MA 575, Linear Models : Homework 2

Question 1

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2) \\
&= \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\bar{x}\sum_{i=1}^{n}x_i \\
&= \sum_{i=1}^{n}(x_i^2-x_i\bar{x}) \\
&= \sum_{i=1}^{n}(x_i-\bar{x})x_i
\end{aligned}$$

Question 2

$$RSS(\beta_0,\beta_1) = \sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))^2$$

$$\begin{aligned}
\Rightarrow \frac{\partial RSS(\beta_0,\beta_1)}{\beta_0} &= \sum_{i=1}^{n}\left(\left.\frac{\partial z^2}{\partial z}\right|_{z=y_i-(\beta_0+\beta_1x_i)}\frac{y_i-(\beta_0+\beta_1x_i)}{\partial\beta_0}\right) \\
&= \sum_{i=1}^{n}2(y_i-(\beta_0+\beta_1x_i))(-1) \\
&= -2\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \frac{\partial RSS(\beta_0,\beta_1)}{\beta_1} &= \sum_{i=1}^{n}\left(\left.\frac{\partial z^2}{\partial z}\right|_{z=y_i-(\beta_0+\beta_1x_i)}\frac{y_i-(\beta_0+\beta_1x_i)}{\partial\beta_1}\right) \\
&= \sum_{i=1}^{n}2(y_i-(\beta_0+\beta_1x_i))(-x_i) \\
&= -2\sum_{i=1}^{n}x_i(y_i-(\beta_0+\beta_1x_i))
\end{aligned}$$

Question 3

$$\begin{aligned} & \frac{\partial RSS(\beta_0,\beta_1)}{\beta_0} && = 0 \\ \Leftrightarrow \quad & -2\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i)) && = 0 \\ \Leftrightarrow \quad & n\bar{y} - n\beta_0 - n\beta_1\bar{x} && = 0 \\ \Leftrightarrow \quad & \beta_0 && = \bar{y} - \beta_1\bar{x} \end{aligned}$$

$$\begin{aligned} & \frac{\partial RSS(\beta_0,\beta_1)}{\beta_1} && = 0 \\ \Leftrightarrow \quad & -2\sum_{i=1}^{n} x_i(y_i - (\beta_0 + \beta_1 x_i)) && = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{n} x_i y_i - n\beta_0\bar{x} - \beta_1\sum_{i=1}^{n} x_i^2 && = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{n} x_i y_i - n(\bar{y} - \beta_1\bar{x})\bar{x} - \beta_1\sum_{i=1}^{n} x_i^2 && = 0 \end{aligned}$$

Therefore,

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} + y_i\bar{x} - \bar{y}\bar{x})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i^2 - x_i\bar{x})} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Hence,

$$\beta_1 = \frac{SXY}{SXX}$$

Problem 1.2

1.2.1

When observing the graph about the dependence of soil temperature on month number, it does not seem that that the temperature is correlated to the month number.

1.2.2

If we redraw the previous graph (using the R-code below) but make sure that the x-axis is at least for times longer than the y-axis, we obtain the graph below. We notice that a correlation now appears between the time of the year and the soil temperature.

Figure 1: Soil temperature vs the month number

```
library(alr3)

# Attach the Mitchell data file
attach(Mitchell)

# Plot the temperature on month number
plot(Mitchell$Month,Mitchell$Temp,xlab="Month after January 1976",ylab="Average Soil Temperature",asp=1/4)
```

Problem 2.1

2.1.1

Let's draw the scatter plot of the weight of our individuals versus their size.

```
#Attach the htwt file
attach(htwt)

#2.1.1 Scatterplot Wt vs Ht
plot(Ht,Wt,xlab="Height",ylab="Weight")
grid()
```

Figure 2: Weight vs Height

From this graph and because the sample's size is small ($n = 10$), it is hard to say if fitting a linear regression is appropriate. However, we can fit one and use statistical tests to answer that question.

2.1.2

To answer the first part of the question, we remind you that :

$$\bar{x} = \sum_{i=1}^{n} x_i; \bar{y} = \sum_{i=1}^{n} y_i$$

$$SXX = \sum_{i=1}^{n} (x_i - \bar{x})^2; SYY = \sum_{i=1}^{n} (y_i - \bar{y})^2; SXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

To comupte these equations, we run the following code.

```
y_bar = mean(Wt)
x_bar = mean(Ht)

SXX = sum((Ht-x_bar)^2)
SXY = sum((Ht-x_bar)*(Wt-y_bar))
SYY = sum((Wt-y_bar)^2)

data.frame(x_bar = x_bar, y_bar=y_bar, SXX = SXX, SXY = SXY, SYY = SYY)

x_bar y_bar     SXX     SXY     SYY
1 165.52 59.47 472.076 274.786 731.961
```

To obtain an estimation of the parameters of the linear regression, we use the formulas derived in Question 3.

```
Beta1 = SXY/SXX
Beta0 = y_bar - Beta1*x_bar

c(Beta0,Beta1)
[1] -36.87588   0.58208

y_predicted = Beta0 + Beta1*Ht

plot(Ht,Wt,xlab="Height",ylab="Weight")
lines(Ht,y_predicted)
grid()
```

Figure 3: Weight vs Height

2.1.3

Reminder :

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2}; \widehat{Var}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right); \widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SXX}; \widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\hat{\sigma}^2 \frac{\bar{x}}{SXX}$$

```
n   = length(Ht)
RSS = sum((y_predicted - Wt)^2)
s2  = RSS/(n-2)

var_Beta0 = s2*(1/n + x_bar^2/SXX)
var_Beta1 = s2/SXX

cov_Beta0_Beta1 = -s2*x_bar/SXX

c(s2,var_Beta0,var_Beta1,cov_Beta0_Beta1)
[1]   71.5016955 4156.7419408    0.1514623  -25.0700325
```

We now have enough material to compute statistical tests and check wether or not a linear regression is appropriate for this sample.

First, we will test the hypothesis $H_0 : \hat{\beta}_0 = 0$ at a significance level $\alpha = 5\%$. We can formulate this problem as follows :

$$\mathbb{P}_{H_0}(rejectH_0) = \mathbb{P}_{H_0}(|\hat{\beta}_0| \geq z_{1-\alpha}) = \alpha \Rightarrow \mathbb{P}_{H_0}(-z_{1-\alpha} \leq \hat{\beta}_0 \leq z_{1-\alpha}) = 1 - \alpha$$

By noticing that :

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right) \right)$$

We can write :

$$\mathbb{P}_{H_0}\left(\frac{-z_{1-\alpha} - \beta_0}{\sigma\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{SXX})}} \leq \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{SXX})}} \leq \frac{\hat{z}_{1-\alpha} - \beta_0}{\sigma\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{SXX})}} \right)$$

Under the hypothesis H_0, we have$\beta_0 = 0$, therefore :

$$\mathbb{P}_{H_0}\left(\frac{-z_{1-\alpha}}{\sigma\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)}} \leq \frac{\hat{\beta}_0}{\sigma\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)}} \leq \frac{z_{1-\alpha}}{\sigma\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)}}\right)$$

In our case, we cannot use he statistic $\frac{\hat{\beta}_0}{\sigma\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)}}$ because σ is an unknown parameter. However, it can be easily proved that :

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2(n-2)}{n-2}$$

Therefore, the statistic :

$$T = \frac{\frac{\hat{\beta}_0}{\sigma\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)}}}{\frac{\hat{\sigma}}{\sigma}} = \frac{\hat{\beta}_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \sim \tau(n-2)$$

Therefore, we would reject the H_0 if :

$$|\beta_0| \geq z_{1-\alpha} = t_{1-\alpha}\sqrt{\widehat{Var}(\hat{\beta}_0)}$$

(where $t_{1-\alpha}$ is the quantile defined such that : $\mathbb{P}(-t_{1-\alpha} \leq T \leq t_{1-\alpha})$.)

or if:

$$\frac{|\hat{\beta}_0|}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \geq t_{1-\alpha}$$

The same logic works for β_1. We will reject the hypothesis $\beta_1 = 0$ if:

$$\frac{|\hat{\beta}_1|}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \geq l_{1-\alpha}$$

where $l_{1-\alpha}$ satisfies $\mathbb{P}\left(-l_{1-\alpha} \leq U = \frac{|\hat{\beta}_1|}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \leq l_{1-\alpha}\right)$.

(U also follows a $\tau(n-2)$ distribution)

```
# t-test Statistics
alpha = 0.05
abs(c(Beta0/sqrt(var_Beta0),Beta1/sqrt(var_Beta1)))
[1] 0.5719603 1.4956517

# t-test
abs(c(Beta0/sqrt(var_Beta0),Beta1/sqrt(var_Beta1)))> qt(1-alpha/2,n-2)
[1] FALSE FALSE
```

In the case of β_0, to compute the two-tailed p-values we need to estimate :

$$\mathbb{P}(T \leq -|t| \cup T \geq |t|)$$

which is equivalent to :

$$p_{val_{\beta_0}} = 2\left(1 - \mathbb{P}\left(T \leq |t|\right)\right) \qquad \text{where} \qquad t \qquad \text{is the estimation of} \qquad \frac{\hat{\beta}_0}{\widehat{Var}(\hat{\beta}_0)}$$

By following the same train of thougts, one can obtain :

$$p_{val_{\beta_1}} = 2\left(1 - \mathbb{P}\left(U \leq |u|\right)\right) \qquad \text{where} \qquad u \qquad \text{is the estimation of} \qquad \frac{\hat{\beta}_1}{\widehat{Var}(\hat{\beta}_1)}$$

```
pval_Beta0 = 2*(1 - pt(abs(Beta0/sqrt(var_Beta0)),n-2))
pval_Beta1 = 2*(1 - pt(abs(Beta1/sqrt(var_Beta1)),n-2))
c(pval_Beta0,pval_Beta1)
[1] 0.5830589 0.1731089
```

2.1.4

Source	df	SS	MS	F	p-value
Regression on Weight	1	159.9474	159.9474	2.236974	0.1731089
Residuals	8	572.0136	71.5017		

Table 1: ANOVA table

Reminder : $p - val = \mathbb{P}(U > F)$ where $U \sim F(1, n-2)$.

In addition, one can notice that $F = \frac{|\hat{\beta}_1|}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = 2.236974$.

Last but not least, the linear regression can be fitted and tested directly using the following R-code :

```
HtWt.lm <- lm(Wt~Ht , data=htwt)  # Fit linear model
summary(HtWt.lm)                  # (t-test)
anova(HtWt.lm)                    # (ANOVA)
```