Wenxiu Yang (B00791125)

Lanchen Jiang (B00805439)

Whether the weight of fish is related to its *Length1*, *Length2*, *Length3*, *Height* and *Width*

December 8, 2020

Abstract

In this project, we analyze whether the factors of *Length1*, *Length2*, *Length3*, *Height*, and *Width* are related to the weight of fish. We use Backward to determine the original model. In order to eliminate heteroskedasticity, we add the $Length3^2$. Moreover, we determine the final model6 by comparing Residuals vs Fitted, Normal Q-Q, and Residuals vs Leverage diagrams. Finally, we found that *Length3* and $Length3^2$ is related to weight of the fish.

Introduction

In the fish market, consumers usually judge the weight of a fish by its body shape, whoever is longer means its weight is heavier than others. But only the length of a fish an effect its weight? Hwang & Choi (2018) studies the weight of fish and fish's area are closely related to linear ones, which means when increase in the area of fish, the fish's weight also increasing. And Miller et al. (2015) posted that the weight of fish is related with the length of fish. Therefore, in our project, we use the *Length1*, *Length2*, *Length3*, height and width of fish to research what factors will affect fish's weight.

Data Description

In our data, we have 160 fishes to be our observations. There are 7 common species for fish market: Bream, Whitefish, Roach, Parkki, Smelt, Pike and Perch. And dependent variable is *Weight* (weight of fish (in grams)), independent variables are *Length1* (length from the nose to the beginning of the tail (in cm)); *Length2* (length from the nose to the notch of the tail (in cm)); *Length3* (length from the nose to the end of the tail (in cm)); *Hight* (maximal height as % of length3); *Width* (maximal width as % of length3).

Before the research, the Scatter Matrix Plot (Figure1) shows *Length1* and *Length2*, *Length2* and *Length3*, *Length1* and *Length3* are linear relation. *Weight* and *Length3* have relationship. And we add a new data for *Pike*, it's *Weight*=1680g, *Length1*=61cm, *Length2*=65cm, *Length3*=70cm, *Hight*=11.238% of length3, and *Width*=8.45% of length3.

Method

Refer to Figure2, because VIF cannot larger than 10; therefore, there is collinearity in *Length1*, *Length2*, and *Length3*. Then, we need to filter the variables to determine the final model.

According to the Figure3, firstly, from the C(p) diagram and the AIC diagram, we can see that when the number of x in the model is 4, 5 and 6, there is a tendency to go up. Hence, the number of x's in the model cannot be 4, 5 and 6. Secondly, from the Adj. R-Square diagram, the number of x's in the model is 2 or 3, not a big difference. Thirdly, from R-Square diagram, the inflection point is when the number of x in the model is 2. Therefore, the number of x in the model should be 2.

Next, we do the Backward:

We are setting $\alpha_{out} = 0.1$, we need to delete the variable if the p-value large than 0.1 and it is the largest p-value.

First, the model1 is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

In the summary result, we find that the *Height* (x_4) with the largest p-value and large than 0.1. Hence, we delete *Height* (x_4) from the model1.

Second, the model2 is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_5 x_5$$

In the summary result, we find that the *Width* (x_5) with the largest p-value and large than 0.1. Hence, we delete *Width* (x_5) from the model2.

Third, the model3 is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

In the summary result, we find that the *Length*2 (x_2) with the largest p-value and large than 0.1. Hence, we delete *Length*2 (x_2) from the model3.

Then, the model4 is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3$$

In the summary result, we find that the *Length1* (x_1) with the largest p-value and large than 0.1. Hence, we delete *Length1* (x_1) from the model4.

Finally, we get model5:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3$$

In the summary result, *Length3*'s p-value is less than 0.1; therefore, we do not delete *Length3* (x_3) from the model.

To summarize, the provisional model is：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3$$

From Figure4, we can see that there is heteroskedasticity in the residuals. Therefore, in order to eliminate heteroskedasticity, we need to add variables.

We add $Length3^2$ in model 6:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_3 + \hat{\beta}_2 x_3^2$$

After adding $Length3^2$, refer to the Residuals vs Fitted diagram in Figure5, we can see that the Residuals vs Fitted line is closer to the line with Residuals equal to 0. In addition, when we compare the Normal Q-Q diagram in Figure6 and Figure7, we find that the residuals in model6 are closer to the normal distribution. Therefore, model6 is better than model5.

Refer to Residuals vs Leverage diagram in Figure8, 145 set of data is an outlier and will affect the trend of the Residuals vs Leverage, so we need to delete 145. After delete 145, the Residuals vs Leverage diagram in Figure9 is better.

The model6 is our final model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_7^2$$

In addition, x_1 is *Parkki* dummy variable, x_2 is *Perch* dummy variable, x_3 is *Pike* dummy variable, x_4 is *Roach* dummy variable, x_5 is *Smelt* dummy variable, x_6 is *Whitefish* dummy variable, and x_7 shows *Length3* and $Length3^2$. According to Figure10, the VIF is less than 10; therefore, there is no collinearity in the final model6.

Result

From Figure2 we can see that since VIF is greater than 10, there is collinearity in *Length1*, *Length2* and *Length3*. In Figure 3, combining the C(p), AIC, Adj. R-Square, and R-Square diagrams show that the number of x in the model is 2. After the Backward, we get the model5. However, Figure4 shows there is heteroskedasticity in the model5; therefore, we have added $Length3^2$ to model5 and obtained model6. By comparing Figure4 and Figure5, it can be seen that the Residuals vs Fitted line in Figure5 is closer to the line with Residuals equal to 0. Similarly, the comparison between Figure6 and Figure7 also shows that the residuals in model6 is closer to the normal distribution, as shown in Figure7. Furthermore, 145 is an outlier in model6. Therefore, after we delete the outlier, as shown in Figure9, the model becomes better. Refer to Figure10, the model6 is no collinearity because VIF is less than 10.

According to Figure11, we can see that in the final model6, the p-value of x_1, x_2, x_3, x_4, x_5, x_6 and x_7 are less than 0.1; therefore, x1, x2, x3, x4, x5, x6, and x7 are all significant. Moreover, F-statistic p-value equal to 0 is less than 0.1, the model is significant. R^2 under the influence of x can explain 96.51% of y. When we eliminate the effect of x, we get adjusted R^2, x still can explain 96.33% of y. Consequently, the above indicators show that model6 can be used. Model6 can explain the relationship between x and y in the dataset.

Conclusion

In conclusion, in the final model, we only keep the *Length3*, because *Length3* is from fish's nose to end of the tail, which is include *Length1* and *Length2*. For *Height* and *Width*, they are having weakly relationship with *Weight*. Therefore, the result of *Length3* and *Weight* have relationship is same with Miller et al. (2015)'s study. In the fish market, there are huge differences in body size between different kinds of fish, some fish have large or many fins, so, it difficult to determine the weight of the fish by observing the width of the fish. We can look at the length of fish's whole length to judge the weight. The longer of the whole fish, the heavier of its weight.

References

J.Puranen. (1917). Fishcatch from http://jse.amstat.org/datasets/fishcatch.txt

K. H. Hwang and J. W. Choi. (2018). Machine Vision Based Weight Prediction for Flatfish. 2018 18th International Conference on Control, Automation and Systems (ICCAS), Daegwallyeong, 2018, pp. 1628-1631.

McKenzie, K., & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. Higher education research & development, 20(1), 21-33.

Miller, S. J., VanGenechten, D. T., & Cichra, C. E. (2015). Length–weight relationships and an evaluation of fish–size and seasonal effects on relative condition (K n) of fishes from the Wekiva River, Florida. Florida Scientist, 1-19.

olsrr from https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html

Pyae, A. (2019, June 13). Fish market. Retrieved December 10, 2020, from https://www.kaggle.com/aungpyaeap/fish-market

Appendix

Figure1

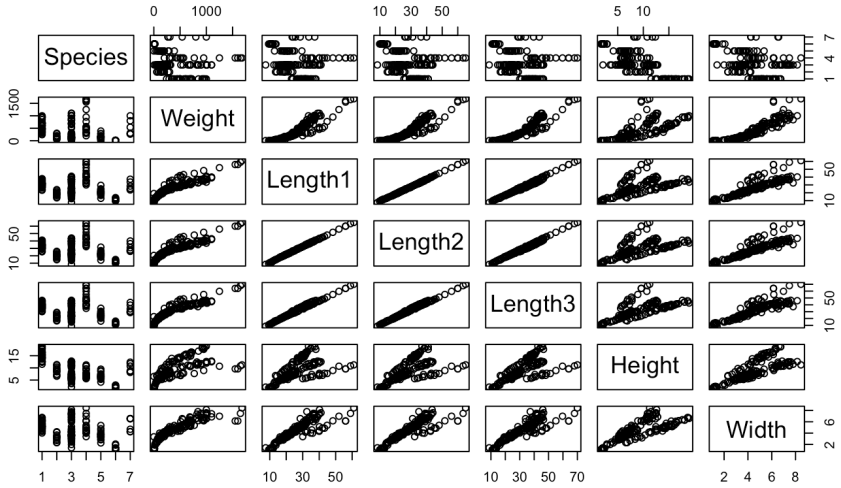

Figure2

	GVIF	Df	GVIF^(1/(2*Df))
Species	1462.63157	6	1.835529
Length1	2539.34515	1	50.391916
Length2	4539.04109	1	67.372406
Length3	2124.66963	1	46.094139
Height	51.43974	1	7.172151
Width	29.63125	1	5.443460

Figure3

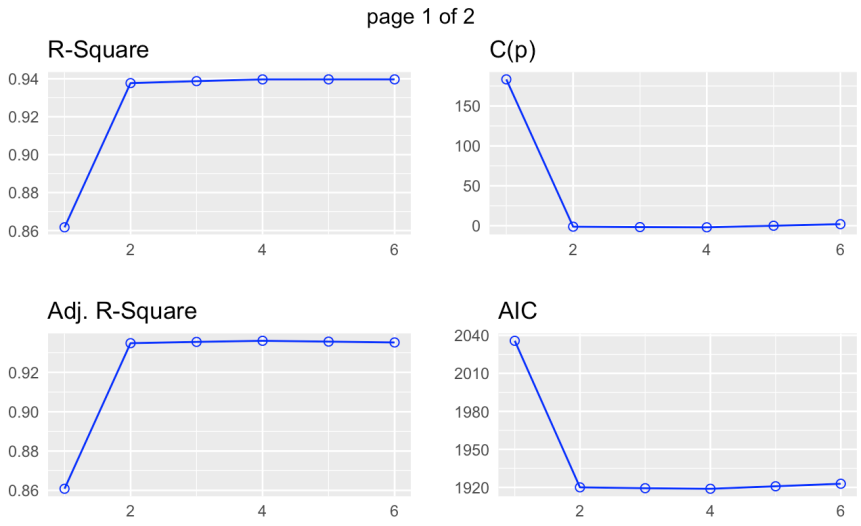

Figure4

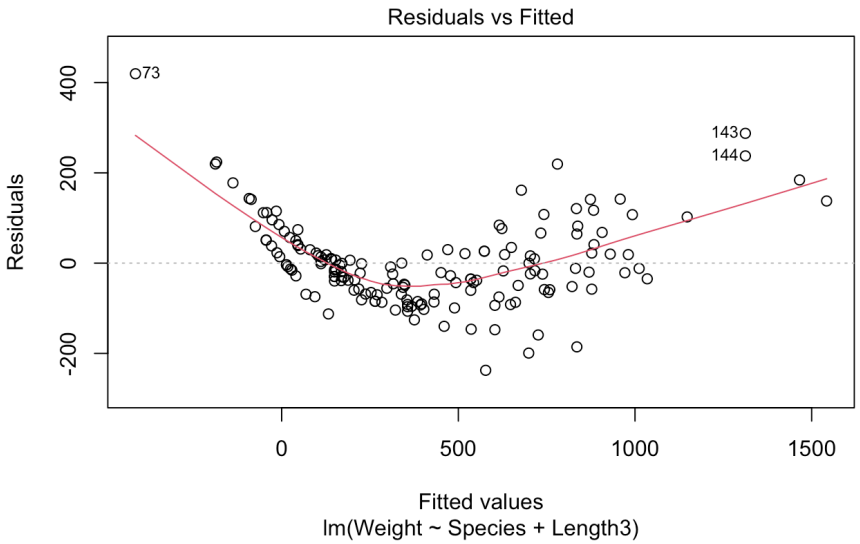

Figure5

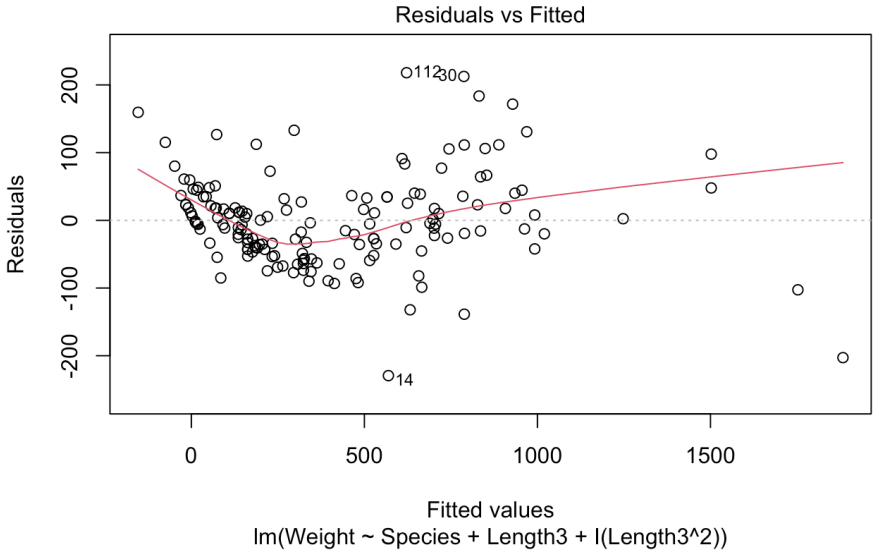

Figure6

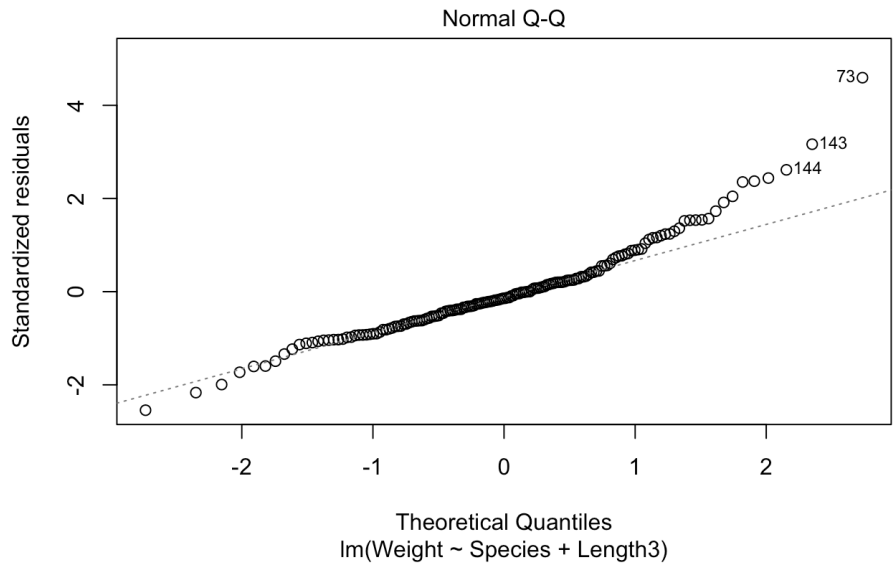

Figure7

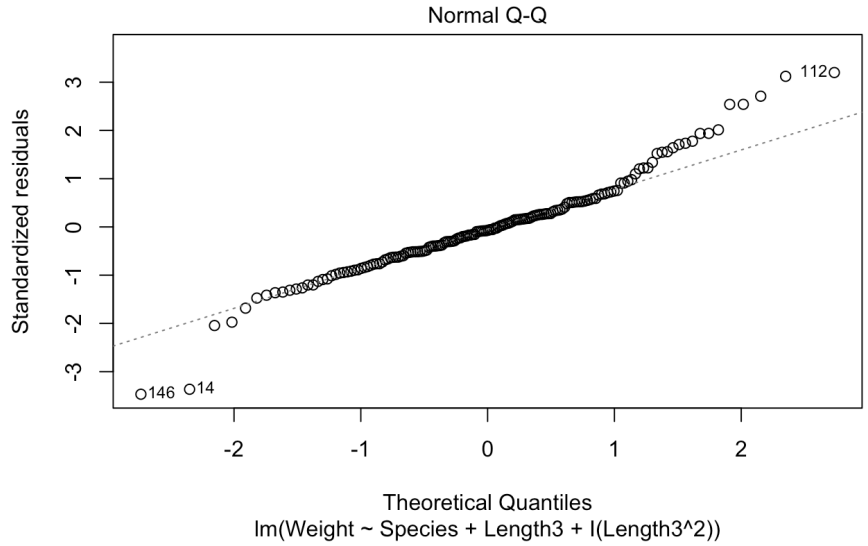

Figure8

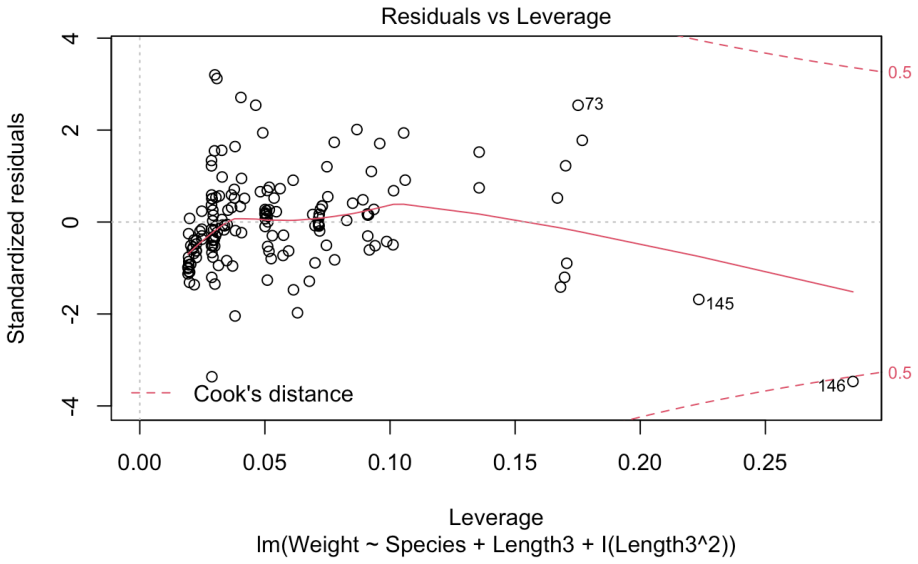

Figure9

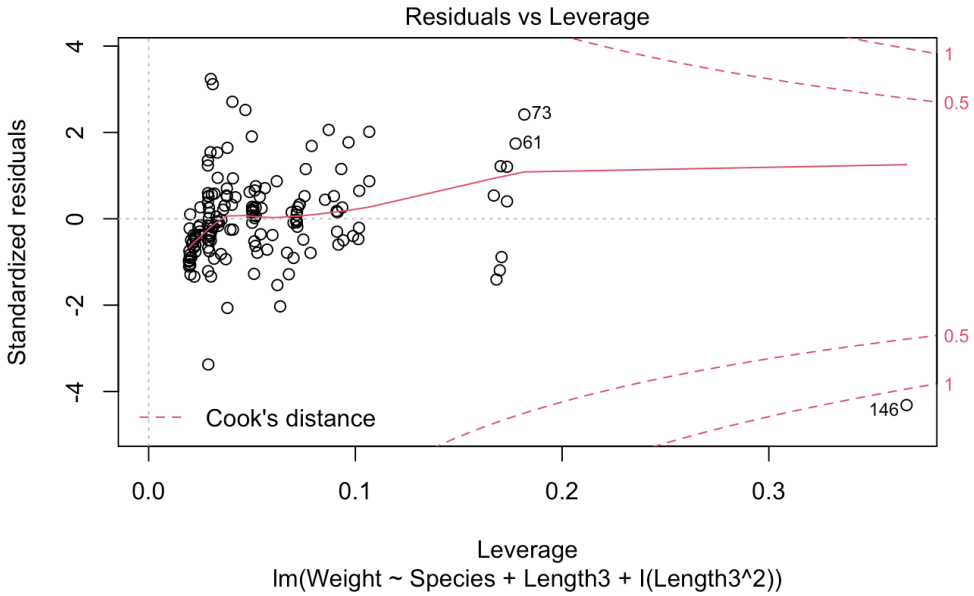

Figure10

	GVIF	Df	GVIF^(1/(2*Df))
Species	4.72389	6	1.138129
Length3	36.44743	1	6.037171
I(Length3^2)	31.77984	1	5.637361

Figure11

```
Call:
lm(formula = Weight ~ Species + Length3 + I(Length3^2), data = Fish_1)

Residuals:
     Min       1Q   Median       3Q      Max
-236.093  -37.657   -4.806   33.891  219.144

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -289.43341   57.31441  -5.050 1.27e-06 ***
SpeciesParkki      58.11337   27.27016   2.131 0.034717 *
SpeciesPerch       52.26232   16.74531   3.121 0.002162 **
SpeciesPike      -403.93594   23.87467 -16.919  < 2e-16 ***
SpeciesRoach       -5.25220   22.22035  -0.236 0.813469
SpeciesSmelt      137.99787   34.69374   3.978 0.000108 ***
SpeciesWhitefish   62.52483   30.53433   2.048 0.042335 *
Length3             6.72071    2.83371   2.372 0.018976 *
I(Length3^2)        0.43653    0.03821  11.423  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.74 on 150 degrees of freedom
Multiple R-squared:  0.9651,    Adjusted R-squared:  0.9633
F-statistic: 518.8 on 8 and 150 DF,  p-value: < 2.2e-16
```