RSFAS ASSIGNMENT COVER SHEET

Australian National University

Submission and assessment is anonymous where appropriate and possible. Please do not write your name on this coversheet.

This coversheet must be attached to the front of your assessment when submitted in hard copy. If you have elected to submit in hard copy rather than Turnitin, you must provide copies of all references included in the assessment item.

Student ID: u7192040

Course Code and Name: 2008 regression modelling

Assignment Number: no.1

Assignment Due Date: 21/04/21

Lecturer: **Xuan Liang**

Tutor: Xian Li

Tutorial number, day and time: Friday 8am-9am

Word Count: 750

I declare that this work:

- upholds the principles of academic integrity, as defined in the ANU Policy: Code of Practice for Students University Academic Misconduct Rules;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;
- is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

Signed: yuyu sui

Dated Submitted: 20/04/21

Q1

a) $Cov(b_0, b_1) = E[b_0 - E(b_0)][b_1 - E(b_1)]$

$= E(\bar{Y} - b_1\bar{X} - E(b_0)][b_1 - E(b_1)]$

$= E[\bar{Y} - b_1\bar{X} - E(\bar{Y} - b_1\bar{X})][b_1 - \beta_1]$

$= E[\bar{Y} - b_1\bar{X} - E(\bar{Y}) - b_1\bar{X}][b_1 - \beta_1]$

$= E[\bar{Y} - b_1\bar{X} - \bar{Y} + \beta_1\bar{X}][b_1 - \beta_1]$

$= E[\beta_1\bar{X} - b_1\bar{X}][b_1 - \beta_1]$

$= E[(\beta_1 - b_1)\bar{X}][b_1 - \beta_1]$

$= E[-\bar{X}(\beta_1 - b_1)^2]$

$= E[-\bar{X}] \cdot E[(\beta_1 - b_1)^2]$

$= -\bar{X} \cdot var(b_1)$

$var(b_1) = var \sum\left(\frac{(x_i - \bar{x})}{S_{xx}} Y_i\right)$

$= \sum \frac{(x_i - \bar{x})^2}{S^2} \cdot var(Y_i) + 2\sum \frac{(x_i - \bar{x})(x_j - \bar{x})}{S^2} Cov(Y_i, Y_j)$

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

$\therefore Cov(b_0, b_1) = -\bar{x} \cdot \frac{\sigma^2}{S} = -\frac{\bar{x}}{S}\sigma^2$

b) ① $x = 1:10$

$\bar{x} = 5.5$

$S_{xx} = (n-1) \cdot S_x^2 = 82.5$

$\because$ random $(n, 0, 2)$ $\therefore \mu = 0 \quad \sigma = 2 \quad \therefore \sigma^2 = 4$

$\therefore Cov(b_0, b_1) = -\frac{\bar{x}}{S} \cdot \sigma^2$

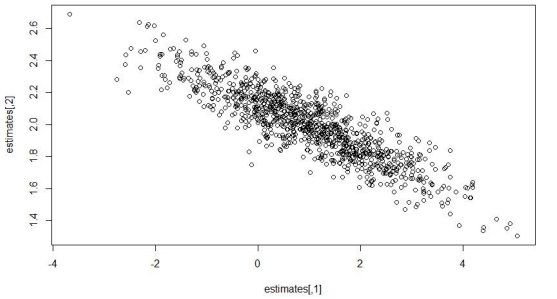
$= -\frac{5.5}{82.5} \times 4 = -0.2667$

② $S_{b_1} = \sigma \sqrt{\frac{1}{(n-1) S_x^2}} = 2 \sqrt{\frac{1}{82.5}} = 0.2202$

$S_{b_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1) S_x^2}} = 2 \sqrt{\frac{1}{10} + \frac{5.5^2}{82.5}} = 1.3663$

$r = \frac{Cov(b_0, b_1)}{S_{b_1} \cdot S_{b_0}} = \frac{-0.2667}{0.2202 \times 1.3663} = -0.8865$

Q1 c) we can see that most observations are concentrated in the center and very dense, which can roughly form a straight line. So b0 and b1 can be correlated. And there is a negative relationship between them.




```
set.seed(7192040)
rnorm(7192040)
x <- 1:10
n <- length(x)
estimates <- matrix(1, 1000, 2)
names(estimates) <- c("b0","b1")
for(r in 1:1000) {
  y <- 1 + 2*x + rnorm(n,0,2)
  estimates[r,] <- lm(y~x)$coefficients
}
plot(estimates)
```

D)

```
> b0=estimates[,1]
> cov(b0,b1)
[1] -0.2545359
> cor(b0,b1)
[1] -0.8757025
```

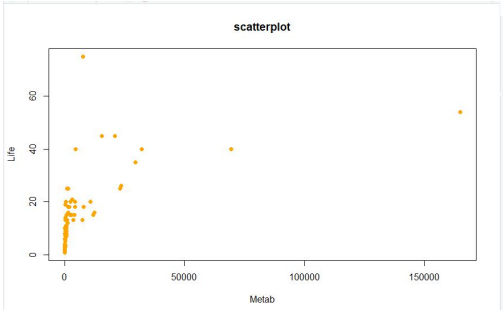
We can find that although cov and cor is not totally equal to each other, but it is roughly equal.

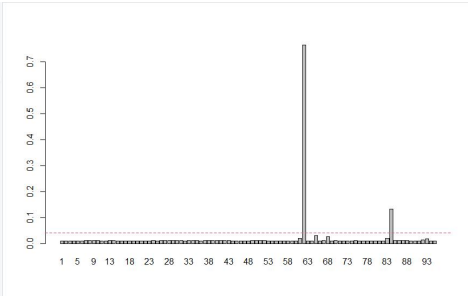
Q2

a) There is 2 high leverage observations, one is 62, the other one is 84.
No.62 name is Asian elephant, and specie is Elephas maximum
No.84 name is Bottle-nosed whale, and specie is Hyperoodon ampullatus.

```
#question 2 (a)
mammal<-read.csv("mammal.csv",header=T)
mammal
attach(mammal)
mammal.lm<-lm(Life~Metab)
summary(mammal.lm)
plot(Metab,Life,xlab="Metab",ylab="Life",main="scatterplot",pch=16,col="orange")
barplot(hatvalues(mammal.lm))
abline(h=4/length(Metab),col=2,lty=2)
which(hatvalues(mammal.lm)>4/length(Metab))
```

```
> summary(mammal.lm)

Call:
lm(formula = Life ~ Metab)

Residuals:
    Min      1Q  Median      3Q     Max
-20.205  -6.885  -2.341   3.598  61.775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.030e+01  1.124e+00   9.161 1.22e-14 ***
Metab       3.873e-04  5.740e-05   6.748 1.26e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```





b) The X and Y coordinates ranges should adjust X~ (0,10000), Y~(0,40)
It needs to hypothetical test that slope is 0 or not, it's like beta1 equal 0 or not
H0: beta1=0
H2: beta1=/0
Through ANOVA, p-valus < 1.262e-09. P-value < 0.05, so we reject H0, so beta1 is not 0.
So slope is not 0, so there is a relationship between Life and Metab.
And through summary, we know that beta1= 3.873e-04 which is above 0, so Life and Metab are positively correlated, the more Metab a specie has, the longer life it has.

```
#(b)
plot(Metab,Life,xlab="Metab",ylab="Life",xlim=c(0,10000),ylim=c(0,40),main="scatterplot",pch=16,col="orange")
abline(mammal.lm, col = "blue", lty =2, lwd = 1.5)
anova(mammal.lm)
summary(mammal.lm)
```

```
> anova(mammal.lm)
Analysis of Variance Table

Response: Life
           Df  Sum Sq Mean Sq F value    Pr(>F)
Metab       1  5083.8  5083.8  45.535 1.262e-09 ***
Residuals 93 10383.1   111.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

D

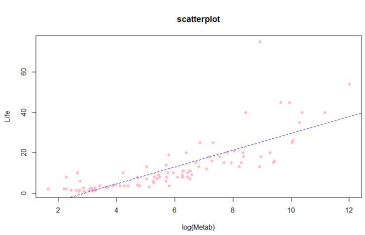
```
> summary(mammal.lm)

Call:
lm(formula = Life ~ Metab)

Residuals:
    Min      1Q  Median      3Q     Max
-20.205  -6.885  -2.341   3.598  61.775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.030e+01  1.124e+00   9.161 1.22e-14 ***
Metab       3.873e-04  5.740e-05   6.748 1.26e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.57 on 93 degrees of freedom
Multiple R-squared:  0.3287,    Adjusted R-squared:  0.3215
F-statistic: 45.53 on 1 and 93 DF,  p-value: 1.262e-09
```

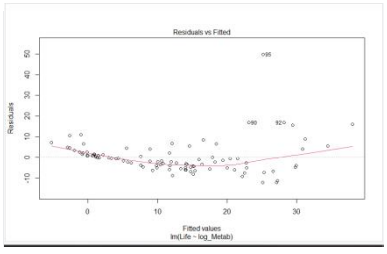
c) We can find that if we take log transformation for x, x and y relationship looks like more linear and have a stronger relationship, so we need to transform x into log(x).

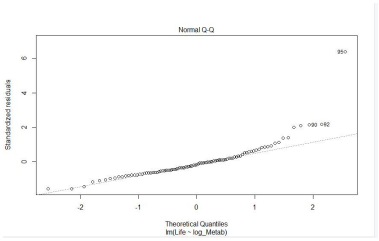
Through fitted value vs residual, we can find that most of the residuals are around zero randomly, and most between (-1,1), only 95th observation is larger than the other residuals.
Through Q-Q plot, most points are close to a heavy tail, except the 95th observation.
So the assumption is true.

```
#(c)
log_Metab<-log(Metab)
plot(log(Metab),Life,xlab="log(Metab)",ylab="Life",main="scatterplot",pch=16,col="pink")
mammal_log.lm=lm(Life~log_Metab)
anova(mammal_log.lm)
summary(mammal_log.lm)
abline(mammal_log.lm, col = "blue", lty =2, lwd = 1.5)
plot(mammal_log.lm,which=1)
plot(mammal_log.lm,which=2)
```

```
> anova(mammal_log.lm)
Analysis of Variance Table

Response: log_Metab
          Df Sum Sq Mean Sq    F value    Pr(>F)
log_Mass   1 978.16  978.16 5.6492e+32 < 2.2e-16 ***
Residuals 93   0.00    0.00
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(mammal_log.lm)

Call:
lm(formula = log_Metab ~ log_Mass)

Residuals:
       Min         1Q     Median         3Q        Max
-1.493e-15 -1.802e-16 -1.213e-16 -4.220e-17  1.238e-14

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) -1.823e-16  1.355e-16 -1.345e+00    0.182
log_Mass     1.000e+00  4.207e-17  2.377e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.316e-15 on 93 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 5.649e+32 on 1 and 93 DF,  p-value: < 2.2e-16
```

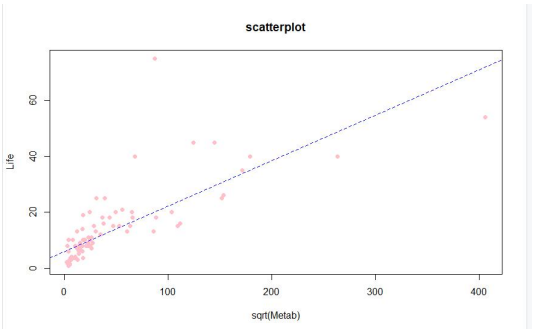


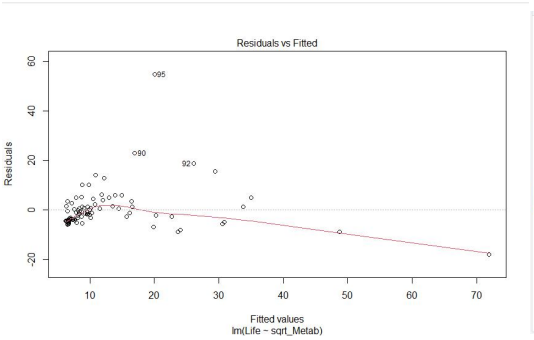


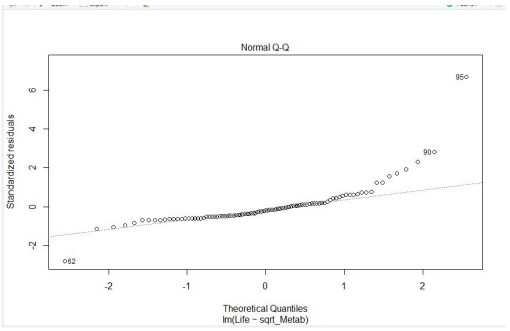


d) The correlations between log(x) and y is 0.7894737 and the correlations between sqrt(x) and y is 0.7661926, due to 0.7894737>0.7661926, and the scatterplots also show that there is a stronger relationship between x and y in log one, so we need to choose log model. The form is Life=-12.1068+ 4.1742 log(Metab)

```
#(d)
sqrt_Metab=sqrt(Metab)
plot(sqrt(Metab),Life,xlab="sqrt(Metab)",ylab="Life",main="scatterplot",pch=16,col="pink")
mammal_sqrt.lm=lm(Life~sqrt_Metab)
abline(mammal_sqrt.lm, col = "blue", lty =2, lwd = 1.5)
cor(sqrt_Metab,Life)
cor(log_Metab,Life)
plot(mammal_sqrt.lm,which=1)
plot(mammal_sqrt.lm,which=2)
```

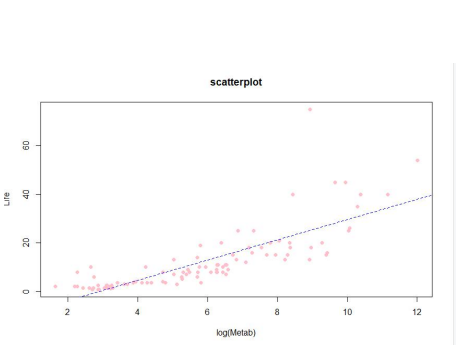


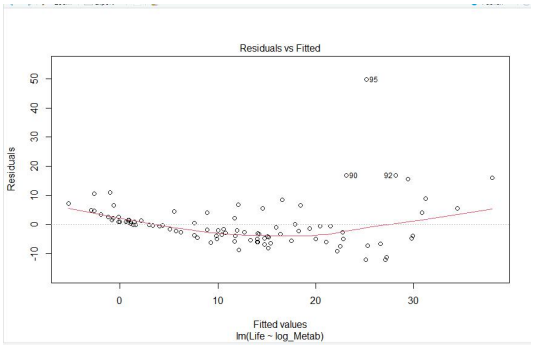


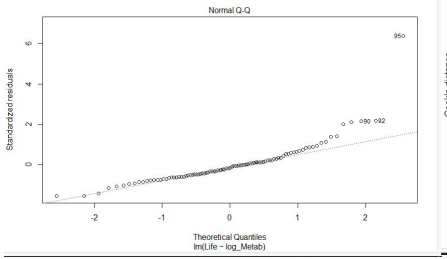


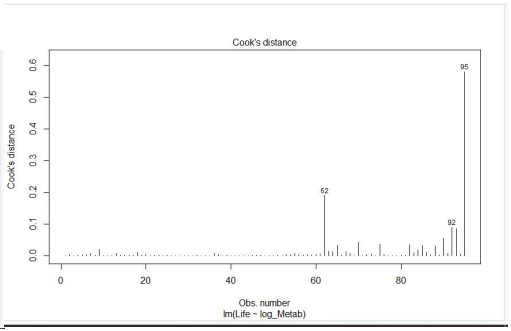
e) Finally, we can find that using the log transformation can make the observations more look like a simple linear regression, so the final simple linear regression model is Life=-12.1068+ 4.1742 log(Metab). And 95th, 90th and 92th points are higher than the normal one in fitted value vs residual, the rest are around zero. The Q-Q plot also shows it is a heavy tight and 62th is much further below the main line and 90th and 95th is much higher than it. And through cook's distance, we can get the 95th is the furthest observation in the data. But overall, most of the data fit this model.

```r
#(e)
log_Metab<-log(Metab)
plot(log(Metab),Life,xlab="log(Metab)",ylab="Life",main="scatterplot",pch=16,col="pink")
mammal_log.lm=lm(Life~log_Metab)
anova(mammal_log.lm)
summary(mammal_log.lm)
abline(mammal_log.lm, col = "blue", lty =2, lwd = 1.5)
plot(mammal_log.lm,which=1)
plot(mammal_log.lm,which=2)
plot(mammal_log.lm,which=4)
```









f) H0: beta1=0

H1: beta1=/0

Through ANOVA, we can get p-value is < 2.2e-16, p-value < 0.05, so reject H0, so slope is not 0.

The 95% confidence interval : lower = 3.505928 ,upper= 4.842445

```r
#(f)
anova(mammal_log.lm)
coef(mammal_log.lm)
b1=coef(mammal_log.lm)[2]
b1
summary(mammal_log.lm)$coef
seb_1=summary(mammal_log.lm)$coef[2,2]
seb_1
mammal_log.lm$df
c(b1-qt(0.975,mammal_log.lm$df)*seb_1,b1+qt(0.975,mammal_log.lm$df)*seb_1)
```

g) In order to conduct the model is significant or not, we need to assume the standard error is independent of Metab. And we assume that ϵ mean is 0 and has constant variance and is in accordance with normal distribution. So we need to test beta1 is 0 or not.

H0: beta1=0

H1: beta1=/0

Through ANOVA, we can get p-value is < 2.2e-16, p-value < 0.05, so reject H0, so beta1 is not 0.

So the conclusion is that model (e) is significant

```
> #(g)
> anova(mammal_log.lm)
Analysis of Variance Table

Response: Life
           Df Sum Sq Mean Sq F value    Pr(>F)
log_Metab  1 9640.0  9640.0  153.86 < 2.2e-16 ***
Residuals 93 5826.8    62.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

h) The 90% confidence interval : lower =2.385637 ,upper= 4.05859

```
#(h)
log_Metab_1=log(8000)
predict(mammal_log.lm, newdata=data.frame(Metab=log_Metab_1), interval="confidence",level=0.9)
predict(mammal_log.lm, newdata=data.frame(Metab=log_Metab_1), interval="prediction",level=0.9)
```

I) Due to question, we know that Metab=a Mass ^ (3/4)

Log(Metab)= 3/4 log(Mass) + log(a)

Log (Metab)=beta1log(Mass)+beta0

H0:beta1=0

H1:beta1 =/ 0

Through summary, we can find that p= <2e-16 , p>0.05, so fail to reject H0, beta1=0. So there is not a strong relationship between log Mass and log Metab.

H0:beta0=0

H1:beta0=1

Through summary, p= 0.182, which is >0.05, so fail to reject H0, sp beta0=0. So the intercept=0, so we couldn't intercept value in this situation.

So the conclusion is that Log(Metab) and Log(Mess) doesn't have a strong linear relationship.

```
#(i)
log_Mass=log(Mass)
mammal_log.lm=lm(log_Metab~log_Mass)
summary(mammal_log.lm)
```

Appendix-Rcode

```
#Q1 c)
set.seed(7192040)
rnorm(7192040)
x <- 1:10
```

```
n <- length(x)
estimates <- matrix(1, 1000, 2)
names(estimates) <- c("b0","b1")
for(r in 1:1000) {
   y <- 1 + 2*x + rnorm(n,0,2)
   estimates[r,] <- lm(y~x)$coefficients
}
plot(estimates)

#Q1 d)
b0=estimates[,1]
b1=estimates[,2]
cov(b0,b1)
cor(b0,b1)

#question 2 (a)
mammal<-read.csv("mammal.csv",header=T)
mammal
attach(mammal)
mammal.lm<-lm(Life~Metab)
summary(mammal.lm)
plot(Metab,Life,xlab="Metab",ylab="Life",main="scatterplot",pch=16,col="orange")
barplot(hatvalues(mammal.lm))
abline(h=4/length(Metab),col=2,lty=2)
which(hatvalues(mammal.lm)>4/length(Metab))

#(b)
plot(Metab,Life,xlab="Metab",ylab="Life",xlim=c(0,10000),ylim=c(0,40),main="scatterplot",pc
h=16,col="orange")
abline(mammal.lm, col = "blue", lty =2, lwd = 1.5)
anova(mammal.lm)
summary(mammal.lm)

#(c)
log_Metab<-log(Metab)
plot(log(Metab),Life,xlab="log(Metab)",ylab="Life",main="scatterplot",pch=16,col="pink")
mammal_log.lm=lm(Life~log_Metab)
anova(mammal_log.lm)
summary(mammal_log.lm)
abline(mammal_log.lm, col = "blue", lty =2, lwd = 1.5)
plot(mammal_log.lm,which=1)
plot(mammal_log.lm,which=2)

#(d)
```

```
sqrt_Metab=sqrt(Metab)
plot(sqrt(Metab),Life,xlab="sqrt(Metab)",ylab="Life",main="scatterplot",pch=16,col="pink")
mammal_sqrt.lm=lm(Life~sqrt_Metab)
abline(mammal_sqrt.lm, col = "blue", lty =2, lwd = 1.5)
cor(sqrt_Metab,Life)
cor(log_Metab,Life)
plot(mammal_sqrt.lm,which=1)
plot(mammal_sqrt.lm,which=2)

#(e)
log_Metab<-log(Metab)
plot(log(Metab),Life,xlab="log(Metab)",ylab="Life",main="scatterplot",pch=16,col="pink")
mammal_log.lm=lm(Life~log_Metab)
anova(mammal_log.lm)
summary(mammal_log.lm)
abline(mammal_log.lm, col = "blue", lty =2, lwd = 1.5)
plot(mammal_log.lm,which=1)
plot(mammal_log.lm,which=2)
plot(mammal_log.lm,which=4)

#(f)
anova(mammal_log.lm)
coef(mammal_log.lm)
b1=coef(mammal_log.lm)[2]
b1
summary(mammal_log.lm)$coef
seb_1=summary(mammal_log.lm)$coef[2,2]
seb_1
mammal_log.lm$df
c(b1-qt(0.975,mammal_log.lm$df)*seb_1,b1+qt(0.975,mammal_log.lm$df)*seb_1)

#(g)
anova(mammal_log.lm)

#(h)
log_Metab_1=log(8000)
predict(mammal_log.lm,newdata=data.frame(Metab=log_Metab_1),interval="confidence",le
vel=0.9)
predict(mammal_log.lm,newdata=data.frame(Metab=log_Metab_1),interval="prediction",lev
el=0.9)

#(i)
log_Mass=log(Mass)
mammal_log.lm=lm(log_Metab~log_Mass)
```

```
summary(mammal_log.lm)
```