Predicting the age of Abalone

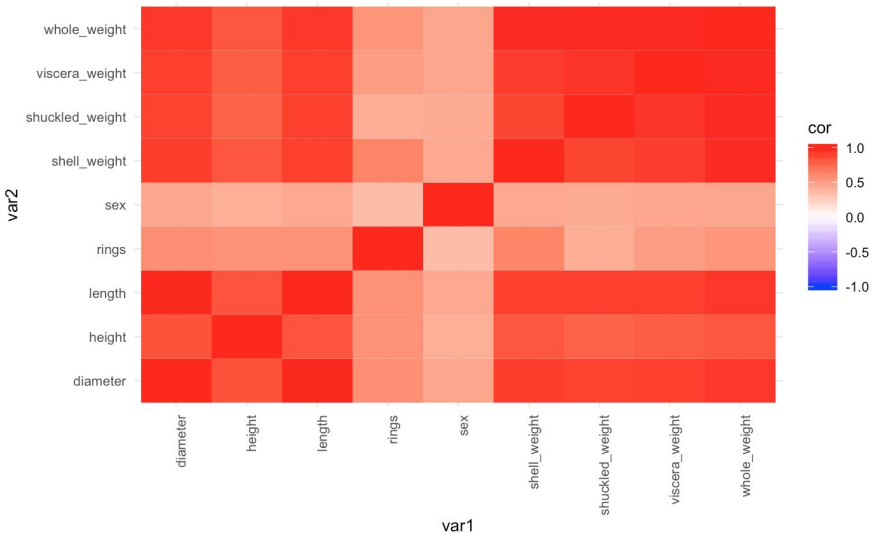
Data Description

The abalone dataset consists of 4177 samples of physical abalone attributes provided by the University of California, Irvine. Each sample contains 9 attributes:

- Sex - M (male), F (female), I (infant)
- Length - longest shell measurement (mm)
- Diameter - perpendicular to the length (mm)
- Height - with meat in shell (mm)
- Whole weight - weight of the whole abalone (g)
- Shucked weight - weight of the abalone meat (g)
- Viscera weight - weight of the internal organs after bleeding (g)
- Shell weight - weight of the shell after being dried (g)
- Rings - number of layers in the shell, adding 1.5 gives the age in years (integer)

The process of determining the age of an abalone is complicated, it involves taking the shell, staining it and then counting the number of layers (rings) using a microscope.

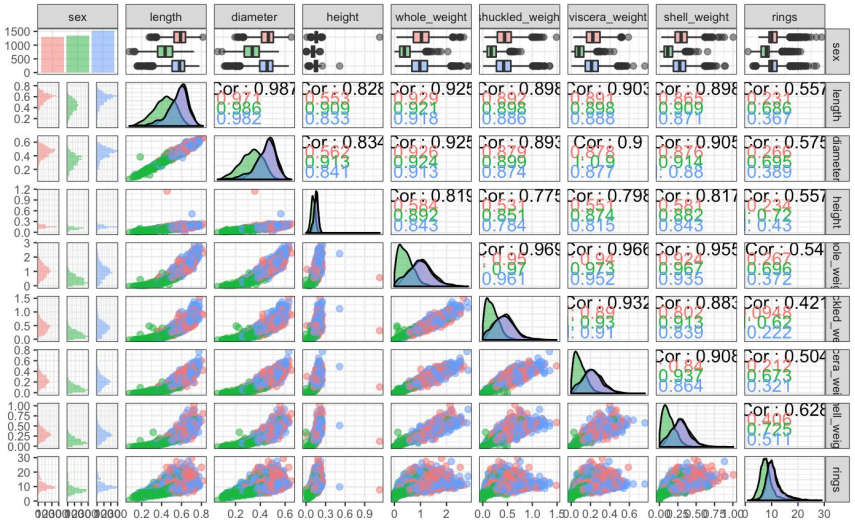
Graphs and Tables

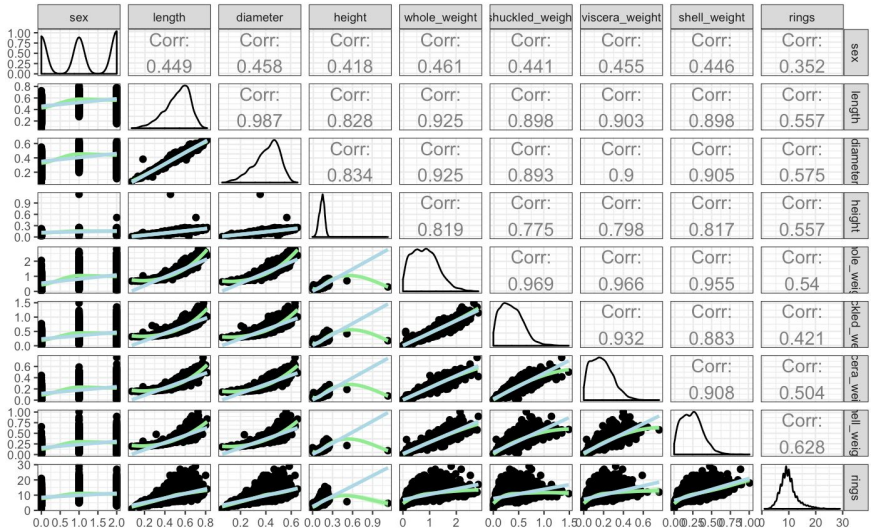


Correlation Heat Map

- Every pair of variables has positive correlation.
- Different weight variable has relatively high correlation with each other.
- Length, height and diameter have high correlation with each other.
- Rings variable has low correlation with anyone

Graphs and Tables





Exploratory Analysis

Assumptions

Assumption 1

Linearity:

All predictor variables have an approximate linear relationship with the response variable

Assumption 2

Collinearity:

No perfect collinearity

Assumption 3

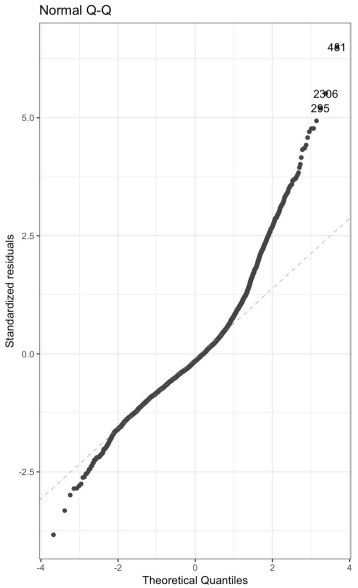
Independence:

All the data is IID

Assumption 4

Normality:

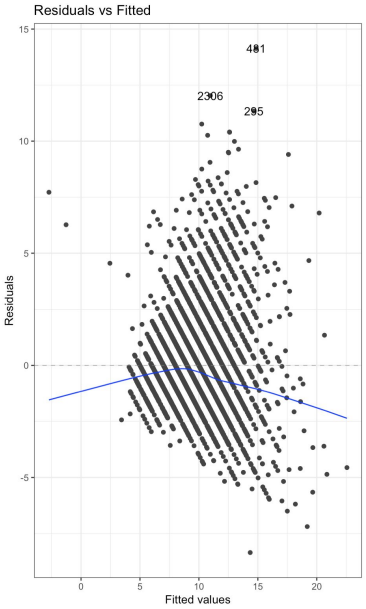
The errors follow a normal distribution



Assumption 5

Homoskedasticity:

The errors have constant variance



Modelling/ Model Selection

Simple Linear Regression - rings ~ height

```
Linear Regression

4173 samples
   1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3755, 3756, 3755, 3755, 3756, 3755, ...
Resampling results:

  RMSE     Rsquared   MAE
  2.55251  0.3752872  1.849458

Tuning parameter 'intercept' was held constant at a value of TRUE
```

MLR including all variables

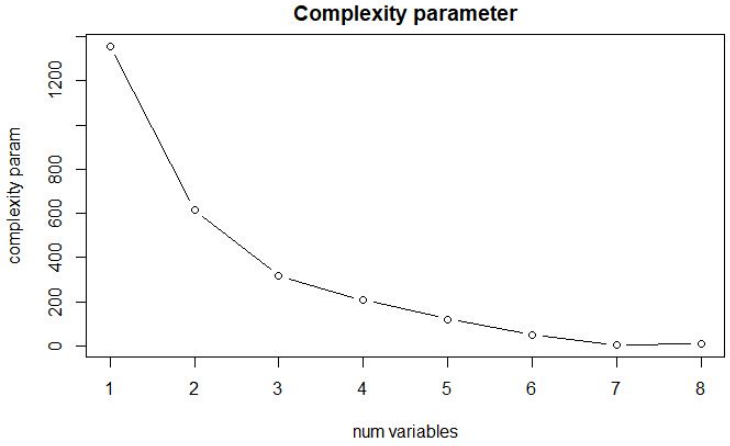
```
Linear Regression

4173 samples
   8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3756, 3756, 3756, 3756, 3756, 3756, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.187342  0.5421348  1.57983

Tuning parameter 'intercept' was held constant at a value of TRUE
```

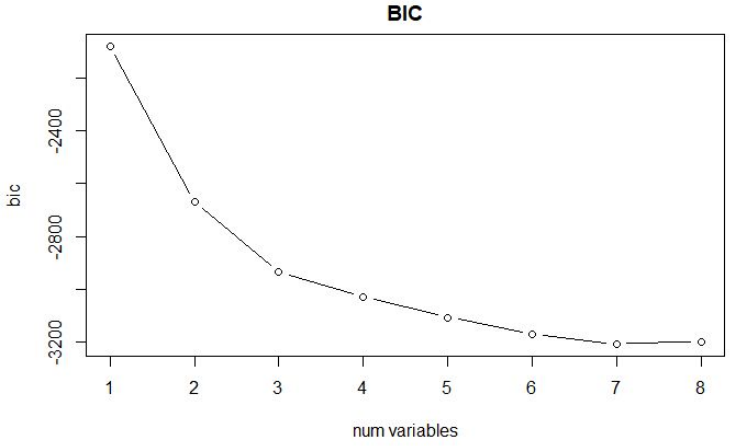
MLR generated using backwards stepwise

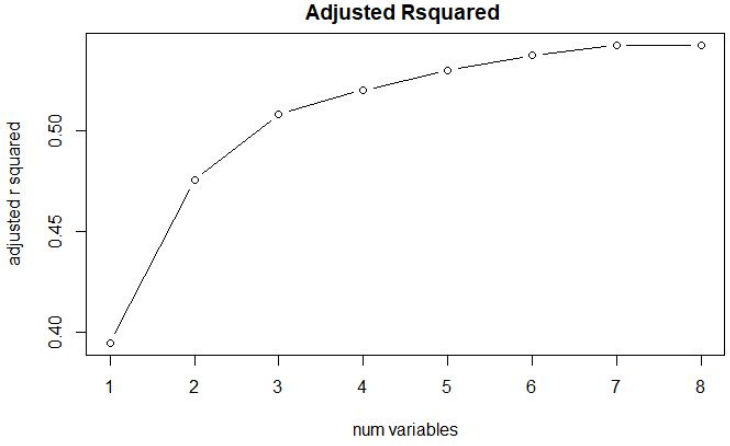
```
Linear Regression

4173 samples
   7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3755, 3756, 3755, 3756, 3757, 3756, ...
Resampling results:

  RMSE      Rsquared  MAE
  2.181897  0.541999  1.576802

Tuning parameter 'intercept' was held constant at a value of TRUE
```

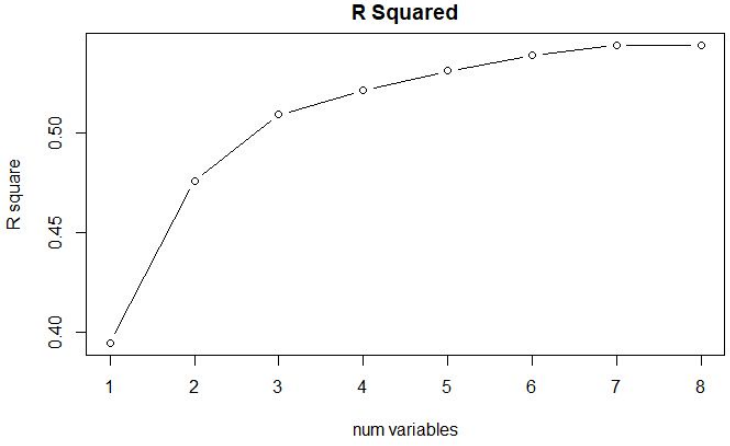
MLR generated by exhaustive search

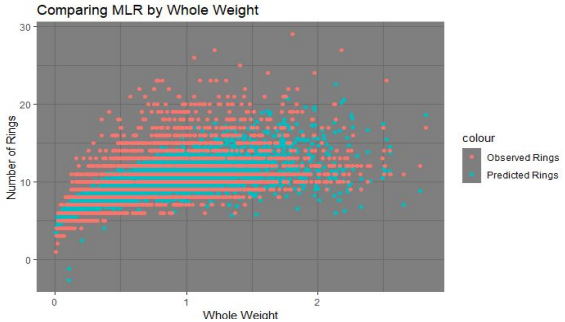
```
Linear Regression

4173 samples
   7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3755, 3755, 3756, 3755, 3756, 3756, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.181785  0.5438199  1.576448

Tuning parameter 'intercept' was held constant at a value of TRUE
```

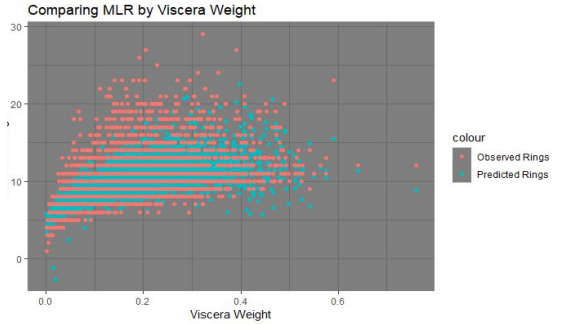

Complexity parameter
complexity param
0
200
400
600
800
1200
1
2
3
4
5
6
7
8
num variables

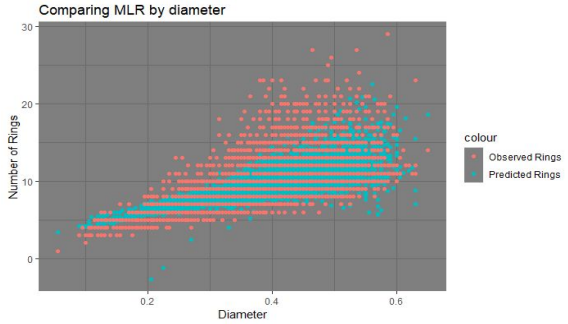

BIC
bic
-3200
-2800
-2400
1
2
3
4
5
6
7
8
num variables

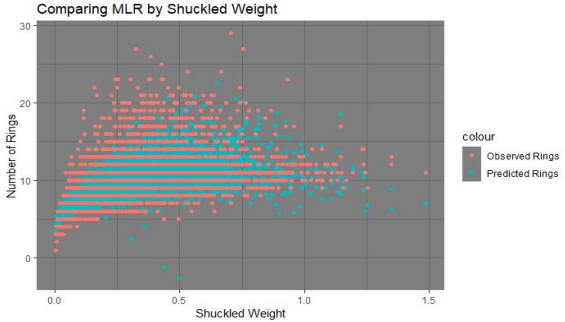

Adjusted Rsquared
adjusted r squared
0.40
0.45
0.50
1
2
3
4
5
6
7
8
num variables

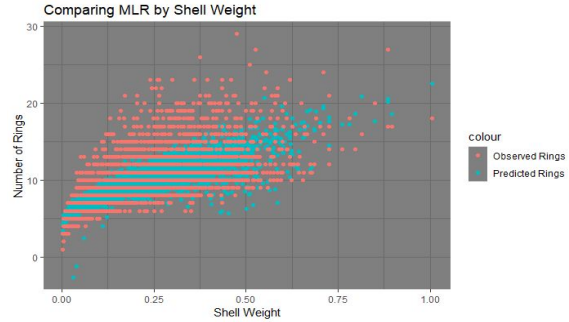

R Squared
R square
0.40
0.45
0.50
1
2
3
4
5
6
7
8
num variables

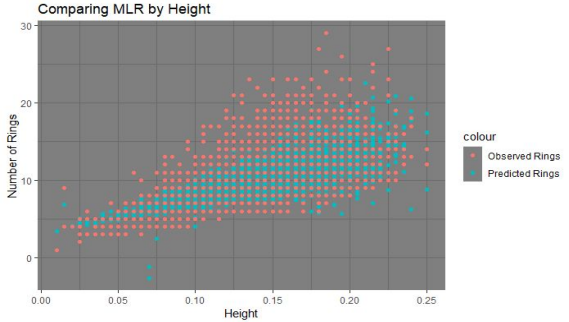

Comparing MLR by Whole Weight
Number of Rings
Whole Weight
colour
Observed Rings
Predicted Rings

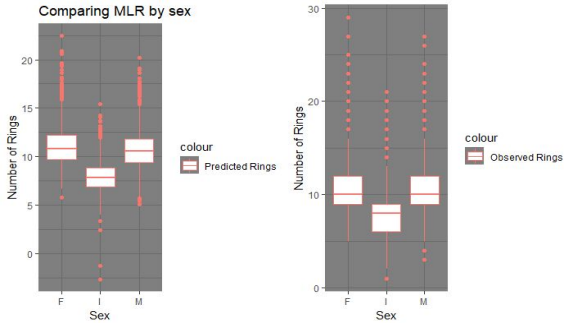

Comparing MLR by Viscera Weight
Viscera Weight
colour
Observed Rings
Predicted Rings


Comparing MLR by diameter
Number of Rings
Diameter
colour
Observed Rings
Predicted Rings


Comparing MLR by Shuckled Weight
Number of Rings
Shuckled Weight
colour
Observed Rings
Predicted Rings


Comparing MLR by Shell Weight
Number of Rings
Shell Weight
colour
Observed Rings
Predicted Rings


Comparing MLR by Height
Number of Rings
Height
colour
Observed Rings
Predicted Rings


Comparing MLR by sex
Number of Rings
Sex
colour
Predicted Rings
Observed Rings

Results

R	R^2	Adjusted R^2	Standard Error
0.7368	0.5429	0.542	2.182

df	F-ratio	p-value
8 and 4167	618.7	$< 2.2e^{-16}$

Results

Age of Abalone =
3.605 – (0.804 x sex I) + (0.051 x sex M) + (8.481 x Diameter) + (21.157 x Height) + (8.875 x Whole weight) – (19.531 x Shuckled weight) – (11.136 x Viscera weight) + (7.899 x Shell weight)

The age of an Abalone should be equal to the number of rings it has +1.5

Model	Estimate	Standard Error	t-value	p-value
Intercept	3.556	0.278	12.803	$< 2e^{-16}$
Sex I	-0.804	0.102	-7.910	$3.27e^{-15}$
Sex M	0.051	0.083	0.617	0.537
Diameter	8.481	1.030	8.230	$2.47e^{-16}$
Height	21.157	2.162	9.784	$< 2e^{-16}$
Whole weight	8.875	0.722	12.300	$< 2e^{-16}$
Shuckled weight	-19.531	0.812	-24.068	$< 2e^{-16}$
Viscera weight	-11.136	1.283	-8.679	$< 2e^{-16}$
Shell weight	7.899	1.126	7.021	$2.56e^{-12}$