STAT 306 Project

Investigation of the Relation Between Variations Concern Hull Geometry Coefficients, the Froude Number and Residuary Resistance

Name: Yuntian Xie	ID: 74376245
Name: Shiyang Zhang	ID: 66326729
Name: Yuan Cao	ID: 85579795
Name: Haneul Kim	ID: 58285446

1. Introduction

1.1 Background and Motivation

The yacht is an essential means of transportation at sea; its lightness, speed, and high flexibility are widely used in different fields such as transportation, military, and recreation. For the general public, yachts are mainly used for racing, cruising and excursions, all of which are related to their fast speed.

In this project, we select the hydrodynamic data of yachts from the UCI machine learning database provided by the Ship Hydromechanics Laboratory of Maritime and Transport Technology Department of the Technical University of Delft as the analysis database. By analyzing the relationship between the explanatory variables and the residuary resistance in table 1, this group hopes to give a best fit linear regression model to help people better determine the speed of a yacht for a given condition. Of course, more factors affect the residuary resistance in the actual yacht operation. However, this project provides a feasible method for determining the value of residuary resistance on a yacht.

1.2 Data Source and Description

The Delft data set comprises 308 full-scale experiments performed at the Delft Ship Hydromechanics Laboratory for that purpose. There are seven different measured variables in the collected data set: 6 explanatory variables and a response variable.

	Variables	**Description**
Explanatory Variable	LC: Longitudinal position of the center of buoyancy	The position of the center of mass of the immersed ship measured along its length.
	PC: Prismatic coefficient	The ratio of the volume of displacement of a ship to that of a prism equal in length to the distance between perpendiculars of the ship and in cross section to that of the immersed mid-ship section.
	LD: Length-displacement ratio	The ratio between the length and the displacement of the vessel.
	BD: Beam-draught ratio	The ratio between the lengths of the widest point at the ship's nominal waterline and the draft of the ship.
	LB: Length-beam ratio	The ratio between the length of the vessel and its beam.

	FR: Froude number	The ratio of the flow inertia to the external field.
Response Variable	RR: Residuary resistance	Residuary resistance per unit weight of displacement

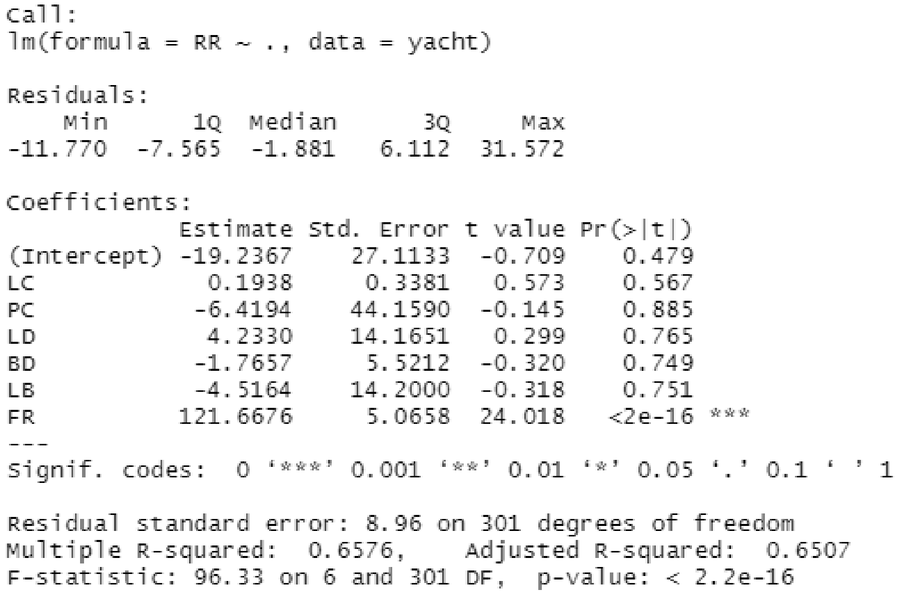
Table 1: Variables and Description

2. Analysis

2.1 Model 1

Using the exhaustive model selection method in r, which compares the possible combination of the various criteria, the model including all explanatory terms was selected as the best one of the linear models, including no quadratic or higher-order terms. This model was named Model 1. Summary of the Model 1 (Figure 1) shows that only FR is significant (P-value <0.05), and the $adjR^2$is 0.6507. Also, a quadratic pattern in the residual plot (Figure 2) differs from the normal error assumption for linear regression is observed.

To sum up, model 1 is not an adequate model. By the pattern of residual plot, adding higher-order explanatory variable terms is considered to improve the model further.



```
Call:
lm(formula = RR ~ ., data = yacht)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.770  -7.565  -1.881   6.112  31.572 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -19.2367    27.1133  -0.709    0.479    
LC            0.1938     0.3381   0.573    0.567    
PC           -6.4194    44.1590  -0.145    0.885    
LD            4.2330    14.1651   0.299    0.765    
BD           -1.7657     5.5212  -0.320    0.749    
LB           -4.5164    14.2000  -0.318    0.751    
FR          121.6676     5.0658  24.018   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.96 on 301 degrees of freedom
Multiple R-squared:  0.6576,	Adjusted R-squared:  0.6507 
F-statistic: 96.33 on 6 and 301 DF,  p-value: < 2.2e-16
```

Figure 1: Summary of Model 1

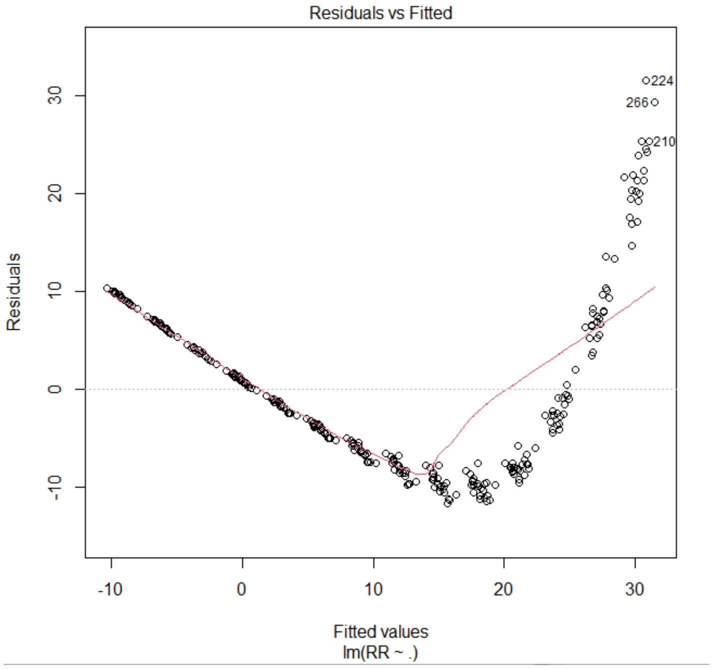


Figure 2: Residuals VS Fitted Value (Model 1)

By using the *ggpair()* function in R, which shows the correlation between each variable, a scatterplot with an obvious pattern between FR term and the RR term is indicated. The pattern between FR term and RR term indicates that adding higher order terms of explanatory variable FR could address the pattern from the residual plot.

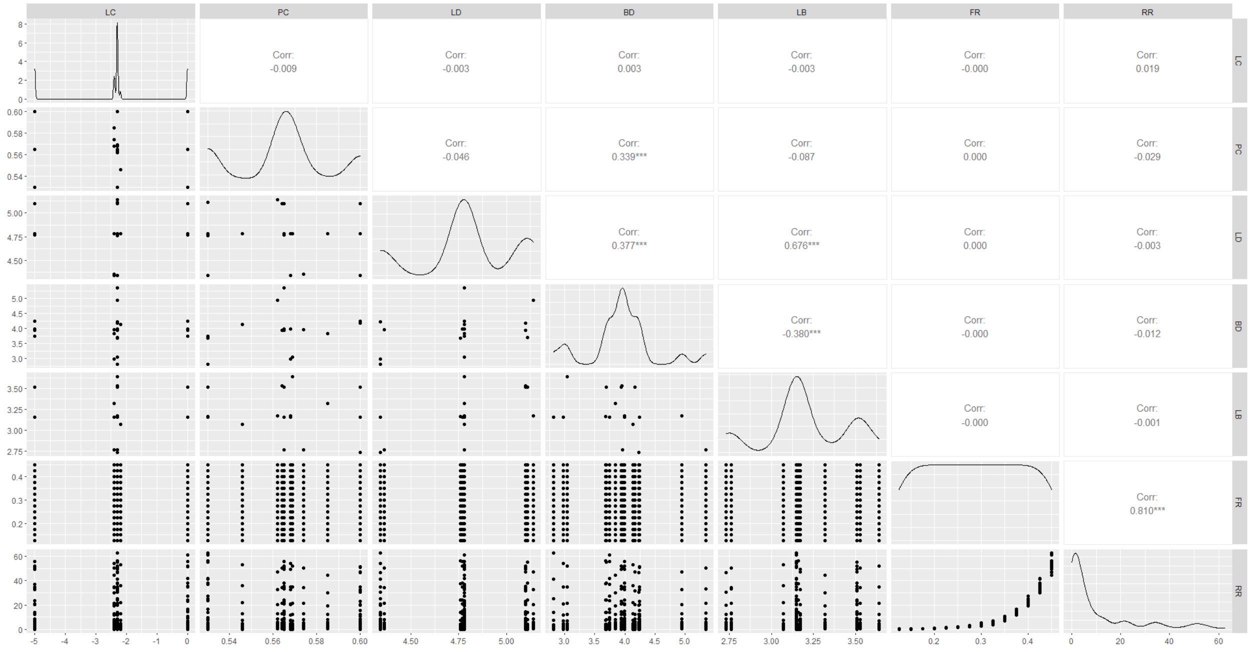
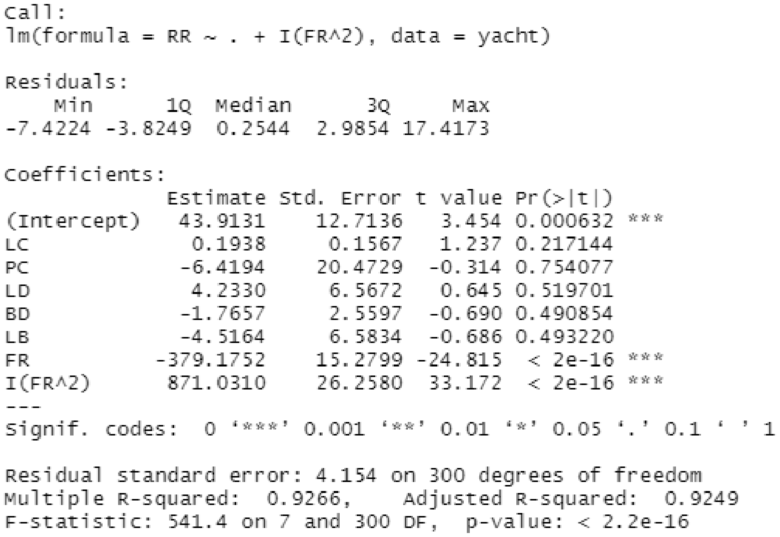


Figure 3: GGpairs plot

2.2 Model 2

The group included the FR2 term to the best linear model from the exhaustive model and got a better $adjR^2$ value 0.9249, which was a significant improvement from the previous model (with $adjR^2$ value 0.6507).

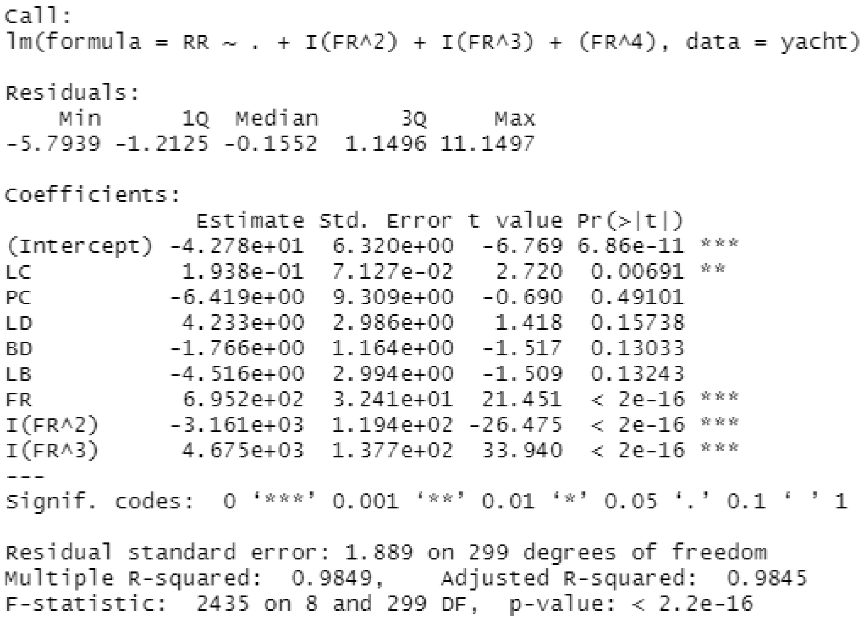


```
Call:
lm(formula = RR ~ . + I(FR^2), data = yacht)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.4224 -3.8249  0.2544  2.9854 17.4173 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   43.9131    12.7136   3.454 0.000632 ***
LC             0.1938     0.1567   1.237 0.217144    
PC            -6.4194    20.4729  -0.314 0.754077    
LD             4.2330     6.5672   0.645 0.519701    
BD            -1.7657     2.5597  -0.690 0.490854    
LB            -4.5164     6.5834  -0.686 0.493220    
FR          -379.1752    15.2799 -24.815  < 2e-16 ***
I(FR^2)      871.0310    26.2580  33.172  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.154 on 300 degrees of freedom
Multiple R-squared:  0.9266,    Adjusted R-squared:  0.9249 
F-statistic: 541.4 on 7 and 300 DF,  p-value: < 2.2e-16
```

Figure 4: Summary for the Model 2

Also, by adding higher-order terms of explanatory variable FR, both adjR2 (increase to 0.9845) and the residual plot was improved. So, the group included up to FR4.

However, adding FR5 to the model did not have a significant impact on theR2 value or the appearance of the residual plot, so the group decided to include up to Fr^4 term.



```
Call:
lm(formula = RR ~ . + I(FR^2) + I(FR^3) + (FR^4), data = yacht)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7939 -1.2125 -0.1552  1.1496 11.1497 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.278e+01  6.320e+00  -6.769 6.86e-11 ***
LC           1.938e-01  7.127e-02   2.720  0.00691 ** 
PC          -6.419e+00  9.309e+00  -0.690  0.49101    
LD           4.233e+00  2.986e+00   1.418  0.15738    
BD          -1.766e+00  1.164e+00  -1.517  0.13033    
LB          -4.516e+00  2.994e+00  -1.509  0.13243    
FR           6.952e+02  3.241e+01  21.451  < 2e-16 ***
I(FR^2)     -3.161e+03  1.194e+02 -26.475  < 2e-16 ***
I(FR^3)      4.675e+03  1.377e+02  33.940  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.889 on 299 degrees of freedom
Multiple R-squared:  0.9849,    Adjusted R-squared:  0.9845 
F-statistic:  2435 on 8 and 299 DF,  p-value: < 2.2e-16
```

Figure 5: Summary for Model 2 involves quadratic terms.

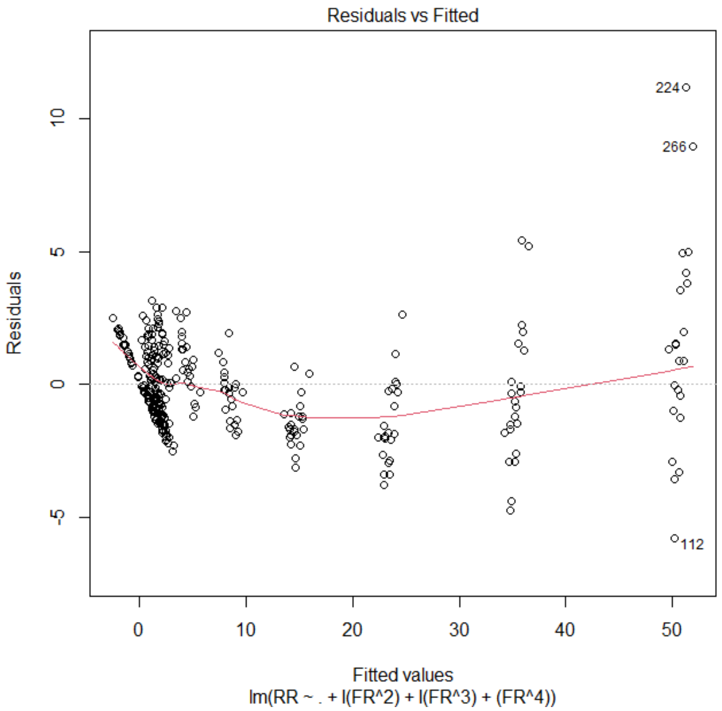
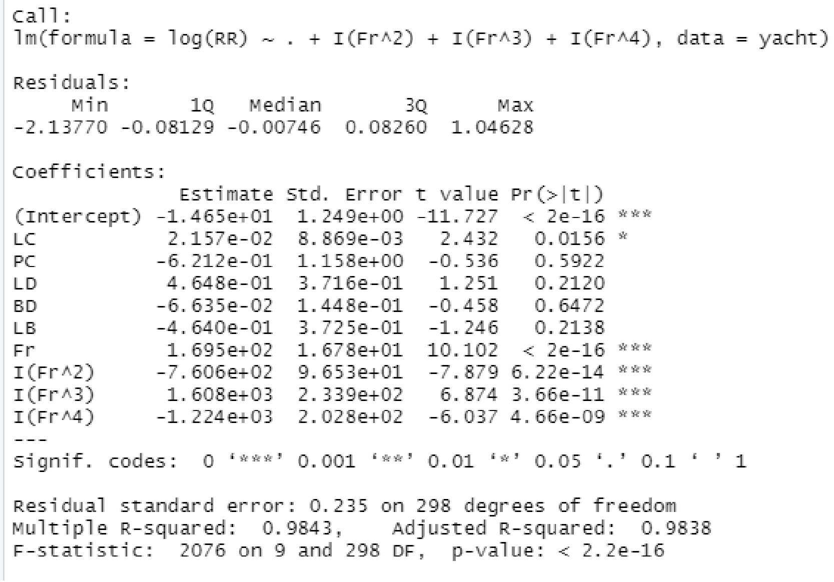


Figure 6: Residual VS Fitted Value plot (Model 2)

2.3 Model 3

From Model 2 it is observed that the overall residual plot (Figure 3) has a pattern, which is condensed residuals at the left side of the plot and spread out residuals at the right side of the plot. This suggests that residuals from Model 2 have a non-constant variance. This means it violates the homoscedasticity assumption for our linear regression. This, by taking log of the response variable, the non-constant variance problem was addressed. The new model is as follows.



```
Call:
lm(formula = log(RR) ~ . + I(Fr^2) + I(Fr^3) + I(Fr^4), data = yacht)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.13770 -0.08129 -0.00746  0.08260  1.04628 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.465e+01  1.249e+00 -11.727  < 2e-16 ***
LC           2.157e-02  8.869e-03   2.432   0.0156 *  
PC          -6.212e-01  1.158e+00  -0.536   0.5922    
LD           4.648e-01  3.716e-01   1.251   0.2120    
BD          -6.635e-02  1.448e-01  -0.458   0.6472    
LB          -4.640e-01  3.725e-01  -1.246   0.2138    
Fr           1.695e+02  1.678e+01  10.102  < 2e-16 ***
I(Fr^2)     -7.606e+02  9.653e+01  -7.879 6.22e-14 ***
I(Fr^3)      1.608e+03  2.339e+02   6.874 3.66e-11 ***
I(Fr^4)     -1.224e+03  2.028e+02  -6.037 4.66e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.235 on 298 degrees of freedom
Multiple R-squared:  0.9843,	Adjusted R-squared:  0.9838 
F-statistic:  2076 on 9 and 298 DF,  p-value: < 2.2e-16
```

Figure 7: Summary for the Model 3

By comparing Model 2 and Model 3, it was noticed that the $adjR^2$ in Model 3 is a little bit lower than in Model 2. However, the residual plot does have significant improvement.

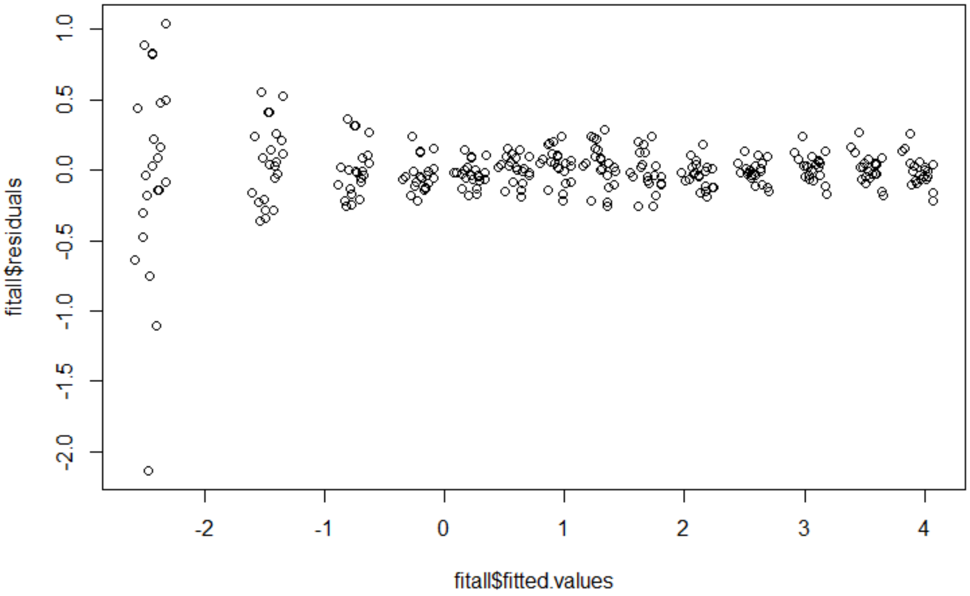
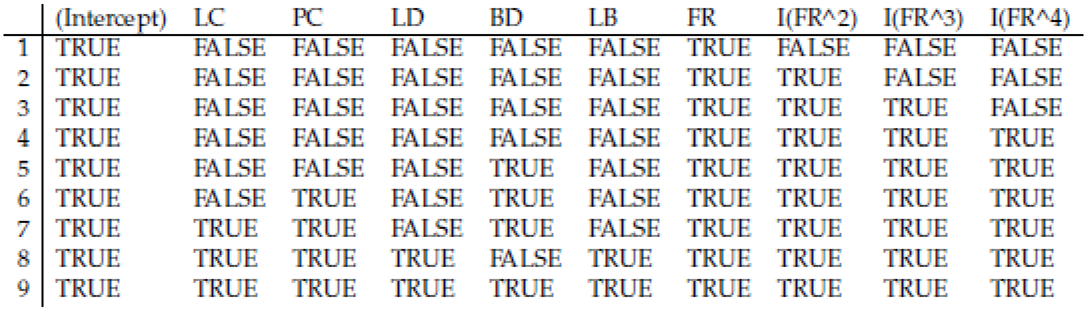


Figure 8: Residual plot for Model 3

However, a model with lesser explanatory variables but still with as many significant parameters as possible is preferred. Hence, the team aimed to improve the p-value of the variables while looking for the model with the least variable.

2.4 Model 4, 5, 6

The team tried to use regsubset() method to find a simple model with a good value of adjR2. This method provides the best model that includes n number of explanatory variables, n ranging from 1 to 9.



	(Intercept)	LC	PC	LD	BD	LB	FR	I(FR^2)	I(FR^3)	I(FR^4)
1	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
3	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
4	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
5	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
6	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
7	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
8	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Figure 9: Best subset selection by using the function *regsubset()* in R

Figure 10 and Figure 11 are two figures based on the value of Cp and $adjR^2$. Recall that in linear regression, the goal is to find the model which has smaller differences between the Cp value and the number of variables.
In Figure 10, a red line was drawn to represent the number of parameters. The closer a Cp value to the red line, the better the model. The team found that the last three points are the best match for the red line. In addition, Figure 11 shows the relationship between $adjR^2$ and the number of parameters. Clearly, the last 3 points have the highest $adjR^2$. Thus, the team decided to fit the model for 7,8,9 parameters, respectively.

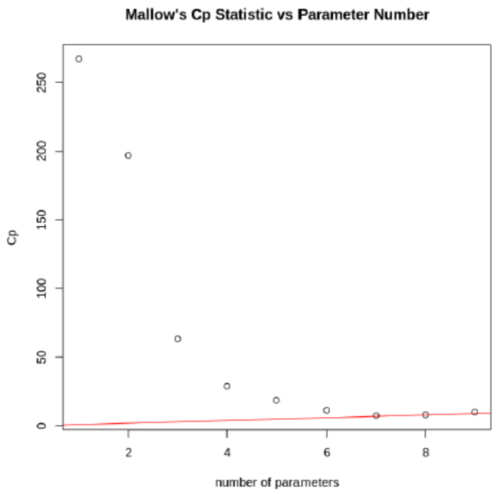


Figure 10: Cp Value VS number of parameters

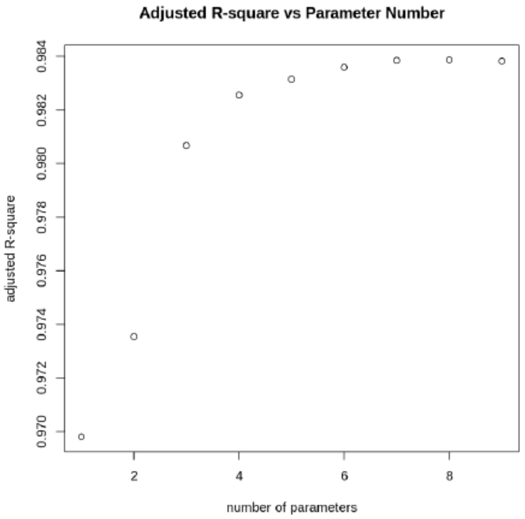


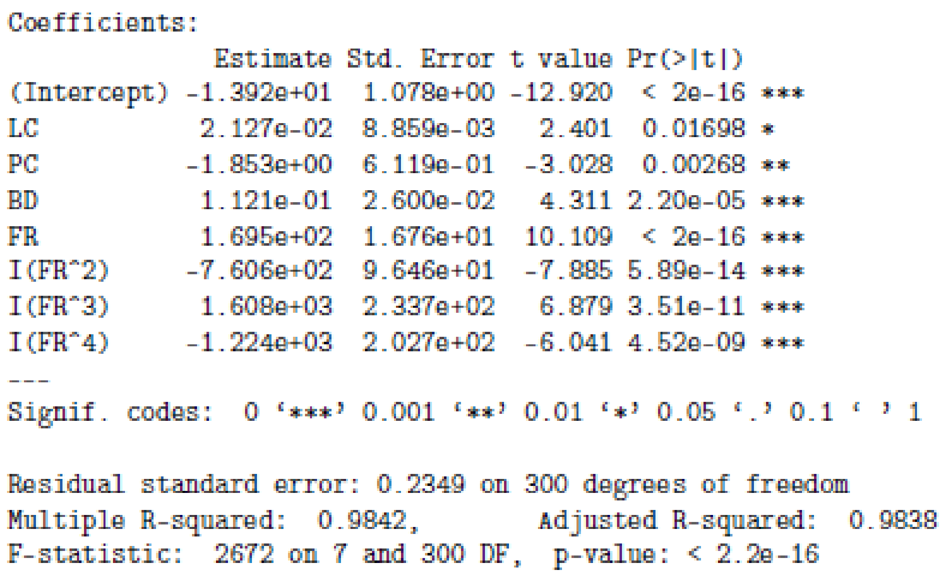
Figure 11: $adjR^2$ VS number of parameters

From Figure 11, it is noticed that the three best $adjR^2$ do not have many differences. So, the team calculates the Akaike information criterion (AIC) in order to make a decision.

Taking a look at Table 2, Model 4 has the smallest AIC, and it is the simplest one (smallest number of parameters) compared with Model 5 and Model 6. Moreover, all parameters in Model 4 are significant (shown in Figure12).

Model	Equation	AIC
4	log(RR)~LC+PC+BD+FR+I(FR^2)+I(FR^3)+I(FR^4)	-8.4467
5	log(RR) ~LC+PC+LD+LB+FR+I(FR^2)+I(FR^3)+I(FR^4)	-7.8473
6	log(RR) ~LC+LD+PC+BD+LB+FR+I(FR^2)+I(FR^3)+I(FR^4)	-6.0641

Table 2: AIC for Model 4, 5, 6



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.392e+01  1.078e+00 -12.920  < 2e-16 ***
LC           2.127e-02  8.859e-03   2.401  0.01698 *  
PC          -1.853e+00  6.119e-01  -3.028  0.00268 ** 
BD           1.121e-01  2.600e-02   4.311 2.20e-05 ***
FR           1.695e+02  1.676e+01  10.109  < 2e-16 ***
I(FR^2)     -7.606e+02  9.646e+01  -7.885 5.89e-14 ***
I(FR^3)      1.608e+03  2.337e+02   6.879 3.51e-11 ***
I(FR^4)     -1.224e+03  2.027e+02  -6.041 4.52e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2349 on 300 degrees of freedom
Multiple R-squared:  0.9842,	Adjusted R-squared:  0.9838 
F-statistic:  2672 on 7 and 300 DF,  p-value: < 2.2e-16
```

Figure 12: Model 4 Regression model(7 parameters)

However, with the residual plot (Figure 13) and QQ plot (Figure 14) of model 4, we can see that model 4 violates the assumption of linear regression that the error terms follow a normal distribution and homoscedasticity. Therefore, we decided to transform explanatory variables to solve the error term problem.

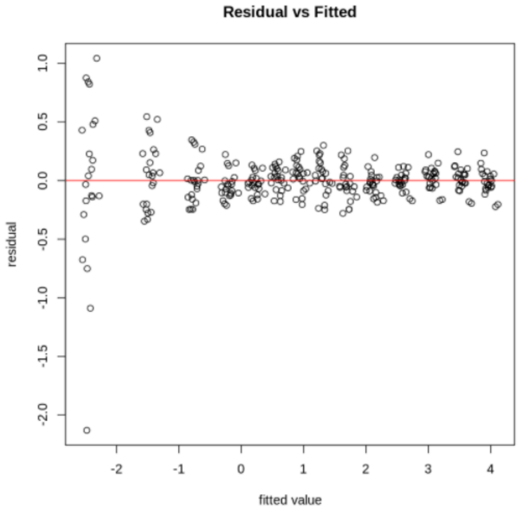


Figure 13: Residual Plot for Model 4

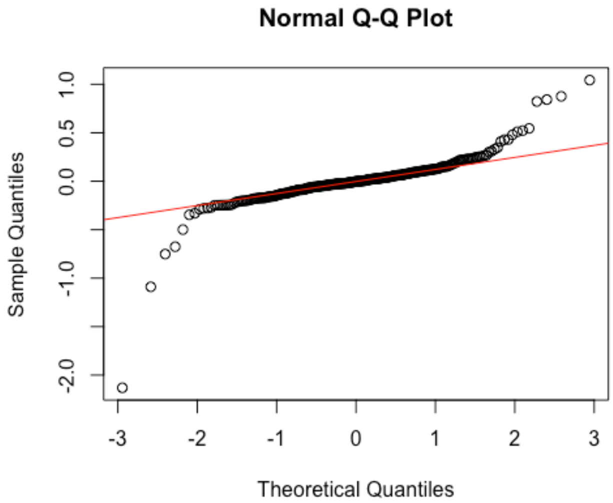


Figure 14: QQ plot for Model 4

2.5 Model 7,8

We add log(PC) and log(BD) terms and then find the best-fitted model using the *regsubset()* method. The mallow's Cp plot(Figure 15) and $adjR^2$ plot(Figure 16) show that the models with eight and nine parameters are probably the best because their $adjR^2$ is large enough, and their mallow's Cp values are closest to their number of parameters.

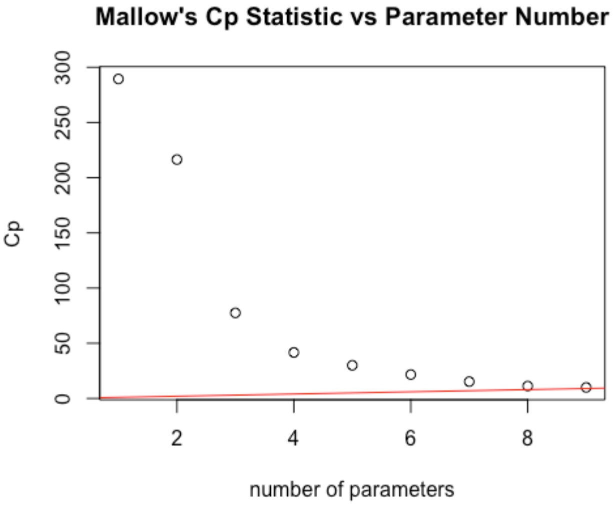


Figure 15: CP Value VS number of parameters

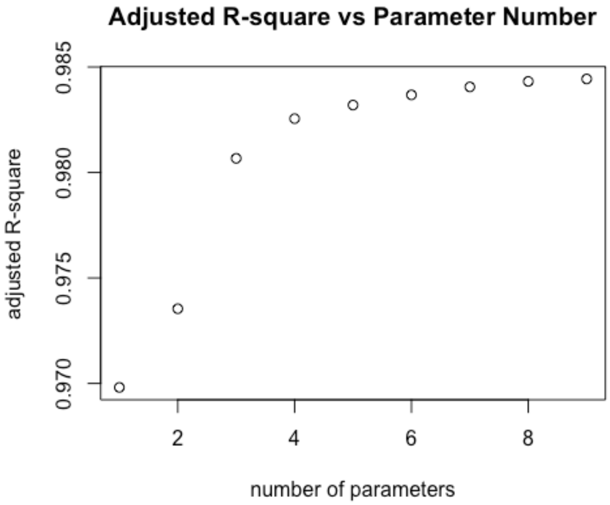
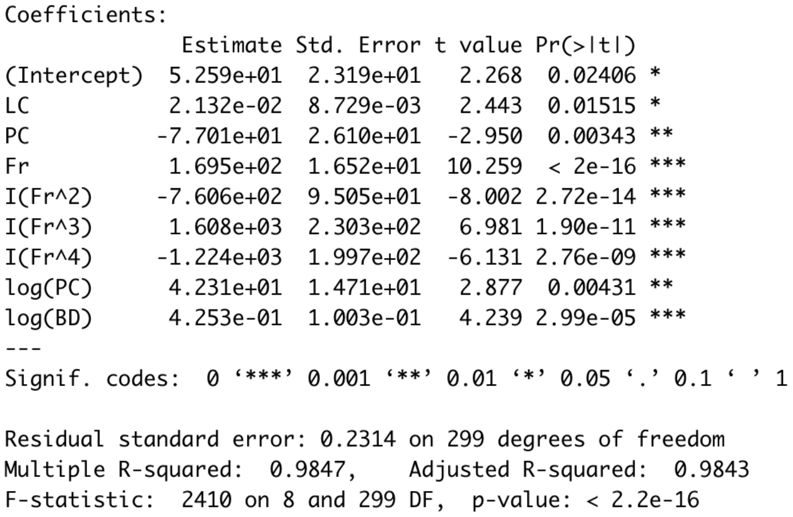


Figure 16: $adjR^2$ VS number of parameters

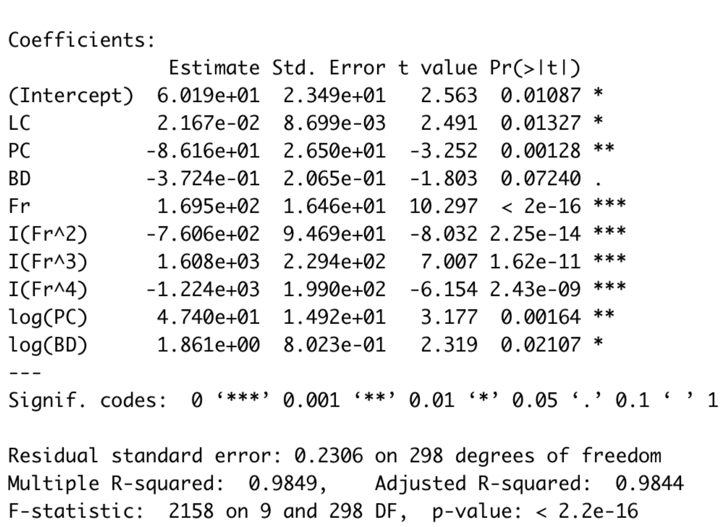
Then based on the information in table 3, we can see that model 8 has the smallest AIC value, so we choose model 8 for further study. However, the summary plot of model 8 (Figure 18) shows that the variable BD is not significant, so we have to remove BD to make a new model, model 7. Although the AIC value of model 7 is not as good as that of model 7, all of its variables are significant (Figure 17), so model seven is the best-fitted model we have chosen so far.

Model Name	Equation	AIC
Model 7	$log(RR) \sim LC + PC + FR + FR^2 + FR^3 + FR^4 + log(PC) + log(BD)$	-16.551
Model 8	$log(RR) \sim$ $LC + PC + BD + FR + FR^2 + FR^3 + FR^4 + log(PC) + log(BD)$	-17.892

Table 3: AIC for Model 7,8



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.259e+01  2.319e+01   2.268  0.02406 *  
LC           2.132e-02  8.729e-03   2.443  0.01515 *  
PC          -7.701e+01  2.610e+01  -2.950  0.00343 ** 
Fr           1.695e+02  1.652e+01  10.259  < 2e-16 ***
I(Fr^2)     -7.606e+02  9.505e+01  -8.002 2.72e-14 ***
I(Fr^3)      1.608e+03  2.303e+02   6.981 1.90e-11 ***
I(Fr^4)     -1.224e+03  1.997e+02  -6.131 2.76e-09 ***
log(PC)      4.231e+01  1.471e+01   2.877  0.00431 ** 
log(BD)      4.253e-01  1.003e-01   4.239 2.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2314 on 299 degrees of freedom
Multiple R-squared:  0.9847,	Adjusted R-squared:  0.9843 
F-statistic:  2410 on 8 and 299 DF,  p-value: < 2.2e-16
```



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.019e+01  2.349e+01   2.563  0.01087 *  
LC           2.167e-02  8.699e-03   2.491  0.01327 *  
PC          -8.616e+01  2.650e+01  -3.252  0.00128 ** 
BD          -3.724e-01  2.065e-01  -1.803  0.07240 .  
Fr           1.695e+02  1.646e+01  10.297  < 2e-16 ***
I(Fr^2)     -7.606e+02  9.469e+01  -8.032 2.25e-14 ***
I(Fr^3)      1.608e+03  2.294e+02   7.007 1.62e-11 ***
I(Fr^4)     -1.224e+03  1.990e+02  -6.154 2.43e-09 ***
log(PC)      4.740e+01  1.492e+01   3.177  0.00164 ** 
log(BD)      1.861e+00  8.023e-01   2.319  0.02107 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2306 on 298 degrees of freedom
Multiple R-squared:  0.9849,	Adjusted R-squared:  0.9844 
F-statistic:  2158 on 9 and 298 DF,  p-value: < 2.2e-16
```

Figure 17: Model 7 Regression model

Figure 18: Model 8 Regression model

However, model 7 does not improve the problem of model 4. With the residual plots of model 7 (Figure 19) and QQ plots (Figure 20), we can see that model 7 still violates the assumption of linear regression that the error terms follow a normal distribution and homoscedasticity.

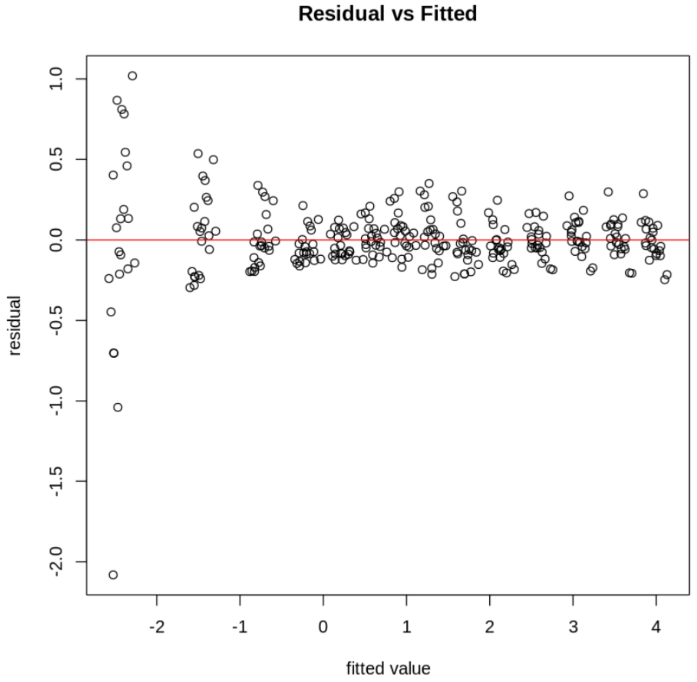


Figure 19: Residual Plot for Model 7

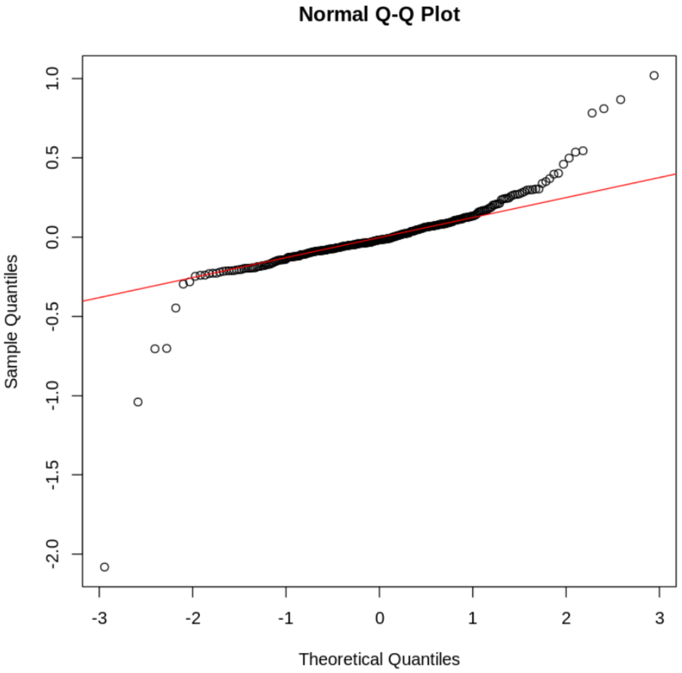


Figure 20: QQ plot for Model 7

3. Discussion

The model best fitting the data given is model 7, its Cp values, $adjR^2$ values, and AIC values are closest to expected. To be specific, for the best model, it needs to have three qualities: 1. Cp value should be close to the number of variables, 2. A large $adjR^2$, 3. A small AIV value. However, model 7 still has some problems. The group noticed that in the residual plot, the pattern still exists. On the leftmost side of the plot, some points exceed the normal range of the residual value. Normal distribution random errors are one of the assumptions for a linear regression model, but after many adjustments, this model still appears to be right-skewed. Also, an ideal qq plot is evenly distributed on the red line, but from the results the group got, some points on the left are obviously lower than the red line, and some points on the right are obviously higher than the red line. In summary, model 7 may not be a true linear regression model. The group proposed the following two possible reasons for the nonlinear model. The first reason is that although this dataset is laboratory data, it is hard to tell that the result is ideal for model fitting. The second reason is that some other factors may affect the resistance that is not considered in this dataset.

Model 7 includes many high-order terms and log terms, which makes the interpretation of the model and the application of the model difficult. Compared with model 4 and 7, the residual plot and the $adjR^2$ values are similar, except that the AIC values are different (model 4 having -8.4467, and model 7 having -16.551). Since AIC value can suggest a different best model from other criteria, model 4 can be selected when the Cp value, $adjR^2$are considered. Also, considering the prediction power of the models, two-fold cross-validation suggests better prediction power of model 4 than model 7. The prediction error from model 4 was 0.1080, whereas model 7 had 0.1280. In addition, model 4 has fewer explanatory variable terms. Therefore, from another point of view, model 4 is also a model that could be considered for this dataset.

In addition, the group can conclude from model 4 and model 7 that they share common significant explanatory variables, LC, PC, FR, and BD terms even if they are in different forms, meaning the response variable changes as these significant terms change.

4. Conclusion

In conclusion, the group found a good model with an acceptable value of Cp, $adjR^2$ and AIC, but its residual plot and QQ plot are not close to the ideal model. Since there is no research on the relationship between the various physical quantities in the data, the interaction terms of the fitted model are not considered. Also, this model is based on real data; some experimental errors or outliers will be accepted.

Reference

Dua, Dheeru and Graff, Casey. “UCI Machine Learning Repository .” University of California, Irvine, School of Information and Computer Sciences, 2017, http://archive.ics.uci.edu/ml.

Baressi Šegota, Sandi & Anđelić, Nikola & Kudlaček, Jan & Čep, Robert. (2020). Artificial neural network for predicting values of residuary resistance per unit weight of displacement. Journal of Maritime & Transportation Science. 57. 10.18048/2019.57.01..