AGE PREDICTION OF ABALONE

GROUP 15

BOJJA CHARITHA REDDY - EE19BTECH11001
MARPU MYTHRI VARSHITHA -EE19BTECH11014
BANOTHU SANA -EE19BTECH11021
LAKKU NALINI AISHWARYA -EE19BTECH11033
JALADURGAM ESHWARI -EE19BTECH11042
KALAPAGOOR MAHIMA -MA19BTECH11006

PROBLEM STATEMENT

ABALONE - A Marine Snail, a Rare and Expensive Species.

Endangered in state, Scientists are trying to increase the population count of this Species.

Age Prediction will be highly useful in experimental Studies.

Cutting the shell through the cone, dyeing it, and counting the number of rings via a microscope are used to assess the age of abalone- A general time consuming process.

THE SOLUTION !!! A Rather simple way to predict age by obtaining a model to predict the age of abalone from physical measurements .

Please click on this to watch the video

EXPLAINED!!

The Physical Measurements of Abalone will predict the no. of rings(Dependent variable). Age of abalone can be determined from the number of it's rings.

Fitting the parameters into a multiple regression model solves the problem. Optimizing the model by performing tests is our target.

The DataSet:-
https://www.dcc.fc.up.pt/~ltorgo/Regression/abalone.tar.gz

PROJECT HIGHLIGHTS

- ADDITIVE MODEL
- ADDITIVE MODEL REMOVING FEW PARAMETERS
- AIC SELECTED MODEL
- NUMBER OF VARIABLES VS AIC VALUES
- ADDITIVE LOG MODEL
- ADDITIVE LOG MODEL WITH AUTOCORRELATION CORRECTION

THE DATA

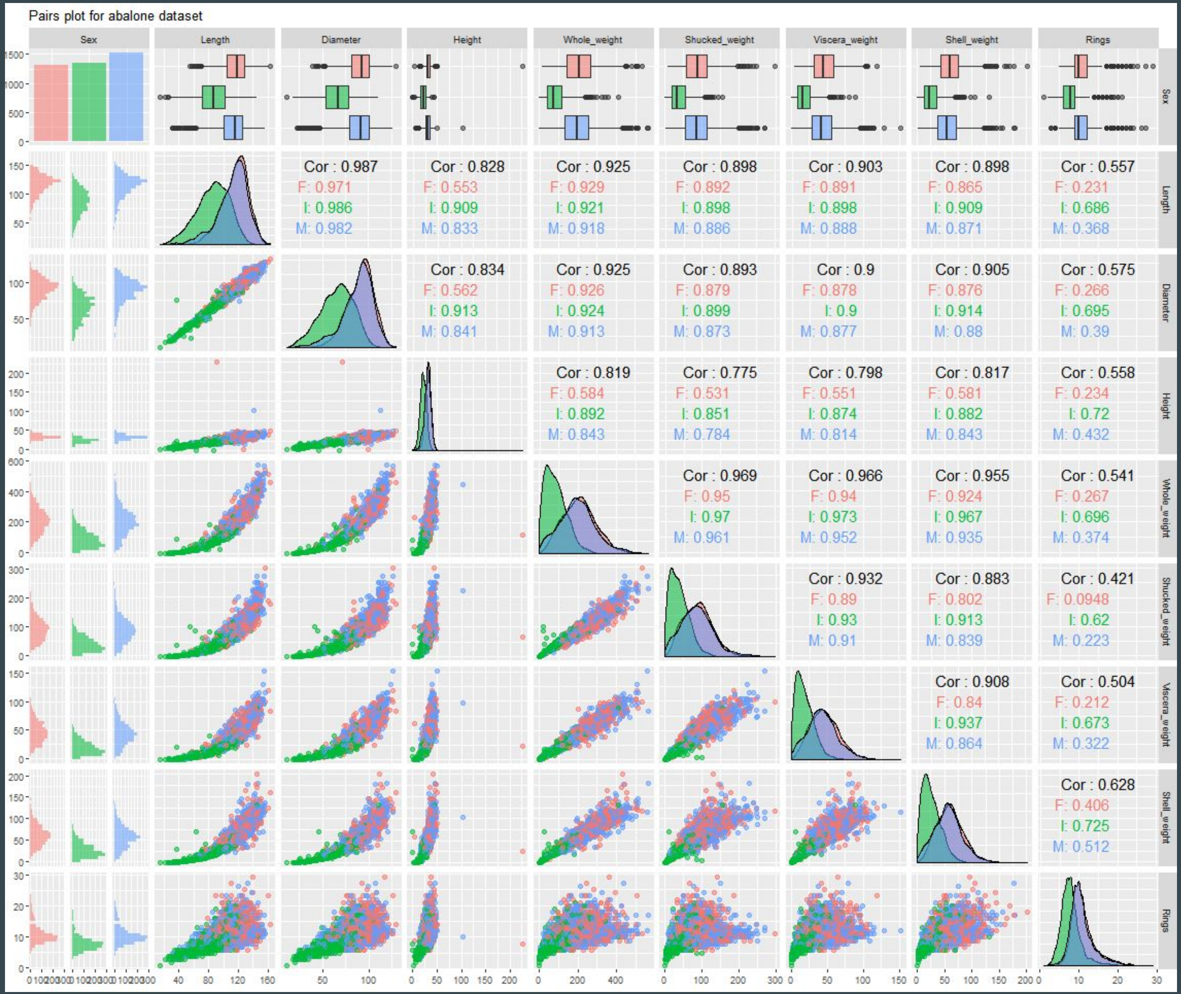
Sex is the categorical variable(M,F,Infant) Length,Diameter,Height, Whole Weight,Shucked Weight,Viscera Weight,Shell Weight,Rings are all numerical variables.

The dataset includes the dependent variable Rings. Rings + 1.5 approximately gives the age of an abalone.

PAIR PLOTS

Building a great scatter-plot matrix

ggpairs(abalone, aes(colour = Sex, alpha = 0.8), title="Pairs plot for abalone dataset") + theme_grey(base_size = 8)



Observations from the pair plot:

- From pair plots, we observe that there is high multicollinearity between the predictors. The correlation between Diameter and Length is extremely high (98.7).
- Whole_weight = Shucked_weight+ Viscera_weight +Shell_weight.
- Whole_weight have high correlation with other weight predictors
- Abalones rings are between 5 and 15 mostly.
- We use gg pairs to see the scatter plots, covariance, and box plots -everything in one big matrix.
- From the plots, we see that plots for male and female are almost the same for every variable. Thus we categorize the abalones into two types-infant (I) and non-infant(NI). And then add it as a new variable.

Calculating Statistical parameters : to achieve basic understanding.

```
> #gives mean,median etc. of each variable
> summary(abalone)
 Sex          Length         Diameter          Height        Whole_weight
 F:1307   Min.   : 15.0   Min.   : 11.00   Min.   :  0.00   Min.   :  0.4
 I:1342   1st Qu.: 90.0   1st Qu.: 70.00   1st Qu.: 23.00   1st Qu.: 88.3
 M:1527   Median :109.0   Median : 85.00   Median : 28.00   Median :159.9
          Mean   :104.8   Mean   : 81.58   Mean   : 27.91   Mean   :165.8
          3rd Qu.:123.0   3rd Qu.: 96.00   3rd Qu.: 33.00   3rd Qu.:230.7
          Max.   :163.0   Max.   :130.00   Max.   :226.00   Max.   :565.1
 Shucked_weight   Viscera_weight    Shell_weight        Rings
 Min.   :  0.20   Min.   :  0.10   Min.   :  0.30   Min.   : 1.000
 1st Qu.: 37.20   1st Qu.: 18.68   1st Qu.: 26.00   1st Qu.: 8.000
 Median : 67.20   Median : 34.20   Median : 46.80   Median : 9.000
 Mean   : 71.88   Mean   : 36.12   Mean   : 47.77   Mean   : 9.932
 3rd Qu.:100.40   3rd Qu.: 50.60   3rd Qu.: 65.80   3rd Qu.:11.000
 Max.   :297.60   Max.   :152.00   Max.   :201.00   Max.   :29.000
```

ADDITIVE MODEL (abalone_add)

- We use a simple additive model involving all the variables.
- After fitting the additive model with all predictors we can see that test statistics showing all variables as significant except 'Length'.
- We see the summary of this model and calculate VIF for this model (shows high values of VIF especially for Whole_weight and diameter, this means multicollinearity is high).

```
> VIF(abalone_add)
                      GVIF Df GVIF^(1/(2*Df))
Sex               1.566331  2        1.118719
Length           40.642565  1        6.375152
Diameter         42.508482  1        6.519853
Height            6.808247  1        2.609262
Whole_weight    110.660026  1       10.519507
Shucked_weight   28.946988  1        5.380240
Viscera_weight   17.242553  1        4.152415
Shell_weight     22.257194  1        4.717753
```

```
lm(formula = Rings ~ Sex + Length + Diameter + Height + Whole_weight +
    Shucked_weight + Viscera_weight + Shell_weight, data = abalone_train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3868 -1.2940 -0.3390  0.9013 12.0768

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.802806   0.353842  10.747  < 2e-16 ***
SexI           -0.744744   0.120331  -6.189 6.90e-10 ***
SexM           -0.007072   0.098555  -0.072  0.94280
Length         -0.006108   0.010664  -0.573  0.56684
Diameter        0.041404   0.013199   3.137  0.00172 **
Height          0.117031   0.013586   8.614  < 2e-16 ***
Whole_weight    0.042376   0.004272   9.918  < 2e-16 ***
Shucked_weight -0.093104   0.004831 -19.271  < 2e-16 ***
Viscera_weight -0.057060   0.007554  -7.554 5.62e-14 ***
Shell_weight    0.044114   0.006789   6.498 9.54e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.159 on 2914 degrees of freedom
Multiple R-squared:  0.5422,    Adjusted R-squared:  0.5408
F-statistic: 383.5 on 9 and 2914 DF,  p-value: < 2.2e-16
```

MULTICOLLINEARITY ISSUE ? , TRY TO REMOVE PARAMETERS BY CHECKING VIF

```r
#Multicollinearity:
library(regclass)
VIF(abalone_add)
#High correlation for whole_Weight and diameter is found



#Partial correlation coefficient between Whole_weight & Rings
#check variability in high collinearity variables
whole_weight_fit <- lm(Whole_weight ~ Sex + Length + Diameter + Height + Shucked_weight + Viscera_weight + Shell_weight, data=abalone_train)

abalone_add_without_whole_weight <- lm(Rings ~ Sex + Length + Diameter + Height
                                       + Shucked_weight + Viscera_weight + Shell_weight,data = abalone_train)
#correlation coefficient
cor(resid(whole_weight_fit),resid(abalone_add_without_whole_weight))

#Variance inflation factor of the additive model without the Whole_weight
VIF(abalone_add_without_whole_weight)
#Partial correlation coefficient between Diameter & Rings(without whole weight)
diameter_fit <- lm(Diameter ~ Sex + Length + Height + Shucked_weight + Viscera_weight + Shell_weight, data=abalone_train)

abalone_add_small <- lm(Rings ~ Sex + Length + Height + Shucked_weight + Viscera_weight + Shell_weight,data = abalone_train)
cor(resid(diameter_fit),resid(abalone_add_small))
VIF(abalone_add_small)
#is smaller for variables than abalone_add model
```

Observations:- Additive Model removing parameters

- We see that covariance between the model without whole_weight and whole_weight is negligible. So, remove whole_weight from the model. (Corr: 0.1807136)
- Similarly, we remove diameter too from the set of predictor variables.(Corr:0.06057918)
- We run VIF on the abalone_add_small model. We can observe that the values reduce greatly after removing the two variables.

```
> VIF(abalone_add_without_whole_weight)
                    GVIF Df GVIF^(1/(2*Df))
Sex             1.563404  2        1.118196
Length         40.630799  1        6.374229
Diameter       42.492299  1        6.518612
Height          6.807620  1        2.609142
Shucked_weight  8.943926  1        2.990640
Viscera_weight 10.736332  1        3.276634
Shell_weight    8.429893  1        2.903428
```

```
> VIF(abalone_add_small)
                    GVIF Df GVIF^(1/(2*Df))
Sex             1.545084  2        1.114905
Length          9.196951  1        3.032648
Height          6.645554  1        2.577897
Shucked_weight  8.939459  1        2.989893
Viscera_weight 10.706342  1        3.272055
Shell_weight    8.137067  1        2.852554
```

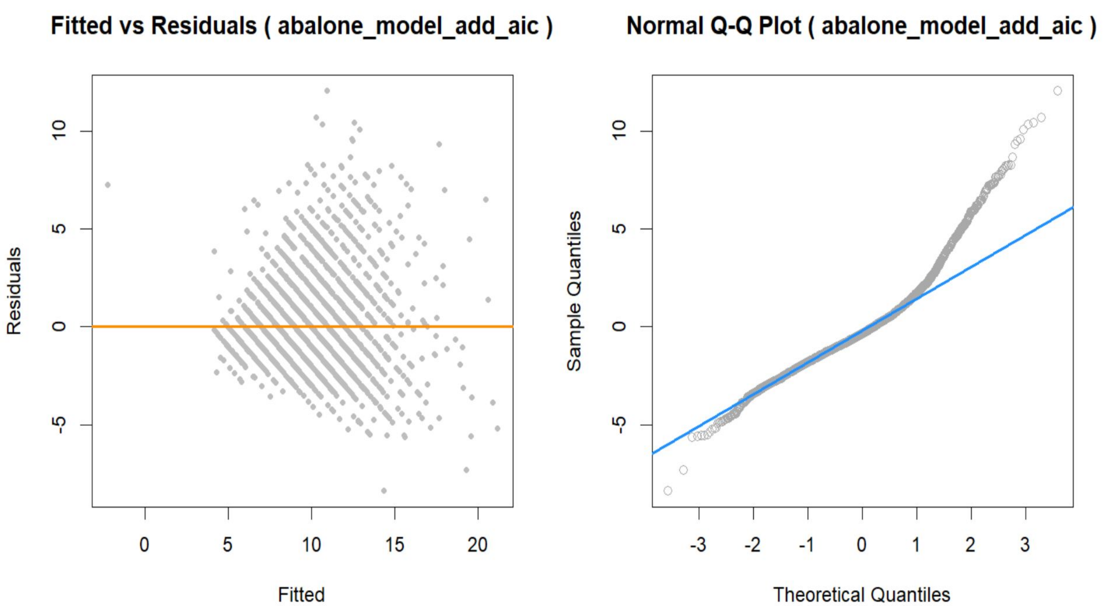
F-test:

```
> #F test to chose b/w Ho =abalone_add_small(without diameter and whole_weight)
> #Ha=abalone_add(all variables)
> anova(abalone_add_small,abalone_add)
Analysis of Variance Table

Model 1: Rings ~ Sex + Length + Height + Shucked_weight + Viscera_weight + 
    Shell_weight
Model 2: Rings ~ Sex + Length + Diameter + Height + Whole_weight + Shucked_weight + 
    Viscera_weight + Shell_weight
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1   2916 14092                                  
2   2914 13582  2    510.25 54.736 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We run an F test where the null hypothesis is that abalone_add_small is the better one, and the alternative hypothesis selects the simple abalone_add.

- The F-test rejects the null hypothesis. So, we reject abalone_add_small and run AIC on abalone_add to select the best model.

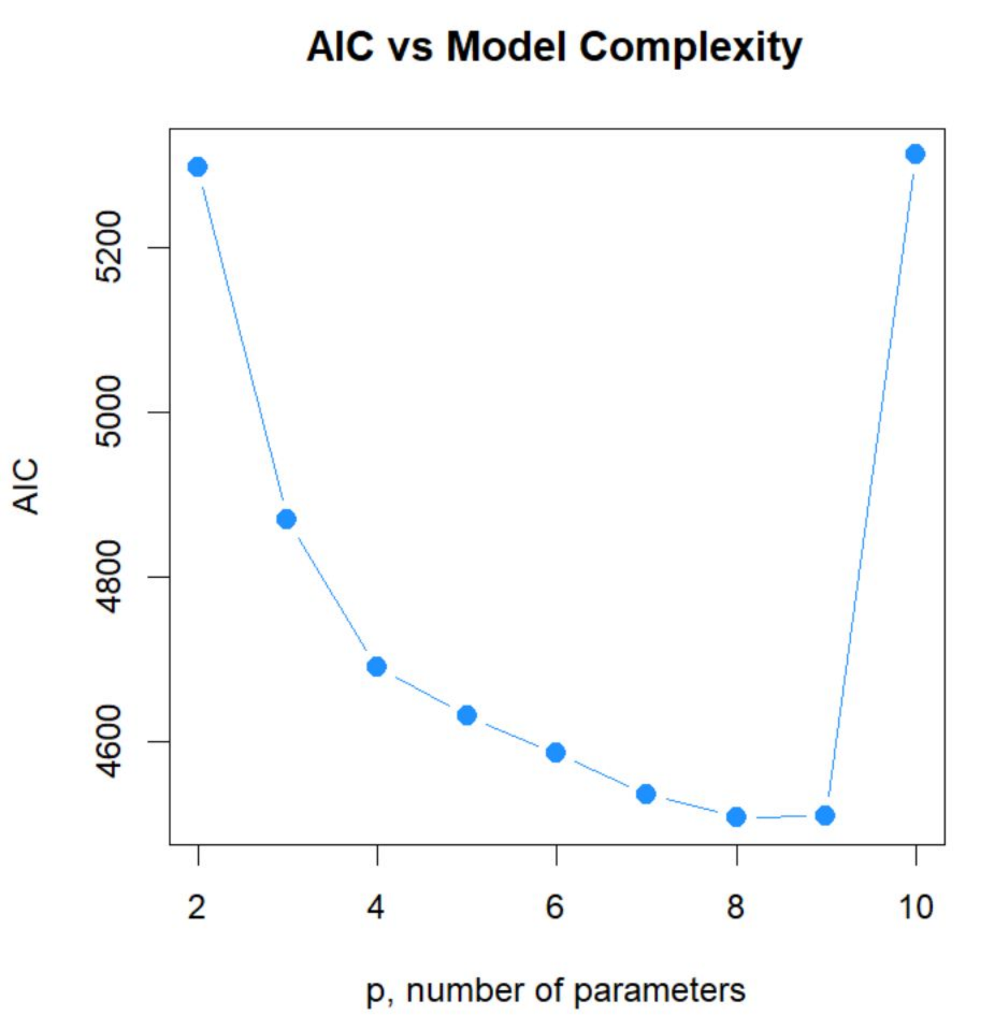
```
Call:
lm(formula = Rings ~ Sex + Diameter + Height + Whole_weight +
    Shucked_weight + Viscera_weight + Shell_weight, data = abalone_train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3757 -1.2928 -0.3372  0.8981 12.0493

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.738800   0.335695  11.137  < 2e-16 ***
SexI           -0.749276   0.120057  -6.241 4.98e-10 ***
SexM           -0.008540   0.098510  -0.087    0.931
Diameter        0.034754   0.006278   5.536 3.37e-08 ***
Height          0.116545   0.013558   8.596  < 2e-16 ***
Whole_weight    0.042418   0.004271   9.931  < 2e-16 ***
Shucked_weight -0.093341   0.004813 -19.394  < 2e-16 ***
Viscera_weight -0.057503   0.007513  -7.654 2.64e-14 ***
Shell_weight    0.044275   0.006782   6.528 7.82e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.159 on 2915 degrees of freedom
Multiple R-squared:  0.5422,    Adjusted R-squared:  0.5409
F-statistic: 431.5 on 8 and 2915 DF,  p-value: < 2.2e-16
```



AIC SELECTED MODEL

AIC selects the model without the length predictor.

We plot the residuals and the qq-plot obtained through this AIC model. We observe that the residuals show a fan-out effect. The qq-plot also deviates from the normal line–especially the head.



NUMBER OF VARIABLES VS. AIC VALUES

We get the summary of RSS(residual sum of squares) and adjusted R^2 of the best models of all sizes.

We calculate AIC values for each model and plot AIC vs the number of variables in the model.

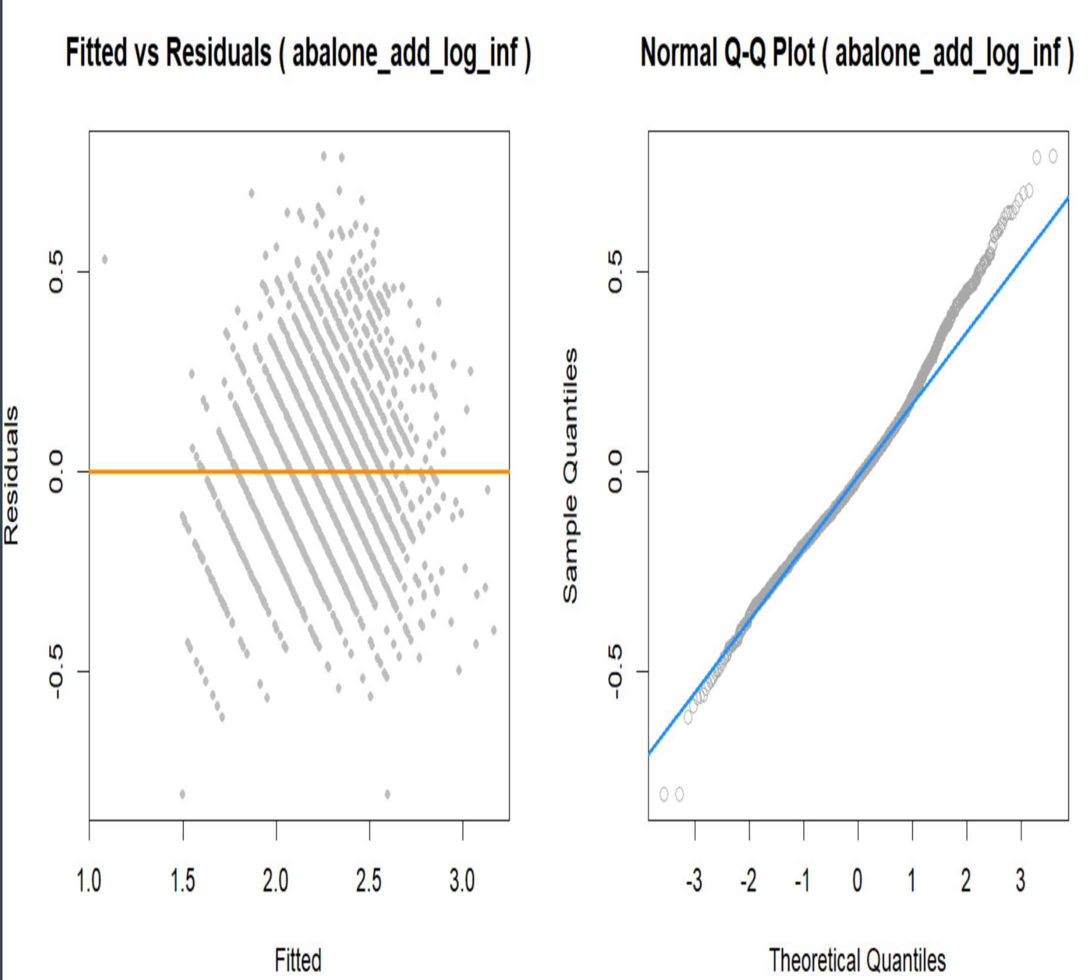
Interestingly, we observe that the AIC value is the lowest when the number of predictor variables is 8.

ADDITIVE LOG MODEL and its coefficients:

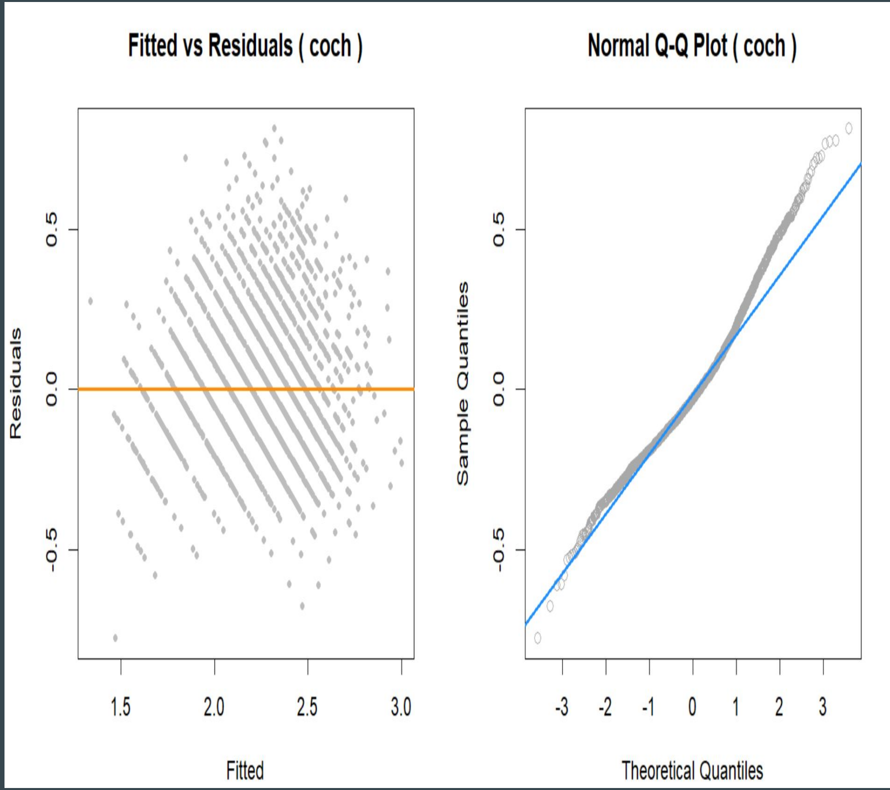
```
> abalone_add_log_inf <- lm(log(Rings) ~ Infant + Length + Diameter + Height + Whole_weight + Shucked_weight + Viscera_weight + Shell_weight,data = abalone_train)
> summary(abalone_add_log_inf)

Call:
lm(formula = log(Rings) ~ Infant + Length + Diameter + Height +
    Whole_weight + Shucked_weight + Viscera_weight + Shell_weight,
    data = abalone_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80673 -0.13114 -0.01389  0.11170  0.78901

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.2585709  0.0297832  42.258  < 2e-16 ***
InfantNI        0.0813365  0.0097255   8.363  < 2e-16 ***
Length          0.0017015  0.0009838   1.729   0.0838 .
Diameter        0.0061998  0.0012170   5.094 3.72e-07 ***
Height          0.0128475  0.0012537  10.248  < 2e-16 ***
Whole_weight    0.0028851  0.0003943   7.318 3.24e-13 ***
Shucked_weight -0.0077561  0.0004449 -17.435  < 2e-16 ***
Viscera_weight -0.0045794  0.0006965  -6.575 5.76e-11 ***
Shell_weight    0.0028414  0.0006265   4.535 5.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1992 on 2915 degrees of freedom
Multiple R-squared:  0.5985,    Adjusted R-squared:  0.5974
F-statistic: 543.2 on 8 and 2915 DF,  p-value: < 2.2e-16
```

ADDITIVE LOG MODEL (abalone_add_log_inf)

The analysis with Infant - I and NI (2 categorical variables) are the same as M, F, I because M and F have almost the same effect. Therefore we replace Sex with the Infant variable in the model. Also to reduce heteroscedasticity, we use log transforming response i.e we use log(Rings). We once again plot the residuals and qq-plot. The residuals look a lot better and the qq-plot is also much closer to the line.



```
> #To address auto-correlation
> coch = cochrane.orcutt(abalone_add_log_inf)
> summary(coch)
Call:
lm(formula = log(Rings) ~ Infant + Length + Diameter + Height +
    Whole_weight + Shucked_weight + Viscera_weight + Shell_weight,
    data = abalone_train)

                  Estimate  Std. Error t value  Pr(>|t|)
(Intercept)     1.20939917  0.03220061  37.558 < 2.2e-16 ***
InfantNI        0.04000283  0.00963053   4.154 3.365e-05 ***
Length          0.00359478  0.00087841   4.092 4.386e-05 ***
Diameter        0.00525555  0.00108914   4.825 1.469e-06 ***
Height          0.00927119  0.00112539   8.238 2.609e-16 ***
Whole_weight    0.00188890  0.00035293   5.352 9.373e-08 ***
Shucked_weight -0.00461313  0.00041822 -11.030 < 2.2e-16 ***
Viscera_weight -0.00402599  0.00061535  -6.543 7.114e-11 ***
Shell_weight    0.00228762  0.00055194   4.145 3.499e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1853 on 2914 degrees of freedom
Multiple R-squared:  0.5273 ,  Adjusted R-squared:  0.526
F-statistic: 406.3 on 8 and 2914 DF,  p-value: < 0e+00

Durbin-Watson statistic
(original):    1.36558 , p-value: 1.181e-66
(transformed): 2.23652 , p-value: 1e+00
```



To address auto - correlation, we perform Durbin Watson test.

```
> #RMSE for training and test data
> kable(log_rmse(abalone_add_log_inf,"Additive Log Model"), digits = 4,format = 'markdown')


|Model              | RMSE.Train| RMSE.Test|
|:------------------|----------:|---------:|
|Additive Log Model |     2.2209|    3.5353|
> kable(log_rmse(coch,"Additive Log Model with auto-correlation correction"), digits = 4,format = 'markdown')


|Model                                              | RMSE.Train| RMSE.Test|
|:--------------------------------------------------|----------:|---------:|
|Additive Log Model with auto-correlation correction |     2.3167|    4.0259|
```

- We run the Durbin-Watson test on the resultant log model. The initial Durbin-Watson factor was around 1.36. After the remedy, it's around 2.23.
- Durbin-Watson values between 1.5-2.5 are considered to be normal.
- We compare the coch model and additive log model using RMSE.
- By comparison, we can observe that additive log model performs better.

```
Durbin-Watson statistic
(original):      1.36558 , p-value: 1.181e-66
(transformed): 2.23652 , p-value: 1e+00
```

Actual.no.of.Rings	Predicted.no.of.Rings	Actual.age.of.abalone	Predicted.age.of.abalone
7	8	8	10
7	5	8	6
9	10	10	12
7	10	8	12
10	15	12	16

```
> #Confidence Interval
> exp(predict(abalone_add_log_inf, newdata=sample,interval="confidence"))
        fit       lwr       upr
1  8.394375  8.245541  8.545896
2  4.682946  4.500320  4.872984
3  9.859051  9.690483 10.030552
4  9.937401  9.620472 10.264770
5 15.303036 14.952243 15.662059
```

Conclusions:

The Additive Log Model is used for prediction.

We observe the confidence intervals for the first 5 predictions. The prediction intervals are in the same range. It has almost constant variance and is much closer to normal(except at the tail and at the head) as compared to other models. The additive log model has the Durbin-Watson factor around 1.36, which means the auto-correlation isn’t that much. But this model shows high multicollinearity.

Thank you

for experiencing the journey of abalone:)