University of British Columbia
STAT306: Finding Relationships in Data

Group 14 Project:
Yacht Hydrodynamics Data Set

Ning Wang (20552618)
Calvin Lefebvre (72429111)
Martin Pi (49981319)
Jaskaran Singh (44025013)

INTRODUCTION

The resistance provided by a ship to move through water can be categorised into two distinct components**:** ***frictional resistance and residual resistance.*** The dataset that we are incorporating for this experiment involves working with residual resistance. In this analysis we wish to explain how residuary resistance per unit weight of displacement is influenced by different factors such as velocity, buoyancy, Froude number, etc.

The Delft dataset comprises 308 full-scale experiments which were performed at the Delft Ship Hydrodynamics to predict the hydrodynamic performance of sailing yachts and estimate the required propulsive power. Essential inputs include the basic hull dimensions and boat velocity. There are 22 different hull forms in these experiments, all of which are derived from a parent that is closely related to Frans Mass' 'Standfast 43'. The variable definitions and attribute information is as follows [1].

Explanatory Variable:

- (V1) Longitudinal position of the center of buoyancy, adimensional.
- (V2) Prismatic coefficient, adimensional.
- (V3) Length-displacement ratio, adimensional.
- (V4) Beam-draught ratio, adimensional.
- (V5) Length-beam ratio, adimensional.
- (V6) Froude number, adimensional.

Response Variable:

- (V7) The measured variable is the residuary resistance per unit weight of displacement.

ANALYSIS

Residual plots

We start our investigation by exploring residual plots and QQ plots to determine the validity of the linear relationship between x and y variables. As seen in the residual plot for lm(y ~ V6) in Figure 1, there is a clear curved pattern, which indicates the linear relationship between x and y is invalid. To solve this problem, we needed to add quadratic terms, V6^2. In Figure 1, the normal qq plot for lm(y ~ V6 + V6^2) shows almost all the points are on the line, except for points at the tail.

After adding the V6^2 term into the model, we can still see there still is a clear curved pattern in Figure 1. To resolve this, we add the V6^3 term into the model. By including V6^2 and V6^3, we can see in Figure 1 there is a clear linear increasing trend for the residuals through increasing the fitted values. Thus, the assumption for a linear relationship between x and y is made.

Lastly, in Figure 1 we see the common variance assumption is invalid, which leads us to transform V7 to log V7. This transformation allows for common variance .

We also performed the residual plots and normal qq plots for the categorical data (V1 to V5) as seen in Figure 2. In the residual plots, we can see the residuals are randomly distributed which are centered at 0. Also, there is no positive or negative correlation between each observation. In these normal QQ plots, we can see most points are not on the line and that it is also heavily tailed. We decided to take the log of V7 to fix this.

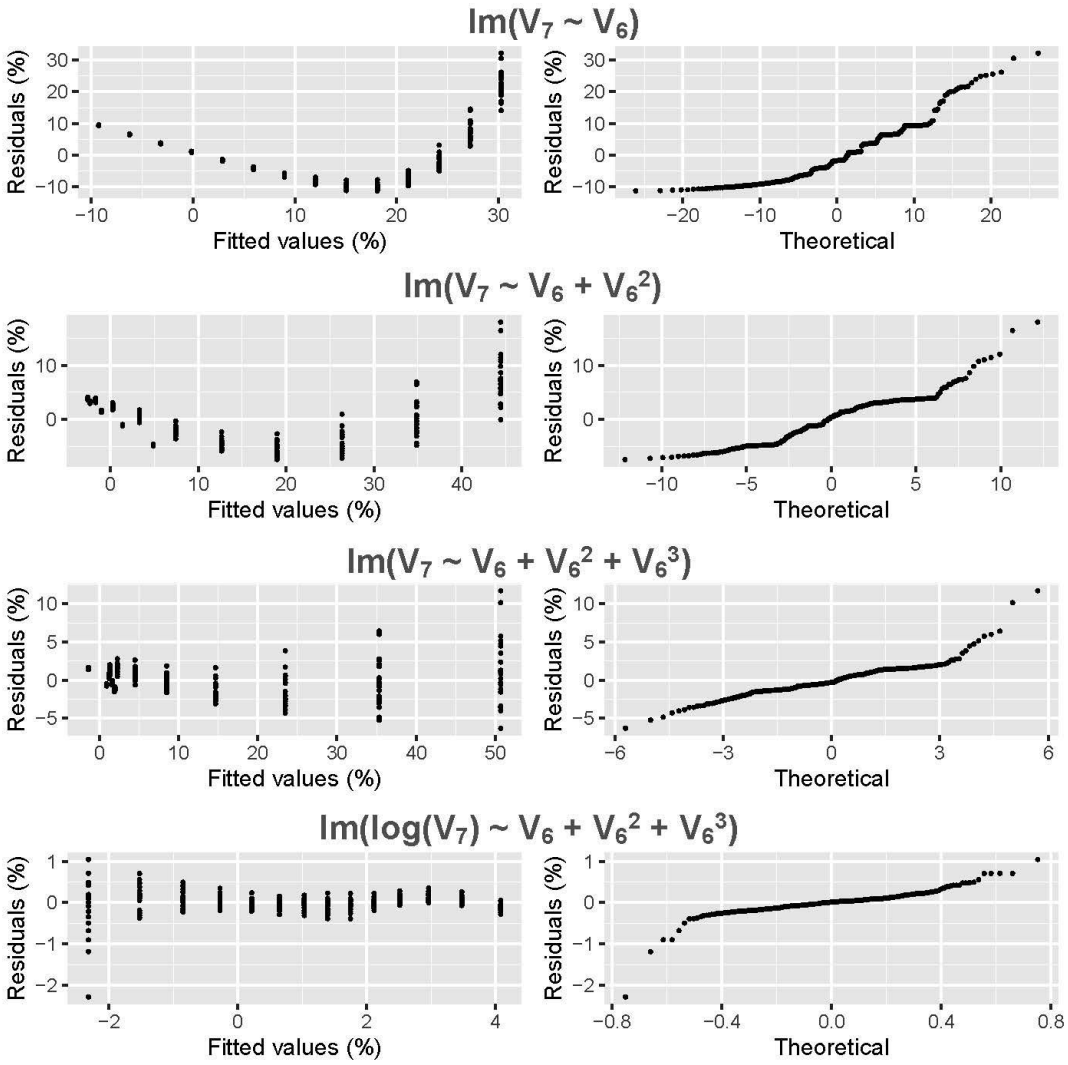


Figure 1: Residual plots (left) and QQ plots (right) for V6 transformation.

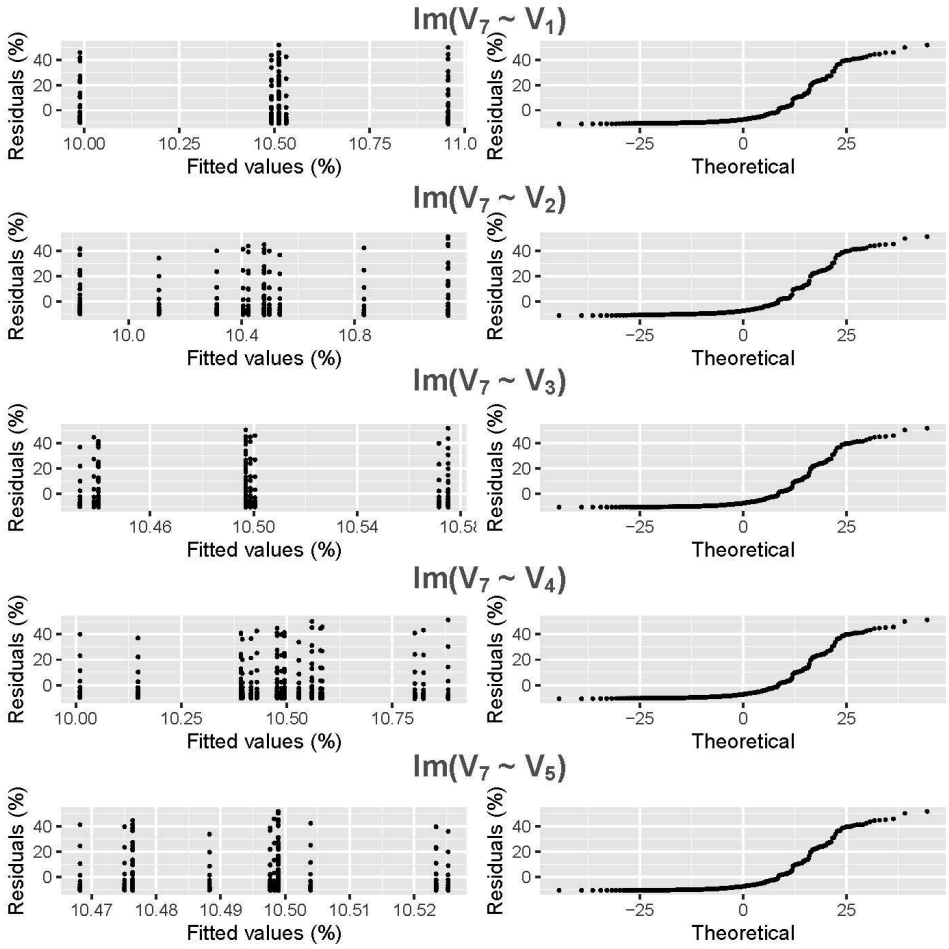


Figure 2: Residual plots (left) and QQ plots (right) for V1-V5

Finding the models

Firstly, we explored using a best subset selection algorithm with exhaustive selection,forward selection and backward selection to choose models with the best 3 to 4 predictors. The subset algorithm calculates adjusted R squared, and Mallows Cp and we calculated additional model performance features, such as: Akaike information criterion (AIC), root mean square error (RMSE), and prediction error. We choose models with the highest adjusted R squared, lowest Mallow's Cp and model simplicity.

Additionally, we investigated if including interaction terms would increase the model performance. To do this, we included interaction terms between only two explanatory variables (V1*V2, V1*V3, … , V5*V6). The top model that includes an interactive term is labeled as "Interaction" Model**,**

$$\log(v_7) = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3 + \beta_4 v_3 v_4$$

seen in Table 1. This model was determined by the exhaustive selection mode.

As mentioned in the above section, we chose to take the log transformation of the response variable. In contrast, we also included the model without this transformation for comparison purposes. This model is labeled as **"Best3withNoLog"** model,

$$v_7 = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3$$

which indicates the best 3 variables using exhaustive selection mode and no log transformation on V7. We also made some other models from the selection:

$$\log(v_7) = \beta_0 + \beta_1 v_4 + \beta_2 v_6 + \beta_3 v_6^2 + \beta_4 v_6^3$$

This model is the best 4 predictors model selected by both exhaustive selection and forward selection.

$$\log(v_7) = \beta_0 + \beta_1 v_3 + \beta_2 v_6 + \beta_3 v_6^2 + \beta_4 v_6^3$$

This model is the best 4 predictors model selected by backward selection.

$$\log(v_7) = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_4 + \beta_4 v_6 + \beta_5 v_6^2 + \beta_6 v_6^3$$

This model has the lowest Mallow's cp in all the selections.

$$\log(v_7) = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \beta_4 v_5 + \beta_5 v_6 + \beta_6 v_6^2 + \beta_7 v_7^3$$

This model has the largest adjusted R squared in all the selections.

The RMSE and prediction error were calculated using training and testing sets. We first calculated the prediction error by splitting the data into 2 approximately equal sized subsets, then fit the model on the first half and compute prediction errors for the other half. Secondly, we calculated the RMSE by randomly sampling 70 percent of data to become the training set, which is used to fit the model, and the remaining 30 percent of data will be used for the testing set. From these two sets,we can calculate the RMSE.

```
> set.seed(569)
> train_ind<-sample.int(nrow(yacht),0.7*nrow(yacht))
> train <- yacht[train_ind , ]
> test <- yacht[-train_ind , ]
> reg_train<-lm(log(V7)~V6+I(V6^2)+I(V6^3),data=train)
> reg2_train<-lm(V7~V6+I(V6^2)+I(V6^3),data=train)
> reg3_train<-lm(log(V7)~V4+V6+I(V6^2)+I(V6^3),data=train) #best 4 from forward selection and exhaustive se
lection
> reg4_train<-lm(log(V7)~V3+V6+I(V6^2)+I(V6^3),data=train)#best 4 from backward selection
> reg5_train<-lm(log(V7) ~V1+V2+V4+V6+I(V6^2)+I(V6^3), data=train) # 6 from min cp
> reg6_train<-lm(log(V7) ~V1+V2+V3+V5+V6+I(V6^2)+I(V6^3), data=train)
> reg7_train<-lm(log(V7)~V6+I(V6^2)+I(V6^3)+V3*V4-V3-V4, data=train)
> sqrt(mean((test$V7-exp(predict(reg_train,newdata = test)))^2))
[1] 3.065069
> sqrt(mean((test$V7-predict(reg2_train,newdata = test))^2))
[1] 1.677241
> sqrt(mean((test$V7-exp(predict(reg3_train,newdata = test)))^2))
[1] 3.442238
> sqrt(mean((test$V7-exp(predict(reg4_train,newdata = test)))^2))
[1] 3.283484
> sqrt(mean((test$V7-exp(predict(reg5_train,newdata = test)))^2))
[1] 3.271653
> sqrt(mean((test$V7-exp(predict(reg6_train,newdata = test)))^2))
[1] 3.242125
> sqrt(mean((test$V7-exp(predict(reg7_train,newdata = test)))^2))
[1] 3.527683
>
```

Figure 3: Code for RMSE method

```
> #prediction error
>
> #best 3 without log
> train <- 1:as.integer(dim(yacht)[1]/2)
> reg1 <- lm(V7~V6+I(V6^2)+I(V6^3), data=yacht[train,])
> error1 <- sum((yacht$V7[-train] - predict(reg1, yacht[-train,]))^2)
> reg2 <- lm(V7~V6+I(V6^2)+I(V6^3), data=yacht[-train,])
> error2 <- sum((yacht$V7[train] - predict(reg2, yacht[train,]))^2)
> error <- (error1 + error2)/dim(yacht)[1]
> error
[1] 4.546575
>
>
>
> #best 3
> train <- 1:as.integer(dim(yacht)[1]/2)
> reg1 <- lm(log(V7)~V6+I(V6^2)+I(V6^3), data=yacht[train,])
> error1 <- sum((yacht$V7[-train] - exp(predict(reg1, yacht[-train,])))^2)
> reg2 <- lm(log(V7)~V6+I(V6^2)+I(V6^3), data=yacht[-train,])
> error2 <- sum((yacht$V7[train] - exp(predict(reg2, yacht[train,])))^2)
> error <- (error1 + error2)/dim(yacht)[1]
> error
[1] 8.293667
>
> #best4 from exhaustive and forward
> train <- 1:as.integer(dim(yacht)[1]/2)
> reg1 <- lm(log(V7)~V4+V6+I(V6^2)+I(V6^3), data=yacht[train,])
> error1 <- sum((yacht$V7[-train] - exp(predict(reg1, yacht[-train,])))^2)
> reg2 <- lm(log(V7)~V4+V6+I(V6^2)+I(V6^3), data=yacht[-train,])
> error2 <- sum((yacht$V7[train] - exp(predict(reg2, yacht[train,])))^2)
> error <- (error1 + error2)/dim(yacht)[1]
> error
[1] 9.690707
>
> #best4 from backward
> train <- 1:as.integer(dim(yacht)[1]/2)
> reg1 <- lm(log(V7)~V3+V6+I(V6^2)+I(V6^3), data=yacht[train,])
> error1 <- sum((yacht$V7[-train] - exp(predict(reg1, yacht[-train,])))^2)
> reg2 <- lm(log(V7)~V3+V6+I(V6^2)+I(V6^3), data=yacht[-train,])
> error2 <- sum((yacht$V7[train] - exp(predict(reg2, yacht[train,])))^2)
> error <- (error1 + error2)/dim(yacht)[1]
> error
[1] 8.879072
>
> #6 min cp
> train <- 1:as.integer(dim(yacht)[1]/2)
> reg1 <- lm(log(V7) ~V1+V2+V4+V6+I(V6^2)+I(V6^3), data=yacht[train,])
> error1 <- sum((yacht$V7[-train] - exp(predict(reg1, yacht[-train,])))^2)
> reg2 <- lm(log(V7) ~V1+V2+V4+V6+I(V6^2)+I(V6^3), data=yacht[-train,])
> error2 <- sum((yacht$V7[train] - exp(predict(reg2, yacht[train,])))^2)
> error <- (error1 + error2)/dim(yacht)[1]
> error
[1] 22.41788
>
> #7 max adjr2
> train <- 1:as.integer(dim(yacht)[1]/2)
> reg1 <- lm(log(V7) ~V1+V2+V3+V5+V6+I(V6^2)+I(V6^3), data=yacht[train,])
> error1 <- sum((yacht$V7[-train] - exp(predict(reg1, yacht[-train,])))^2)
> reg2 <- lm(log(V7) ~V1+V2+V3+V5+V6+I(V6^2)+I(V6^3), data=yacht[-train,])
> error2 <- sum((yacht$V7[train] - exp(predict(reg2, yacht[train,])))^2)
> error <- (error1 + error2)/dim(yacht)[1]
> error
[1] 20.04742
>
> #with interaction from exhaustive selection
> train <- 1:as.integer(dim(yacht)[1]/2)
> reg1 <- lm(log(V7) ~V6+I(V6^2)+I(V6^3)+V3*V4-V3-V4, data=yacht[train,])
> error1 <- sum((yacht$V7[-train] - exp(predict(reg1, yacht[-train,])))^2)
> reg2 <- lm(log(V7) ~V6+I(V6^2)+I(V6^3)+V3*V4-V3-V4, data=yacht[-train,])
> error2 <- sum((yacht$V7[train] - exp(predict(reg2, yacht[train,])))^2)
> error <- (error1 + error2)/dim(yacht)[1]
> error
[1] 9.895813
```

Figure 4: Code for Prediction error

Comparing the models

Ideally, the optimal model would have the highest adjusted R squared, the Mallow's Cp should be close to the number of linear regression parameters and the lowest AIC value, prediction error and RMSE.

As seen in Table 2, we can clearly infer that the **"Best3",**

$$\log(v_7) = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3$$

model (the best 3 variables using exhaustive selection mode and log y) and "Best3withNoLog" model has the lowest prediction error and RMSE than the other models. These two models do not have the lowest Mallow's cp and AIC, but we propose that the reason is the other models could be overfitted and overly complicated models with too many variables. This could cause these models to lose the prediction power and cause respectively poor performance. Finally, we decided to choose the "Best3" model because it has the lowest prediction error and RMSE, the other model performance values are good and without log transformed response variable the common variance assumption is invalid. In order to perform linear regression, we have to keep all assumptions valid, including common variance assumption.

Model Name	Model Formulae
Model Best3	$\log(v_7) = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3$
Model Best3withNoLog	$v_7 = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3$
Model Best4ForwardExhaustive	$\log(v_7) = \beta_0 + \beta_1 v_4 + \beta_2 v_6 + \beta_3 v_6^2 + \beta_4 v_6^3$
Model Best4Backward	$\log(v_7) = \beta_0 + \beta_1 v_3 + \beta_2 v_6 + \beta_3 v_6^2 + \beta_4 v_6^3$
Model LowestCP	$\log(v_7) = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_4 + \beta_4 v_6 + \beta_5 v_6^2 + \beta_6 v_6^3$
Model HighestAdjustedRSquare	$\log(v_7) = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \beta_4 v_5 + \beta_5 v_6 + \beta_6 v_6^2 + \beta_7 v_7^3$
Model Interaction	$\log(v_7) = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3 + \beta_4 v_3 v_4$

Table 1: Model Name and Model Formulae

Model Formulae	Adjusted R^2	CP	AIC	Prediction Error	RMSE
$\log(v_7) = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3$	0.9806705	24.787724	42.95161	8.293667	3.065069
$v_7 = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3$	0.9834056	24.996555	1292.359	4.546575	1.677241
$\log(v_7) = \beta_0 + \beta_1 v_4 + \beta_2 v_6 + \beta_3 v_6^2 + \beta_4 v_6^3$	0.9812589	9.547201	34.41452	9.690707	3.442238
$\log(v_7) = \beta_0 + \beta_1 v_3 + \beta_2 v_6 + \beta_3 v_6^2 + \beta_4 v_6^3$	0.9808176	11.622507	41.58377	8.879072	3.283484
$\log(v_7) = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_4 + \beta_4 v_6 + \beta_5 v_6^2 + \beta_6 v_6^3$	0.9819435	6.402748	24.9129	22.41788	3.271653
$\log(v_7) = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_3 + \beta_4 v_5 + \beta_5 v_6 + \beta_6 v_6^2 + \beta_7 v_7^3$	0.9819566	7.187637	25.66454	20.04742	3.242125
$\log(v_7) = \beta_0 + \beta_1 v_6 + \beta_2 v_6^2 + \beta_3 v_6^3 + \beta_4 v_3 v_4$	0.9812965	46.317	33.79683	9.895813	3.527683

Table 2: A summary of model formulae and computed adjusted R square,Mallow CP, Akaike Information Criterion (AIC), Prediction Error, Root Mean Square Error values

Final model

As mentioned, we have chosen our best model to be “Best3”, which includes V6, V6^2, V6^3. This is supported by correlation plots in Figure 5.

The correlation plots between all the variables shows there is a strong correlation between V6 and V7. After we transform V7 to log V7, the correlation still looks strong. We also can see that the correlation between V7 and other variables except V6 is not that strong. This phenomenon explains why V6 is dominant in the model as taking V7 for the response variable.

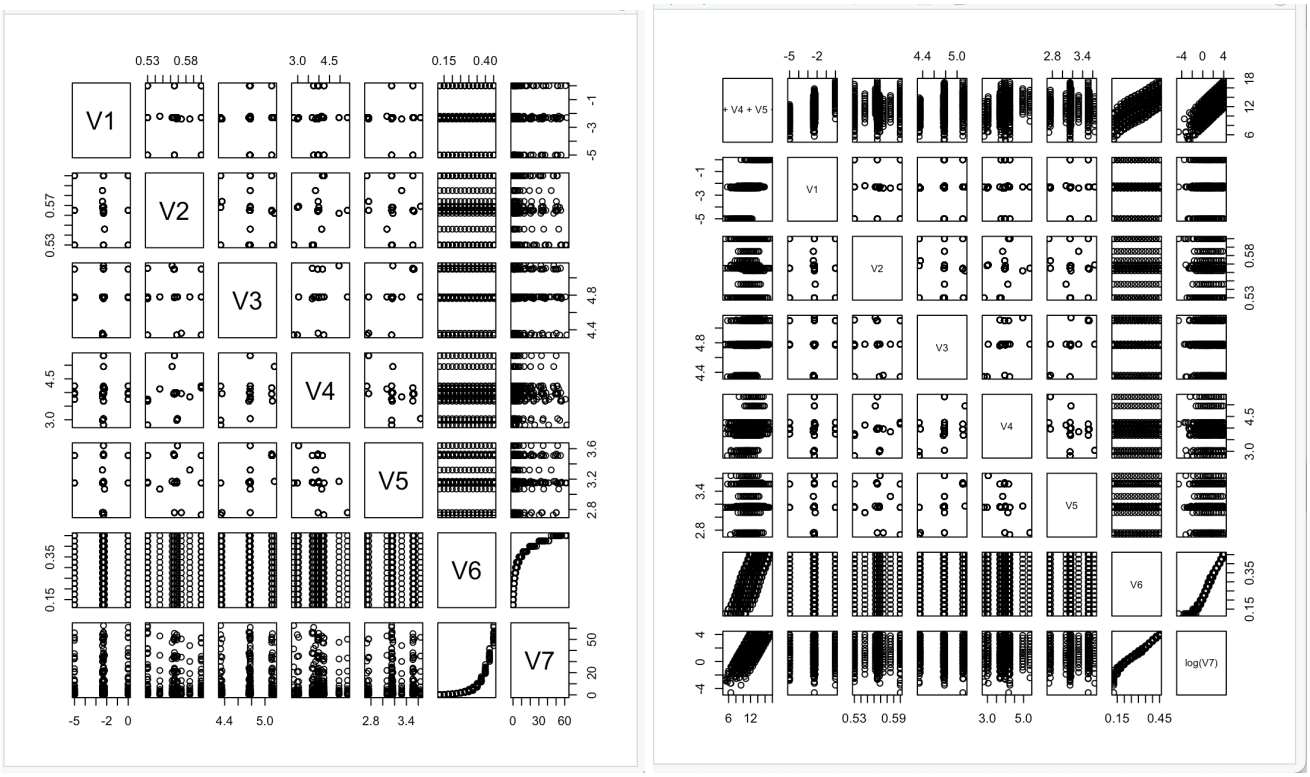
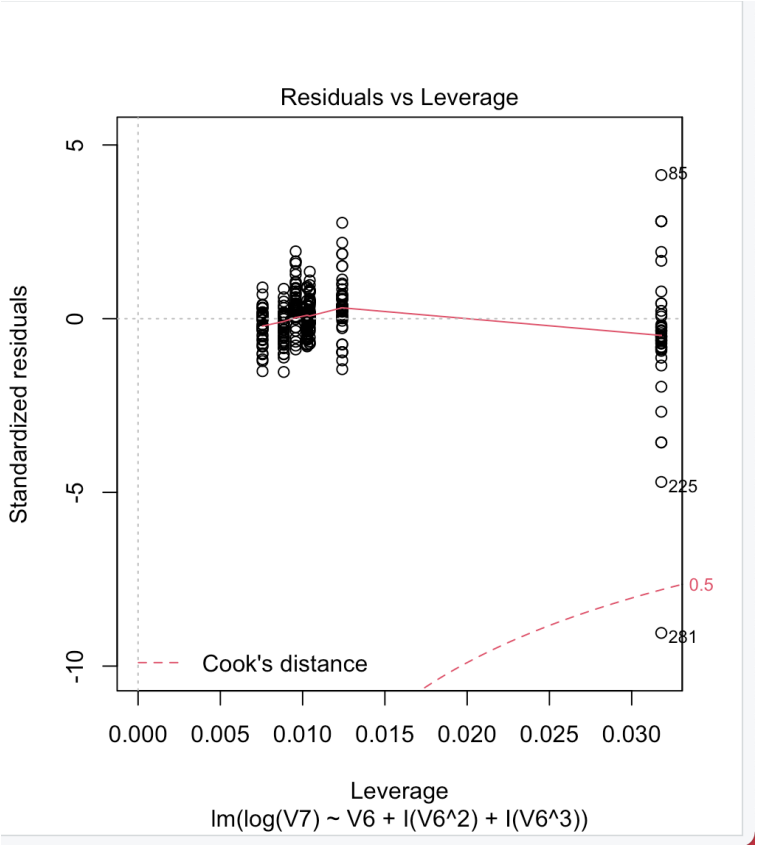


Figure 5: Correlation plots between V1 to V7 and Correlation plots between V1 to log(V7).

In our final step, we talked about which method could help us to improve our final model. We identify any influential points by creating a residuals vs leverage plot. This plot displays how much of an improvement would be made if a proposed influential point was removed.




```
> summary(lm(log(V7)~V6+I(V6^2)+I(V6^3),data=yacht[-281,]))

Call:
lm(formula = log(V7) ~ V6 + I(V6^2) + I(V6^3), data = yacht[-281,
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-1.26339 -0.10120  0.01010  0.09245  0.97020

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.2497     0.3195  -25.82   <2e-16 ***
V6            66.2877     3.8034   17.43   <2e-16 ***
I(V6^2)     -168.7266    13.9920  -12.06   <2e-16 ***
I(V6^3)      182.7646    16.1247   11.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.22 on 303 degrees of freedom
Multiple R-squared:  0.9856,    Adjusted R-squared:  0.9854
F-statistic:  6895 on 3 and 303 DF,  p-value: < 2.2e-16

> summary(lm(log(V7)~V6+I(V6^2)+I(V6^3),data=yacht))

Call:
lm(formula = log(V7) ~ V6 + I(V6^2) + I(V6^3), data = yacht)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28690 -0.09902  0.01018  0.09868  1.04531

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.7162     0.3695  -23.59   <2e-16 ***
V6            71.1627     4.4088   16.14   <2e-16 ***
I(V6^2)     -184.7861    16.2431  -11.38   <2e-16 ***
I(V6^3)      199.6060    18.7382   10.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2569 on 304 degrees of freedom
Multiple R-squared:  0.9809,    Adjusted R-squared:  0.9807
F-statistic:  5193 on 3 and 304 DF,  p-value: < 2.2e-16
```

Figure 6: Residuals vs Leverage plot and summaries for both removing and keeping the influential points

As seen in Figure 4, the observation #281, in the right bottom corner, falls outside of the red dashed line. This indicates that it is an influential point. After removing the influential point, we expect the models' fit will improve. This results in the adjusted R squared slightly increasing from 0.9807 to 0.9854. Since this influential point is not necessarily an error and does not significantly improve the model, we decide to keep this point.

CONCLUSION

We have fitted different models to investigate the relations between the response variable and its explanatory variables. With the above methods and analysis shown above, we selected **"Best3"** as the best predictor model among the 7 high potential of the best models. We identify there exists a high association between Froude number (V6) and residuary resistance per unit weight of displacement (V7); therefore, the final model only contains Froude number and its quadratic terms.

REFERENCES

[1] “UCI Machine Learning Repository: Yacht Hydrodynamics Data Set”. 7 Dec. 2021, 4:16 p.m., https://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics