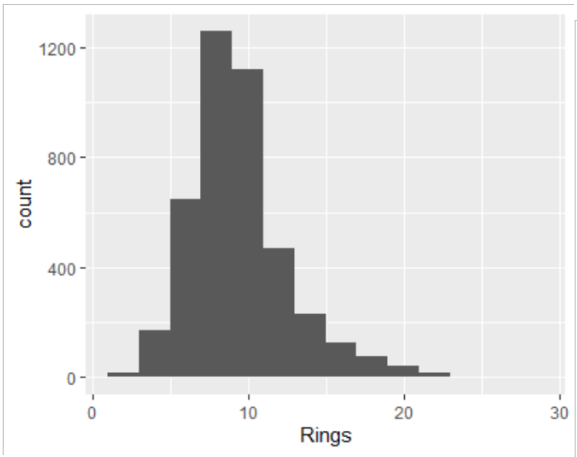
Goal

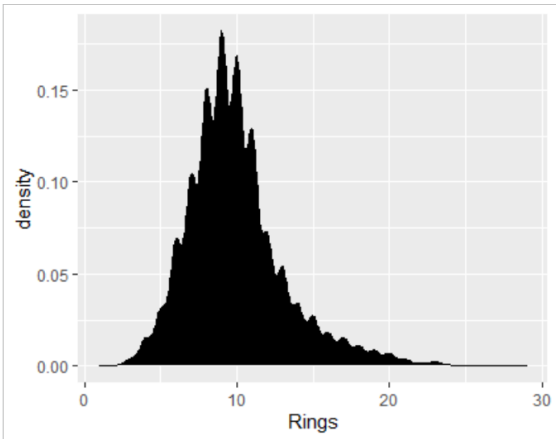
The goal of this project is to develop a model to predict the age of abalone from physical measurements of the shell. The age of abalone here is obtained from the number of rings on the shell. So the response variable of interest here is number of rings treated as a continuous value.

Initial Exploratory Data Analysis (EDA)

Before running any model, we first do a little some data exploration through graphs and correlation checks. The aim of this step is to understand the data better and visually inspect the relationship between variables.

First let's start with our response variable: Rings





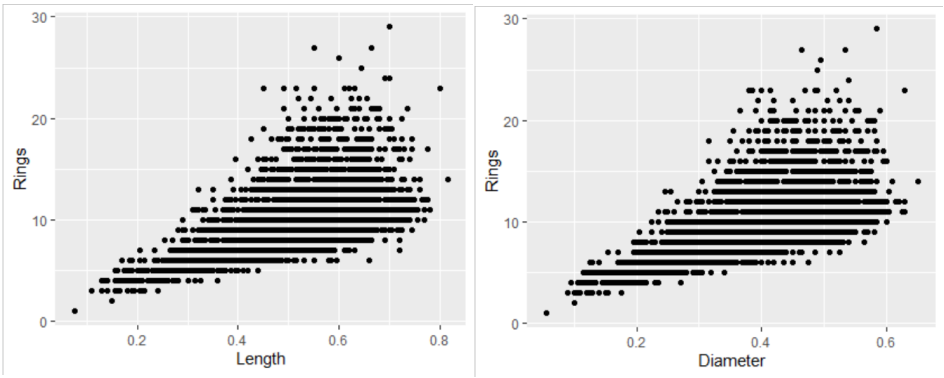
The histogram and density plot of Rings show that it is reasonable to make a starting assumption that rings could be normally distributed. Therefore, I will treat Rings as a continuous variable and perform regression analysis. But it is important to note that this alone DOES NOT establish normality of the response variable. We need to evaluate the residual plots (QQ plots) of any model built.

There are 8 predictor variables of which Sex is Categorical with 3 levels. The remaining 7 are numerical variables. Below is the correlation matrix which shows the strength and direction of linear relationship between these variables. I have also included response Rings in this correlation check.

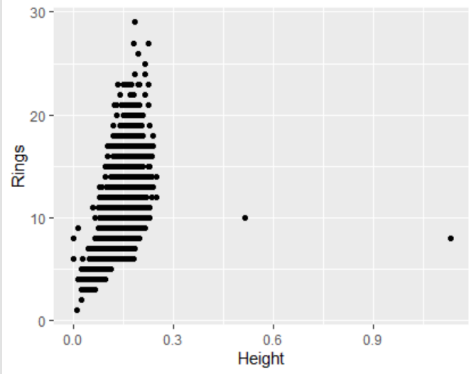
	Length	Diameter	Height	Whole	Shucked	Viscera	Shell	Rings
Length	1.0000000	0.9868116	0.8275536	0.9252612	0.8979137	0.9030177	0.8977056	0.5567196
Diameter	0.9868116	1.0000000	0.8336837	0.9254521	0.8931625	0.8997244	0.9053298	0.5746599
Height	0.8275536	0.8336837	1.0000000	0.8192208	0.7749723	0.7983193	0.8173380	0.5574673
Whole	0.9252612	0.9254521	0.8192208	1.0000000	0.9694055	0.9663751	0.9553554	0.5403897
Shucked	0.8979137	0.8931625	0.7749723	0.9694055	1.0000000	0.9319613	0.8826171	0.4208837
Viscera	0.9030177	0.8997244	0.7983193	0.9663751	0.9319613	1.0000000	0.9076563	0.5038192
Shell	0.8977056	0.9053298	0.8173380	0.9553554	0.8826171	0.9076563	1.0000000	0.6275740
Rings	0.5567196	0.5746599	0.5574673	0.5403897	0.4208837	0.5038192	0.6275740	1.0000000

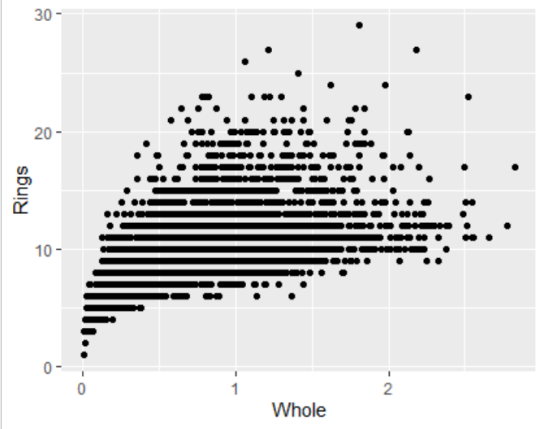
From the above table there seems to be a high correlation among many of the predictors.
'Length' & 'Diameter' have .986 correlation (which is very high!).
'Whole' & 'Viscera' have high correlation (>0.9) with many other predictors (and each other).
From this we infer multicollinearity will definitely be an issue in any modelling, so might need to do variable selection.

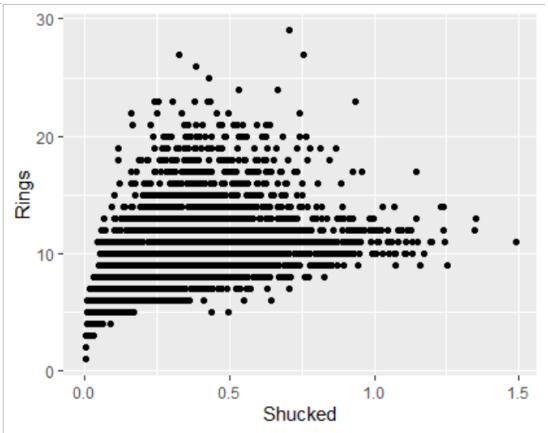
Next, I looked at scatterplots between these numerical predictors and the response variable:



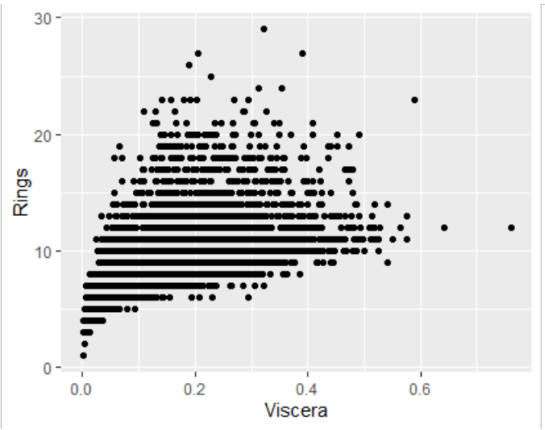
Both these scatterplots looks same, so it is likely we need to include only Length or Diameter. There seems to be somewhat linear relationship here.

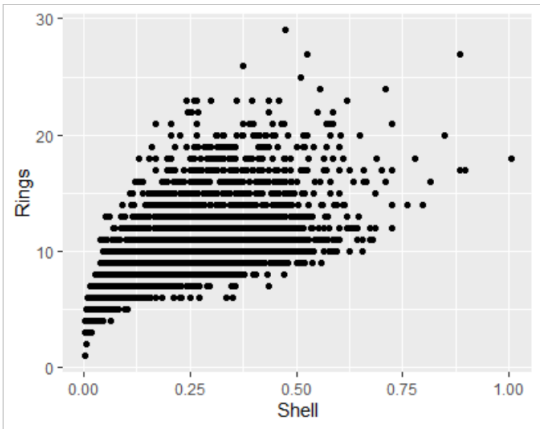






These two plots also look very similar.

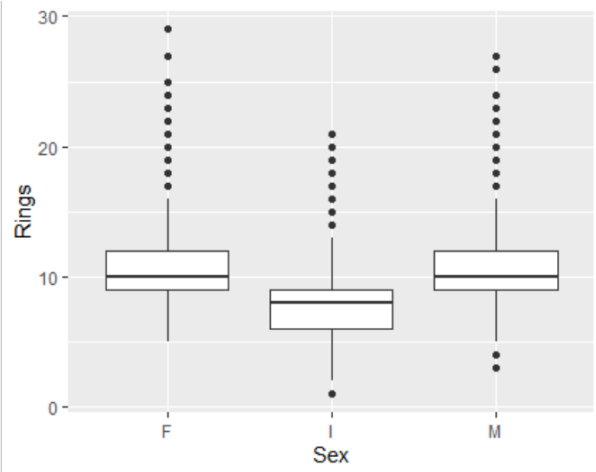




These two plots also look very similar.

From all these plots I get the idea that some of the predictors might not contribute anything extra to a model.

Finally, looking at boxplot of Sex vs Rings:



Looks like F & M are similar, but infant category is different. We will come back to this point when looking at the models.

Train-Test Split

I am splitting the data into 70% training data vs 30% testing data. I will not use the testing data when fitting any model, only to evaluate the MSE. I have set the seed in R using "set.seed(123)" so that these results can be reproduced.

Initial Linear Model

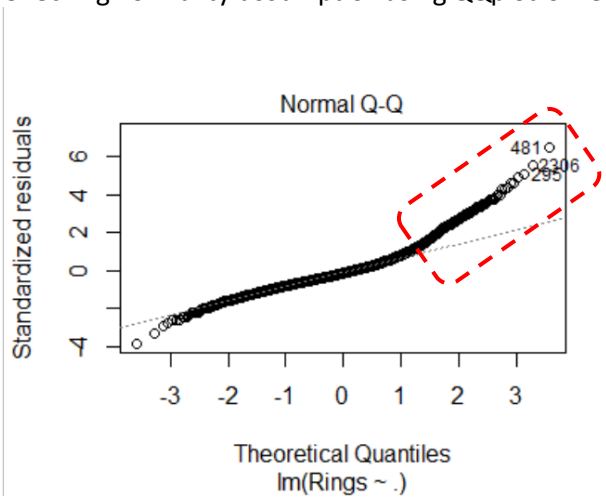
The initial linear model contains all the predictors. The model gives a training MSE = 4.765 and a test MSE = 5.2345. This seems like a good performance but let's make sure to check the model diagnostics before saying that this is a good model or making inferences about the beta estimates.

Checking for multicollinearity using Variance Inflation Factor (VIF):

```
      SexI       SexM     Length   Diameter     Height      Whole    Shucked
  1.993242   1.398279  39.575404  41.700366   6.763715 107.138849  28.616992
    Viscera      Shell
 17.498761  20.599650
```

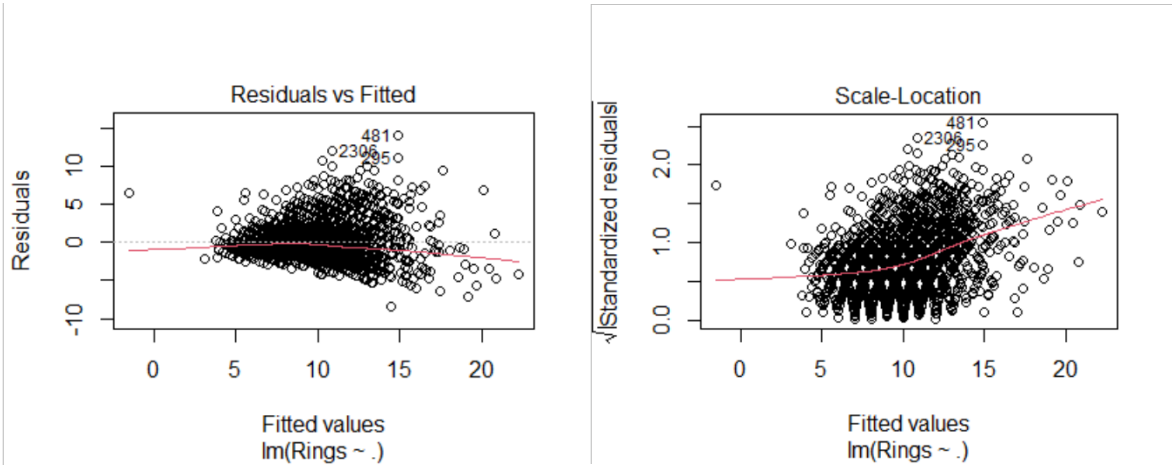
Usually, a VIF value > 10 indicates multicollinearity could be an issue. Almost all the numerical predictors have VIF>10. This confirms what we suspected earlier from scatterplots and correlation matrix. Multicollinearity is definitely an issue!

Checking normality assumption using QQplot of residuals:



Very big deviation for larger values of residuals. This is serious violation of the assumption that the response variable is normally distributed. This can be possibly be corrected with appropriate transformations. We will come back to this later.

Checking for constant of error terms:



This could also be an issue; we will keep monitoring this for successive models to see whether it improves.

So, we see that in spite of low MSE, there are serious violation of linear model assumptions. Therefore, we cannot use this model or interpret anything from this model.

Variable reduction to correct multicollinearity

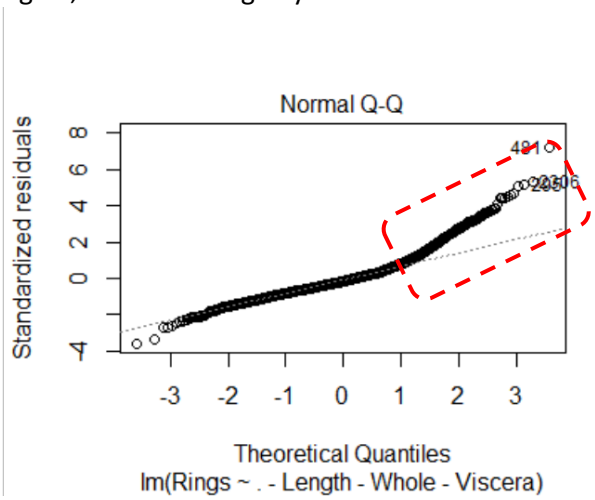
I will first correct the issues of multicollinearity using backward elimination:

Using General F test I first drop 'Length'. Then I check VIF and see that multicollinearity is still an issue. I drop 'Whole' next and then 'Viscera' on next iteration each time checking if VIF > 10 for any predictor.

Finally, I get a model in which VIF condition is not violated. This is "mod4" in my code.

```
> vif(mod4) #multicollinearity not an issue!
    SexI     SexM Diameter   Height  Shucked    Shell
1.956377 1.393733 9.067828 6.603715 5.832181 7.285516
```

Again, before making any inference about this model I will check the diagnostic plots:



Normality assumption is still violated! To correct that we look at possible transformations.

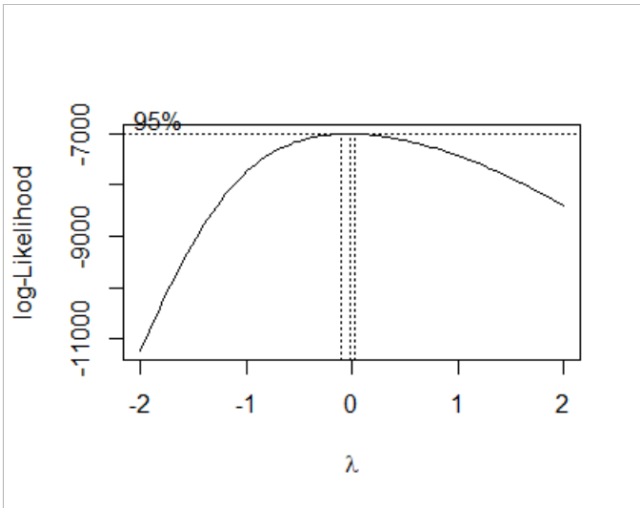
BoxCox Procedure

I used boxcox procedure to check which possible transformation will result in highest log likelihood. The boxcox transformation is

y^lambda = b0 + b1x1 + b2x2 + …. (for lambda not = 0)

Log(y) = b0 + b1x1 + b2x2 + …. (for lambda = 0)

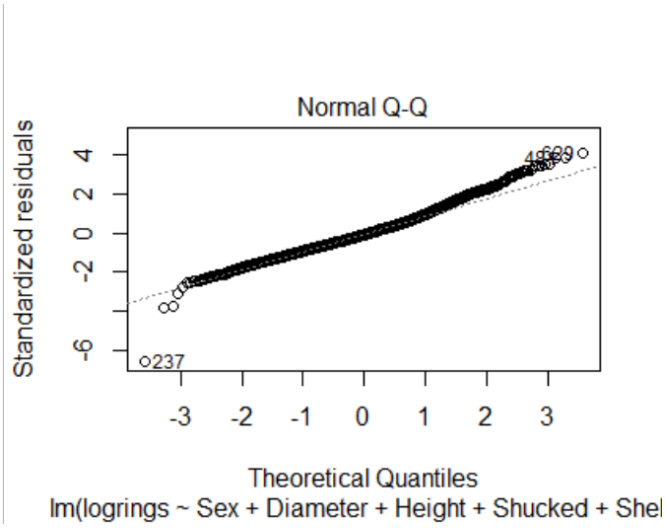
The plot seen below:



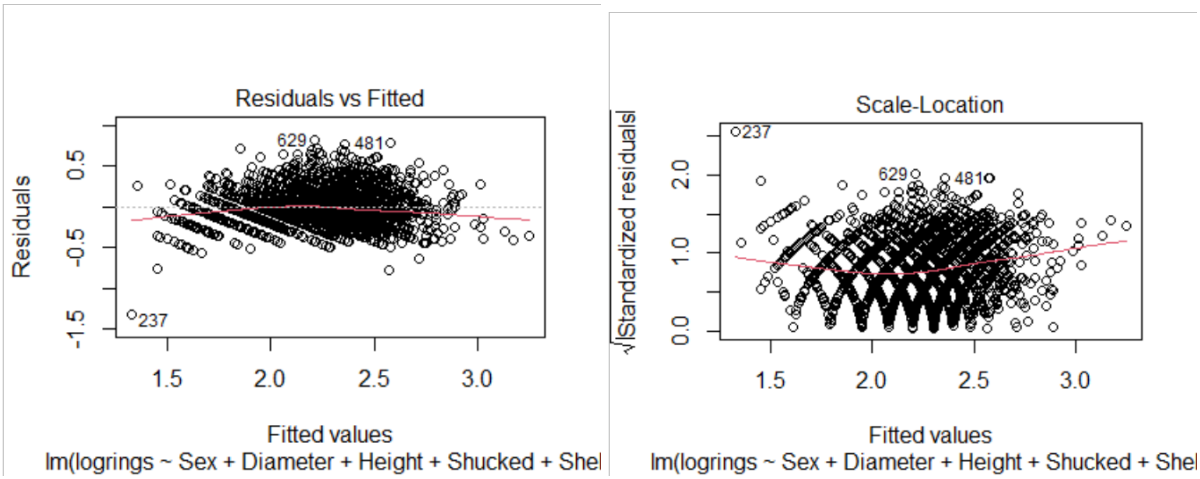
Lambda=0 (very close to zero) maximizes the log likelihood. Therefore, logarithmic transformation is appropriate.

Modelling Log(Rings)

Applying log transformation on Rings and treating this as the response variable we get “mod4log”. The VIF diagnostics are good, lets look at the residual diagnostic plots.



This looks much better!! There may be few smaller values which are outliers, but overall this looks relatively normal.



The constant variance assumption looks alright. Maybe it could be improved little, but overall its seem okay.

Now that the assumptions are validated, lets take a closer look at the fitted model:

```
Call:
lm(formula = logrings ~ Sex + Diameter + Height + Shucked + Shell,
    data = abalone_train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32571 -0.13510 -0.01746  0.11395  0.83121

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.279662   0.030659  41.738  < 2e-16 ***
SexI        -0.081966   0.011284  -7.264  4.8e-13 ***
SexM         0.013265   0.009232   1.437    0.151
Diameter     1.799508   0.114689  15.690  < 2e-16 ***
Height       2.853106   0.253330  11.262  < 2e-16 ***
Shucked     -1.133998   0.041673 -27.212  < 2e-16 ***
Shell        1.098662   0.073264  14.996  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2035 on 2916 degrees of freedom
Multiple R-squared:  0.6011,    Adjusted R-squared:  0.6003
F-statistic: 732.4 on 6 and 2916 DF,  p-value: < 2.2e-16
```

SexM has high p-value, therefore it is not significantly different that SexF (which is used for base comparison by R). This means that the association of SexF & SexM on response variable logrings is relatively same.
All other predictors have a significant p-value and thus do have a significant association with the response variable. The fitted model using the above beta estimates is:
(For eg. The beta estimates can be interpreted as follows: one unit increase in Diameter will correspond to a 1.799 increase in logrings.)

Logrings = 1.279 + SexI(-0.0819) + Diameter*(1.799) + Height*(2.853)*
+ Schuked(-1.1339) + Shell*(1.098)* (EQ 1)

SexI is an indicator variable that takes value 1, when the observation is Infant. The model reduces to:

Logrings = 1.1971 + Diameter(1.799) + Height*(2.853) + Schuked*(-1.1339) + Shell*(1.098)*

So, we see that the effect of logrings reduces by 0.0819 for infants. Please note is in log scale, the actual effect on response “Rings” can be calculated by transforming back.

Similarly, when the observation is Female or Male, SexI=0 and (EQ 1) reduces to:

Logrings = 1.279 + Diameter(1.799) + Height*(2.853) + Schuked*(-1.1339) + Shell*(1.098)*

I tested out if any square term of predictor can be added to model to give a better fit or cleaner residual plots, but not much improvement there.

Also, I tried out weighted linear regression to see if I can better fix the constant variance. But not much improvement here either.

So, my "mod4log" is a final model. Evaluating the MSE on training and test data revealed:

Training MSE = 0.04133
Test MSE = 0.04744

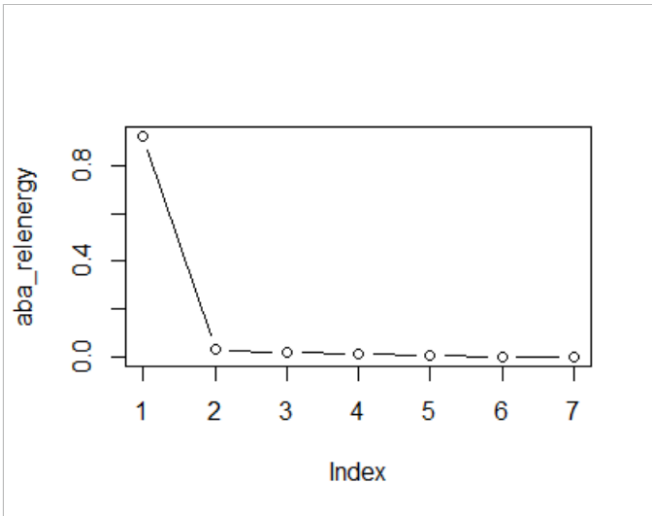
Although these are lower due to it being on log scale, it is quite a low error rate and since the model residual diagnostics are fine it is a good model.

PCA Regression

Another idea I had is to use PCA since we have many highly correlated predictors. I will form PC's and treat them as my new predictors.

I am running PCA on correlation matrix of the numerical predictors. The reason to use correlation instead of covariance matrix is that these predictors are not all on the same scale.

Once I perform eigen decomposition of correlation matrix, I plot the relative energy of each eigenvalue using scree plot.



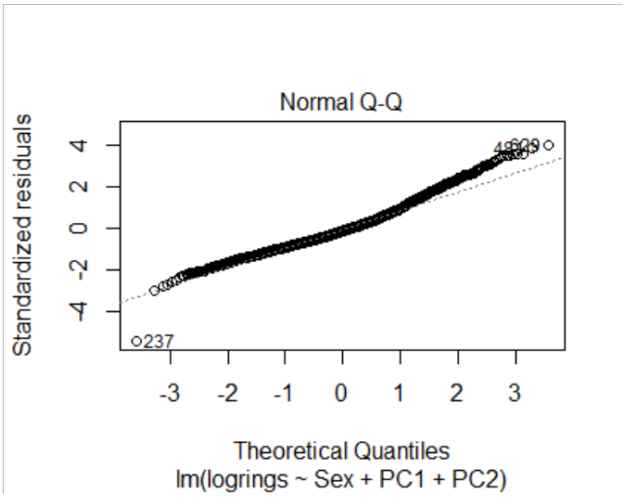
The relative energies are listed below

```
> aba_relenergy
[1] 0.9234600475 0.0305248650 0.0203922759 0.0137921798 0.0089810571 0.0018736933
[7] 0.0009758814
```

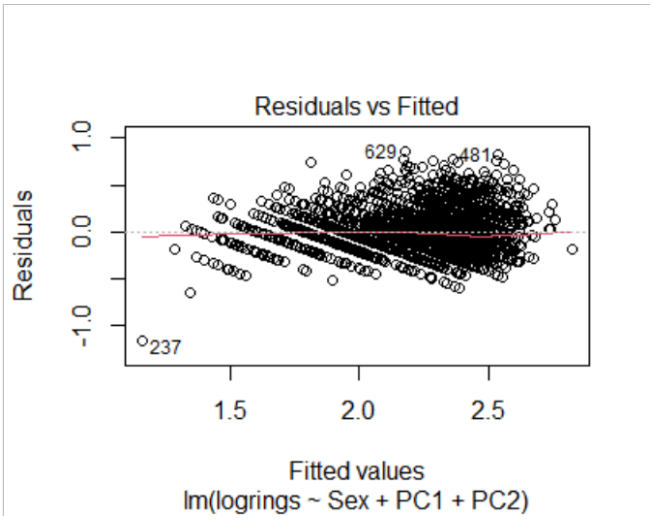
Though the 1st eigenvector contributes 92.3% of total variance, I will use the first 2 eigenvalues. This contributes 95.4% of total variance.

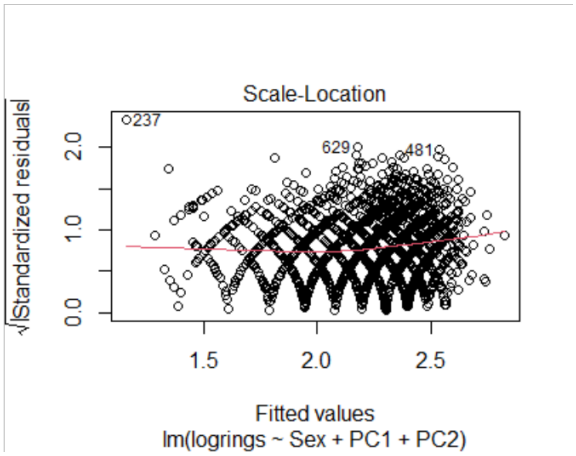
Next, we project the numeric predictors into 2D space using the first 2 eigenvectors.

Now I run a regression model with logrings as response and sex, PC1, PC2 as the predictor variables. Looking at the residual plots:



This plot looks even better than mod4log. Very few values deviate from the dotted line.





Even these pots look good. Constant variance assumption is validated.

Looking at the fitted model:

```
Call:
lm(formula = logrings ~ Sex + PC1 + PC2, data = abalone_train_PC)

Residuals:
     Min       1Q   Median       3Q      Max
-1.16170 -0.14569 -0.02842  0.11649  0.86408

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.072694   0.031512  34.041   <2e-16 ***
SexI        -0.101264   0.011857  -8.540   <2e-16 ***
SexM         0.001303   0.009719   0.134    0.893
PC1         -1.439793   0.036214 -39.758   <2e-16 ***
PC2          2.236238   0.071451  31.297   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2148 on 2918 degrees of freedom
Multiple R-squared:  0.5555,    Adjusted R-squared:  0.5549
F-statistic: 911.8 on 4 and 2918 DF,  p-value: < 2.2e-16
```

SexM has high p-value, therefore it is not significantly different that SexF (which is used for base comparison by R). This means that the association of SexF & SexM on response variable logrings is relatively same. This is same result as before.

The fitted model using the above beta estimates is:

Logrings = 1.0729 + SexI(-0.10126) + PC1*(-1.43979) + PC2*(2.2362)* (EQ 2)

SexI is an indicator variable that takes value 1, when the observation is Infant. The model reduces to:

Logrings = 0.97164 + + PC1(-1.43979) + PC2*(2.2362)*

So, we see that the effect of logrings reduces by 0.10126 for infants. Please note is in log scale, the actual effect on response "Rings" can be calculated by transforming back.

Similarly, when the observation is Female or Male, SexI=0 and (EQ 2) reduces to:

Logrings = 1.0729 + PC1(-1.43979) + PC2*(2.2362)*

For this model evaluating the MSE on training and test data revealed:

Training MSE = 0.04605
Test MSE = 0.048267

These are also low! (on log scale). It is slightly higher than mod4log but not much different.

Result

Important thing to note when using PCA regression is that it is difficult to interpret the PC's as they are linear combinations of many variables that doesn't have physical meaning by combing. (eg. Height + Weight doesn't have any practical meaning). If the client's goal is purely prediction and he is not interested in understanding the relationship between predictors and response I would recommend the PCA model as I am not discarding any predictor variables, and by using just 2 PC's I account for 95.4% of total variance.
However, if the client is interested in understanding the relationship between predictors and response in addition to prediction, then I would recommend the previous model (mod4log).