STAT3340
Regression Analysis
Report

Dataset 3 (Fish Market)

Group 14

Group members:		
Name:	Banner number:	Email:
Kexin Hu	B00781616	kx536478@dal.ca
Yue Ding	B00731454	yz307495@dal.ca
Lefu Xie	B00807655	lf402690@dal.ca

Abstract:

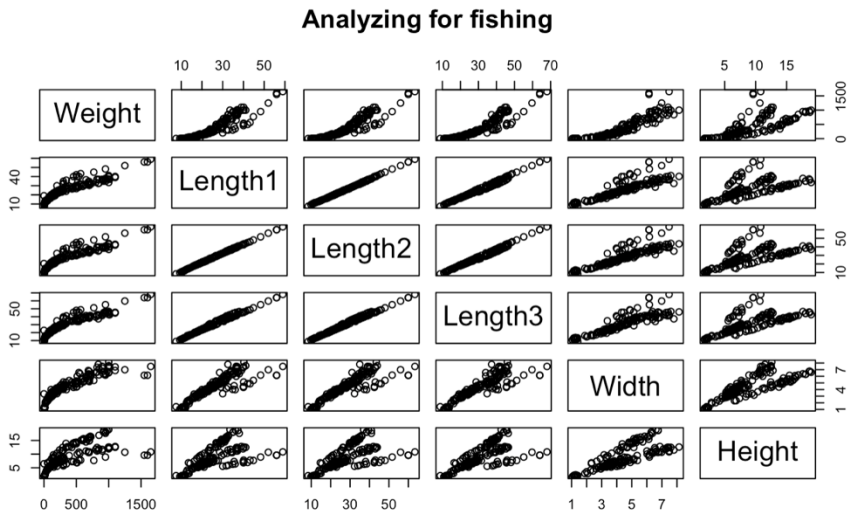
- Our aim is to find the variable that mainly affect the weight of the fish, and then to eliminate the independent variables which have less influence.
- Use Matrix Scatter Plot to find the relationship between predictors and the response variable.
  Use Correlation Coefficient Matrix to find the stength of Corr.
  Observed the interquartile range to add an additional data point to smelt.
  Use "olsrr" package to find out all of the model.
- Width's VIF>10, indicating that the relationship between it and the independent variables which is in other models has some influence on the overall model, but associated the regression coefficients, this value of VIF is tolerable.

Introduction:

In this semester, we learnt about Linear Regression. In this study, our group are going to use R and present a series of models to analyze the Multiple Linear Regression and find the relationship between the variables and the weight of the fish. Then, we need eliminate the independent variables which have less influence on the weight of the fish. The dataset we used to analyzed is from Kaggle., Meanwhile we will add "("Smelt", 19.8, 12.0, 13.5, 14.4, 2.35, 1.401)" as one new additional data point.

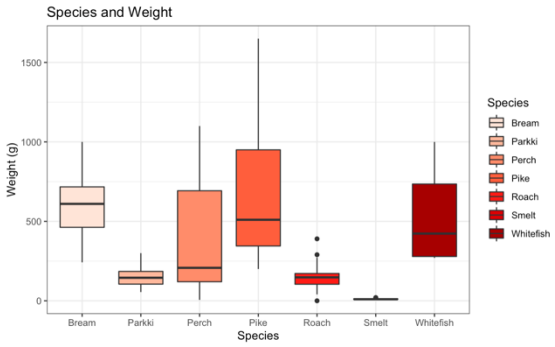
Data Description:

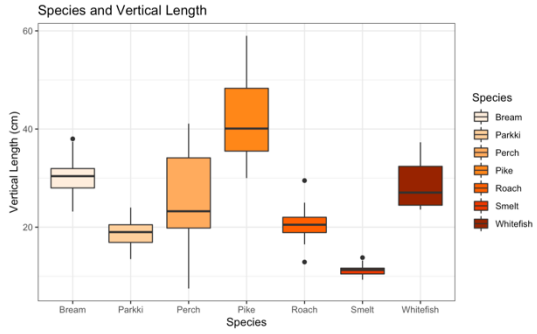
I. Matrix Scatter Plot

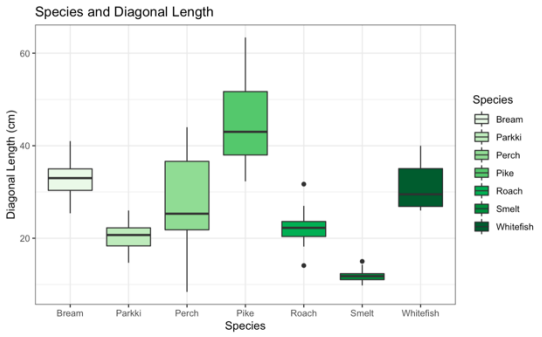


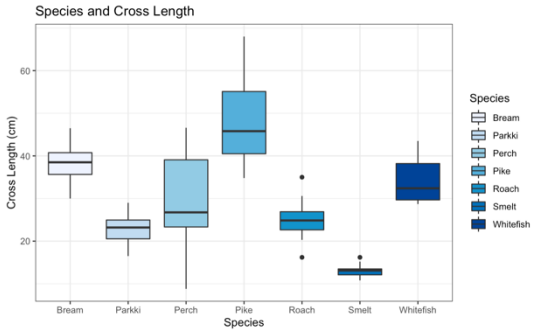
According to the diagram above, we can see that there is linear relationship between the data of Vertical Length and Diagonal Length, Vertical Length and Cross Length, Diagonal Length and Cross Length, while there is no obvious relationship between the predictors and the response

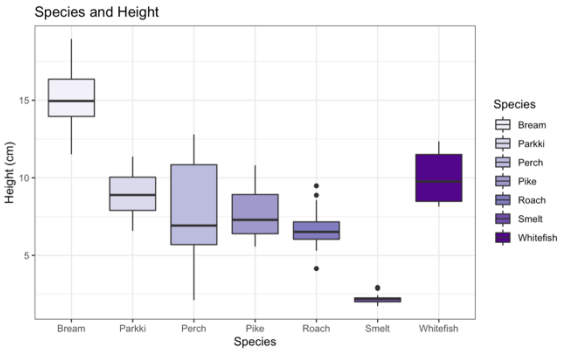
variable.

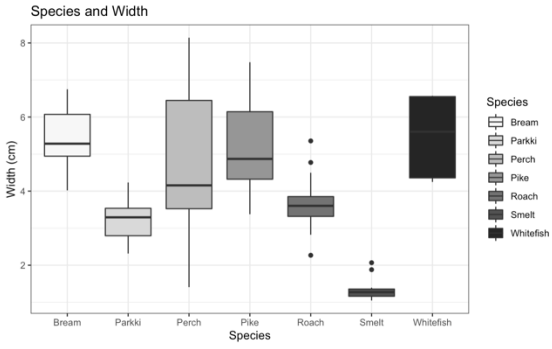












II. Correlation Coefficient Matrix

	Weight	Length1	Length2	Length3	Height	Width
Weight	1.0000000	0.9157117	0.9186177	0.9230436	0.7243453	0.8865066
Length1	0.9157117	1.0000000	0.9995173	0.9920310	0.6253779	0.8670497
Length2	0.9186177	0.9995173	1.0000000	0.9941026	0.6404408	0.8735467
Length3	0.9230436	0.9920310	0.9941026	1.0000000	0.7034089	0.8785202
Height	0.7243453	0.6253779	0.6404408	0.7034089	1.0000000	0.7928810
Width	0.8865066	0.8670497	0.8735467	0.8785202	0.7928810	1.0000000

Correlation coefficient

Corr between Length1 and Length2 is 0.9995173, which is very close to 1, it means that there is strong correlation between the data of Vertical Length and Diagonal Length.

Corr between Length1 and Height is 0.6253779, which is the least correlation coefficient, and the correlation between the data of Vertical Length and Height is relatively weaker than others but the correlation is still significant for 0.6253779 >> 0.

III. Additional Data Point

```
row<- c("Smelt", 19.8, 12.0, 13.5, 14.4, 2.35, 1.401)
Fish<-rbind(Fish,row)
Fish
```

Species <fctr>	Weight <chr>	Length1 <chr>	Length2 <chr>	Length3 <chr>	Height <chr>	Width <chr>
Smelt	8.7	10.8	11.3	12.6	1.9782	1.2852
Smelt	10	11.3	11.8	13.1	2.2139	1.2838
Smelt	9.9	11.3	11.8	13.1	2.2139	1.1659
Smelt	9.8	11.4	12	13.2	2.2044	1.1484
Smelt	12.2	11.5	12.2	13.4	2.0904	1.3936
Smelt	13.4	11.7	12.4	13.5	2.43	1.269
Smelt	12.2	12.1	13	13.8	2.277	1.2558
Smelt	19.7	13.2	14.3	15.2	2.8728	2.0672
Smelt	19.9	13.8	15	16.2	2.9322	1.8792
Smelt	19.8	12	13.5	14.4	2.35	1.401

151–160 of 160 rows Previous 1 … 11 12 13 14 15 16 Next

The additional data point we want to add is ("Smelt 19.8 12.0 13.5 14.4 2.35 1.401"). From the box plots above, we can find the maximum, minimum, average and interquartile range of each variables against species. It is obvious that the variables of perch have the largest interquartile range among the fishes, while smelt is supposed to be the smallest fish for its interquartile range is the smallest and it has the shortest length in all of the plots.
Therefore, we decide to add an additional data point to smelt. Each of the numbers we come up with are in the interquartile range, which makes them relatively normal points.

Methods:

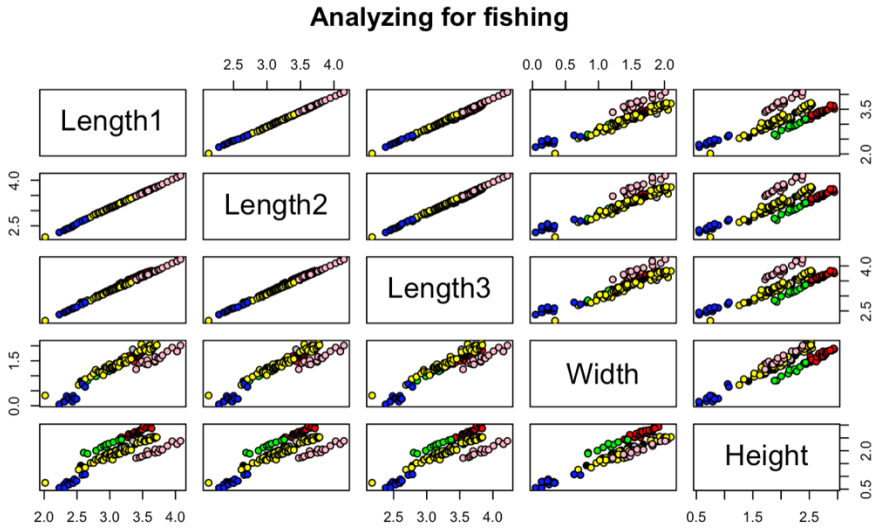
I. Model Selection

1) Analyze data and find all possible models

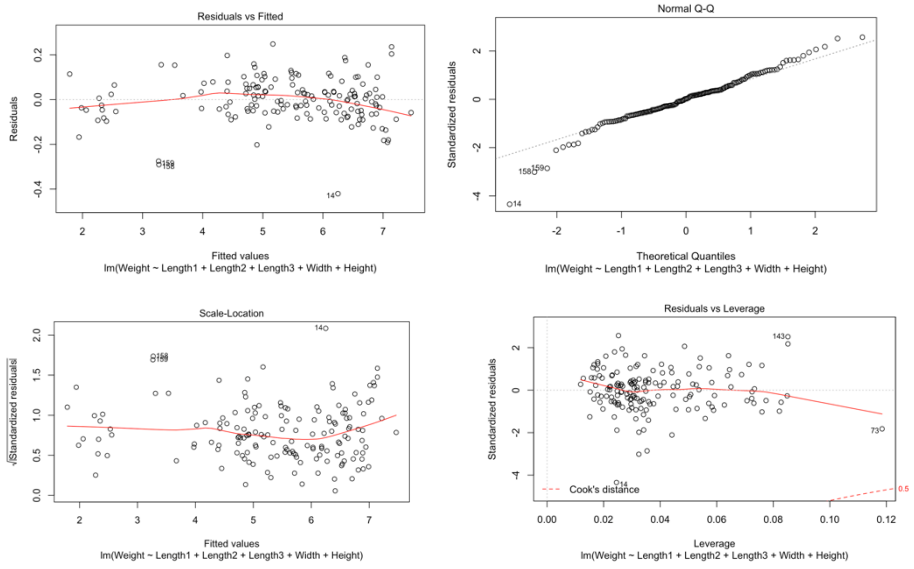
[Analyze dataset]:

For this section, “olsrr” package is used to find all possible models. Since the Scatter plot illustrates that the linear relationships between Length1&Width, Length2&Width, Length3&Width and Height and Width are not obvious. So, “log” is used to adjust the model so that the relationship between each pair of them are more linear. (Note: Since the dependent variable of the 41st data point is 0, which means this data point cannot

be transformed with “log”, so it is deleted.)



This Scatter Plot shows the dataset which is transferred with “log”. It displays the relationship of each pair is more linear.

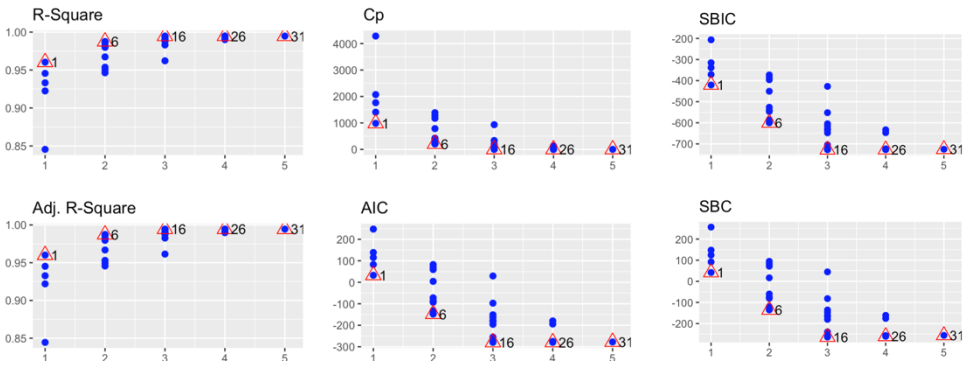


These are the dataset’s Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage graphs.

[Select models]:

Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
4	1	Width	0.9603424	0.9600882	984.961206
3	1	Length3	0.9456076	0.9452590	1408.141791
2	1	Length2	0.9331642	0.9327358	1765.515144
1	1	Length1	0.9224887	0.9219918	2072.114770
5	1	Height	0.8454845	0.8444940	4283.665306
13	2	Length3 Width	0.9875028	0.9873415	206.918461
12	2	Length2 Height	0.9873830	0.9872202	210.357359
9	2	Length1 Height	0.9861964	0.9860183	244.438230
11	2	Length2 Width	0.9823008	0.9820724	356.319345
8	2	Length1 Width	0.9815885	0.9813509	376.776401
14	2	Length3 Height	0.9798518	0.9795918	426.653351
15	2	Width Height	0.9672893	0.9668672	787.447720
6	2	Length1 Length2	0.9536849	0.9530873	1178.162165
7	2	Length1 Length3	0.9502331	0.9495909	1277.298660
10	2	Length2 Length3	0.9463592	0.9456670	1388.557728
24	3	Length2 Width Height	0.9946394	0.9945350	3.956279
21	3	Length1 Width Height	0.9945395	0.9944331	6.825705
25	3	Length3 Width Height	0.9937502	0.9936284	29.493488
23	3	Length2 Length3 Height	0.9908918	0.9907144	111.585037
19	3	Length1 Length3 Width	0.9900393	0.9898452	136.071298
22	3	Length2 Length3 Width	0.9897683	0.9895690	143.853082
20	3	Length1 Length3 Height	0.9887827	0.9885642	172.158539
18	3	Length1 Length2 Height	0.9879166	0.9876812	197.034788
17	3	Length1 Length2 Width	0.9830074	0.9826764	338.025625
16	3	Length1 Length2 Length3	0.9621552	0.9614180	936.896835
30	4	Length2 Length3 Width Height	0.9946960	0.9945574	4.329056
28	4	Length1 Length2 Width Height	0.9946437	0.9945036	5.833322
29	4	Length1 Length3 Width Height	0.9945561	0.9944138	8.348601
27	4	Length1 Length2 Length3 Height	0.9909167	0.9906793	112.869984
26	4	Length1 Length2 Length3 Width	0.9900450	0.9897848	137.904906
31	5	Length1 Length2 Length3 Width Height	0.9947075	0.9945334	6.000000

31 rows



Analyzation:

[1] R-Square: R-Square expresses the portion that independent variables can explain the dependent variable. From the R-Square graph, we could discover that as the number of independent variables increases, the value of R-Square increases. In some ways, more independent variables, more portion of the dependent variable can be explained.

[2] Cp: Small Cp are desirable. So, model 16, 17, 26, 27, 31 are desirable.

model 16: Weight = Length2 + Width + Height
model 17: Weight = Length1 + Width + Height
model 26: Weight = Length2 + Length3 + Width +Height
model 27: Weight = Length1 + Length2 + Width + Height
model 31: Weight = Length1 + Length2 + Length3 + Width + Height

[3] Adj. R-Square: Since Adjusted R-Square could determine there are any unmeaningful independent variables added. Since there is no difference between R-Square and Adj. R-Square from these graphs, there is no useless independent variables in this model; the value of Adj. R-Square is not affected by the number of independent variables.

[4] AIC: The equation of AIC is AIC = −2 ln (L) + 2 p, L is likelihood function, p is num of parameters. When the complexity of the model increases (p increases), the likelihood function(L) increases so that AIC decreases, but when p is too big, L increases slowly so that cause AIC to increase. The complexity of the model may cause overfitting, so our purpose is to choose the model with a small AIC, which means the model is more accurate. So, choosing the model with a small AIC not only improves the model fit but also introduces a penalty term to make the model parameters(p) as few as possible, which helps reduce the possibility of overfitting.

Therefore, the model 16, 26, 31 look good.

model 16: Weight = Length2 + Width + Height

model 26: Weight = Length2 + Length3 + Width +Height

model 31: Weight = Length1 + Length2 + Length3 + Width + Height

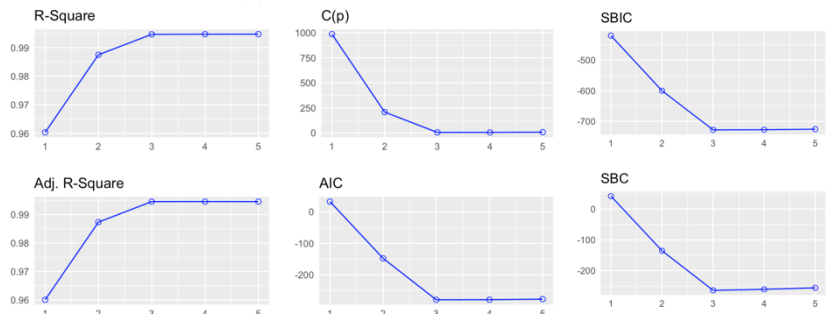
2) Best subset regression

[Select best models]:

```
-------------------------------------------------
Model Index    Predictors
-------------------------------------------------
     1         Width
     2         Length2 Height
     3         Length2 Width Height
     4         Length2 Length3 Width Height
     5         Length1 Length2 Length3 Width Height
-------------------------------------------------

                                                Subsets Regression Summary
-------------------------------------------------------------------------------------------------------------------------------------------
                        Adj.        Pred
Model    R-Square    R-Square    R-Square      C(p)         AIC         SBIC          SBC        MSEP       FPE       HSP       APC
-------------------------------------------------------------------------------------------------------------------------------------------
  1        0.9603      0.9600      0.9592    966.2303     35.3710    -420.7572     44.5778    11.3408    0.0722     5e-04    0.0407
  2        0.9876      0.9874      0.9871    197.0037   -147.7428    -602.8352   -135.4672     3.5630    0.0228     1e-04    0.0129
  3        0.9945      0.9944      0.9942      4.2034   -275.0428    -726.0707   -259.6982     1.5902    0.0103     1e-04    0.0058
  4        0.9946      0.9944      0.9942      4.1121   -275.1998    -726.0407   -256.7864     1.5790    0.0102     1e-04    0.0058
  5        0.9946      0.9944      0.9941      6.0000   -273.3162    -724.0712   -251.8339     1.5883    0.0104     1e-04    0.0058
-------------------------------------------------------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria

    mindex       n      predictors          rsquare           adjr          predrsq
Min.   :1   Min.   :1   Length:5          Min.   :0.9603   Min.   :0.9600   Min.   :0.9592
1st Qu.:2   1st Qu.:2   Class :character   1st Qu.:0.9876   1st Qu.:0.9874   1st Qu.:0.9871
Median :3   Median :3   Mode  :character   Median :0.9945   Median :0.9944   Median :0.9941
Mean   :3   Mean   :3                      Mean   :0.9863   Mean   :0.9861   Mean   :0.9857
3rd Qu.:4   3rd Qu.:4                      3rd Qu.:0.9946   3rd Qu.:0.9944   3rd Qu.:0.9942
Max.   :5   Max.   :5                      Max.   :0.9946   Max.   :0.9944   Max.   :0.9942
      cp               aic              sbic              sbc               msep
Min.   :  4.112   Min.   :-275.20   Min.   :-726.1   Min.   :-259.70   Min.   : 1.579
1st Qu.:  4.203   1st Qu.:-275.04   1st Qu.:-726.0   1st Qu.:-256.79   1st Qu.: 1.588
Median :  6.000   Median :-273.32   Median :-724.1   Median :-251.83   Median : 1.590
Mean   :235.510   Mean   :-187.19   Mean   :-640.0   Mean   :-171.84   Mean   : 3.932
3rd Qu.:197.004   3rd Qu.:-147.74   3rd Qu.:-602.8   3rd Qu.:-135.47   3rd Qu.: 3.563
Max.   :966.230   Max.   :  35.37   Max.   :-420.8   Max.   :  44.58   Max.   :11.341
     fpe               apc               hsp
Min.   :0.01024   Min.   :0.005776   Min.   :6.490e-05
1st Qu.:0.01025   1st Qu.:0.005781   1st Qu.:6.494e-05
Median :0.01036   Median :0.005845   Median :6.571e-05
Mean   :0.02518   Mean   :0.014201   Mean   :1.595e-04
3rd Qu.:0.02283   3rd Qu.:0.012874   3rd Qu.:1.446e-04
Max.   :0.07222   Max.   :0.040728   Max.   :4.572e-04
```



Analyzation:

The function ols_step_best_subset(model) helps us select some models. Also, from the summary result, we could discover that model 3,4,5 has smaller AIC and C(p), which means these models more suitable. Although those graphs illustrate that two independent variables should be better, based on these aspects, model 3 is selected. Because model 3 has the lowest AIC and C(p), it has 3 independent variables

model 3: Weight = Length2 + Width + Height

model 4: Weight = Length2 + Length3 + Width + Height

model 5: Weight = Length1 + Length2 + Length3 + Width + Height

3) Backward

(Note: Based on the estimated results of the previous two steps, the alpha_out is assumed to 0.001.)

Reason for choosing Backward Elimination: Forward Selection is not enough accurate; Stepwise Regression is too complex. Generally, Stepwise Regression's result is same with Backward Elimination.

Step 1: Assuming that the model starts with all variables included, that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

```
Call:
lm(formula = Weight ~ Length1 + Length2 + Length3 + Width + Height,
    data = Fish)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41800 -0.05675  0.00261  0.05611  0.24259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.05407    0.16120 -12.743  < 2e-16 ***
Length1     -0.22420    0.66977  -0.335  0.73827
Length2      2.27192    0.73461   3.093  0.00236 **
Length3     -0.52538    0.38625  -1.360  0.17577
Width        0.80848    0.08131   9.943  < 2e-16 ***
Height       0.67394    0.05988  11.256  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09994 on 153 degrees of freedom
Multiple R-squared:  0.9946,    Adjusted R-squared:  0.9944
F-statistic:  5616 on 5 and 153 DF,  p-value: < 2.2e-16
```

Among the outputs, Length1 has the largest p value ($0.73827 > 0.001$), so remove Length1.

Step 2: Continue backward elimination with

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

```
Call:
lm(formula = Weight ~ Length2 + Length3 + Width + Height, data = Fish)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41843 -0.05637  0.00149  0.05507  0.24769

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.01963    0.12373 -16.323  < 2e-16 ***
Length2      2.06251    0.38397   5.372 2.83e-07 ***
Length3     -0.54909    0.37860  -1.450    0.149
Width        0.81076    0.08079  10.035  < 2e-16 ***
Height       0.67968    0.05720  11.882  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09965 on 154 degrees of freedom
Multiple R-squared:  0.9946,    Adjusted R-squared:  0.9944
F-statistic:  7060 on 4 and 154 DF,  p-value: < 2.2e-16
```

Among the outputs, Length3 has the largest p value ($0.149 > 0.001$), so remove Length3.

Step 3: Continue backward elimination with

$$y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5$$

```
Call:
lm(formula = Weight ~ Length2 + Width + Height, data = Fish)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42825 -0.05882  0.00187  0.05675  0.26233

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.02657    0.12407  -16.33   <2e-16 ***
Length2      1.51117    0.05427   27.85   <2e-16 ***
Width        0.88386    0.06337   13.95   <2e-16 ***
Height       0.61196    0.03316   18.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1 on 155 degrees of freedom
Multiple R-squared:  0.9945,    Adjusted R-squared:  0.9944
F-statistic:  9346 on 3 and 155 DF,  p-value: < 2.2e-16
```

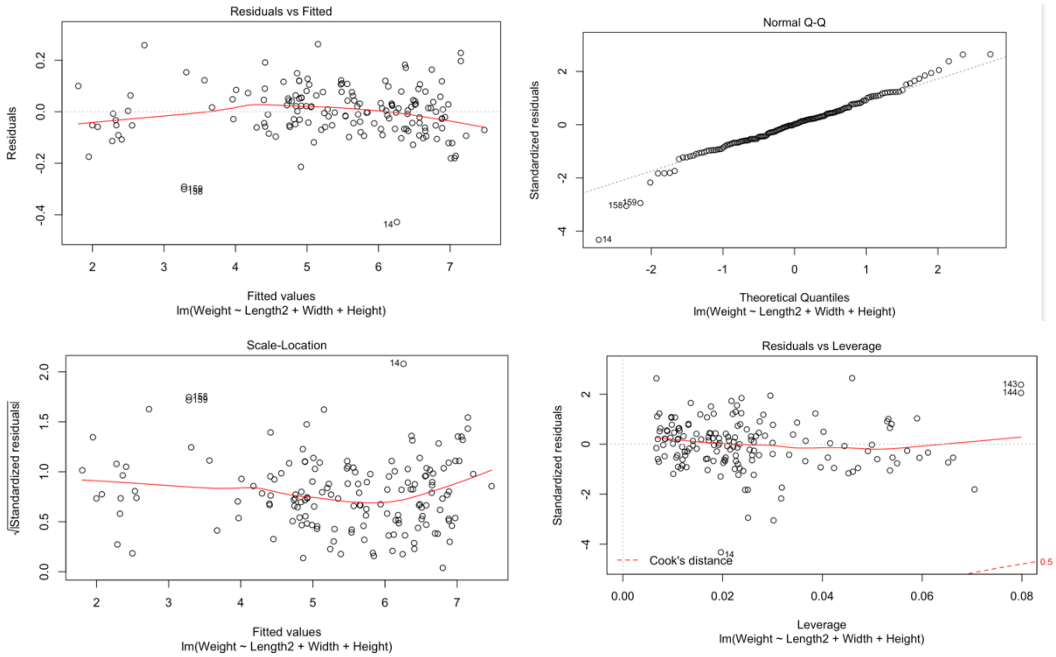
Among the outputs, there is no independent variable's p values less than 0.001.
In summary outputs, all p values are smaller than 0.001, so Length2, Width and Height can remain in model.

So far, backward elimination ends, and we get the final model:

$$y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5$$

II. Adjust Model with the last step's result model

[Gauss-Markov Assumption]:



Analyzation:

[1] Residuals vs Fitted: We could discover that not all residuals locate beside 0. So, the distribution of residuals is not relatively uniform.

[2] Normal Q-Q: From the trend of distribution, we could discover that not all points follow the normal distribution, it needs adjustment.

[3] Scale-Location: Since the variances of each residuals are not equal, they are spread, which means this model needs adjustment.

[4] Residuals vs Leverage: There are some outliers in it.

[VIF]:

```
   Length2     Width    Height
7.730794 14.298660  5.666522
```

Since the VIF of Width is 14.298660, which is large than 10, the associated regression coefficients are estimated poorly because of the multicollinearity. However, this project is used to find the best model which is good fit. So, the value of VIF is tolerable.

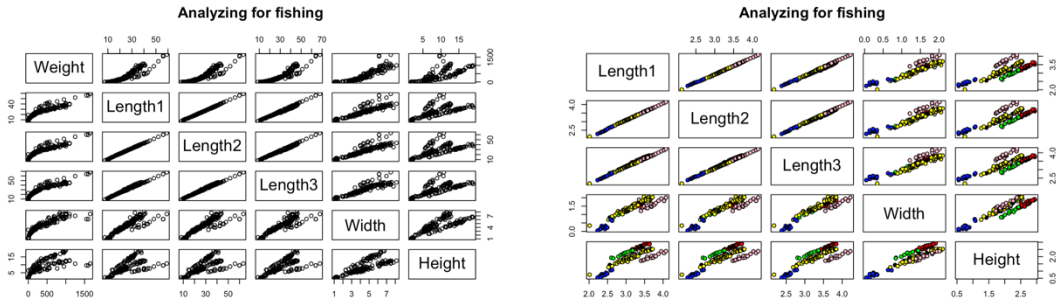
Results:

- As the independent variable increase, R-Square also increases. Which means the more explanatory the linear model is.
- Adjusted R-Square indicates how well terms fit a curve or line. From the plot we provided, we can know that Adjusted R-Square can be equal to the R-Square, and the Adjusted R-Square increase, the more useful variables we add.
- You can see the model in our backward, the p-value of those independent variables all smaller than the value our predicted alpha_out, so there are no independent variables that we can eliminate. The final model is

$$y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5$$

- The 4 plots: Residual vs Fitted, Normal Q-Q, Scale-Location and Residuals vs Leverage. We have the original one in the part of Methods, we can see the points more fitted and evenly distribute in these four plots, which means the multicollinearity improves.
- In the end, although our width's VIF > 10 indicating that it does have a multicollinearity problem with the independent variables in other models and has some influence on the overall models. It is tolerable based on the purpose of the analysis.

Conclusion:



These two Scatter plots are the before linearization and after mineralization. The 1st plot is the original graph, only a part shows the linear relationship, so we add "log" to each independent variable and dependent variable to transfer it and make the model become more linear. From the graph, we could discover the points of the 2nd plot spread as a line and more linear. Then, using the backward elimination to get an appropriate model.

$$y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5$$

The width's VIF>10, which indicates that the model has multiple linear problems, but it can be tolerated for our research purposes.

Appendix:

Appendix A: Fish.csv
Appendix B: 3340-Project(Group14).Rmd