STA 206 FQ 2021- Final Project

Predicting the age of abalone from physical measurements using Linear Regression

a) Saurabh Pattabiraman Sivakumar (sausiva@ucdavis.edu)

b) Sudheesh Kumar Ethirajan (sethirajan@ucdavis.edu)

Abstract

Best linear regression model to predict the age/ rings of abalone was found using a four-step methodology. Exploratory data analysis was performed on the entire abalone dataset. Preliminary model fit indicated the full first order model was insufficient to explain the variation in the abalone dataset. Moreover, from Box-cox procedure, a log transformation of the response variable is needed. Greedy search strategy, stepwise regression algorithm was implemented to find the best model containing the interaction terms according to AIC and BIC criterion. Two-way interaction model was found to give best model metrics. Ridge regression was also performed on the dataset due to high multicollinearity and the two models were compared and their metrics were found to be similar. All the analysis, model building, and the associated plots were made using professional level data analysis software, RStudio.

---

1. Introduction

The Abalone data contains various measurements collected on a sample of 4,177 abalone. From the documentation [1], we see that there are nine variables. The names of the variables along with other relevant information are summarized in table 1 (see Appendix 1). According to the documentation, the database was first owned by the Department of Primary Industry and Fisheries, Tasmania. In particular, the Marine Resources Division of the Marine Research Laboratories - Taroona. And it was donated to UC Irvine's Machine Learning Repository [1] by Sam Waugh who was associated with the Department of Computer Science at the University of Tasmania in December 1995.

It appears that the data was originally created for the study of the Population Biology of Abalone in Tasmania [2, 3]. Furthermore, the dataset has been used by many different individuals and organizations for exploring data science/data analysis concepts. The Statistical Consulting Group at the San Diego State University has written a post on performing linear regressions [4] in R using this dataset.

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope - a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.

In this project, using various statistical concepts [5] and tools, a good linear model with optimal bias- variance tradeoff and good predictive power is desired. The methods section gives an outline on the framework used to find the good model.

2. Methods

The model building steps for the abalone dataset consists of mainly four steps,

a) Exploratory data analysis

Exploratory Data Analysis refers [6] to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights as possible from it. It is all about making sense of data in hand, before starting to build a model. It is mainly used to identify the type of each variable in the dataset, study the distribution of each variable such as symmetric or skewed and finally the relationship among the variables in the dataset.

The distribution of quantitative variables in the abalone dataset were studied with histograms and boxplots and for categorical variable with pie chart. Further, scatter plot matrix, Pearson Correlation heatmap and side-by-side box plots were used to get insights about the relationships among the variables in the abalone dataset.

b) Preliminary model investigation

The abalone dataset was split into train and validation by 80:20. A full first order model was fitted to the abalone training dataset. From the Residuals, and the Normal Q-Q plot of the initial fit model- for nonlinearity in the data, for departure from Normality, and for non-constant error variance were checked. Also, from the Box-Cox plot, required transformation of the response variable was found. Moreover, the residuals and the Variance inflation factors (if greater than 10) indicated high multicollinearity. Interaction terms were added in the full model and Ridge regression was performed (in the end) to reduce the effect of multicollinearity. At the end of this step, required transformations, potential pool of predictor variables was obtained.

c) Model selection

The goal of the model selection was to get a good model with little bias and small variability. Full models were assumed to be correct models. To ensure a good model was chosen with optimal bias-variance tradeoff, Akaike's information criterion , AIC and Bayesian information criterion, BIC were used. Models obtained with BIC criterion are smaller (less complex) than with AIC criterion as BIC penalized the model complexity more than AIC.

The number of possible models grew as 2^{P-1}, where P-1 is the number of potential predictor variables. The full first order model alone had 9 predictor variables (including dummy variables for categorical variable). Upon including possible two-way interaction terms, the potential pool of predictor variables became very large. It is not recommended and unwise to do an exhaustive search for the best model. As a result, greedy search technique was implemented using the stepAIC function in R library MASS, starting with only intercept term in the model.

Forward stepwise procedure often works better than forward selection when there is high multicollinearity among the potential predictor variables. Backward procedures are not good when the number of potential predictor variables is large. The method was also repeated for three-way interaction terms. Finally, the best model was chosen based on AIC and BIC values, and on the complexity of the model.

d) Model diagnostics and validation

Mean-squared prediction errors were calculated on the best model for both the training and the test dataset. Cook's distance was plotted for the final model and influential outlier cases were identified. Bonferroni's procedure was used to find outliers in the response variable and were removed. From residuals versus leverage plot, outliers in predictor variables were identified and were also removed. The best model obtained using the training dataset was refit using the test dataset and then the values of R_a^2 and $MSPE_v$ were compared. Finally, the entire data was refit on the best model to give the final model.

3. Results and Discussions

The dataset contained one categorical variable, age (V1) and seven quantitative variables i.e., length (V2), diameter (V3), height (V4), whole weight (V5), shucked weight (V6), viscera weight (V7) and shell weight (V8). From the pie chart using ggplot [7] of the categorical variable, sex (figure 1), three classes with distributions of 31% male, 32% infant and the remaining as female were found. Histograms were plotted for all the seven quantitative variables i.e., V2 to V8 (figure 2). Length (V2) and diameter (V3) were left skewed, and the rest were right skewed and in particular, height (V4) was heavily right skewed (due to outliers). Further, Boxplots by sex (categorical) level were plotted (figure 3) for all the quantitative variables i.e., V2 to V8. Mean and standard deviation of both male and female were similar while for the infant was lower for each quantitative variable. Significant outliers were visible in the Boxplot of height (V4) in both male and female.

From the histogram of response variable, rings (V9), a right skewed distribution (figure 4) was identified. Also, the distribution of rings (V9) is similar for both male and female with mean around 10. This can be seen from the Boxplot of rings (V9) by categorical variable, sex (V1) level (figure 5). From the scatter plot matrix (figure 6) and Pearson Correlation heat-map (figure 7), strong correlation between the quantitative predictor variables i.e., V2 to V8 were seen and further these variables were weakly correlated with the response variable, rings. However, all the relationships were positively correlated.

A full first order model, M_1 was fit using the training dataset during the preliminary investigation. The summary statistics of the model, M_1 can be found in appendix 2. From the residuals (figure 8) and Q-Q plot (figure 9), strong evidence of non-linearity, and unequal variance were found. Also, the distribution was right skewed. Based on these observations, a linear fit of the full first order model was insufficient. Subsequently, based on the Box-cox plot (figure 10), the response variable, rings (V9), was transformed to log(V9) and the full first order model was refit (M_2, summary statistics can be found in appendix 2) again with the dataset.

Residuals (figure 11) and Q-Q plot (figure 12) showed the log transform improved the model statistics- R_a^2 improved from 0.5375 (M_1) to 0.6025 (M_2). Also, the Q-Q plot moved closer towards a normal distribution, but it is still skewed right and from the residuals versus predictor variables (figure 13) and figure 11, there was no clear evidence of nonlinearity. Scatter plot matrix (figure 14) and the variance inflation factors (table 2) indicates strong multicollinearity among the predictor variables. However, it should not result in significant impact in the fit and model metrics. It is expected in experimental dataset. Upon adding interaction terms, the multicollinearity can be addressed in the fit model.

A stepwise regression method using AIC and BIC criterions for two-way interaction terms found models M_3 and M_4 respectively, and for three-way interactions terms, M_5 and M_4 were found. Note, BIC criterion found the same model (M_4) in both cases (two-way or three-way interactions). Since the calculated model metrics (shown below) such as R_a^2 and $MSPE_v$ are not increasing significantly from two-way interaction to three-way interaction model, interactions higher than three-way were not considered and these are also practically insignificant.

Model	**Model Description**	**# regression coefficients, p**	**R_a^2**	**$MSPE_v$ value**
M_3	2-way Interaction model from AIC	22	0.6644	0.0424
M_4	2-way Interaction model from BIC	17	0.6622	0.0424
M_5	3-way Interaction model from AIC	27	0.6649	0.0430
M_4	3-way Interaction model from BIC	17	0.6622	0.0424

As expected, BIC criterion penalized the model more than the AIC criterion for both the cases (two-way and three-way interaction terms). Based on the principle of parsimony ("Occam's Razor"), model M_4 was chosen as the best model (shown below).

$V_9 \sim V_4 + V_8 + V_6 + V_5 + V_1 + V_7 + V_3 + V_2 + V_5{:}V_1 + V_4{:}V_3 + V_6{:}V_3 + V_8{:}V_5 + V_3{:}V_2 + V_6{:}V_2$

The summary statistics, residuals (figure 15), Q-Q plot (figure 16), Parity plot of training dataset (figure 17) and parity plot of test dataset (figure 18) of model M_4 are mentioned in appendix 1. MSPE values of M_4 on train dataset is 0.03476 and on test dataset is 0.04244. The values are similar which implies M_4 does not overfit the data.

Bonferroni's procedure was used to find cases 2184, 3087, and 237 as outliers in the response variable and were removed. The model M_4 was refit again and compared with the fitted values using all cases (figure 19). As the differences in figure 19 are very minimal, the above three cases (2184, 3087, and 237) were retained.

From Cook's distance (figure 20) and residuals versus leverage plot (figure 21), three outliers in predictor variables, cases 1211, 3977, and 237 were identified and removed. The model M_4 was refit again and compared with the fitted values using all cases (figure 22). As the differences in figure 22 are very minimal, the above three cases (1211, 3977, and 237) were retained.

Model M_4 was fit with the test dataset (figure 23) and MSE and R_a^2 metrics were calculated and compared (below) with that of train dataset. This again proves that model M_4 has an optimal bias-variance tradeoff.

Model M_4	**MSE**	**R_a^2**
Fit with train dataset	0.0348	0.6622
Fit with test dataset	0.0351	0.6287

Finally, the entire data was refit on the best model to give the final model M_5 (shown below).

$V_9 \sim 0.66845 + 1.48826V_4 + 2.15489V_8 - 4.65554V_6 + 1.01272V_5 - 0.21842V_{1I} - 0.03911V_{1M} - 0.56923V_7 + 4.76804V_3 + 2.63116V_2 + 0.23007V_5{:}V_{1I} + 0.03720V_5{:}V_{1M} - 2.44257V_4{:}V_3 + 3.64780V_6{:}V_3 - 0.85595V_8{:}V_5 - 9.86191V_3{:}V_2 + 2.34191V_6{:}V_2$

$R_a^2 = 0.65$; MSE = 0.036

Complete model statistics, ANOVA table, residuals (figure 24), Parity plot of fitted and true values (figure 25) are mentioned in appendix 1.

As the data had high multicollinearity, ridge regression was performed additionally on the train dataset. The tuning parameter, λ was found to be 0.4418. Parity plot of train and test dataset of ridge regressed model (figure 26) was compared with the parity plot of the OLS model M_4 (figures 17 and 18) and found to be similar. $MSPE_v$ of the ridged regressed model on the test dataset was found to be 0.0457. This value is very similar to the $MSPE_v$ of M_4 model. Thus, even though multicollinearity is high, there is a lot of data hence both the fits for interaction model and ridge regressed model are fairly close and both methods help to decrease this effect.

Conclusions

As discussed in the above sections, firstly exploratory data analysis was carried out on the entire dataset. Subsequently, a preliminary first order full model was fit on the training dataset and from the Box-Cox procedure, the response variable, rings (V9) was transformed into log(V9). High multicollinearity was observed in the model and hence two-way and three-way interaction terms were considered in the model selection procedure during stepwise regression. Best model, M_4 was found to be two-way interaction model from the BIC criterion. Influential outliers were found both in predictor variables and the response variable and were removed. However, the fit model with all cases except the outliers weren't much different from the model fit with all cases. Thus, the influential cases were retained. Final model, M_5 was obtained from model M_4 but using the entire abalone dataset. Further, Ridge regression was performed, and the model obtained was similar to that of OLS model M_4. Thus, both the methods help in decreasing the multicollinearity.

The final model has metrics of $R_a^2 = 0.65$ and MSE = 0.036. To improve the model further, non-linear regression can be considered or even a Neural network (data driven model) approach can be considered as the abalone dataset is considerably large. If prior domain knowledge is available, then it needs to be incorporated into the model building.

Appendices

Appendix 1 contains important figures and tables which are referenced in the report.

Appendix 2 (attached separately) contains .Rmd file used for the analysis of the abalone dataset.

a. Appendix 1

Categorical Variable

Sex	Count
Female	1307
Infant	1342
Male	1528
Total	4177

Quantitative Variables

Column	Minimum	Maximum	Mean	Median
Length (V2)	0.075	0.815	0.524	0.545
Diameter (V3)	0.0550	0.6500	0.4079	0.425
Height (V4)	0.0000	1.1300	0.1395	0.1400
Whole weight (V5)	0.0020	2.8255	0.8287	0.7995
Shucked weight (V6)	0.0010	1.4880	0.3594	0.3360
Viscera weight (V7)	0.0005	0.7600	0.1806	0.1710
Shell weight (V8)	0.0015	1.0050	0.2388	0.2340
Rings (V9)	1.000	29.000	9.934	9.000

Table 1: Summary statistics of Abalone dataset

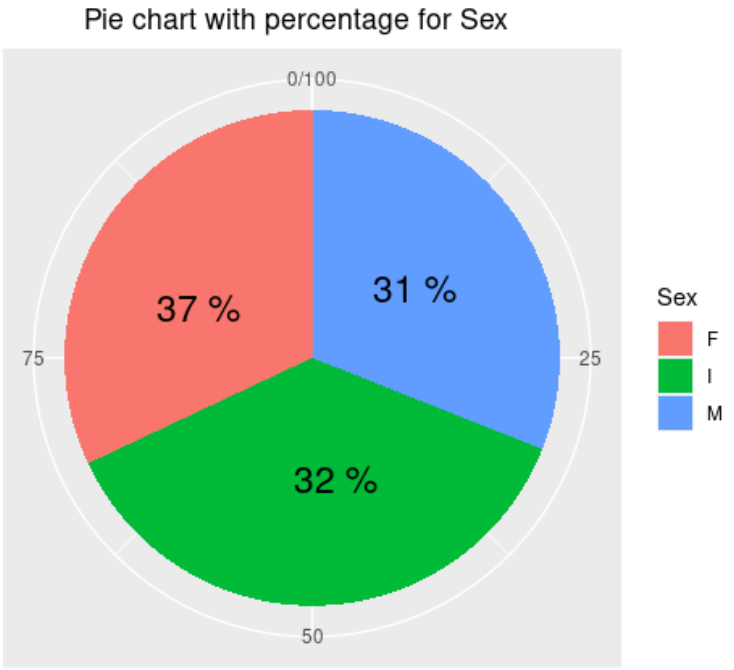

Figure 1: Pie chart for categorical variable, sex

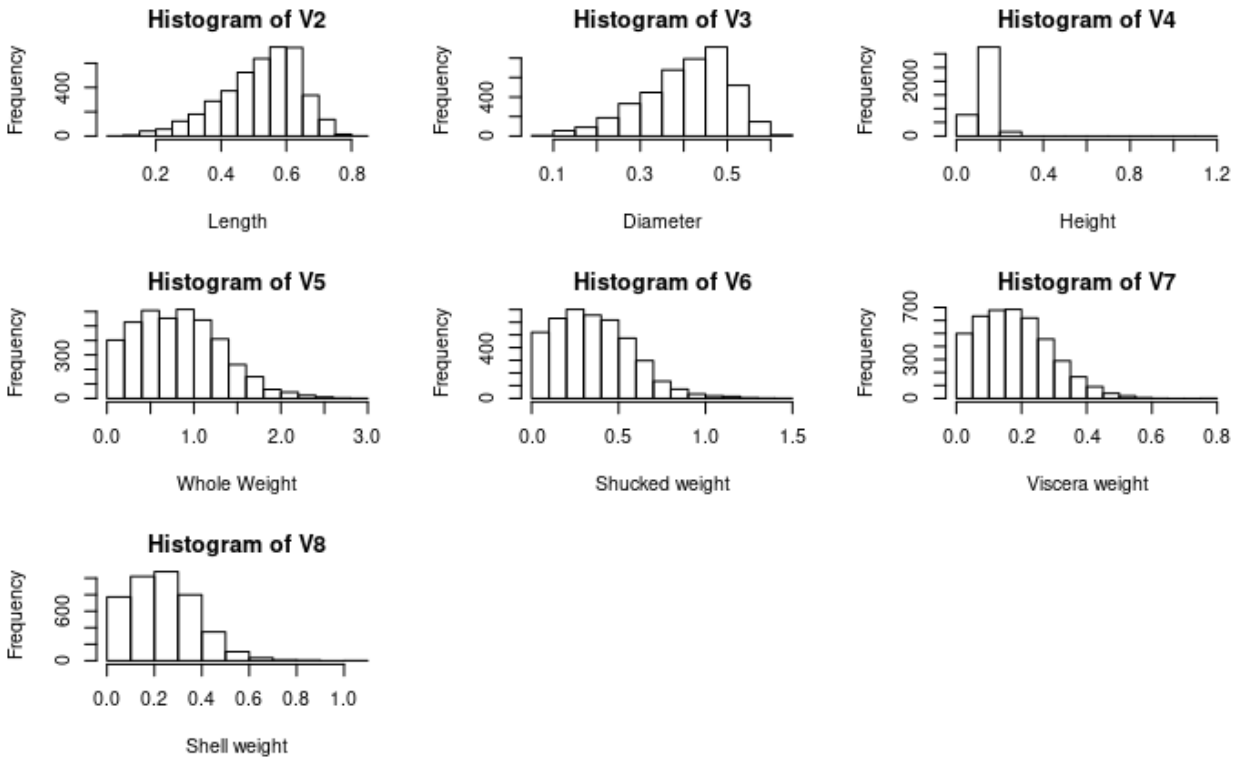

Figure 2: Histograms of quantitative variables

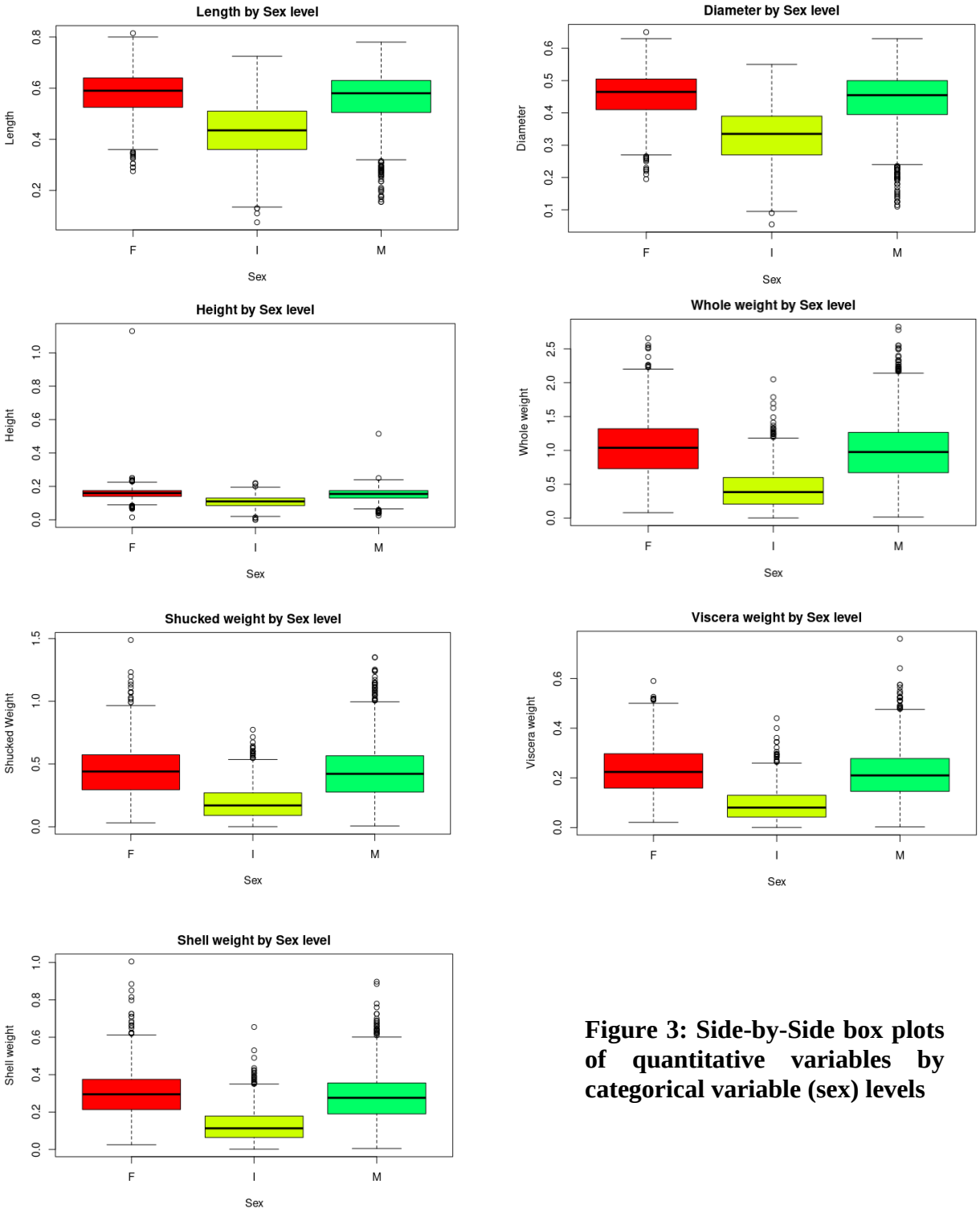

Figure 3: Side-by-Side box plots of quantitative variables by categorical variable (sex) levels

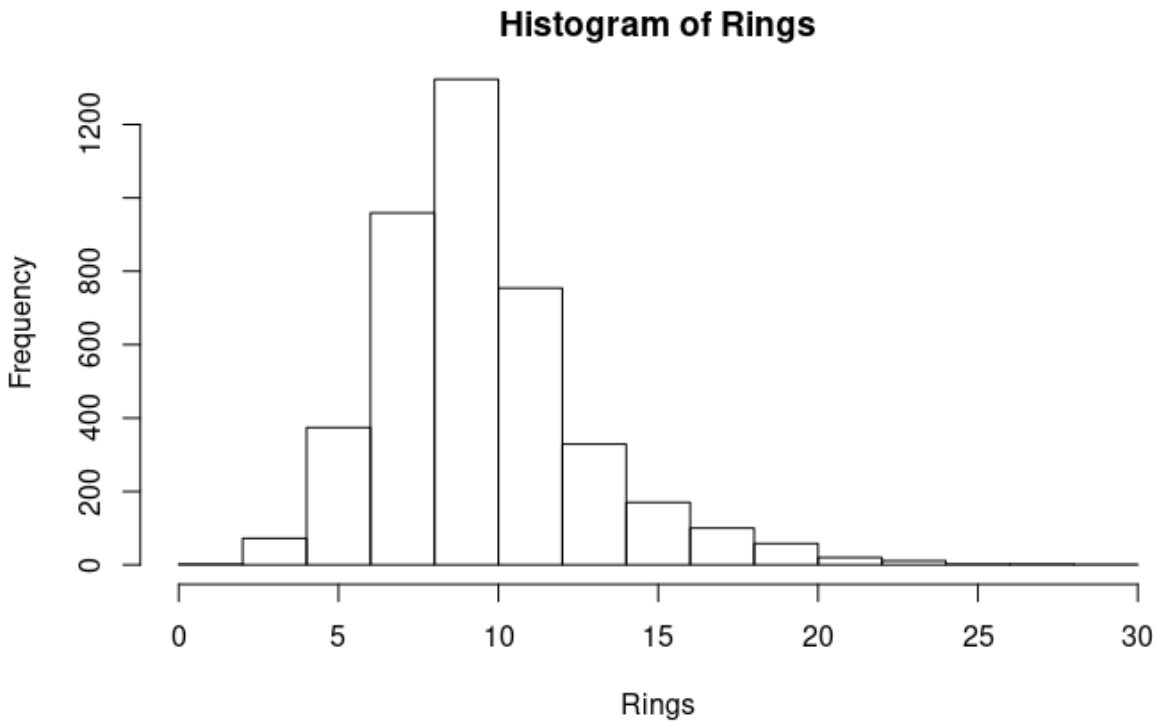

Figure 4: Histogram of Response variable, rings

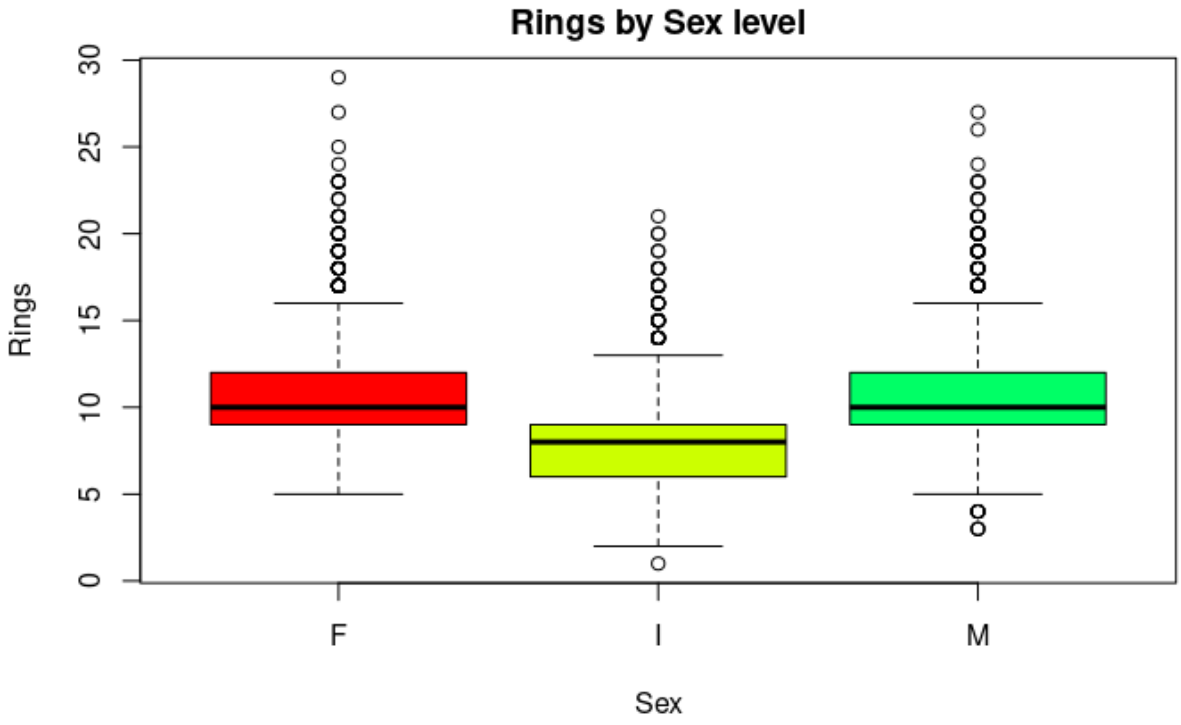

Figure 5: Box plot of response variable, rings by categorical variable (sex) levels

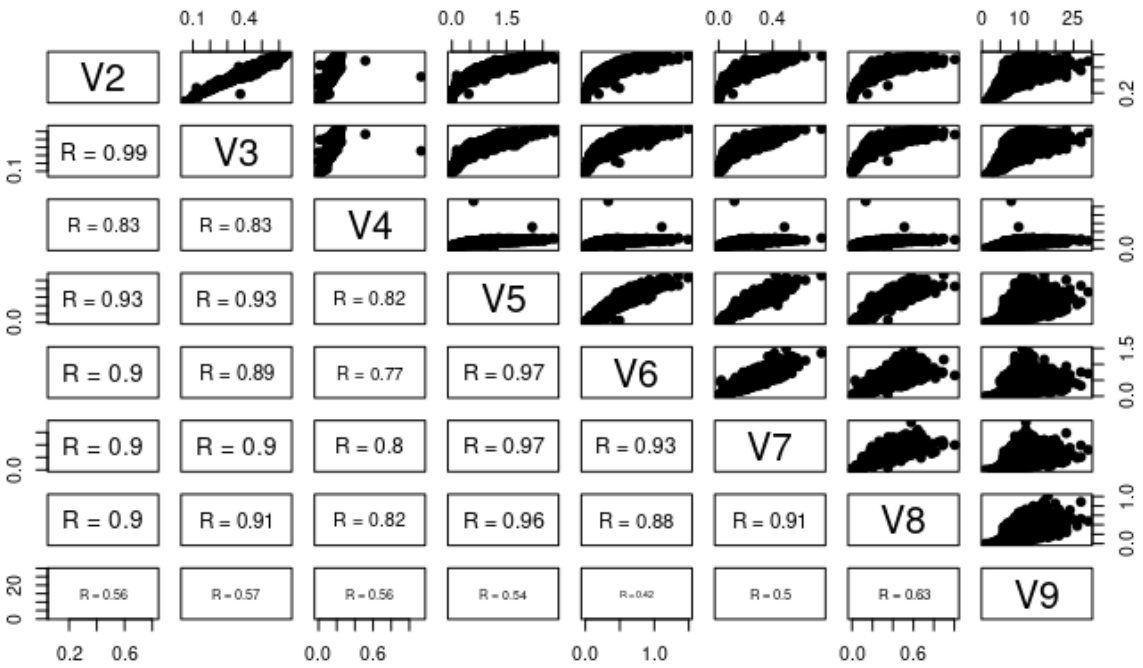

Figure 6: Scatter plot matrix of Abalone dataset

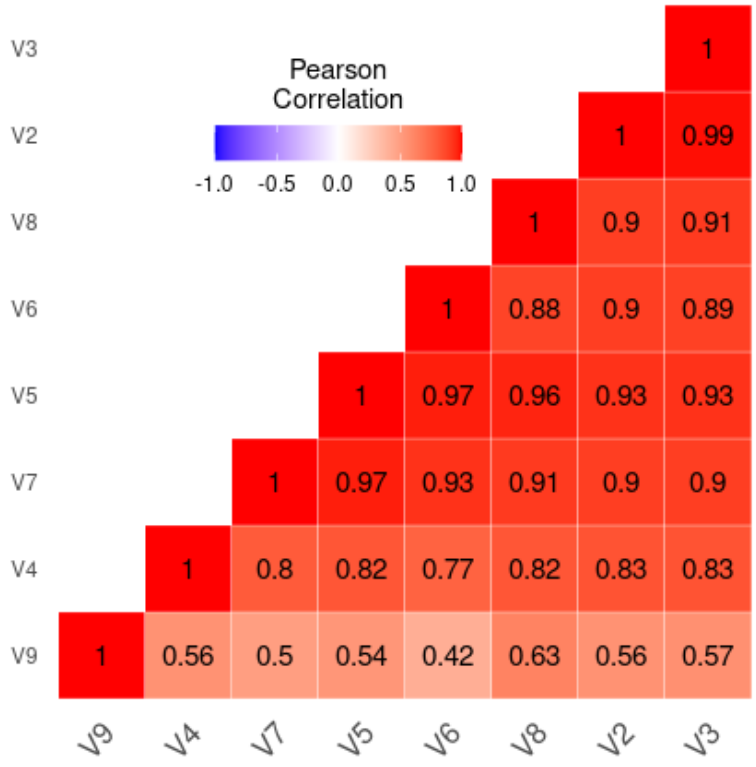

Figure 7: Heatmap of Pearson Correlation coefficients

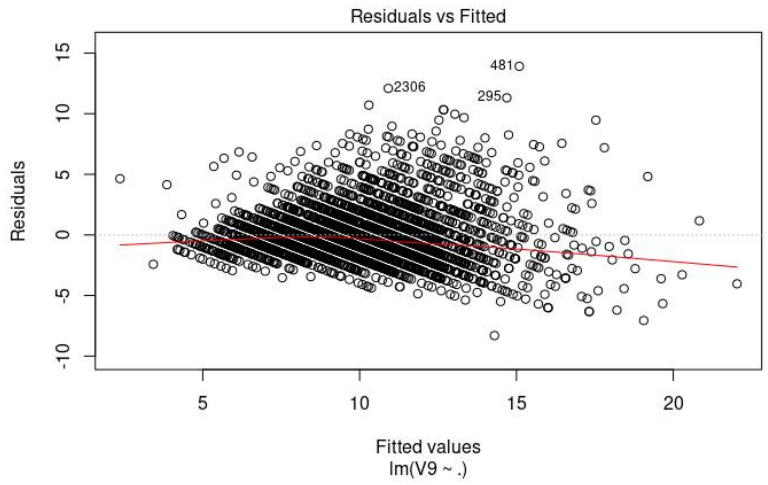

Figure 8: Residuals vs Fitted values plot of full first order model, M_1

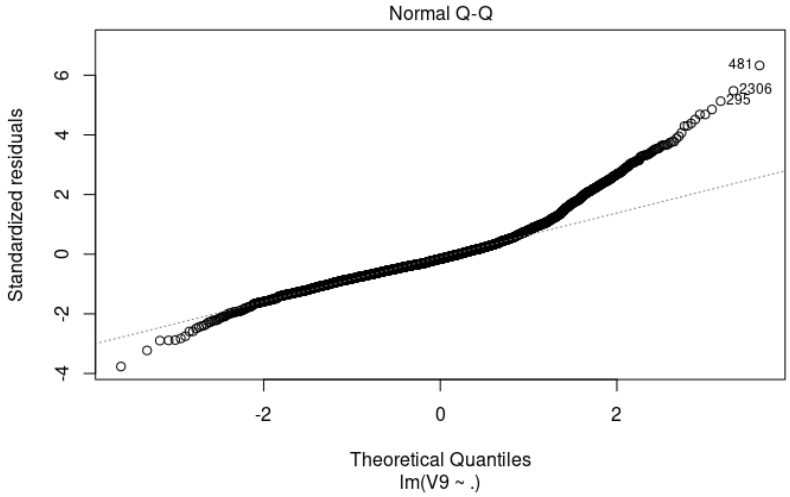

Figure 9: Normal Q-Q plot of full first order model, M_1

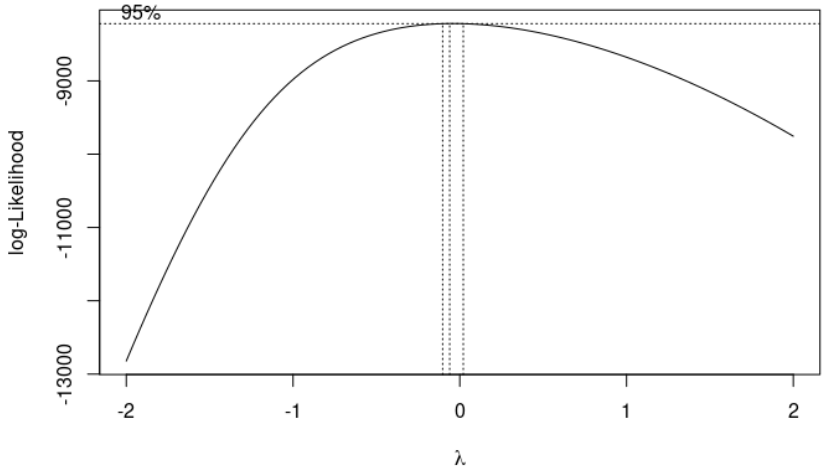

Figure 10: Box-Cox plot of full first order model, M_1 (suggests log transformation)

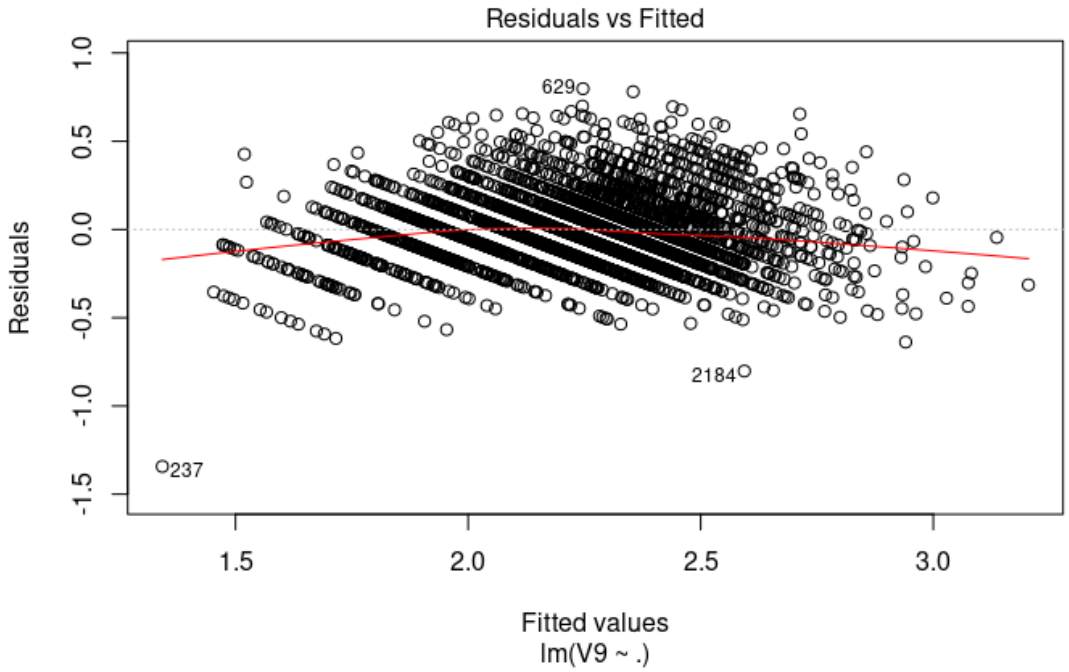

Figure 11: Residuals vs Fitted values plot of full first order model, M_2 (with log transform)

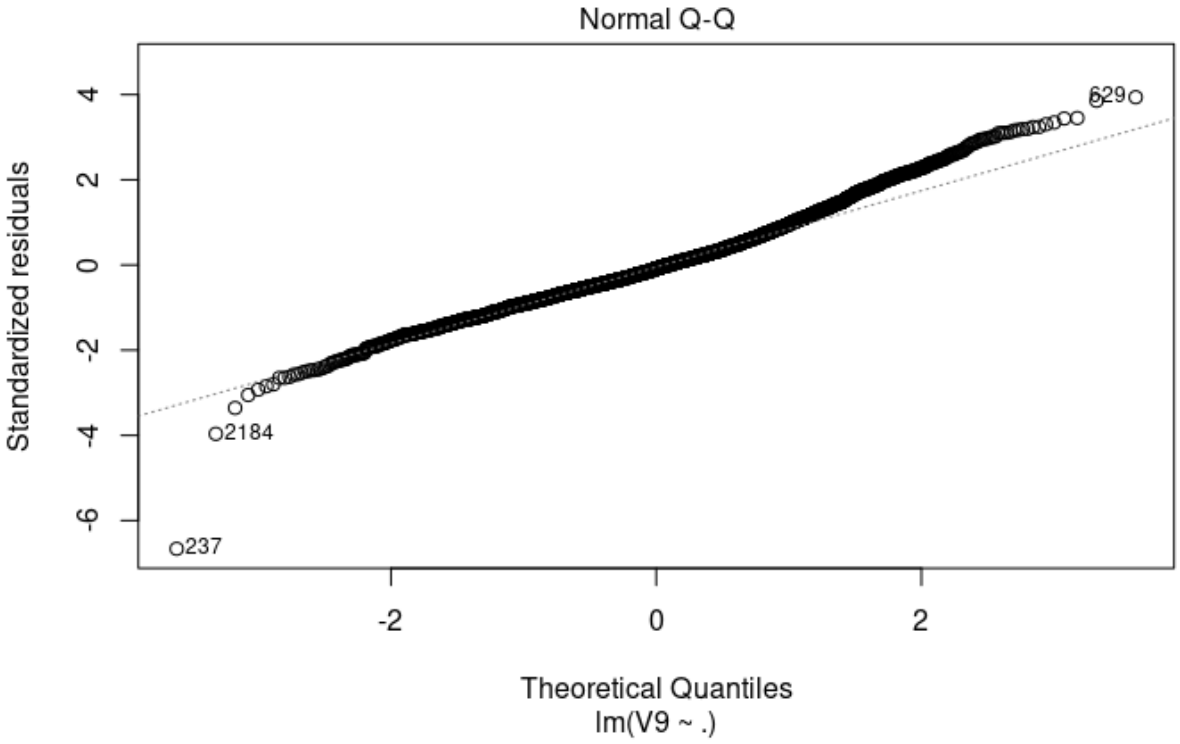

Figure 12: Normal Q-Q plot of full first order model, M_2 (with log transform)

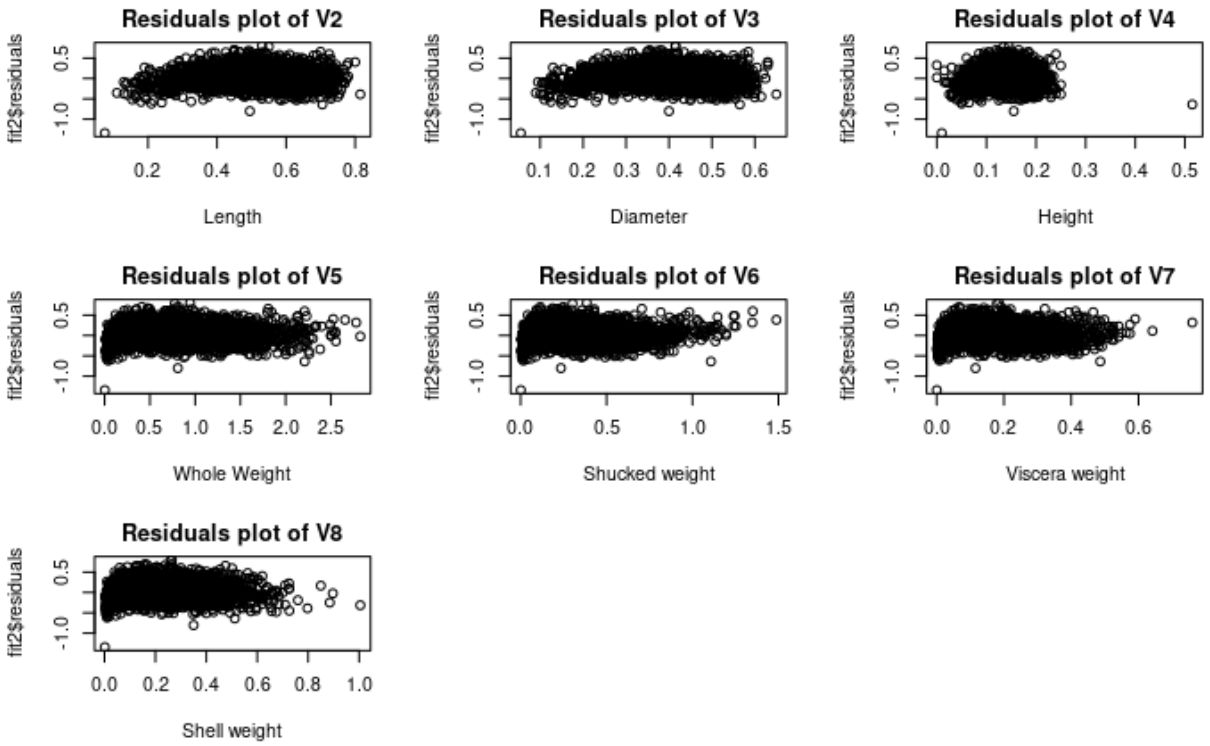

Figure 13: Residuals of model M_2 plot vs all quantitative predictors

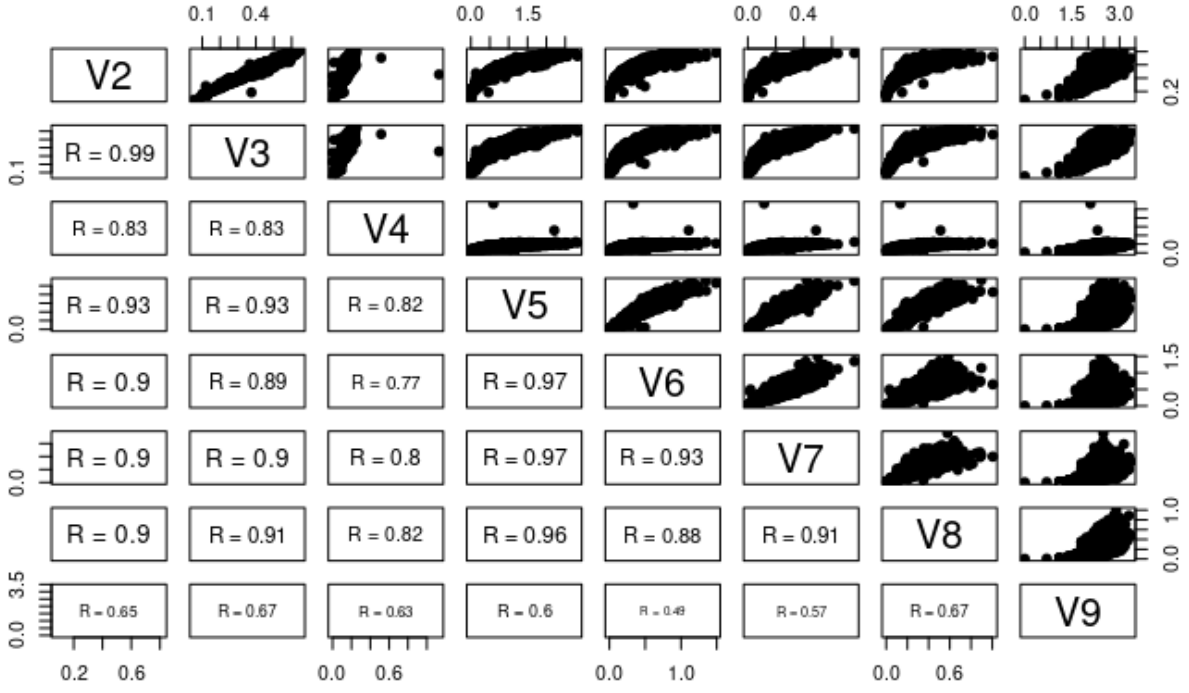

Figure 14: Scatter plot matrix of model M_2 (after log transformation in rings)

V2	V3	V4	V5	V6	V7	V8	V9
40.8281	**42.5038**	**3.6327**	**111.9135**	**31.7346**	**17.5065**	**21.4165**	**2.4123**

Table 2: Variance Inflation Factors of model M_2

Model M_4 (best model) summary statistics

Call:

lm(formula = V9 ~ V4 + V8 + V6 + V5 + V1 + V7 + V3 + V2 + V5:V1 +

V4:V3 + V6:V3 + V8:V5 + V3:V2 + V6:V2, data = abalone.train_transformed)

Residuals:

Min	1Q	Median	3Q	Max
-1.03094	-0.12154	-0.01399	0.10249	0.81871

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.53385	0.06822	7.825	6.75e-15	***
V4	4.68779	1.01834	4.603	4.31e-06	***
V8	2.19866	0.16745	13.130	< 2e-16	***
V6	-5.77126	0.28101	-20.538	< 2e-16	***
V5	1.19510	0.08356	14.302	< 2e-16	***
V1I	-0.21984	0.02253	-9.758	< 2e-16	***
V1M	-0.04272	0.01983	-2.154	0.031273	*
V7	-0.69395	0.12642	-5.489	4.34e-08	***
V3	5.05623	0.47913	10.553	< 2e-16	***
V2	2.43785	0.43018	5.667	1.58e-08	***
V5:V1I	0.24210	0.03005	8.056	1.09e-15	***
V5:V1M	0.03902	0.01762	2.215	0.026823	*
V4:V3	-8.25983	2.21539	-3.728	0.000196	***
V6:V3	4.99473	0.96972	5.151	2.75e-07	***
V8:V5	-1.05564	0.09395	-11.236	< 2e-16	***

V3:V2 -9.83030 0.82805 -11.872 < 2e-16 ***

V6:V2 2.84852 0.76934 3.703 0.000217 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1869 on 3324 degrees of freedom

Multiple R-squared: 0.6638, Adjusted R-squared: 0.6622

F-statistic: 410.2 on 16 and 3324 DF, p-value: < 2.2e-16

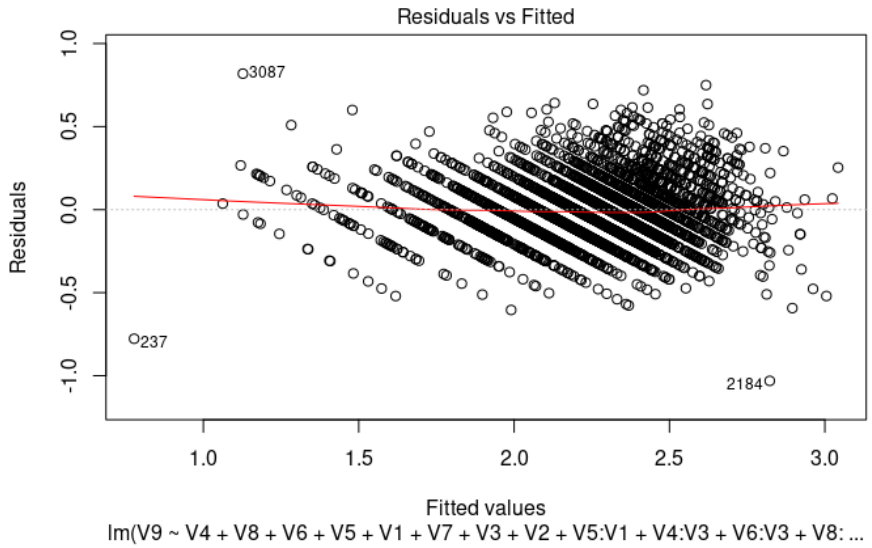

Figure 15: Residuals vs Fitted values of Best Model

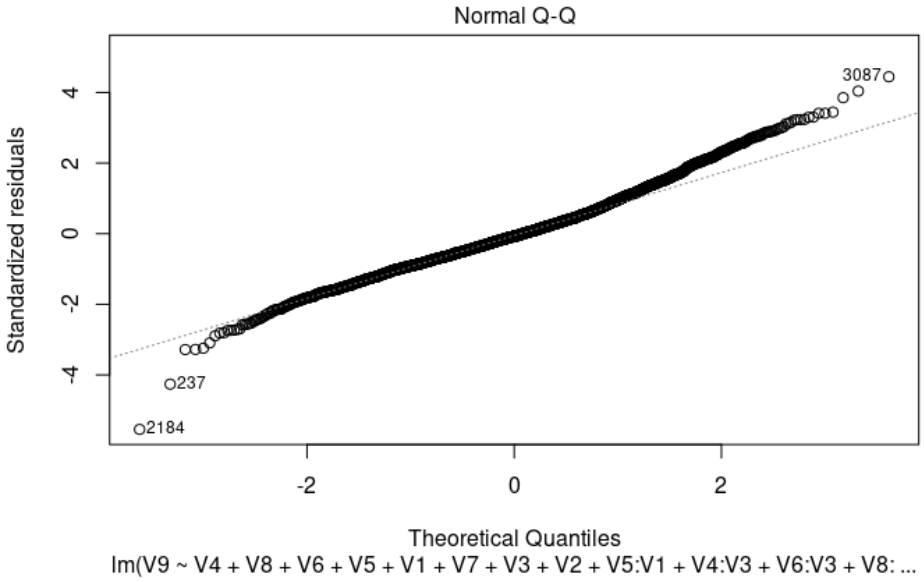

Figure 16: Normal Q-Q plot of Best Model

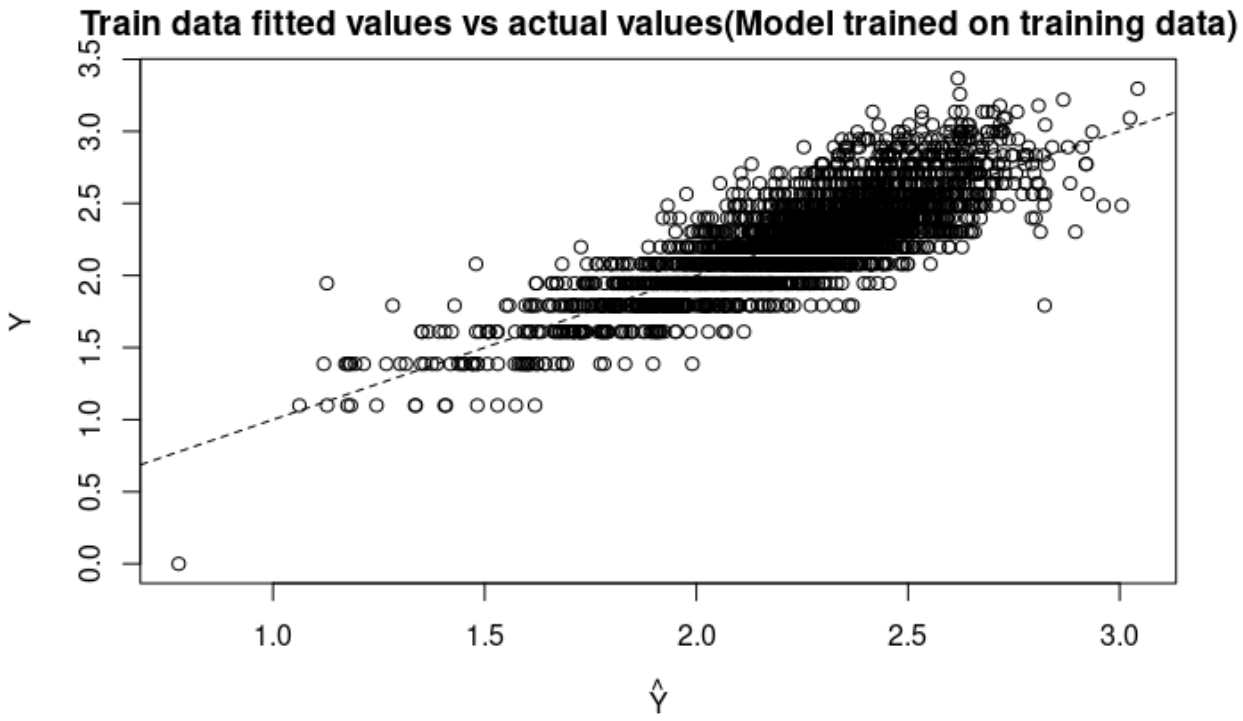

Figure 17: Parity plot of Best Model on train dataset

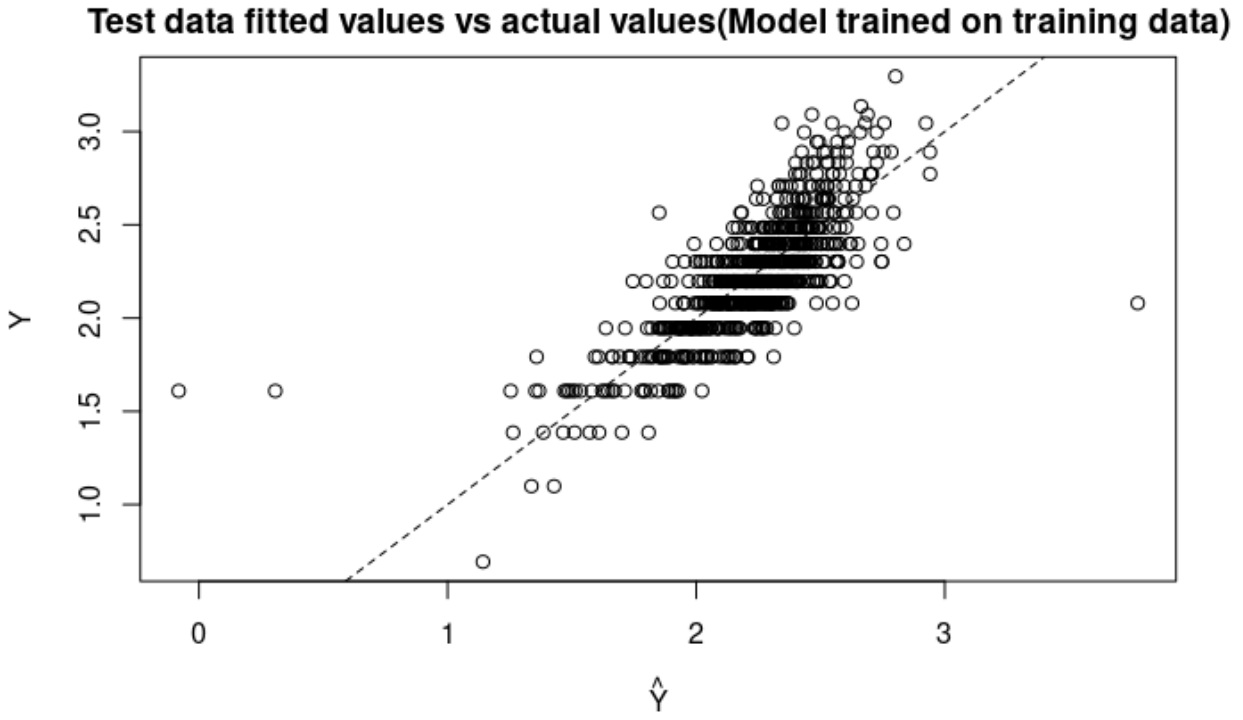

Figure 18: Parity plot of Best Model on test dataset

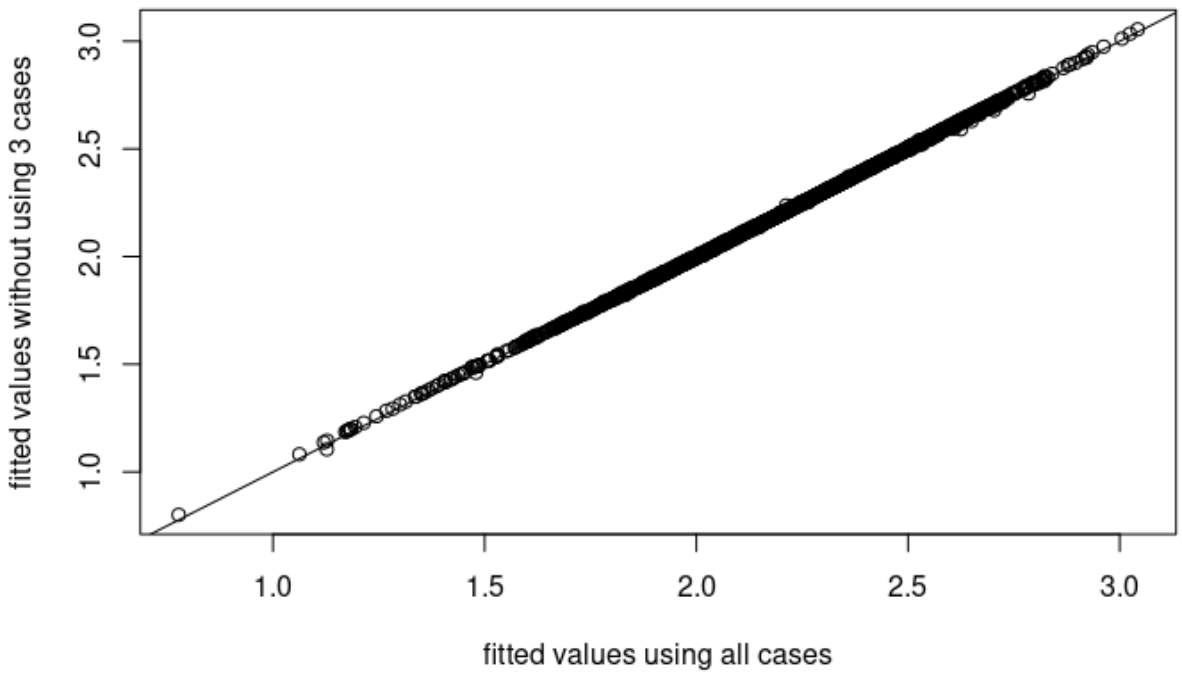

Figure 19: Fitted values without cases 2184, 3087, and 237 vs fitted values for all cases

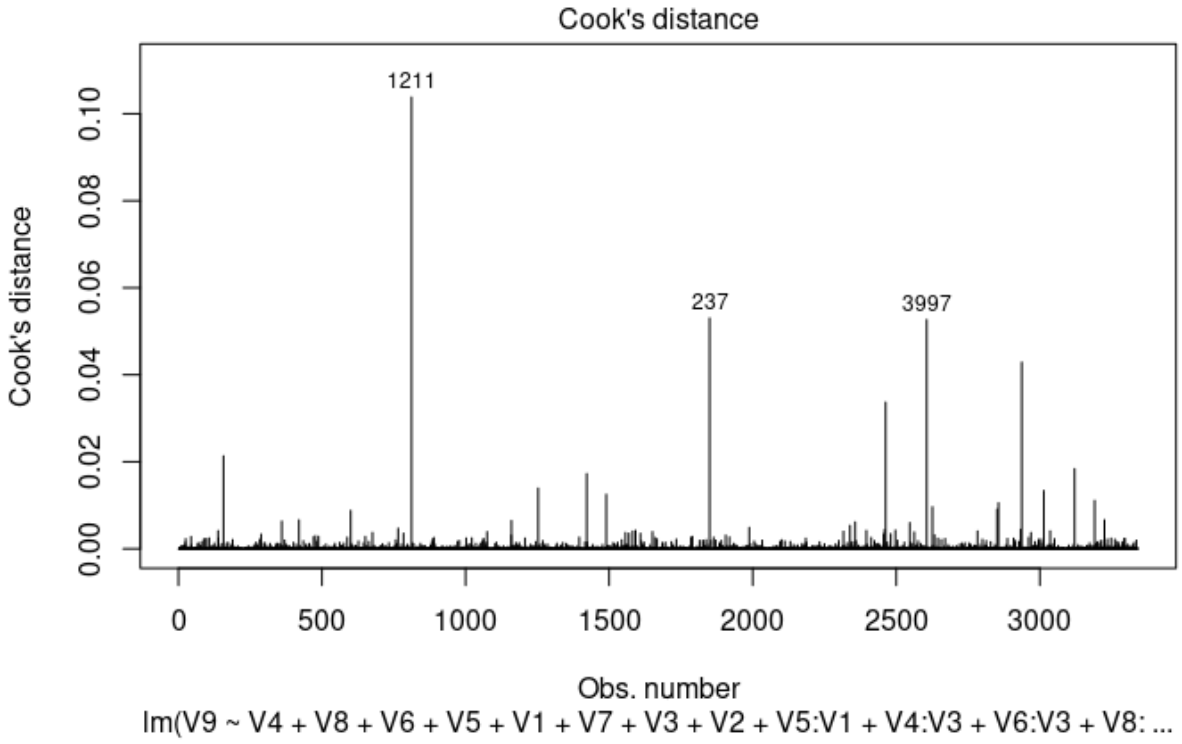

Figure 20: Cook's distance plot of Best Model (to identify influential outlier cases)

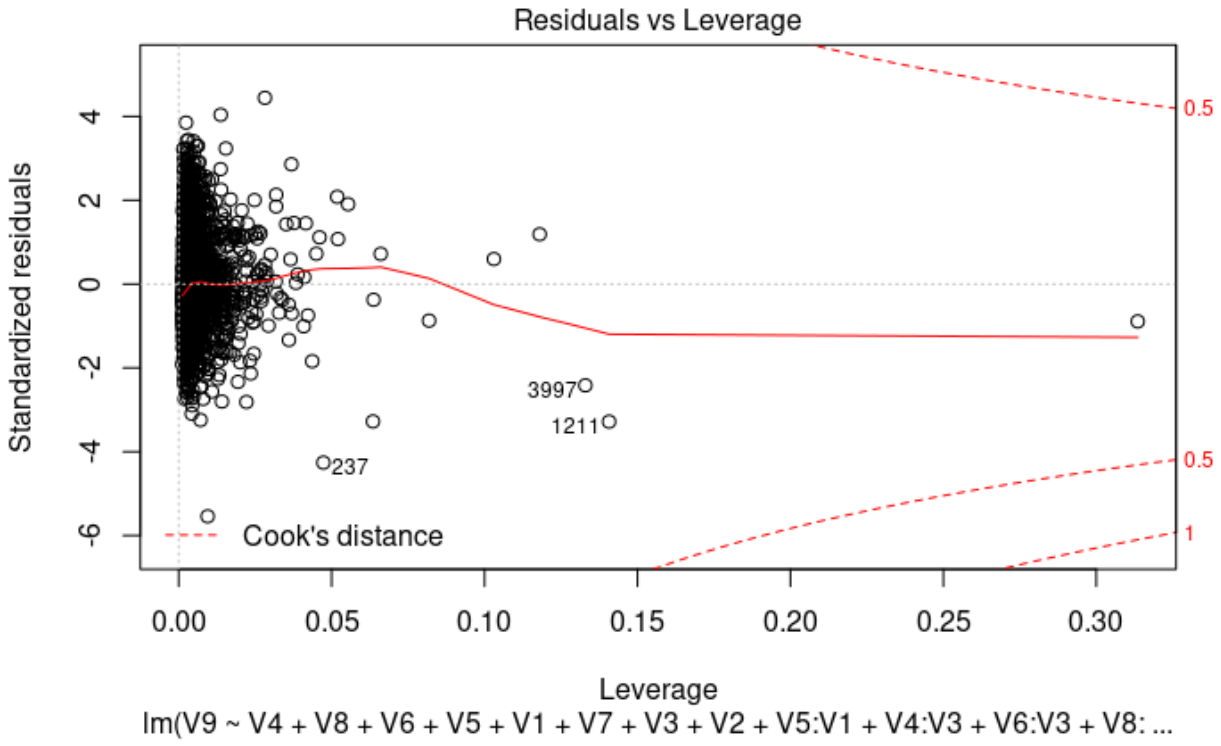

Figure 21: Residuals vs Leverage plot of Best Model (to identify influential outlier cases)

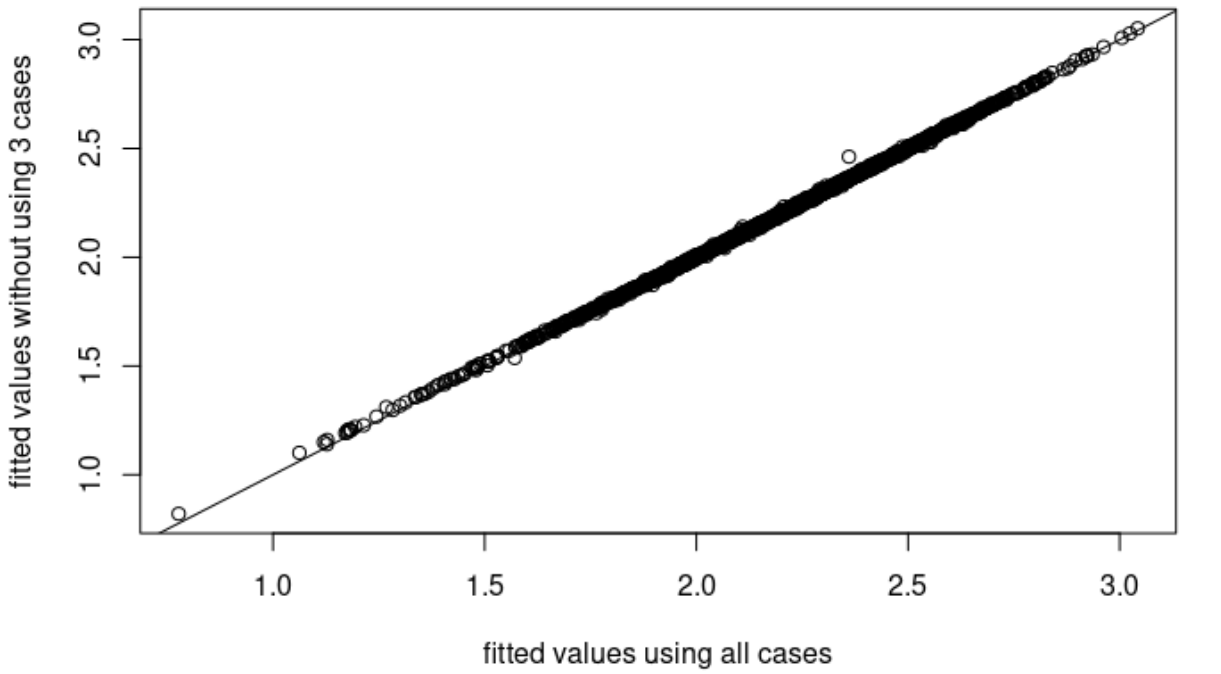

Figure 22: Fitted values without cases 1211, 3977, and 237 vs fitted values for all cases

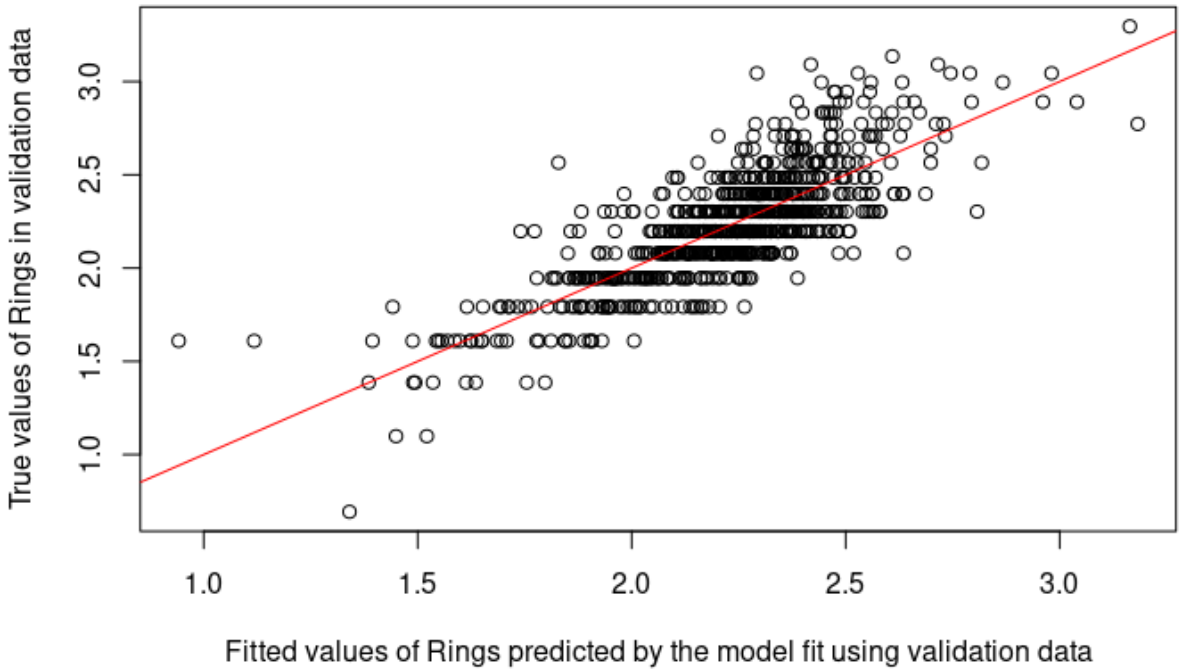

Figure 23: Parity plot of best model fitted with test dataset

Model M_5 (final model) summary statistics

Call:

lm(formula = V9 ~ V4 + V8 + V6 + V5 + V1 + V7 + V3 + V2 + V5:V1 +
V4:V3 + V6:V3 + V8:V5 + V3:V2 + V6:V2, data = abalone.transformed)

Residuals:

Min	1Q	Median	3Q	Max
-0.99121	-0.12282	-0.01488	0.10294	1.37543

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.66845	0.06061	11.028	< 2e-16	***
V4	1.48826	0.68195	2.182	0.02914	*
V8	2.15489	0.15083	14.287	< 2e-16	***
V6	-4.65554	0.23200	-20.067	< 2e-16	***

V5 1.01272 0.07126 14.211 < 2e-16 ***

V1I -0.21842 0.02024 -10.789 < 2e-16 ***

V1M -0.03911 0.01803 -2.169 0.03011 *

V7 -0.56923 0.11348 -5.016 5.49e-07 ***

V3 4.76804 0.44324 10.757 < 2e-16 ***

V2 2.63116 0.37284 7.057 1.98e-12 ***

V5:V1I 0.23007 0.02698 8.529 < 2e-16 ***

V5:V1M 0.03720 0.01616 2.303 0.02134 *

V4:V3 -2.44257 1.68342 -1.451 0.14687

V6:V3 3.64780 0.89925 4.057 5.07e-05 ***

V8:V5 -0.85595 0.08255 -10.369 < 2e-16 ***

V3:V2 -9.86191 0.64155 -15.372 < 2e-16 ***

V6:V2 2.34191 0.71468 3.277 0.00106 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.189 on 4160 degrees of freedom

Multiple R-squared: 0.6513, Adjusted R-squared: 0.65

F-statistic: 485.7 on 16 and 4160 DF, p-value: < 2.2e-16

Model M_5 (final model) ANOVA

Response: V9

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
V4	1	166.702	166.702	4665.311	< 2.2e-16 ***
V8	1	30.988	30.988	867.239	< 2.2e-16 ***
V6	1	24.291	24.291	679.800	< 2.2e-16 ***
V5	1	4.156	4.156	116.317	< 2.2e-16 ***
V1	2	8.022	4.011	112.247	< 2.2e-16 ***
V7	1	0.904	0.904	25.303	5.105e-07 ***

V3 1 19.917 19.917 557.391 < 2.2e-16 ***

V2 1 0.418 0.418 11.697 0.0006322 ***

V5:V1 2 6.212 3.106 86.927 < 2.2e-16 ***

V4:V3 1 5.910 5.910 165.393 < 2.2e-16 ***

V6:V3 1 0.724 0.724 20.259 6.950e-06 ***

V8:V5 1 0.993 0.993 27.777 1.430e-07 ***

V3:V2 1 8.068 8.068 225.782 < 2.2e-16 ***

V6:V2 1 0.384 0.384 10.738 0.0010584 **

Residuals 4160 148.647 0.036

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

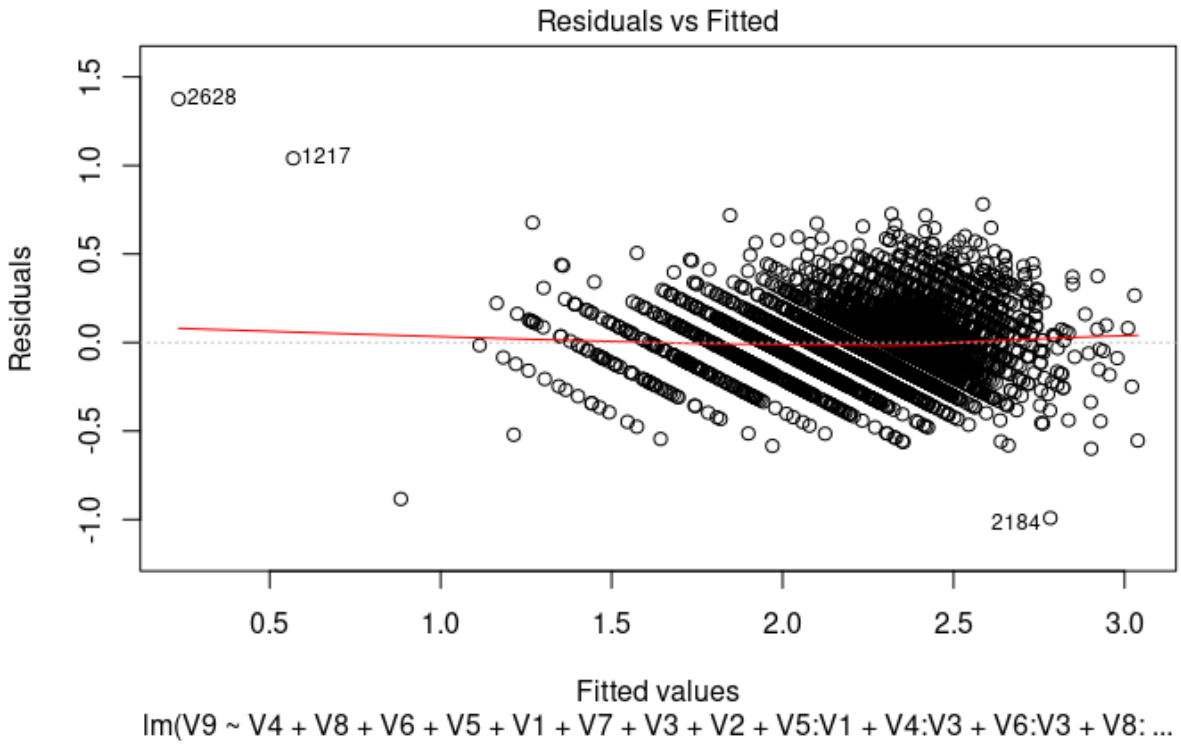

Figure 24: Residuals plot of best model fitted on entire abalone dataset

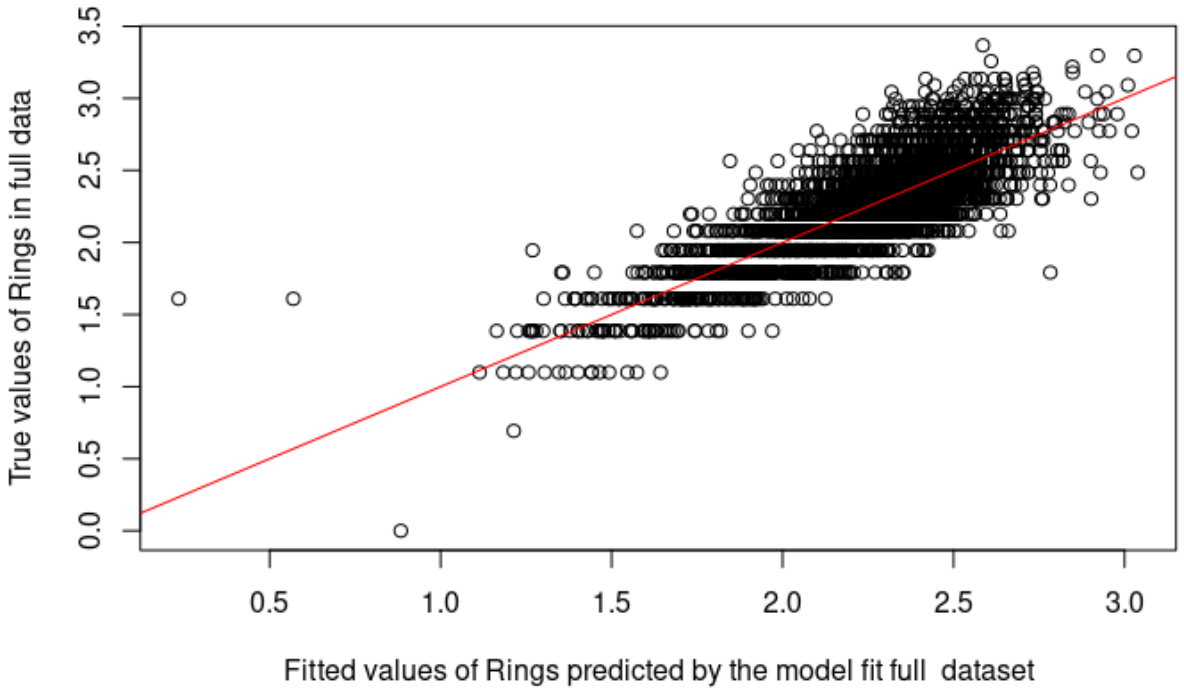

Figure 25: Parity plot of best model fitted on entire abalone dataset

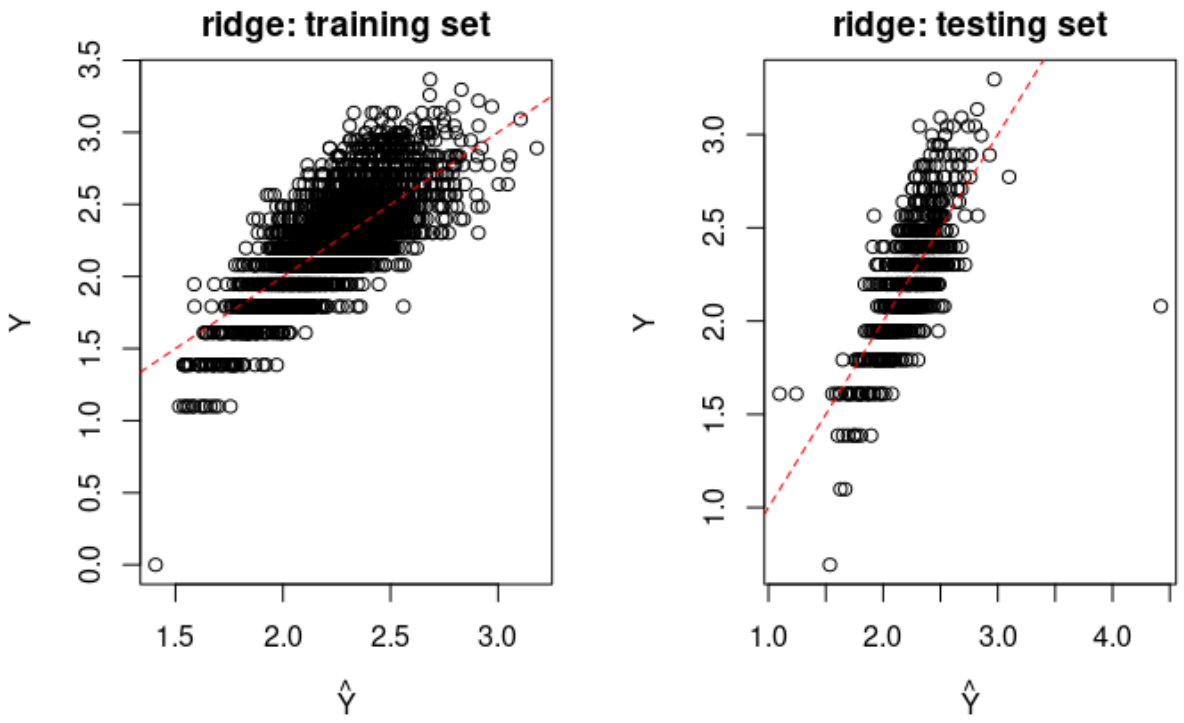

Figure 26: Ridge regressed model on train and test dataset

References

1. http://archive.ics.uci.edu/ml/datasets/Abalone?pagewanted=all
2. Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)
3. McShane, P.E. and M.G. Smith 1992. Shell growth checks are unreliable indicators of age of the abalone Haliotis rubra (Mollusca : Gastropoda). Australian Journal of Marine and Freshwater Research 43: 1215- 1219.
4. https://scg.sdsu.edu/linear-regression-in-r-abalone-dataset/
5. Applied Linear Statistical Models by Michael Kutner, Christopher Nachtsheim, John Neter, and William Li
6. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
7. https://ggplot2.tidyverse.org/reference/ggplot.html