STAT3008 Exercise 3 Solutions

(2011-2012 2nd Semester)

Q1.

(i) Problem 2.4.2

$$E(Dheight \mid Mheight) = \hat{\alpha} + \hat{\beta}_1(Mheight - \overline{Mheight})$$

$\overline{Mheight} = 62.45$, $\hat{\beta}_1 = 0.542$

$\hat{\alpha} = \hat{\beta}_0 + \hat{\beta}_1 \overline{Mheight} = 29.917 + 0.542 * 62.45 = 63.765$.

β_1 is the expected change of height of daughter for an unit change of height of mother.

If $\beta_1 = 1$, that means on average the height of mother is the same as that of daughter.

If $\beta_1 > 1$, that means on average the height of daughter is higher than that of mother. This tells us that daughters are likely taller than their mothers. If $\beta_1 < 1$, that means on average the height of daughter is smaller than that of mother. This tells us that daughters are likely shorter than their mothers

R Codes:

```
library(alr3)
data(heights)
fit=lm(Dheight~Mheight,data=heights)
conf.intervals(fit,level=0.99)
                  0.5  %       99.5  %
(Intercept)    25.7324151  34.1024585
Mheight         0.4747836   0.6087104
```

(ii) Problem 2.4.3

R Codes:

```
predict(fit,data.frame(Mheight=64),interval="prediction",level=.99)
          fit        lwr        upr
[1,] 64.58925 58.74045 70.43805
```

$\widetilde{Dh}\Big|_{Mh=64} = 64.59$

The 99% C.I. of the required prediction is [58.74045, 70.43805].

(iii)

$$Mheight = -\frac{\beta_0}{\beta_1} + \frac{1}{\beta_1} E(Dheight \mid Mheight)$$

$$Mheight = -55.1974 + 1.8450 E(Dheight \mid Mheight)$$

(iv)

R Codes:

```
fit=lm(Mheight~Dheight,data=heights)
```

$$E(Mheight \mid Dheight) = 34.1167 + 0.4445 * Dheight$$

They are not that same.

Q2.

Problem 2.7

2.7.1

$$RSS = \sum (y_i - \beta_1 x_i)^2$$

$$\frac{\partial RSS}{\partial \beta_1} = -2\sum x_i (y_i - \beta_1 x_i)$$

$$\left.\frac{\partial RSS}{\partial \beta_1}\right|_{\beta_1 = \widehat{\beta}_1} = -2\sum x_i (y_i - \widehat{\beta}_1 x_i) = 0$$

Therefore, $\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$

$$E(\widehat{\beta}_1) = E\left(\frac{\sum x_i y_i}{\sum x_i^2}\right) = \frac{\sum x_i E(y_i)}{\sum x_i^2}$$

$$= \frac{\sum x_i (\beta_1 x_i)}{\sum x_i^2} = \beta_1 \frac{\sum x_i^2}{\sum x_i^2} = \beta_1$$

Thus, $\widehat{\beta}_1$ is unbiased for β_1.

$$Var(\widehat{\beta}_1) = Var\left(\frac{\sum x_i y_i}{\sum x_i^2}\right) = \frac{\sum x_i^2 Var(y_i)}{(\sum x_i^2)^2}$$

$$= \frac{\sum x_i^2 \sigma^2}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2}$$

$$\widehat{RSS} = \sum (y_i - \hat{\beta}_1 x_i)^2$$

$$= \sum y_i^2 + \hat{\beta}_1^2 \sum x_i^2 - 2\hat{\beta}_1 \sum x_i y_i$$

$$= \sum y_i^2 - \frac{(\sum x_i y_i)^2}{\sum x_i^2}$$

As there is only one parameter estimate in the regression, the degree of freedom of the estimated variance is (n-1).

So $\widehat{\sigma^2} = \dfrac{\sum y_i^2 - \dfrac{(\sum x_i y_i)^2}{\sum x_i^2}}{n-1}$

2.7.2

For 2.16, $\widehat{RSS_1} = SYY - \dfrac{SXY^2}{SXX}$. For 2.30, $\widehat{RSS_2} = \sum y_i^2 - \dfrac{(\sum x_i y_i)^2}{\sum x_i^2}$.

The ANOVA Table is as follows:

Source	**Df**	**SS**	**MS**	**F**
Regression	**1**	$\widehat{RSS_2}$ - $\widehat{RSS_1}$	$\widehat{RSS_2}$ - $\widehat{RSS_1}$	$\left(\widehat{RSS_2} - \widehat{RSS_1}\right) \Big/ \dfrac{\widehat{RSS_1}}{n-2}$
Residuals	**n-2**	$\widehat{RSS_1}$	$\dfrac{\widehat{RSS_1}}{n-2}$	
Total	**n-1**	$\widehat{RSS_2}$		

R codes

```
library(data)
data(snake)
m2=lm(Y~X-1,data=snake)
m1= update(m2,~.+1)
m1
anova(m2,m1)
```

You can finally find that $t^2 = F$.

2.7.3

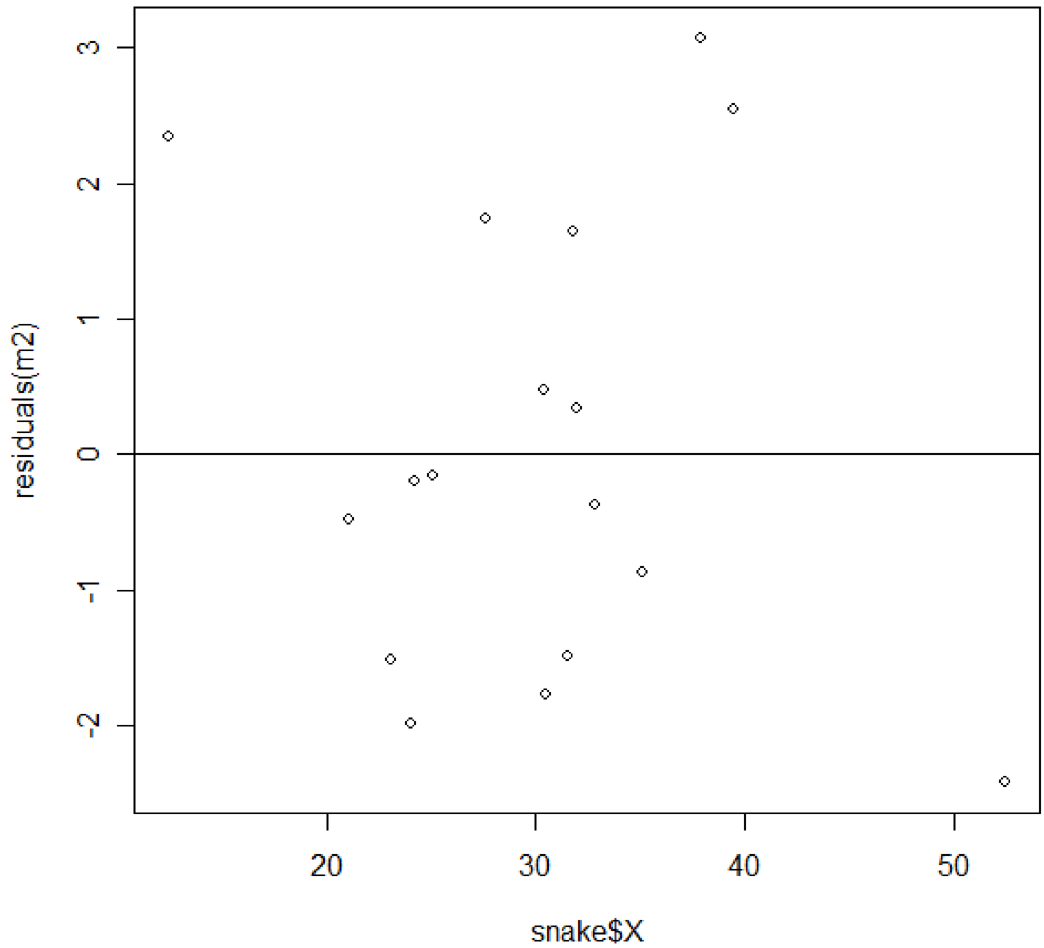
R codes

```
library(data)
data(snake)
m2=lm(Y~X-1,data=snake)
m1= update(m2,~.+1)
summary(m1)
anova(m2,m1)
```

2.7.4

R Codes:

```
plot(snake$X,residuals(m2))
abline(h=0)
```



The residual plot is okay. Only few points are very far away from zero.

Q3.

Problem 2.8

2.8.1

$$E(Y|X) = \beta_0 + \beta_1 x$$
$$= \beta_0 + \frac{\beta_1}{c}(cx)$$

Therefore, β_1 will change to $\frac{\beta_1}{c}$. However, other terms like $\widehat{\sigma^2}$, R^2 and t-statistic of $\beta_1 = 0$ won't change.

2.8.2

$$E(Y|X) = \beta_0 + \beta_1 x$$
$$dE(Y|X) = d\beta_0 + d\beta_1 x$$
$$E(dY|X) = d\beta_0 + d\beta_1 x$$

Therefore, β_1 will change to $d\beta_1$, β_0 will change to $d\beta_0$. Also, $\widehat{\sigma^2}$ will change to $d^2\widehat{\sigma^2}$. However, R^2 and t-statistic will be unchanged.

Q4.

Problem 2.10

2.10.1

R Codes:

```
data(MWwords)
s=MWwords$HamiltonRank<=50
m1=lm(logb(Hamilton,2) ~
logb(HamiltonRank,2),MWwords,subset=s)
summary(m1)
plot(logb(MWwords$HamiltonRank[s],2),logb(MWwords$Hamilto
n[s],2))
abline(m1)
```

$\hat{\beta}_0 = 6.883$

$\hat{\beta}_1 = -1.008$

The use of base-2 logarithms is not relevant because it only changes the intercept but not the slope.

2.10.2

Testing $b = 1$ is equivalent to testing $\beta_1 = -1$. Therefore, the test statistic is

$\frac{\hat{\beta}_1 - (-1)}{se(\hat{\beta}_1)}$. Thus, it is -0.59925. The critical value is 2.0106 compared to t distribution with degree of freedom 48. Thus, 0.59925<2.0106. As a result, there is no strong evidence against $b = 1$.

Q5.

Problem 2.12

2.12.1

```
library(alr3)
data(oldfaith)
m1=lm(Interval~Duration,data=oldfaith)
summary(m1)
```

The prediction equation is $E(Interval|Duration) = 33.9878 + 0.1769 \times Duration$.

2.12.2

R Codes:

```
predict(m1,data.frame(Duration=c(250)),interval="confidence")
```

Therefore, the confidence interval of the fitted value $E(Interval|Duration = 250)$ is:

[77.36915, 79.03794].

2.12.3

R Codes:

```
predict(m1,data.frame(Duration=c(250)),interval="prediction")
```

This question asks you about prediction but not fitted value. Therefore, the 95% C.I. of the prediction $Interval|Duration = 250$ is: [66.35401, 90.05307].

2.12.4
It is a problem of prediction. It is asking about the 0.90 quantile of the distribution of the prediction $Interval \mid Duration = 250$. If we can find the upper bound of a 0.80(100%) prediction interval, that is the required answer. Thus, the quantile is 85.9356.
R Codes:
predict(m1,data.frame(Duration=c(250)),interval="prediction",level=0.80)

Q6.

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{\sum (x_i - \bar{x}) y_i}{SXX}$$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov(\frac{\sum y_i}{n}, \frac{\sum (x_i - \bar{x}) y_i}{SXX})$$

$$= \sum Cov(\frac{y_i}{n}, \frac{(x_i - \bar{x}) y_i}{SXX})$$

$$= \sum \frac{(x_i - \bar{x})}{nSXX} Cov(y_i, y_i)$$

$$= \sigma^2 \sum \frac{(x_i - \bar{x})}{nSXX} = 0$$

Q7.

$$X^T X = \begin{pmatrix} 7 & 21 \\ 21 & 107 \end{pmatrix}$$

$$XX^T = \begin{pmatrix} 10 & 7 & 16 & 4 & 7 & 25 & 1 \\ 7 & 5 & 11 & 3 & 5 & 17 & 1 \\ 16 & 11 & 26 & 6 & 11 & 41 & 1 \\ 4 & 3 & 6 & 2 & 3 & 9 & 1 \\ 7 & 5 & 11 & 3 & 5 & 17 & 1 \\ 25 & 17 & 41 & 9 & 17 & 65 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$(X^T X)^{-1} = \begin{pmatrix} 0.3474 & -0.0682 \\ -0.0682 & 0.0227 \end{pmatrix}$$

$tr(X^T X)$=114

$tr(XX^T)$=114