Predicting The Age of Abalone

Simon Atoyire and Viren Gadkari

December 06, 2023

Introduction

Data Description

Determining the age of an abalone can be a very time-consuming task. It involves cutting the shell through the cone, staining it, and counting the number of rings through a microscope. The age of an abalone is determined by then taking the number of rings and adding 1.5. This process as it is can be something researchers would wish automate. Thus, we consider the task of building a predictive model that can take features of an abalone, and attempt to predict the age with high

accuracy. The following report explores this problem, with a description of our initial exploratory data analysis, the proposed full model, and the models that are generated from model selection procedures. In the end we determine a final model to aid researchers in predicting the age of a model. The dataset that we used can be found at kaggle.com, which consists of 4177 abalones with 8 different features measured for each of them. The 8 predictors in the original dataset are the Sex, Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, Shell Weight. The response variable is Rings, however, for the purposes of interpretability, we create a variable “Age” which is the Rings+1.5, and we treat this as our new response.

Clarification of Original Predictors

We start by clarifying the context of some of the predictors:

Sex - M for male, F for female, and I for infant

Length - The longest shell measurement, in millimeters

Diameter - The shell measurement perpendicular to the length, in millimeters

Height - The height with meat in the shell, in millimeters

Whole Weight - The weight of the whole abalone in grams

Shucked Weight - The weight of the meat in grams

Viscera Weight - The gut weight after bleeding in grams

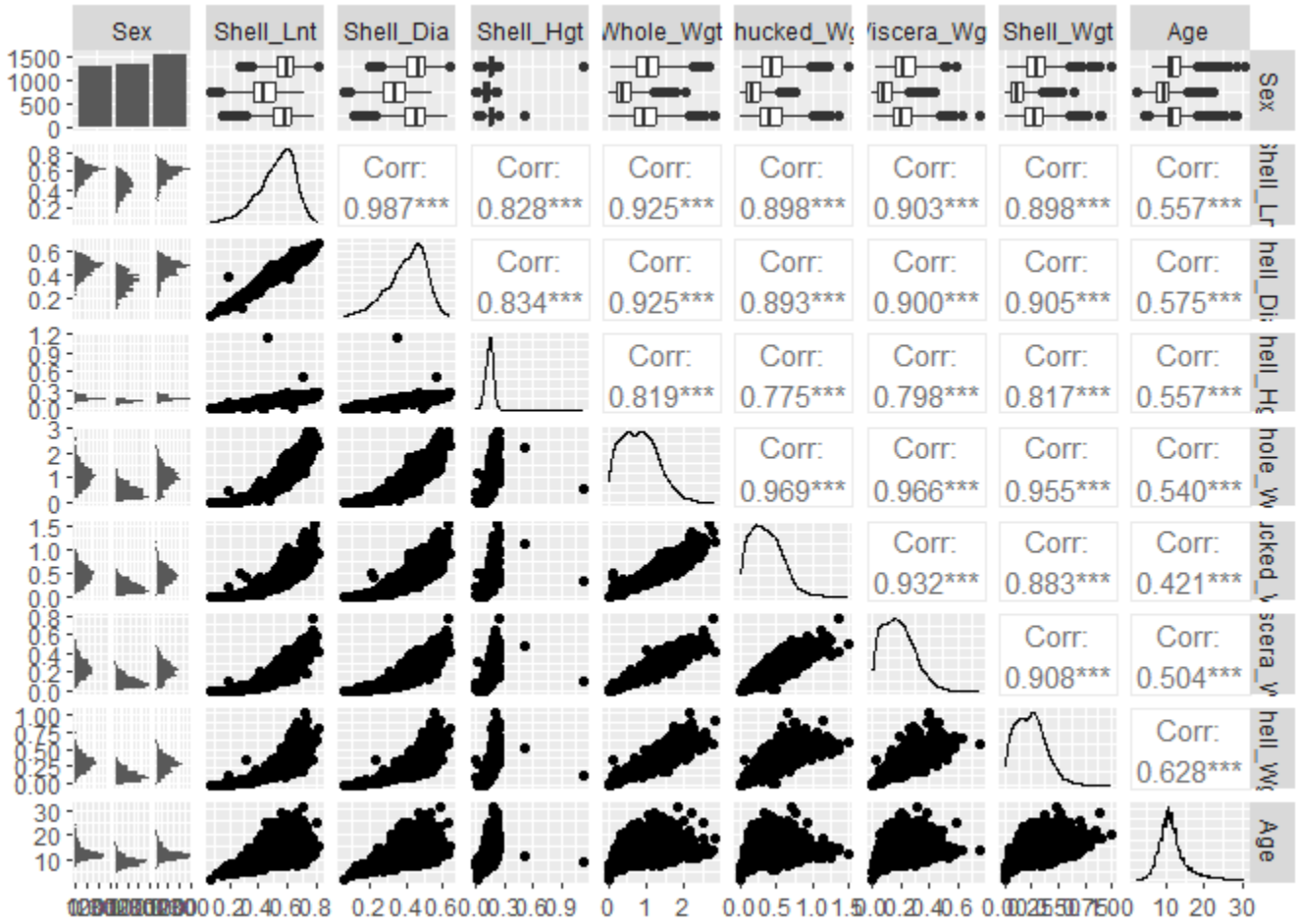
Shell Weight - The shell weight after being dried in grams

Just from the description of these predictors, we believe that these set of predictors could be highly correlated to each other, and could introduce some multicollinearity into a future model. We will verify this with our exploratory data analysis in the next section.

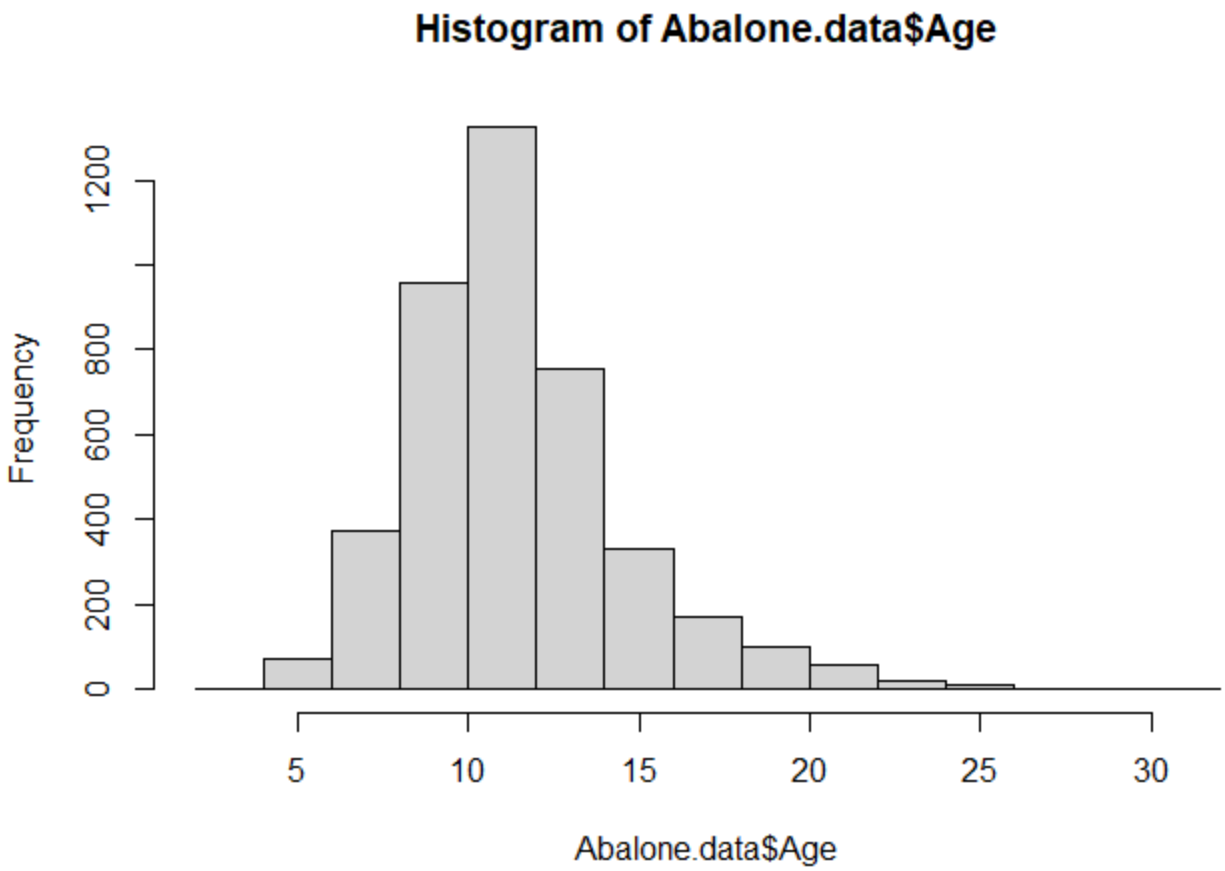
EXPLORATORY DATA ANALYSIS

Exploring the Raw Data

```
## [1] 4177   10
```



The scatter plot matrix above shows the relationship between all the study variables. It is observed that all the predictor variables and the response variables. More importantly, there appears to be high correlation among all the predictor variables which suggests multicollinearity. Thus there is the need to deal with this issue to obtain a robust predictive model. The response variable (Age) is also found to be right skewed and this is more evident in the histogram below. The skewness of the response variable and the issue of multicollinearity requires that the data is transformed to overcome these issues,

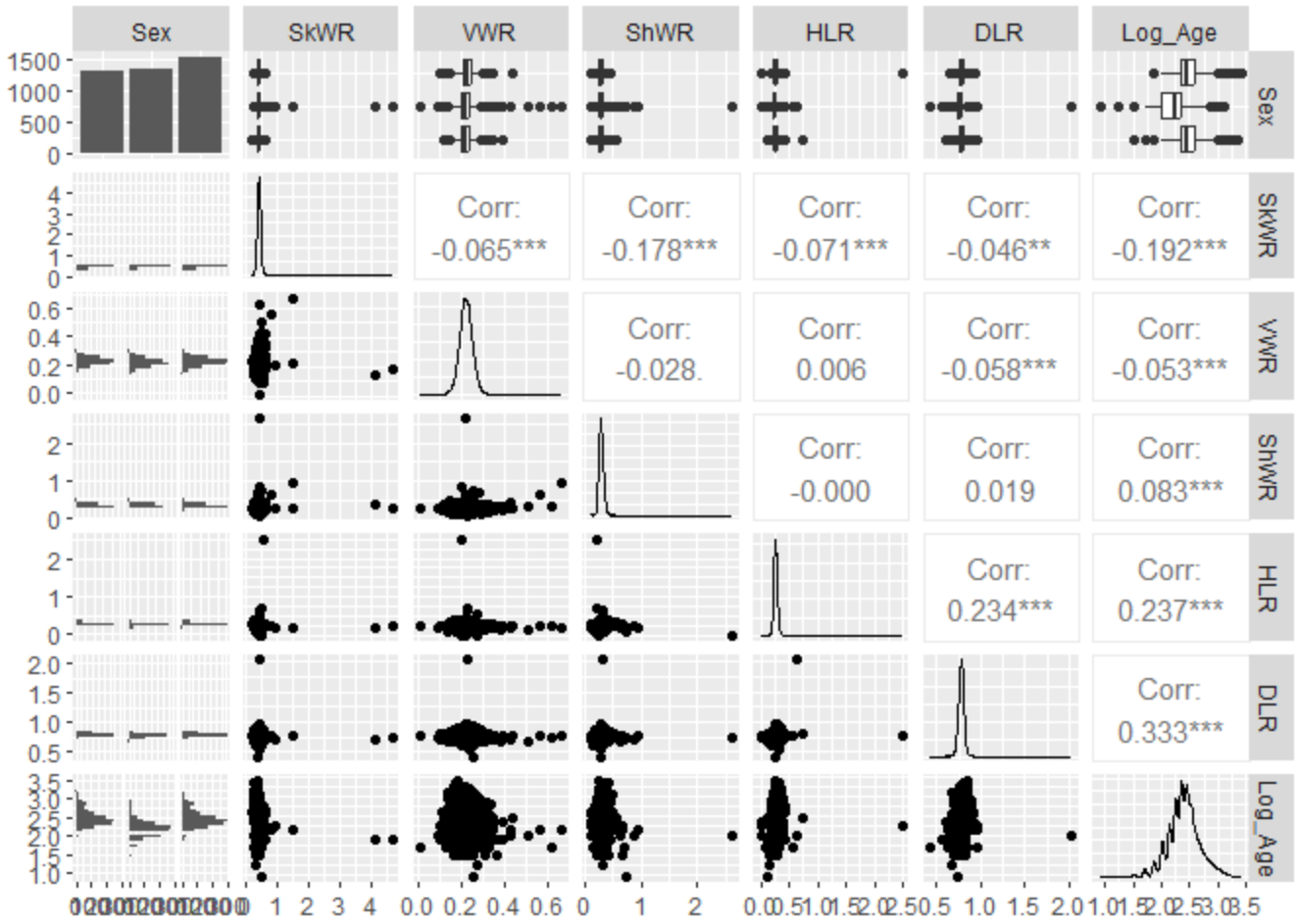


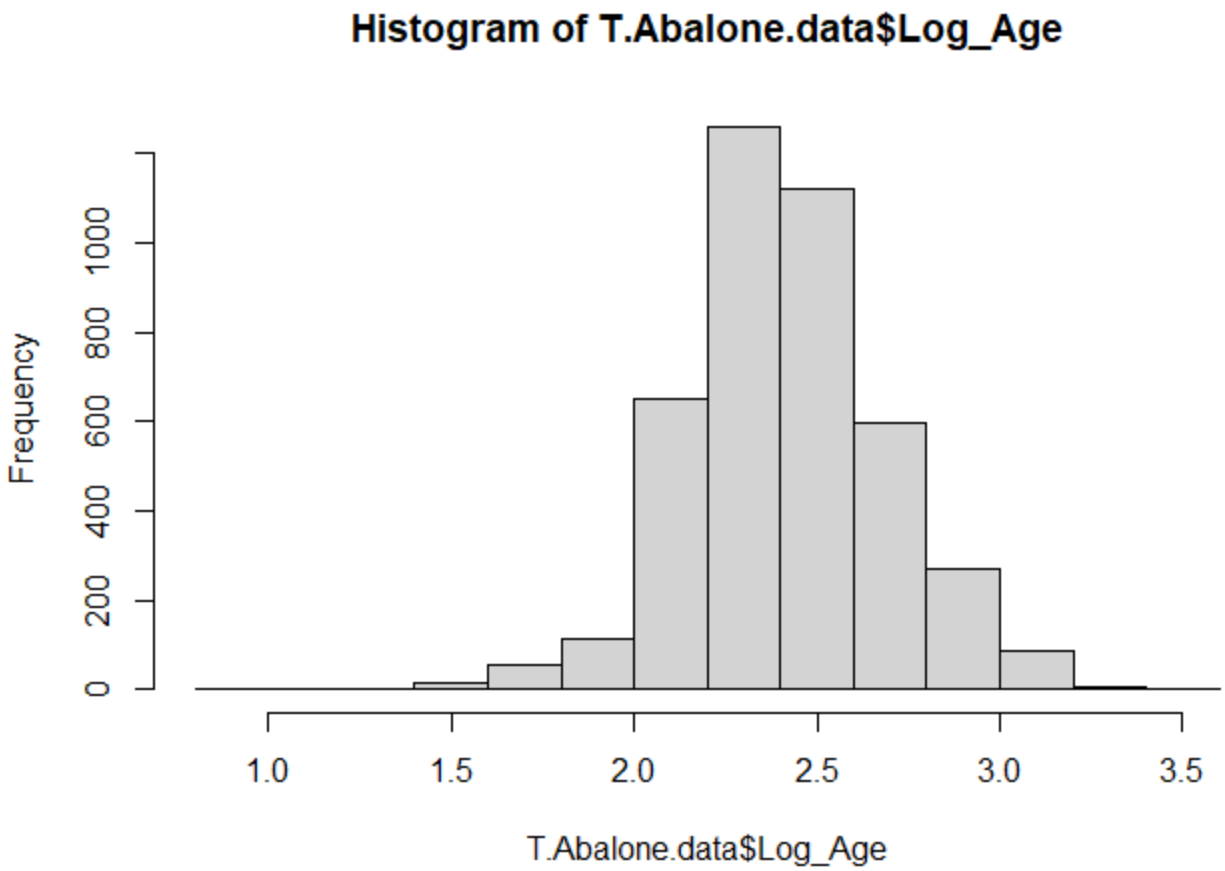
Addressing Multicollinearity and Normality Through Data Transformation.

To deal with the skewness in the response variable, the Age variable was transformed to log(Age) and this was found to be approximately normal. Also, all eight predictor variables were also transformed by standardizing then to remove the correlation between them. Doing so, six predictors were obtained including 'Sex'and five new predictors computed below:

\(Shucked_Wgt-to-Whole_Wgt ratio (SkWR) = Shucked_Wgt/Whole_Wgt\) \(Viscera_Wgt-to-Whole_Wgt ratio (VWR) = Viscera_Wgt/Whole_Wgt\) \(Shell_Wgt-to-Whole_Wgt ratio (ShWR) = Shell_Wgt/Whole_Wgt\) \(Shell_Hgt-to-Shell_Lnt ratio (HLR) = Shell_Hgt/Shell_Lnt\) \(Shell_Dia-to-Shell_Lnt (DLR) = Shell_Dia/Shell_Lnt\)

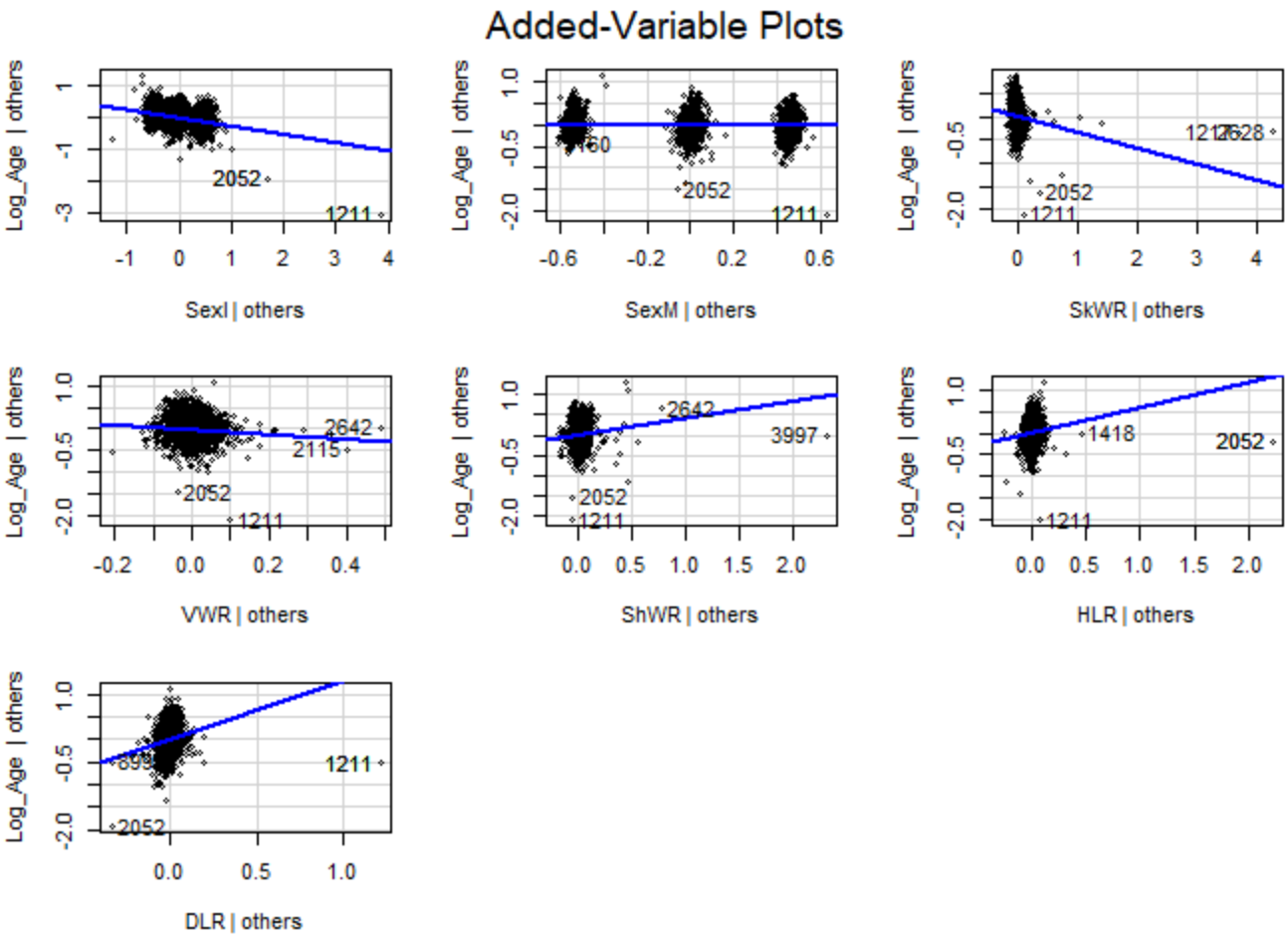
```
## [1] 4177   16
```


Sex
SkWR
VWR
ShWR
HLR
DLR
Log_Age
Corr:
-0.065***
Corr:
-0.178***
Corr:
-0.071***
Corr:
-0.046**
Corr:
-0.192***
Corr:
-0.028.
Corr:
0.006
Corr:
-0.058***
Corr:
-0.053***
Corr:
-0.000
Corr:
0.019
Corr:
0.083***
Corr:
0.234***
Corr:
0.237***
Corr:
0.333***



From the scatter plot matrix, it was observed that the multicollinearity was corrected after the predictors were transformed. The histogram also showed that transforming the “Age” variable to “Log(Age)”, resulted in the response variable following the normal distribution. Thus from here, predictors in this work refers to the transformed predictors (SkWR, VWR, ShWR, HLR, DLR) and Sex which serve as our main effect predictors.

Scatter plots between Age and each main effect predictor variable accounting for the other predictoss.

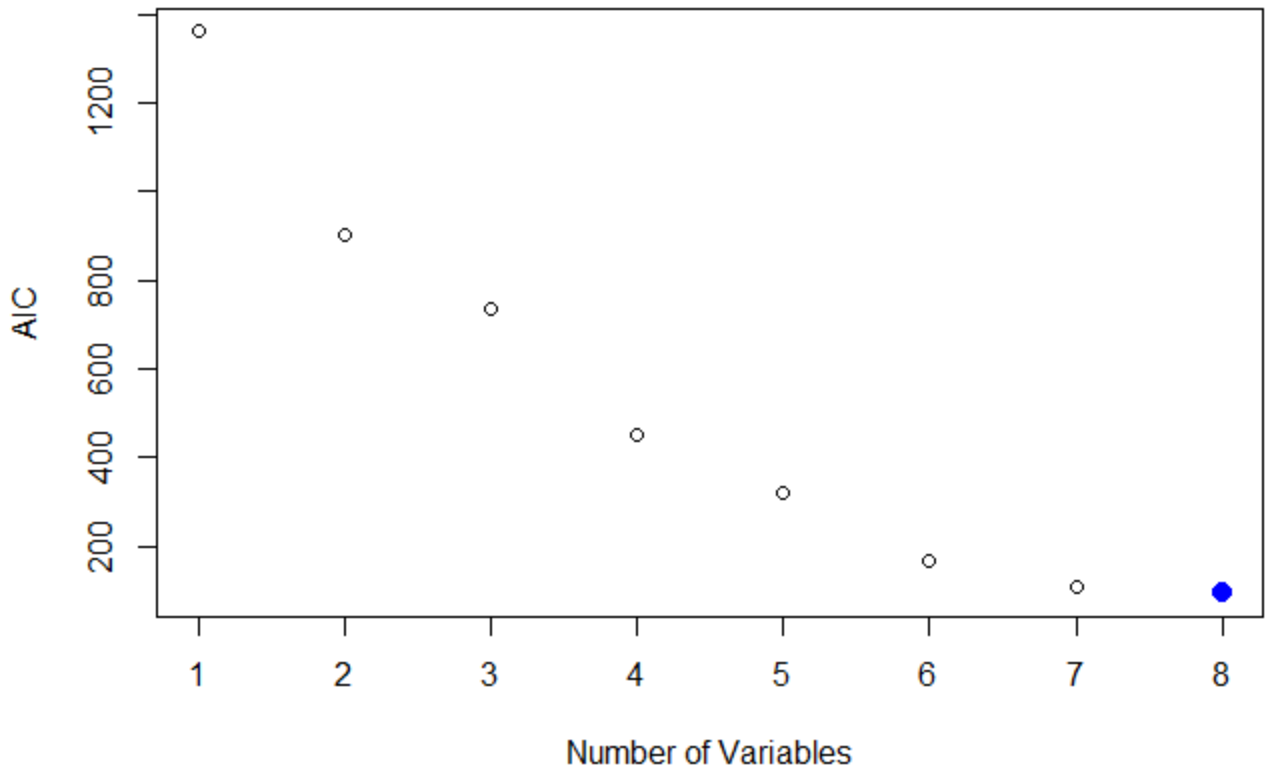


These plots show the relationship between Age and each of the predictor variables conditioned on the other variables. We can see that all the predictor variables show some kind of relationship with the response variable 'Log(Age)' except for sexM which does not really show a significant relationship provided that the other variables are maintained as predictors in the model.

DATA ANALYSIS

Best Subset Selection Using AIC

```
## [1] "which"   "rsq"     "rss"     "adjr2"   "cp"      "bic"     "outmat" "obj"
## [1] 8
```

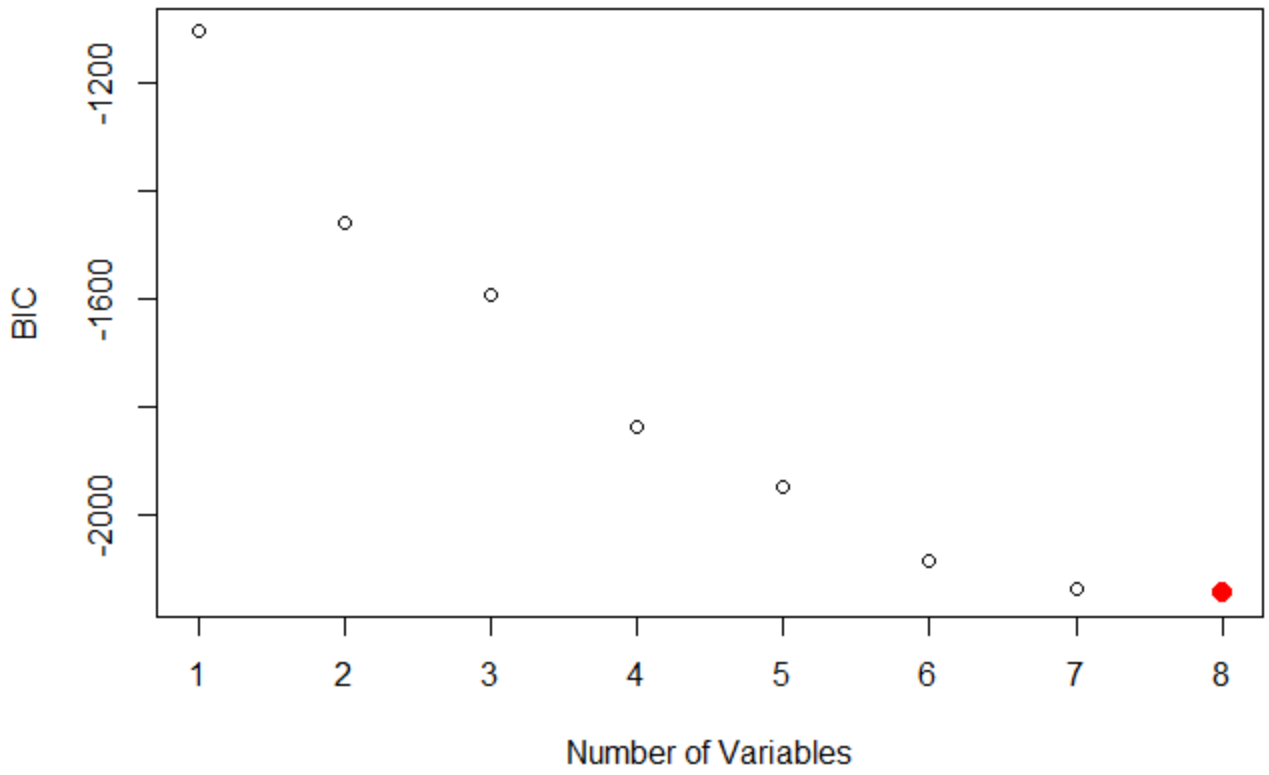



```
## (Intercept)        SexI         HLR         DLR   SexI:SkWR    SexI:DLR
##  -0.3949168  -1.4893406   9.8468363   3.3848308   1.3855930   0.8778160
##    SkWR:HLR     VWR:HLR     HLR:DLR
##  -6.1383952  -2.8774465  -7.3116607
```

The AIC criterion suggests the 8 predictor model above as the best model for predicting Log(Age) of Abalones.

Best Subset Selection Using BIC

```
## [1] 8
```




```
## (Intercept)        SexI         HLR         DLR   SexI:SkWR    SexI:DLR
##  -0.3949168  -1.4893406   9.8468363   3.3848308   1.3855930   0.8778160
##    SkWR:HLR     VWR:HLR     HLR:DLR
##  -6.1383952  -2.8774465  -7.3116607
```

The BIC criterion also suggests the same 8 predictor model suggested by the AIC criterion as the best model for predicting the Log(Age) of Abalones.

Test for Model Quality

To examine whether the model selected using the AIC/BIC criteria does a good job predicting log of Abalones' age, the model is compared to 2 other models: model with only the main effect predictors and the full model with all the 2-factor interactions using the holdout test and the K-fold Cross validation.

Holdout Test

```
## [1] "BIC Model test error:  0.212755131122793"
```

```
## [1] "Main effects Model test error:  0.231745430507723"
## [1] "Full Model test error:  0.250073275203075"
```

The RMSEs from the ho;dpout test show that the model suggested by the BIC criterion has the smallest test error (0.213) compared to 0.231 and 0.25 for the main effects model and the full model with all 2-factor interactions respectively. Thus, the BIC model is the best among the three tested.

Test Using K-Fold Cross Validation.

```
## [1] "BIC Model test error:  0.209899747881466"
## [1] "Main effects Model test error:  0.233812482652544"
## [1] "Full Model test error:  0.209396058503645"
```

The RMSEs from the 'K-fold cross validation test' show that the model suggested by the BIC criterion has the smallest test error (0.210) compared to 0.232 and 0.22 for the main effects model and the full model with all 2-factor interactions respectively. Thus,again, the BIC model is the best among the three tested.

TEST FOR ROBUSTNESS

Using the "forward method" to confirm the best subset selected by the "BIC" criterion.

```
## Subset selection object
## Call: regsubsets.formula(Log_Age ~ (Sex + SkWR + VWR + ShWR + HLR +
##     DLR)^2, data = T.Abalone.data, method = "forward")
## 27 Variables  (and intercept)
##           Forced in Forced out
## SexI          FALSE      FALSE
## SexM          FALSE      FALSE
## SkWR          FALSE      FALSE
## VWR           FALSE      FALSE
## ShWR          FALSE      FALSE
## HLR           FALSE      FALSE
## DLR           FALSE      FALSE
## SexI:SkWR     FALSE      FALSE
## SexM:SkWR     FALSE      FALSE
## SexI:VWR      FALSE      FALSE
## SexM:VWR      FALSE      FALSE
## SexI:ShWR     FALSE      FALSE
## SexM:ShWR     FALSE      FALSE
## SexI:HLR      FALSE      FALSE
## SexM:HLR      FALSE      FALSE
## SexI:DLR      FALSE      FALSE
## SexM:DLR      FALSE      FALSE
## SkWR:VWR      FALSE      FALSE
```

```
## SkWR:ShWR     FALSE      FALSE
## SkWR:HLR      FALSE      FALSE
## SkWR:DLR      FALSE      FALSE
## VWR:ShWR      FALSE      FALSE
## VWR:HLR       FALSE      FALSE
## VWR:DLR       FALSE      FALSE
## ShWR:HLR      FALSE      FALSE
## ShWR:DLR      FALSE      FALSE
## HLR:DLR       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          SexI SexM SkWR VWR ShWR HLR DLR SexI:SkWR SexM:SkWR SexI:VWR
SexM:VWR
## 1  ( 1 ) "*"  " "  " "  " " " " " " " " " "       " "       " "      " "
## 2  ( 1 ) "*"  " "  " "  " " " " " " " " " "       " "       " "      " "
## 3  ( 1 ) "*"  " "  " "  " " " " " " "*" " "       " "       " "      " "
## 4  ( 1 ) "*"  " "  " "  " " " " " " "*" " "       " "       " "      " "
## 5  ( 1 ) "*"  " "  " "  " " " " " " "*" "*"       " "       " "      " "
## 6  ( 1 ) "*"  " "  " "  " " " " "*" "*" "*"       " "       " "      " "
## 7  ( 1 ) "*"  " "  " "  " " " " "*" "*" "*"       " "       " "      " "
## 8  ( 1 ) "*"  " "  " "  " " " " "*" "*" "*"       " "       " "      " "
##          SexI:ShWR SexM:ShWR SexI:HLR SexM:HLR SexI:DLR SexM:DLR SkWR:VWR
## 1  ( 1 ) " "       " "       " "      " "      " "      " "      " "
## 2  ( 1 ) " "       " "       " "      " "      " "      " "      " "
## 3  ( 1 ) " "       " "       " "      " "      " "      " "      " "
## 4  ( 1 ) " "       " "       " "      " "      " "      " "      " "
## 5  ( 1 ) " "       " "       " "      " "      " "      " "      " "
## 6  ( 1 ) " "       " "       " "      " "      " "      " "      " "
## 7  ( 1 ) " "       " "       " "      " "      " "      " "      " "
## 8  ( 1 ) " "       " "       " "      " "      " "      " "      " "
##          SkWR:ShWR SkWR:HLR SkWR:DLR VWR:ShWR VWR:HLR VWR:DLR ShWR:HLR
ShWR:DLR
## 1  ( 1 ) " "       " "      " "      " "      " "     " "     " "      " "
## 2  ( 1 ) " "       " "      " "      " "      " "     " "     "*"      " "
## 3  ( 1 ) " "       " "      " "      " "      " "     " "     "*"      " "
## 4  ( 1 ) " "       "*"      " "      " "      " "     " "     "*"      " "
## 5  ( 1 ) " "       "*"      " "      " "      " "     " "     "*"      " "
## 6  ( 1 ) " "       "*"      " "      " "      " "     " "     "*"      " "
## 7  ( 1 ) " "       "*"      " "      " "      " "     " "     "*"      " "
## 8  ( 1 ) " "       "*"      " "      " "      "*"     " "     "*"      " "
##          HLR:DLR
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

To perform a robustness check, the forward method was use to select the best subset to further confirm the model selected by the BIC method the best and it also suggested the same model as the BIC model below:

$\log(Age) = \beta_{0} + \beta_{1}SexInfant + \beta_{2}HLR + \beta_{3}DLR + \beta_{4}SexInfant*SkWR + \beta_{5}SexInfant*DLR + \beta_{6}SkWR*HLR + \beta_{7}VWR*HLR + \beta_{8}HLR*DLR$

Conclusion

The initial exploratory data analysis pointed to some flaws in the original set of predictors. The predictors were highly correlated with one another, and we addressed this by creating a set of new predictors that were derived from the existing predictors. More specifically, we considered a set of “ratio” predictors that would try to combine the information through a ratio. We noticed that this set of predictors were uncorrelated with one another, resulting in a set of predictors that mitigated multicollinearity. Furthermore, to verify the functional form of our model we perform a log transformation on the age to get a response which looks approximately normal.

After model selection and validation, we propose the final model:

$\log(Age) = \beta_{0} + \beta_{1}SexInfant + \beta_{2}HLR + \beta_{3}DLR + \beta_{4}SexInfant*SkWR + \beta_{5}SexInfant*DLR + \beta_{6}SkWR*HLR + \beta_{7}VWR*HLR + \beta_{8}HLR*DLR$

This model was validated on a 70% Training, 30% testing split, and had the lowest RMSE compared to the two other models, which were the full models with the main effects only, and the model with the main effects and two factor interactions. We used best subset selection to select the best model size, which resulted in the proposed model, and was selected using BIC.

Appendix

Main Effects Model: $\log(Age) = \beta_{0} + \beta_{1}Sex + \beta_{2}HLR + \beta_{3}DLR + \beta_{4}SkWR + \beta_{5}ShWR + \beta_{6}VWR$

Full Model: $\log(Age) = \beta_{0} + \beta_{1}Sex + \beta_{2}HLR + \beta_{3}DLR + \beta_{4}SkWR + \beta_{5}ShWR + \beta_{6}VWR+ \beta_{7}$ Sex:HLR + β_{8}Sex:DLR + β_{9}Sex:SkWR + β_{10}Sex:ShWR + β_{11}Sex:VWR + β_{12}HLR:DLR + β_{13}HLR:SkWR + β_{14}HLR:ShWR + β_{15}HLR:VWR + β_{16}DLR:SkWR + β_{17}DLR:ShWR +

$$\beta_{18}DLR:VWR + \beta_{19}SkWR:ShWR + \beta_{20}SkWR:VWR + \beta_{21}ShWR:VWR$$