\hat{\beta}_1)]^2 = \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2} + \beta_1^2$,

$$E[\sum (y_i - \hat{y}_i)^2] = [\beta_1^2 \sum (x_i - \bar{x})^2 + (n-1)\sigma^2] - [\frac{\sigma^2}{\sum (x_i - \bar{x})^2} + \beta_1^2] \sum (x_i - \bar{x})^2$$

$$E[\sum (y_i - \hat{y}_i)^2] = \beta_1^2 \sum (x_i - \bar{x})^2 + (n-1)\sigma^2 - \sigma^2 - \beta_1^2 \sum (x_i - \bar{x})^2$$

$$E[\sum (y_i - \hat{y}_i)^2] = (n-2)\sigma^2$$

Therefore, $\hat{\sigma}^2 = E[\dfrac{\sum (y_i - \hat{y}_i)^2}{n-2}] = \sigma^2$.