STAT4038 Assignment 1

QUESTION 1

(a) Plot Shield against Weight (which means that Shield should be the response variable on the Y or vertical axis and Weight should be the explanatory variable on the X or horizontal axis). Use the identify() function in R to identify any unusual data points on the plot. Discuss why you chose these observation(s) as being unusual.

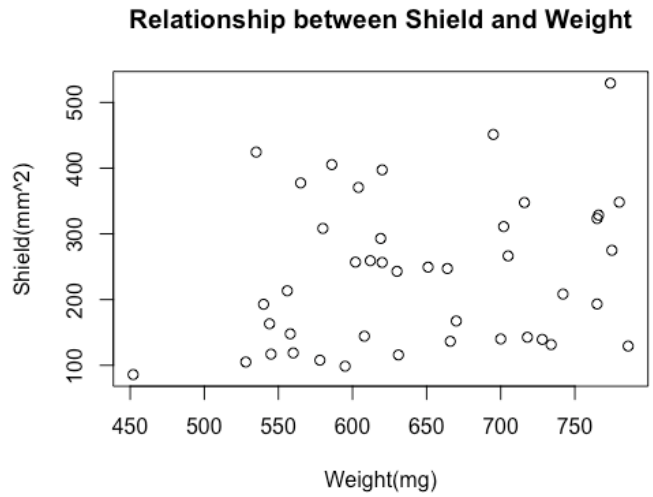

Figure 1: Shield vs Weight

```
> identify(Weight,Shield)
[1] 27
```

Overall, the plot does not show a strong linear relationship. I chose the left bottom point as the unusual observation, which is the 27^{th} observation. Because this point is $(sheild, Weight) = (85.9,452)$, which is far away from the other observations.

(b) Is there a significant correlation between Weight and Shield? Use the cor.test() function to conduct a suitable hypothesis test. Clearly specify the hypotheses you are testing and present and interpret the results.

```
> cor.test(Weight, Shield)
        Pearson's product-moment correlation
data:  Weight and Shield
t = 1.5793, df = 41, p-value = 0.122
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06559203  0.50359325
sample estimates:
    cor
0.2394694
```

Test $H_0: \rho = 0$ vs $H_A: \rho \neq 0$, where ρ is the correlation between Weight and Shield. $t_{95} = 1.5793$, $p > 0.05$, so cannot reject H_0. Conclude that ρ is NOT significantly different from 0. The observed sample correlation $r = 0.2394694$ also suggests the correlation between Weight and Shield is not strong.

(c) Experiment with applying natural log transformations (to the base e, which is the default for the log() function in R) and square root transformations to one or both of Weight and Shield, and repeat the analysis in parts (a) and (b). Do NOT show all of your results, just pick whichever one you think is the best choice of scale for the two variables and show and discuss the results for your chosen combination.

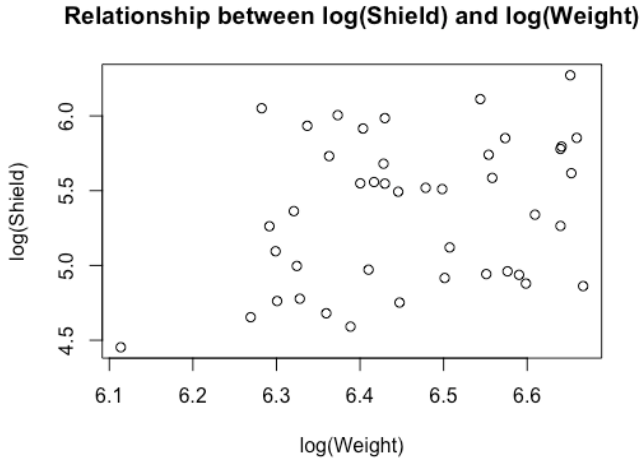

Figure 2: log(Shield) vs log(Weight)

```
>identify(log(Weight), log(Shield))
[1] 27
> cor.test(log(Weight), log(Shield))
        Pearson's product-moment correlation
data:  log(Weight) and log(Shield)
t = 1.9709, df = 41, p-value = 0.05552
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.006763403  0.546257547
sample estimates:
    cor
0.294178
```

I finally choose log($Shield$) vs log($Weight$). Because it has the largest t-score for correlation. For unusual observation, I still choose the left bottom observation (27^{th}) because it is still far away from the other observations.

For correlation, Test $H_0: \rho = 0$ vs $H_A: \rho \neq 0$, where ρ is the correlation between $\log(Weight)$ and $\log(Shield)$.
$t_{95} = 1.9709, p = 0.05552 > 0.05$, so still cannot reject H_0. Conclude that ρ is NOT significantly different from 0. The observed sample correlation $r = 0.294178$ suggests the correlation between $\log(Weight)$ and $\log(Shield)$ is not strong.
In conclusion, even choosing the combination which having the largest t-score, these two values still do not show significant correlation. So we can conclude that Shield and Weight do not have significant correlation.

(d) Fit a simple linear regression (SLR) model with your chosen transformation of Shield as the response variable and your chosen transformation of Weight as the predictor. Construct a plot of the residuals against the fitted values, a normal Q-Q plot of the residuals, a bar plot of the leverages for each observation and a bar plot of Cook's distances for each observation. Use these plots to comment on the model assumptions and on any unusual data points.

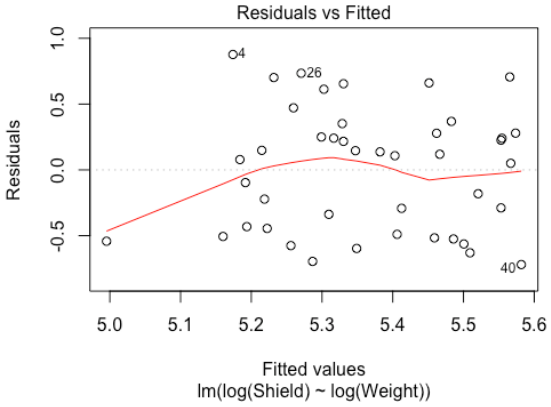

Figure 3: Residual plot for log(Shield)~log(Weight)

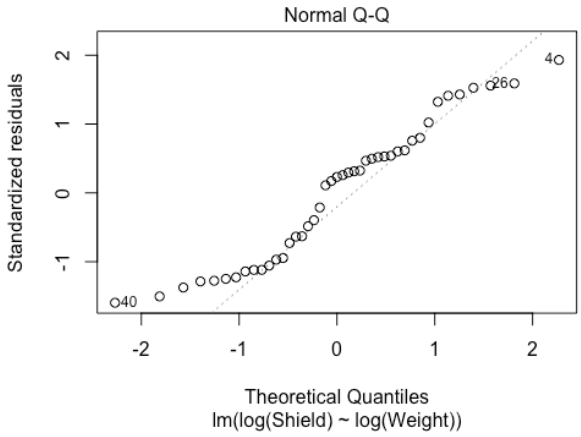

Figure 4: Normal Q-Q plot for log(Shield)~log(Weight)

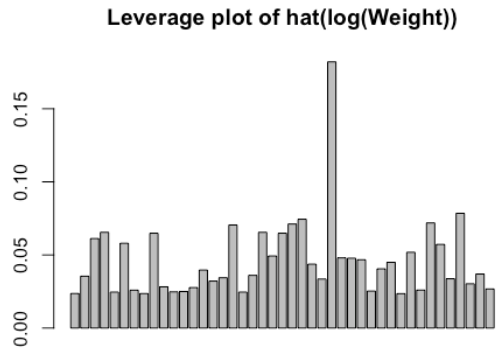

Figure 5: Leverage plot for log(Shield)~log(Weight)

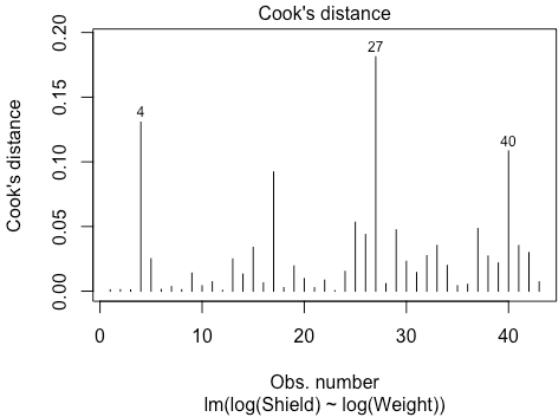

Figure 6: Cook's distance for log(Shield)~log(Weight)

The residual plot shows a possible outlier in the bottom left corner. Apart from that, the residual plot does not show other problems with the assumptions. In other word, residual looks independent and has constant variance.
The normal Q-Q plot shows the distribution is light tailed, which means the data points are less spread out than would have been expected if they were truly normal distributed.
The leverage plot shows that there is one point having very high leverage, which means $\log(Weight)$ of this data point is far from the sample average. And from Cook's distance, it can be found that several unusual points (e.g. point 4, 27, 40) have heavier influence to the model than other points, which can be further investigated.

(e) Produce the ANOVA (Analysis of Variance) table for the SLR model in part (d) and interpret the results of the F test. What is the coefficient of determination for this model and how should you interpret this summary measure?

```
> anova(moorhen.lm)
Analysis of Variance Table
Response: log(Shield)
             Df Sum Sq Mean Sq F value  Pr(>F)
log(Weight)  1 0.8569 0.85689  3.8843 0.05552 .
Residuals   41 9.0447 0.22060
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0: \frac{\sigma^2_{model}}{\sigma^2_{error}} = 1 \qquad H_A = \frac{\sigma^2_{model}}{\sigma^2_{error}} > 1$
$F_{1,95} = 3.8843$, $p = 0.05552 > 0.05$, so cannot reject H_0. Conclude the variance explained by the model is NOT large compared to the error variance, which means the model involving Weight is NOT explaining a significant proportion of variability in Shield.
It can be found that p-value for F-test is equal to the p-value for t-test on correlation in (b). Actually for simple linear regression, these two tests are equivalent.

The coefficient of determination $R^2 = \frac{SSR}{SSR+SSE} = \frac{0.8569}{0.8569+9.0447} = 0.08654 = 8.654\%$. The R^2 value means only 8.654% of variation of the response is explained by the model. Hence this model is not good for explaining variation.

(f) What are the estimated coefficients of the SLR model in part (d) and the standard errors associated with these coefficients? Interpret the values of these estimated coefficients and perform t-tests to test whether or not these coefficients differ significantly from zero. What do you conclude as a result of these t-tests?

```
> summary(moorhen.lm)

Call:
lm(formula = log(Shield) ~ log(Weight))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.7196 -0.4674  0.1076  0.2787  0.8769

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4844     3.4757  -0.427   0.6716
log(Weight)   1.0599     0.5378   1.971   0.0555 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4697 on 41 degrees of freedom
Multiple R-squared:  0.08654,    Adjusted R-squared:  0.06426
F-statistic: 3.884 on 1 and 41 DF,  p-value: 0.05552
```

Model:

$$\log(Shield) = \beta_0 + \beta_1 \log(Weight) + \varepsilon \quad \varepsilon \sim i.i.d.\ N(0, \sigma^2)$$

$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$
$t_{95} = 1.971, p = 0.0555 > 0.05$, so cannot reject H_0. Conclude that the slope coefficient of $\log(Weight)$ is NOT significantly different from 0, implying that there is no significant linear relationship between $\log(Sheild)$ and $\log(Weight)$. Notice that this test is also equivalent to the tests in (b) and (e) for simple linear regression model.

$H_0: \beta_0 = 0 \quad H_1: \beta_0 \neq 0$
$t_{95} = -0.427, p = 0.6716 > 0.05$, so cannot reject H_0. Conclude that the intercept is not significantly different from 0.

In conclusion, according to previous two t-tests, both slope and intercept of the simple linear regression model are NOT significantly different from 0, which means there is no significant linear relationship between $\log(Sheild)$ and $\log(Weight)$.

(g) Repeat part (a) and again plot Shield against Weight, but this time extend both X and Y axes to include the origin. Now include the transformed SLR model from part (d) as a curve on your plot and also include the untransformed SLR of Shield against Weight as a line on the plot. Use different line types for the two curves and also include an appropriate legend on the plot. What are you overall conclusions about the relationship between Weight and Shield, and the broader research questions discussed in the second paragraph of this question?

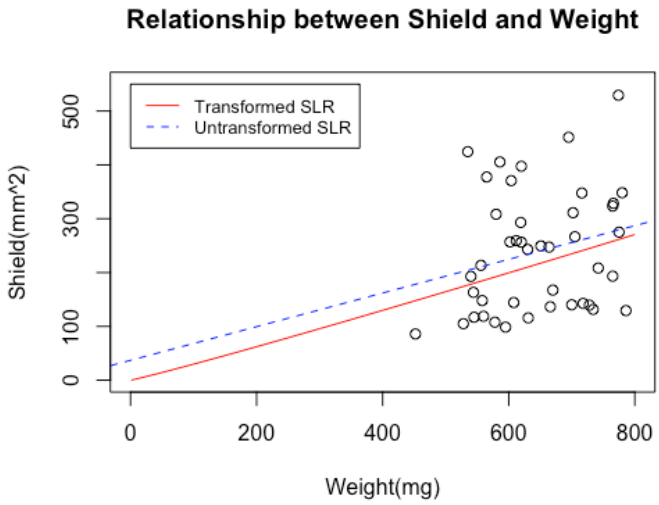

Figure 7: Transformed SLR and untransformed SLR

Overall, both transformed and untransformed SLR plots show that there is no strong linear relationship between Weight and Shield.

The broader research question can be interpreted as: whether bird's status is related to their overall size. And this research uses bird's shield to indicate bird's status, and use bird's weight to indicate bird's size. Maybe the second indicator is not properly chosen. The weight of the bird may not be a good indicator for bird's size. For instance, there might be some small and fat birds and some big and thin birds. I suggest trying to measure the beak-to-tail length of birds might be proper to indicate bird's size. In other words, replacing Weight by other measurement for size, like beak-to-tail length, may help show the relationship between moorhen's status and size.

QUESTION2:

(a) Plot body.fat against BMI. Describe the correlation shown in the plot. Would you expect a simple linear regression model to be a reasonable model for the relationship shown in the plot?

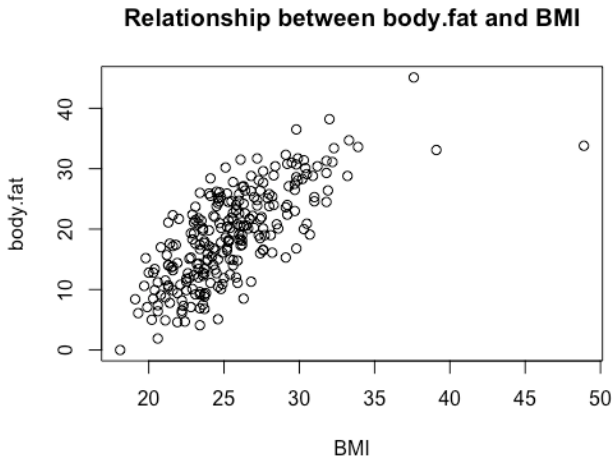

Figure 8: body.fat vs BMI

```
> cor.test(BMI,body.fat)
        Pearson's product-moment correlation
data:  BMI and body.fat
t = 16.789, df = 250, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6641703 0.7812826
sample estimates:
    cor
0.7279942
```

Test $H_0: \rho = 0$ vs $H_A: \rho \neq 0$, where ρ is correlation between body.fat and BMI.
$t_{95} = 16.789, p \ll 0.05$, so reject H_0 in favour of H_A, and conclude that ρ is significantly different from zero. The observed sample correlation $r = 0.7279942$ suggests a strong positive correlation between body.fat and BMI.
Because the correlation shows there is a positive relationship between body.fat and BMI, and the data points show a linear trend apart from some outliers. I would expect a simple linear regression model is a reasonable model for the relationship.

(b) Fit a simple linear regression (SLR) model with body.fat as the response variable and BMI as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points.

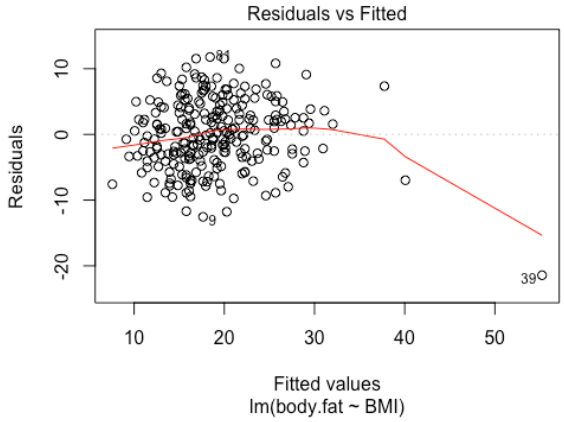

Figure 9: Residual plot for body.fat~BMI

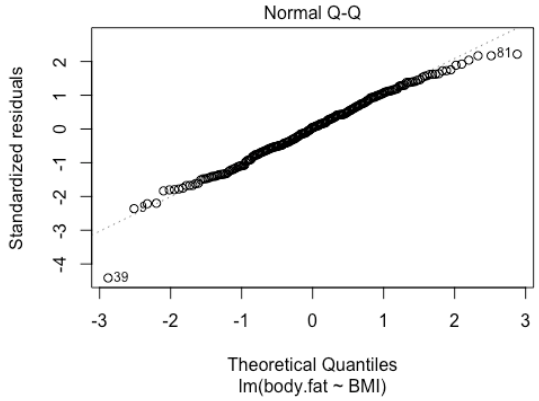

Figure 10: Normal Q-Q plot for body.fat~BMI

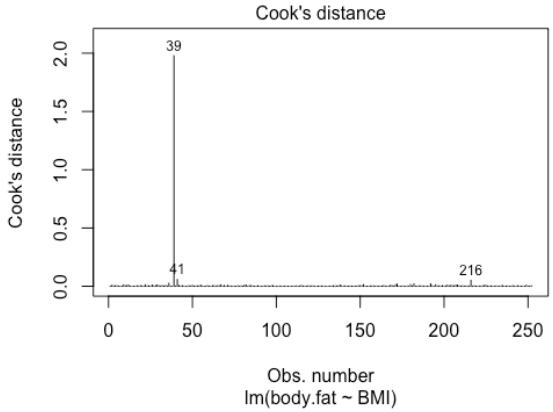

Figure 11: Cook's distance for body.fat~BMI

The "Residual vs Fitted" plot clearly shows the problem of an outlier (data point 39). Data point 39 has low $body.fat$ values with high BMI, which leads to a very low residual in the plot. Apart from that, we notice the sample variance doesn't change much and data points do not show obvious patterns (data is roughly identical and independent). Therefore, no other problems are identified.

The Normal Q-Q plot shows the data is roughly normal distributed. However, point 39 is a potential outlier, which should be investigated.

The plot of Cook's distance shows the point 39 has much larger cook's distance than the other data. Since point 39 has large leverage and residual, it has a large influence on regression model. Point 39 is an outlier and should be investigated.

(c) A natural log (to the base *e*) transformation (to one or both of the response and predictor variables) is often used to adjust the scale of the variables prior to fitting an SLR model. Now fit another SLR model with body.fat as the response variable and log(BMI) as the predictor. What would be the problem with also applying a log transformation to the response variable? Check the same plots you produced for the earlier model in part (b). Are the same problems still apparent?

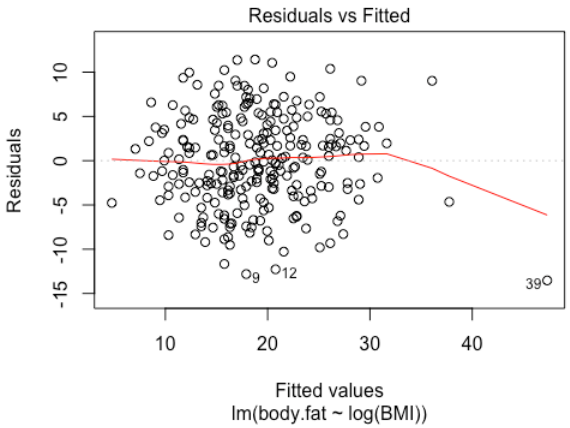

Figure 11: Residual plot for body.fat~log(BMI)

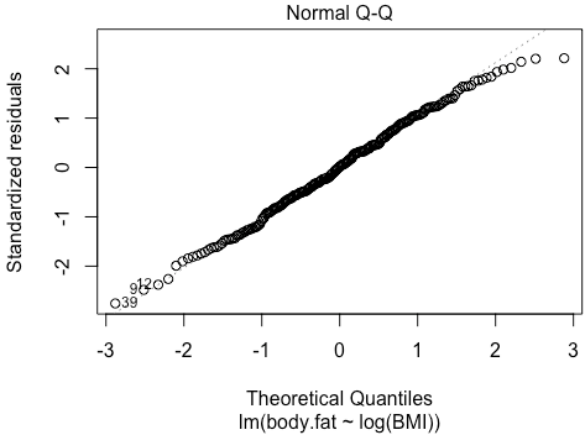

Figure 12: Normal Q-Q plot for body.fat~log(BMI)

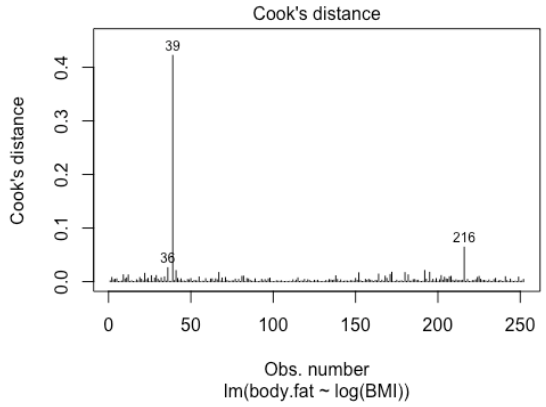

Figure 13: Cook's distance for body.fat~log(BMI)

```
> fat.loglog.lm <- lm(log(body.fat) ~ log(BMI))
Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
  NA/NaN/Inf in 'y'
```

When applying logarithm on $body.fat$, an error occurred as shown above. After checking the data in 'y' ($body.fat$), it can be found that one entry (the 182th row) has a value of 0, therefore its logarithm is not defined, which is the reason of this error information. Also, this entry can be checked by:

```
> body.fat[c(182)]
[1] 0
```

There are slightly differences between these plots. In residual plot, point 39 (a potential outlier) moves closer towards the other data points. In normal Q-Q plot, point 39 (a potential outlier) is closer to the normal distribution, and the Cook's distance for others data (other than point 39) looks slightly more obvious on the graph. The reason of these is the logarithm function. However, overall I think the problem about potential outlier (point 39) is still apparent by inspecting these three graphs.

(d) Produce the ANOVA table and the table of the estimated coefficients for the transformed SLR model in part (c). Interpret the values of the estimated coefficients for this SLR model and the results of the overall F test and the t-tests on the estimated coefficients.

```
> anova(fat.log.lm)
Analysis of Variance Table

Response: body.fat
          Df Sum Sq Mean Sq F value    Pr(>F)
log(BMI)   1 8386.5  8386.5  313.28 < 2.2e-16 ***
Residuals 250 6692.6    26.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0: \frac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1 \quad H_A: \frac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1$

$F_{1,250} = 313.28,\ p \ll 0.05$, so reject H_0 in favour of H_A, and conclude that the variance explained by the model is large compared to the error variance, which means that the model involving $body.fat$ is explaining a significant proportion of the variability in $\log(BMI)$.

```
> summary(fat.log.lm)

Call:
lm(formula = body.fat ~ log(BMI))

Residuals:
    Min      1Q  Median      3Q     Max
-13.5263  -3.3776   0.0751   3.8273  11.4306

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -119.233      7.813  -15.26   <2e-16 ***
log(BMI)     42.820      2.419   17.70   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.174 on 250 degrees of freedom
Multiple R-squared:  0.5562,  Adjusted R-squared:  0.5544
F-statistic: 313.3 on 1 and 250 DF,  p-value: < 2.2e-16
```

Model: $body.fat = \beta_0 + \beta_1 \log(BMI) + \varepsilon \quad \varepsilon \sim i.i.d.N(0, \sigma^2)$

$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$
$t_{95} = 17.70$, $p \ll 0.05$, so reject H_0 in favour of H_A. Conclude that the slope coefficient of $\log(BMI)$ is significantly different from 0, implying that there is a significant linear relationship between $\log(BMI)$ and $body.fat$. Notice that this test is also equivalent to the F-tests above.

$H_0: \beta_0 = 0 \quad H_1: \beta_0 \neq 0$
$t_{95} = -15.26$, $p \ll 0.05$, so reject H_0 in favour of H_A. Conclude that the intercept is significantly different from 0.

(e) Body mass index values less than 18.5 are typically categorised as "underweight"; from 18.5 to 25 as "normal", 25 to 30 as "overweight" and over 30 as "obese". Use the transformed SLR model from part (c) to predict the body.fat percentage for groups of males with typical BMI values 17.25 ("moderately underweight"), 21.75 ("normal"), 27.5 ("overweight") and 32.5 ("moderately obese"), respectively. Find 95% confidence intervals for these predictions. Do you think this SLR model is a good model for making these predictions?

```
> log_BMI = log(BMI)
> fat_log.lm <- lm(body.fat~log_BMI)
> bmi_predict <- c(17.25, 21.75, 27.5, 32.5)
> bmi_log <- log(bmi_predict)
> confidence <- predict(fat_log.lm,
+ newdata=data.frame(log_BMI=bmi_log), interval='confidence')
> confidence
       fit       lwr      upr
1  2.709611  0.7930648  4.626157
2 12.635297 11.6845100 13.586085
3 22.679621 21.9145334 23.444709
4 29.832835 28.4611122 31.204557
```

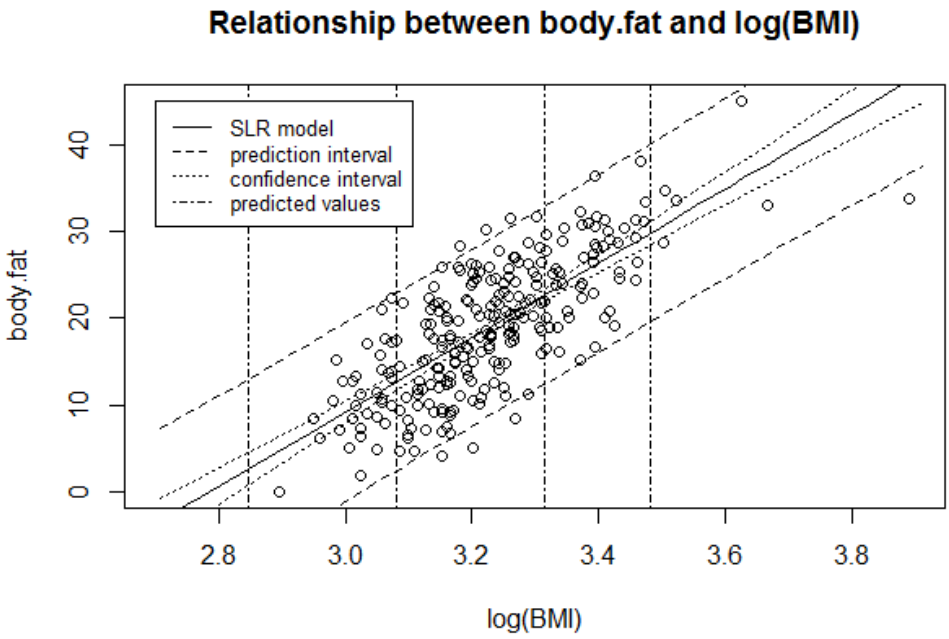

Figure 14: prediction interval and confidence interval on body.fat~log(BMI)

The 95% confidence intervals for these predictions are shown above.

Overall, this SLR model is a good predictive model.
Given a testing value x and a SLR model, confidence interval is the interval of the mean value of the responses from the SLR model, while prediction interval is the interval of the single responses from the SLR model. In this prediction problem, we are asked to predict $body.fat$ percentage for a group of males given typical BMI. Therefore, confidence interval is the appropriate interval to use.
The confidence interval values and the plot indicate that there is a significant linear relationship between $body.fat$ and $\log(BMI)$. From the plot, it can be seen that for a group of data points around the predicted values, the average $body.fat$ percentage will locate in the confidence interval. However, notice that for the 'moderately underweight' and 'moderately obese' groups, the confidence interval is wider and may not be accurate due to the lack of data. But personally, I think it is acceptable for SLR model.
In conclusion, this SLR model is good for predicting this task.

Appendix

```
# Code for Assignment1 STAT4038 - Dingying Li - U5493820

# Q1 A Plot shield against weight
moorhen<-read.csv('moorhen.csv',header = T)
attach(moorhen)
plot(Weight,Shield,xlab='Weight(mg)', ylab='Shield(mm^2)', main='Relationship between Shield and
Weight')
identify(Weight,Shield)

# Q1 B
cor.test(Weight, Shield)
#cor.test(Shield, Weight)

# Q1 C
log(Weight)
log(Shield)
plot(log(Weight), log(Shield), main='Relationship between log(Shield) and log(Weight)')
identify(log(Weight), log(Shield))

#plot(log(Weight[-c(27)]), log(Shield[-c(27)]))
cor.test(log(Weight), log(Shield)) # t = 1.9709, max!
cor.test((Weight), log(Shield))
cor.test(log(Weight), (Shield))
cor.test(sqrt(Weight), sqrt(Shield))
cor.test((Weight), sqrt(Shield))
cor.test(sqrt(Weight), (Shield))

# Q1 D
moorhen.lm <-lm(log(Shield)~log(Weight))
summary(moorhen.lm)
abline(moorhen.lm$coefficients)
plot(moorhen.lm, which=1)
plot(moorhen.lm, which=2)
plot(moorhen.lm, which=4)
barplot(hat(log(Weight)), main='Leverage plot of hat(log(Weight))')

plot(log(Weight), moorhen.lm$residuals)

# Q1 E
anova(moorhen.lm)

# Q1 F
summary(moorhen.lm)

# Q1 G
moorhen_non.lm <- lm(Shield ~ Weight)
plot(Weight,Shield, xlim=c(0,800), ylim=c(0,550), main='Relationship between Shield and Weight',
xlab='Weight(mg)', ylab='Shield(mm^2)')
arr <- c(1:800) # array for lines plot
pred <- predict(moorhen.lm, newdata=data.frame(Weight=arr))
```

```
lines(arr,exp(pred[order(pred)]),col='red')
abline(moorhen_non.lm$coefficients, col='blue', lty=2)
legend(0,550,legend=c('Transformed SLR','Untransformed SLR'),
     col=c('red','blue'), lty=1:2, box.lty=1, cex=0.8)

#Q2 A
fat<-read.csv('fat.csv',header=TRUE)
attach(fat)
plot(BMI,body.fat,main='Relationship between body.fat and BMI')
cor.test(BMI,body.fat)

#Q2 B
fat.lm<- lm(body.fat~BMI)
plot(fat.lm)
plot(fat.lm, which=1)
plot(fat.lm, which=2)
plot(fat.lm, which=4) # cook's distance


#Q2 C
fat.log.lm <- lm(body.fat~log(BMI))
plot(fat.log.lm, which=1)
plot(fat.log.lm, which=2)
plot(fat.log.lm, which=4)
fat.loglog.lm <- lm(log(body.fat) ~ log(BMI))
(body.fat)

#Q2 D
anova(fat.log.lm)
summary(fat.log.lm)

#Q2 E
plot(log(BMI),body.fat,main='Relationship between body.fat and log(BMI)')

log_BMI = log(BMI)
fat_log.lm <- lm(body.fat~log_BMI)
abline(fat.log.lm$coefficients)
bmi_predict <- c(17.25, 21.75, 27.5, 32.5)
bmi_log <- log(bmi_predict)
confidence <- predict(fat_log.lm,
newdata=data.frame(log_BMI=bmi_log), interval='confidence')
confidence

arr_bmi <- c(15:50)
confidence_plot <- predict(fat_log.lm,
newdata=data.frame(log_BMI=log(arr_bmi)),interval='confidence')
prediction_plot <- predict(fat_log.lm,
newdata=data.frame(log_BMI=log(arr_bmi)),interval='prediction')
lines(log(arr_bmi),prediction_plot[,1],lty=1)
lines(log(arr_bmi),prediction_plot[,2],lty=2)
lines(log(arr_bmi),prediction_plot[,3],lty=2)
```

```
lines(log(arr_bmi),confidence_plot[,2],lty=3)
lines(log(arr_bmi),confidence_plot[,3],lty=3)
abline(v=log(bmi_predict),lty=4)
legend(2.7,45,legend=c('SLR model','prediction interval','confidence interval','predicted values'),
    col='black',lty=c(1,2,3,4), box.lty=1, cex=0.8)
```