Exercise 1

In this exercise, we will solely refer to the dataset `mussels.csv`.

(a) In this task, we aim to find the best linear model that explains the relationship between Mass (response variable) and other variables. Firstly, we define a linear model that consider all possible predictors and we call it `modelfull`, and then apply `step(modelfull)` to obtain a new model with minimized AIC.

```
> modelfull <- lm(Mass ~ Height*Width*Length + I(Height^2) + I(Width^2) + I(Length^2))
> modelstep <- step(modelfull)
> modelstep
lm(formula = Mass ~ Height + Width + Length + I(Width^2) + I(Length^2) +
  Height:Width + Height:Length)
```

In high school science, we learnt that mass is directly proportional to volume of a matter. We will now define another linear model as follows.

```
modelscience <- lm(Mass~Height:Width:Length)
```

To compare these models, we will use Akaike Information Criterion (AIC) and the predicted R-squared R^2_{pred} as shown and given during the practical session.

```
> pred.R2(modelstep)
[1] 0.9458487
> pred.R2(modelscience)
[1] 0.9559329
> AIC(modelstep)
[1] 722.1254
> AIC(modelscience)
[1] 730.6663
```

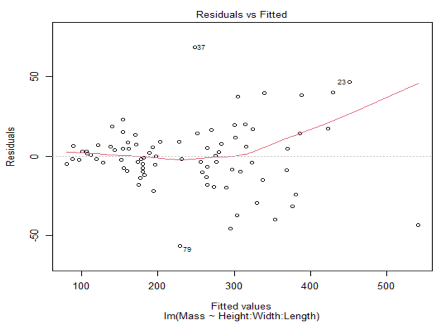
The tests above did not provide a clear cut which one the best to choose since AIC favors `modelstep` but R^2_{pred} shows otherwise. However, `modelstep` has a problem with the standard error of individual estimators. Some of them will be shown below.

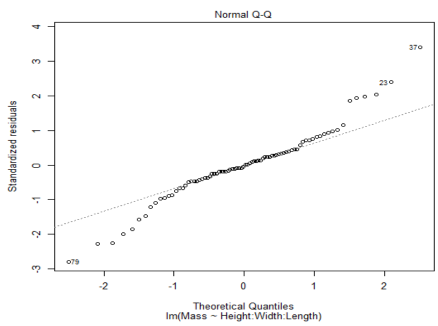
```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.134797  79.204160   0.557 0.579053
Height        11.127855   2.850436   3.904 0.000207 ***
Width        -16.937851   5.730847  -2.956 0.004185 **
```

Comparing with the estimators in `modelscience`

```
                     Estimate  Std. Error t value Pr(>|t|)
(Intercept)          5.567e+01 4.785e+00   11.63   <2e-16 ***
Height:Width:Length  1.421e-04 3.292e-06   43.17   <2e-16 ***
```

We decide to choose modelscience over modelstep based on the evidence above. Checking the model assumptions, we will run `plot(modelscience)` to obtain the diagnostic plots.





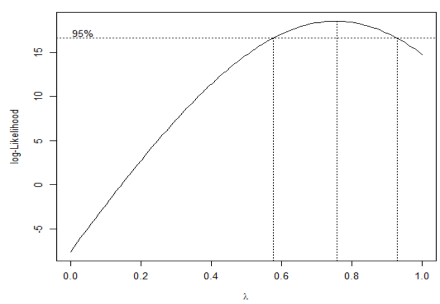


Figure 1: Two diagnostic plots and BoxCox figure of modelscience

The diagnostic plots suggest the linear model has an issue to show the normality of the errors. We immediately apply Box-Cox transformation to fix this issue with $\lambda = 0.5$ as the figure suggests. Our new linear model is now

```
modelbest <- lm(((Mass^0.5 - 1)/0.5)~Height:Width:Length)
```

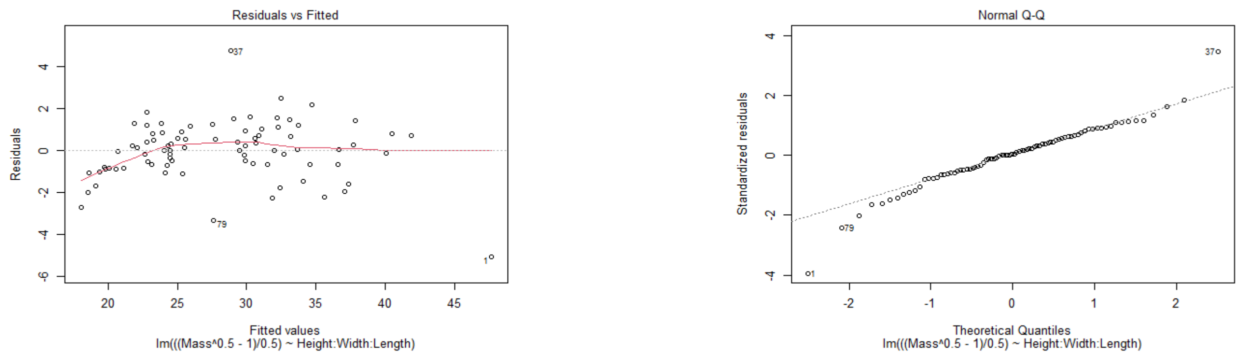


Figure 2: Two of the diagnostic plots of modelbest

The deviation of the errors on the Normal probability plot has evidently improved after the Box-Cox transformation. However, the fit might be problematic for small values of the fitted values.

(b) Given a new observation, the prediction are as follows

```
> newdata <- data.frame(Height = 100, Width = 100, Length = 200)
> ypred <- predict(modelbest, newdata, interval = "prediction", level=0.95)
> (ypred*0.5 + 1)^2
       fit      lwr      upr
1 337.5484 288.4245 390.5333
```

Since we transformed the response variable, therefore extra arithmetic work needs to be done to get the intervals in terms of Mass.

Exercise 2

For this exercise, we will carry out partial F-test on two given linear models of the dataset `ex2.csv`.

```
model1 <- lm(y ~ x1 + x2)
model2 <- lm(y ~ x1 + x2 + I(x1^2) + I(x2^2) + x1:x2)
```

(a) Let the hypothesis be as follows:

$$H_0 : \text{Data are generated from model1} \quad H_1 : \text{Data are generated from model2}$$

```
> anova(model1, model2)
Analysis of Variance Table
Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + I(x1^2) + I(x2^2) + x1:x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     77 1565.7
2     74 1479.7  3    85.989 1.4335 0.2399
```

From the result above, we have enough evidence to accept the null hypothesis H_0 at 5% significance level.

(b) The diagnostic plots shows some issue on the normality of errors of the model and also highlights some potential outliers.

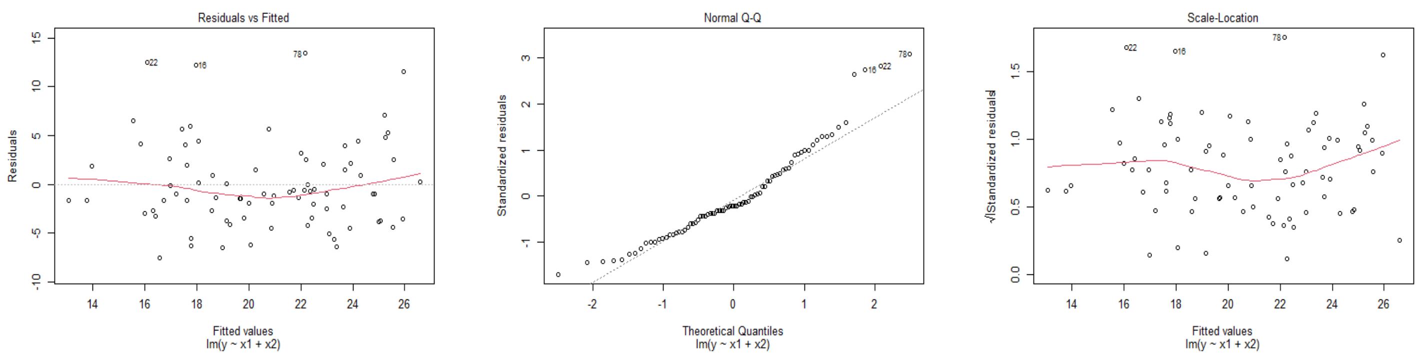


Figure 3: Three diagnostic plots and BoxCox figure of model1