Using Abalone's Physical Features to Predict its Age

Chengyuan Zhang*

School of data science, The Chinese University of HongKong, ShenZhen, China

*Corresponding author: 119010424@link.cuhk.edu.cn

Abstract. Abalone is one of the most delicious and highly-prized seafood around the world, its deliciousness also makes the abalone industry a non-negligible part of the global economic circle, a lot of people and countries rely on abalone for their lives and economy. Therefore, it means a lot for us to study about abalone and it's population. However, as a necessary step when studying about abalone, getting the age of abalone is a very complicated and time-consuming task. That's why we need a model to help us predict the age of abalone according to it's physical measurements which are easy to acquire. The project considered three models, linear regression model, polynomial regression model and Random Forest. 10-fold cross validation is used to compute the mean square error, multiple R square is also considered when evaluating the models. In the results, the polynomial model is the best model among three models, with lowest mean square error and largest R-square. The research provides us with a model to get the age of abalone in an easy and convenient way, which makes the study about abalone more convenient and thus be beneficial for the development of abalone industry.

Keywords: Abalone, Linear Regression, Polynomial Regression.

1. Introduction

1.1. Background

According to a research conducted by Peter A. Cook(2014), Global fish production has increased a lot in the past 20 years, and it continues to outpace world population growth [1]. As one of the most famous and highly prized seafood delicacies, the total volume of worldwide abalone market has also increased significantly especially in the past few years, and abalone industry has became an important part in world's economy. However, in the last century, abalone fisheries suffered a huge decline. According to the research conducted by Warwick J. Nash, Tracy L. Sellers, Simon R. Talbot, Andrew J. Cawthorn and Wes B. Ford (1994), the decline in abalone catches has had a huge impact on people in related industries, many people lost their incomes because of the decline of abalone [2]. After that, artificial farming of abalone became more and more important, according to a research about abalone industry conducted by Ke(2013), "In modern times, due to the intensification of natural environmental changes, abalone's natural resources are decreasing day by day, and the market is in short supply, resulting in rising prices, so abalone artificial farming came into being. Among the 66 species of abalone in the world, more than 10 species have been cultivated in captivity" [3].

Therefore, it's important for us to learn more about abalone and its population in order to keep the stable development of abalone industry. When we wants to learn about abalone, getting the age of it is a necessary step. However, getting the age of abalone is a complicated and time-consuming task, that's why we need a model to help us predict the age of abalone.

1.2. Related research

In a research conducted by Liu and Hou(2016), they chose the same data set with this research to do discriminant analysis and cluster analysis, and as the result, they found that there are significant difference between these 8 variables[4], which means the variables contained in the data set are reasonable and suitable to fit the model.

1.3. Objection

In this research, the goal is using eight physical measurements of abalone, the Sex, Length, Height, Diamater, Wholoe weight, shucked weight, viscera weight and shell weight to to construct a regression model to help us make a prediction of the abalone's age. If an accurate model can be constructed, scientist can save a lot of time and resource when they are studying about abalone.

2. Methodology

2.1. Source of Data

The data comes from UCI, they are collected and donated on 1995-12-01. The data consists of 4177 instances with 9 variables, for the first 8 variables, they each represents a physical measurement of abalone. The last variable is the rings, which equals to the age of abalone minus 1.5 according to the definition of abalone's age.

2.1.1 Dependent Variable

The dependent variable which we are going to make prediction is the age of abalone, it's a newly constructed variable which is formed by adding 1.5 to the Rings of abalone

2.1.2 Independent Variables

There are eight independent variables. First is the Sex, abalone has three types of sex - Male, Female and Infant. The second one is the length, which is the longest shell measurement of abalone. The third one is the Diameter, and it is perpendicular to the length. The fourth one is the height, it is measured with meat in the shell. Then it is the whole weight of abalone. The next one is the shucked weight, which is the weight of the meat. Then the viscera weight, which is the gut weight after bleeding. The last independent variable is the shell weight after being dried.

2.2. Data processing

First, constructing a new variable to represent the age of abalone, the new variable age is constructed by rings plus 1.5, after this step, the rings can be deleted. Then divides the dataset into two parts randomly, the first part is training part, which contains 80% of the data. The second part is testing part, it contains the rest 20% of the data. The training data is used to build and select the model, and the testing data is used to evaluate the effectiveness of the chosen model.

2.3. Machine learning method

R studio is used to construct the model.

The first model considered is the linear regression model. It's the most basic and common regression model, but it's reliable in reality according to a research conducted by Xing(2007)[5]. By using the functions "glm()" and "lm()", p-values of variables and the covariance table are checked and being used to adjust the model. New variables may be necessary to construct in order to deal with the correlation problem between variables. As for the variables whose p-value are significantly large, they may be removed to improve the accuracy of the model. The mean square error will decrease first and then increase as we remove those variables with large p-value, so when the mse reaches the lowest point, variables would remain unchanged.

The second model considered is the polynomial regression model. It should be better than linear regression model. According to a research about polynomial regression conducted by Wang and Fu (2004), polynomial regression model is more suitable for a real life problem because for most variables, there is a power multiplication relationship between them[6]. Therefore, polynomial regression model was expected to get a better result in the research. the functions used are also glm() and lm(), some variables are turned to the second or third or even fourth power to decrease the mean square error. Same as the linear regression, new variables will be constructed base on the covariance table and variables may be removed if they have large p-

value. Adjust the variables until the mean square error reaches the lowest point, and record the multiple R square.

The last model considered is random forest, random forest model has its unique advantage, according to Li(2013), "Random forest does not need to worry about the multivariate collinearity problem faced by general regression analysis, and does not need to make variable selection. Existing random forest packages give the importance of all variables. In addition, random forests facilitate the calculation of nonlinear effects of variables and can reflect the interaction between variables"[7]. Randomforest package is used to construct the model, with ntree equals to 500 and mtry equals to 3. Adjusting the variables according to the importance plot.

2.4. Evaluation metric

For the linear regression model and polynomial model, multiple R square and 10-fold cross validation mean square error are used to evaluate the model. As for the random forest model, only mean square error is considered. According to Miles(2005), "R square represents the proportion of variance in the outcome variable which is explained by the predictor variables in the sample"[8]. The mean square error measures the error of using the model to make a prediction, according to Allen(1971), "Mean square error is believed to be more meaningful than the commonly used criterion, the residual sum of squares."[9]. Using k-fold cross validation can decrease the randomness of the MSE, according to Berrar(2018), "Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters. Ten-fold stratified cross-validation is often applied in practice. "[10] . Therefore, 10-fold cross validation is chosen for the research.

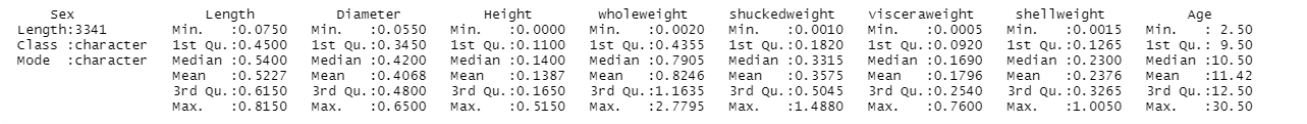
In conclusion, the larger the R square, the smaller the MSE, indicating the better the model.

3. Results and Discussion

3.1. Data Visualization

Table 1. Summary of the Variables

Name	Type	Unit	Description
Sex	nominal	--	M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried



Sex	Length	Diameter	Height	wholeweight	shuckedweight	visceraweight	shellweight	Age
Length:3341	Min. :0.0750	Min. :0.0550	Min. :0.0000	Min. :0.0020	Min. :0.0010	Min. :0.0005	Min. :0.0015	Min. : 2.50
Class :character	1st Qu.:0.4500	1st Qu.:0.3450	1st Qu.:0.1100	1st Qu.:0.4355	1st Qu.:0.1820	1st Qu.:0.0920	1st Qu.:0.1265	1st Qu.: 9.50
Mode :character	Median :0.5400	Median :0.4200	Median :0.1400	Median :0.7905	Median :0.3315	Median :0.1690	Median :0.2300	Median :10.50
	Mean :0.5227	Mean :0.4068	Mean :0.1387	Mean :0.8246	Mean :0.3575	Mean :0.1796	Mean :0.2376	Mean :11.42
	3rd Qu.:0.6150	3rd Qu.:0.4800	3rd Qu.:0.1650	3rd Qu.:1.1635	3rd Qu.:0.5045	3rd Qu.:0.2540	3rd Qu.:0.3265	3rd Qu.:12.50
	Max. :0.8150	Max. :0.6500	Max. :0.5150	Max. :2.7795	Max. :1.4880	Max. :0.7600	Max. :1.0050	Max. :30.50

Fig. 1 Basic distribution of the data (Photo credit: Original)

As shown in the Table 1 and Fig. 1, the summary of the data gives tells us the basic distribution of variables.

3.2. Variable Selection and correlation analysis

```
Call:
glm(formula = Age ~ ., data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9533  -1.3160  -0.3275   0.8819  14.3552

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.21390    0.32702  15.943  < 2e-16 ***
SexI          -0.91010    0.11297  -8.056 1.08e-15 ***
SexM          -0.07089    0.09255  -0.766 0.443755
Length        -0.34182    1.99532  -0.171 0.863988
Diameter       8.55059    2.46296   3.472 0.000524 ***
Height        21.71907    2.41407   8.997  < 2e-16 ***
wholeweight    7.52607    0.80011   9.406  < 2e-16 ***
shuckedweight -18.32787    0.89956 -20.374  < 2e-16 ***
visceraweight  -9.30065    1.42550  -6.524 7.86e-11 ***
shellweight    9.40449    1.25838   7.473 9.93e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.716761)

    Null deviance: 34584  on 3340  degrees of freedom
Residual deviance: 15712  on 3331  degrees of freedom
AIC: 14676
```

Fig. 2 Result of the initial linear regression (Photo credit: Original)

First, as shown in the Fig. 2, fit the model with its original variables in order to do a basic variable selection. According to the result, sex Male has a significantly large p-value, which means male is basically the same as female when making a prediction of the age. Therefore, Male and Female can be treated as the same in the model. To make it convenient, replace all the Male in the data with Female.

Then check the covariance table, and we found there are significant correlation between some variables such as length and diameter, height and diameter, viscera weight and whole weight and so on. To deal with the correlation problem, new variables such as length*diameter and viscera weight* whole weight should be constructed to make the model more accurate.

3.3. Model training and evaluation

3.3.1 Linear regression model

The first model considered is Linear Regression Model, which is one of the most basic one. First, directly fit the linear model, and use 10-fold cross validation to calculate the mean square error.

```
[1] 4.753203

Call:
glm(formula = Age ~ ., data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9533  -1.3160  -0.3275   0.8819  14.3552

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.21390    0.32702  15.943  < 2e-16 ***
SexI          -0.91010    0.11297  -8.056 1.08e-15 ***
SexM          -0.07089    0.09255  -0.766 0.443755
Length        -0.34182    1.99532  -0.171 0.863988
Diameter       8.55059    2.46296   3.472 0.000524 ***
Height        21.71907    2.41407   8.997  < 2e-16 ***
wholeweight    7.52607    0.80011   9.406  < 2e-16 ***
shuckedweight -18.32787    0.89956 -20.374  < 2e-16 ***
visceraweight  -9.30065    1.42550  -6.524 7.86e-11 ***
shellweight    9.40449    1.25838   7.473 9.93e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.716761)

    Null deviance: 34584  on 3340  degrees of freedom
Residual deviance: 15712  on 3331  degrees of freedom
AIC: 14676

Number of Fisher Scoring iterations: 2
```

Fig. 3 The mean square error of the initial fit (Photo credit: Original)

As shown in the Fig. 3, the Sex male has a very large p value just like discussed in 3.2. Then checked the covariance table, and it's found that there are correlation between variables, so then construct new variables by multiplying those correlated variables to solve the problem.

```
set.seed(1)
ntrain_data<-train_data
ntrain_data[,10]<-train_data$wholeweight*train_data$shuckedweight
ntrain_data[,11]<-train_data$wholeweight*train_data$visceraweight
ntrain_data[,12]<-train_data$wholeweight*train_data$shellweight
ntrain_data[,13]<-train_data$shuckedweight*train_data$visceraweight
ntrain_data[,14]<-train_data$shuckedweight*train_data$shellweight
ntrain_data[,15]<-train_data$visceraweight*train_data$shellweight
ntrain_data[,16]<-train_data$wholeweight*train_data$shuckedweight*train_data$visceraweight
ntrain_data[,17]<-train_data$wholeweight*train_data$shuckedweight*train_data$shellweight
ntrain_data[,18]<-train_data$shellweight*train_data$shuckedweight*train_data$visceraweight
ntrain_data[,19]<-train_data$wholeweight*train_data$shellweight*train_data$visceraweight
ntrain_data[,20]<-train_data$wholeweight*train_data$shuckedweight*train_data$visceraweight*train_data$shellweight
ntrain_data[,21]<-train_data$Length*train_data$Diameter
ntrain_data[,22]<-train_data$Height*train_data$Diameter
cv.error.10=rep(0,10)
for (i in 1:10){
lm.fit=glm(Age~.,data=ntrain_data)
cv.error.10[i]=cv.glm(ntrain_data, lm.fit, K=10)$delta[1]
}
mse=mean(cv.error.10)
mse
summary(lm.fit)
```

Fig. 4 Constructing new variables for correlation problem (Photo credit: Original)

As shown in the Fig. 4, 13 new variables have been constructed for possible correlation. After the construction of new variables, fit the linear regression model again.

```
[1] 4.512397

Call:
glm(formula = Age ~ ., data = ntrain_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.0068  -1.2416  -0.2710   0.8722  14.9159

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.73342    0.88452   5.351 9.32e-08 ***
SexI          -0.70675    0.11362  -6.220 5.58e-10 ***
SexM          -0.08690    0.08964  -0.970 0.332353
Length        -7.56233    4.09693  -1.846 0.065003 .
Diameter      10.95683    4.18613   2.617 0.008900 **
Height        36.31055   11.28592   3.217 0.001306 **
wholeweight    2.27432    3.38168   0.673 0.501286
shuckedweight -12.61642    3.22298  -3.915 9.24e-05 ***
visceraweight   6.81970    8.35269   0.816 0.414291
shellweight    45.14410    5.54031   8.148 5.17e-16 ***
...10          19.65814    5.31320   3.700 0.000219 ***
...11          -1.60778    9.72335  -0.165 0.868676
...12          12.73823    7.55017   1.687 0.091670 .
...13         -37.88173   22.53413  -1.681 0.092841 .
...14        -114.35684   18.71000  -6.112 1.10e-09 ***
...15         -55.96691   33.74852  -1.658 0.097341 .
...16         -14.37548   11.16156  -1.288 0.197855
...17          -3.51523    8.70986  -0.404 0.686539
...18         293.75256   68.84302   4.267 2.04e-05 ***
...19         -30.16896   17.18503  -1.756 0.079260 .
...20         -25.03154   13.98733  -1.790 0.073612 .
...21           0.49728    8.57118   0.058 0.953738
...22         -56.03795   24.48193  -2.289 0.022145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.405646)

    Null deviance: 34584  on 3340  degrees of freedom
Residual deviance: 14618  on 3318  degrees of freedom
AIC: 14461

Number of Fisher Scoring iterations: 2
```

Fig. 5 The mse after adding the variables (Photo credit: Original)

As shown in the Fig. 5, the mean square error has decreased significantly, but there are some newly constructed variables that have large p-value. Then remove the variables with significantly large p-value. The first one chosen to remove is the SexM, like the discussion before, replace all male with female.

```
[1] 4.510448

Call:
glm(formula = Age ~ ., data = ntrain_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.0388  -1.2414  -0.2724   0.8738  14.9525

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.6968     0.8837   5.315 1.14e-07 ***
SexI            -0.6560     0.1009  -6.504 8.98e-11 ***
Length          -7.7083     4.0941  -1.883 0.059821 .
Diameter        10.9192     4.1859   2.609 0.009133 **
Height          36.4625    11.2847   3.231 0.001245 **
wholeweight      2.3061     3.3815   0.682 0.495309
shuckedweight  -12.6359     3.2229  -3.921 9.01e-05 ***
visceraweight    6.9831     8.3509   0.836 0.403095
shellweight     45.1333     5.5403   8.146 5.25e-16 ***
...10           19.6779     5.3131   3.704 0.000216 ***
...11           -1.8575     9.7199  -0.191 0.848459
...12           12.7059     7.5500   1.683 0.092490 .
...13          -38.1548    22.5322  -1.693 0.090483 .
...14         -114.6957    18.7066  -6.131 9.74e-10 ***
...15          -55.6399    33.7465  -1.649 0.099292 .
...16          -14.1708    11.1595  -1.270 0.204229
...17           -3.4175     8.7092  -0.392 0.694790
...18          294.4376    68.8388   4.277 1.95e-05 ***
...19          -29.9918    17.1839  -1.745 0.081018 .
...20          -25.4933    13.9791  -1.824 0.068291 .
...21            0.8380     8.5639   0.098 0.922055
...22          -56.3826    24.4791  -2.303 0.021324 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 6 Result of the model after replacing male with female (Photo credit: Original)

As shown in the Fig. 6, it can be seen that the mean square error decreased slightly, but there are still many variables with p-value larger than 0.05. Next, remove the variable with the largest p-value one at a time, and check the 10-fold cross validation mean square error every time a variable is removed, stop removing variable until the mean square error begin to increase, and that should be the final linear regression model.

```r
set.seed(1)
newtrain_data<-select(ntrain_data,c(-21,-17,-15,-16,-11,-7,-12,-20))
cv.error.10=rep(0,10)
for (i in 1:10){
lm.fit=glm(Age~.,data=newtrain_data)
cv.error.10[i]=cv.glm(newtrain_data, lm.fit, K=10)$delta[1]
}
mse=mean(cv.error.10)
mse
summary(lm.fit)
```

```
[1] 4.480284

Call:
glm(formula = Age ~ ., data = newtrain_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.4777  -1.2585  -0.2709   0.8707  15.1817

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.28254    0.59337   7.217 6.54e-13 ***
SexI          -0.65180    0.09913  -6.575 5.62e-11 ***
Length        -7.41916    2.10585  -3.523 0.000432 ***
Diameter      13.23968    3.17786   4.166 3.18e-05 ***
Height        44.00530    6.65892   6.608 4.50e-11 ***
wholeweight    4.71724    1.31581   3.585 0.000342 ***
shuckedweight -15.79767    1.66231  -9.503  < 2e-16 ***
shellweight    34.93903    2.89646  12.063  < 2e-16 ***
...10          10.78154    1.41032   7.645 2.72e-14 ***
...13         -25.02219    4.42745  -5.652 1.72e-08 ***
...14         -49.96441    5.81697  -8.589  < 2e-16 ***
...18          79.20914   16.72285   4.737 2.26e-06 ***
...19         -20.49349    4.64710  -4.410 1.07e-05 ***
...22         -72.44806   14.68602  -4.933 8.49e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 4.42757)

    Null deviance: 34584  on 3340  degrees of freedom
Residual deviance: 14731  on 3327  degrees of freedom
AIC: 14468

Number of Fisher Scoring iterations: 2
```

Fig. 7 Final version of linear model (Photo credit: Original)

The final version of linear regression is shown in the Fig. 7. Therefore, for the linear regression model, the lowest mean square error is 4.480284, and after getting the final model, use the function "lm()" to refit the model in order to get the multiple R square.

```r
nlm.fit=lm(Age~.,data=newtrain_data)
summary(nlm.fit)
```

```
Call:
lm(formula = Age ~ ., data = newtrain_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4777 -1.2585 -0.2709  0.8707 15.1817

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.28254    0.59337   7.217 6.54e-13 ***
SexI            -0.65180    0.09913  -6.575 5.62e-11 ***
Length          -7.41916    2.10585  -3.523 0.000432 ***
Diameter        13.23968    3.17786   4.166 3.18e-05 ***
Height          44.00530    6.65892   6.608 4.50e-11 ***
wholeweight      4.71724    1.31581   3.585 0.000342 ***
shuckedweight -15.79767    1.66231  -9.503  < 2e-16 ***
shellweight     34.93903    2.89646  12.063  < 2e-16 ***
...10           10.78154    1.41032   7.645 2.72e-14 ***
...13          -25.02219    4.42745  -5.652 1.72e-08 ***
...14          -49.96441    5.81697  -8.589  < 2e-16 ***
...18           79.20914   16.72285   4.737 2.26e-06 ***
...19          -20.49349    4.64710  -4.410 1.07e-05 ***
...22          -72.44806   14.68602  -4.933 8.49e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.104 on 3327 degrees of freedom
Multiple R-squared:  0.5741,    Adjusted R-squared:  0.5724
F-statistic: 344.9 on 13 and 3327 DF,  p-value: < 2.2e-16
```

Fig. 8 Result of the lm() function (Photo credit: Original)

The result of lm() is shown in the Fig. 8, and the multiple R square for the best linear model is 0.5741, which is not a bad result for a real life problem.

In conclusion, for the linear regression model, the mean square error is 4.480284, multiple R square is 0.5741, and the equation of the model is Age= -0.65*Sex(equals to 1 if it's Infant) - 7.42*Length + 13.24*Diameter + 44.01*Height + 4.72*Wholeweight - 15.80*shuckedweight + 34.94*shellweight + 10.78*whole*shucked - 25.02*shucked*viscera - 49.96*shucked*shell + 79.21*shell*shucked*viscera - 20.49*whole*shell*viscera - 72.45*Height*Diameter + 4.28 .

3.3.2 Polynomial regression model

Second model considered is the polynomial mode, it's similar to the linear regression model but it's more practical in the real world. First, use glm function to fit the model, turn all variables to second power, and calculate the 10-fold cross validation mean square error.

```r
set.seed(1)
cv.error.10=rep(0,10)
for (i in 1:10){
poly.fit=glm(Age~Sex+poly(Length,2)+poly(Diameter,2)+poly(Height,2)+poly(wholeweight,2)+poly(shuckedweight,2)+poly(viscerweight,2)+poly(shellweight,2),data=train_data)
cv.error.10[i]=cv.glm(train_data, poly.fit, K=10)$delta[1]
}
mean(cv.error.10)
summary(poly.fit)
```

```
[1] 4.472435

Call:
glm(formula = Age ~ Sex + poly(Length, 2) + poly(Diameter, 2) +
    poly(Height, 2) + poly(wholeweight, 2) + poly(shuckedweight,
    2) + poly(viscerweight, 2) + poly(shellweight, 2), data = train_data)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-7.2560  -1.2675  -0.2794   0.8920  15.6868

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.69008    0.06945 168.333  < 2e-16 ***
SexI                      -0.71379    0.11283  -6.326 2.85e-10 ***
SexM                      -0.09712    0.08958  -1.084   0.2784
poly(Length, 2)1         -36.56289   15.62091  -2.341   0.0193 *
poly(Length, 2)2         -20.35606    8.86707  -2.296   0.0218 *
poly(Diameter, 2)1        23.23508   15.43229   1.506   0.1323
poly(Diameter, 2)2        -6.17545    8.99388  -0.687   0.4924
poly(Height, 2)1          29.92885    6.04294   4.953 7.68e-07 ***
poly(Height, 2)2          -6.55848    2.75797  -2.378   0.0175 *
poly(wholeweight, 2)1    315.25496   26.16457  12.049  < 2e-16 ***
poly(wholeweight, 2)2    -64.67951   11.73551  -5.511 3.83e-08 ***
poly(shuckedweight, 2)1 -278.12200   12.79937 -21.729  < 2e-16 ***
poly(shuckedweight, 2)2   61.33628    6.66699   9.200  < 2e-16 ***
poly(viscerweight, 2)1   -64.44991   10.49623  -6.140 9.21e-10 ***
poly(viscerweight, 2)2    16.05768    5.81714   2.760   0.0058 **
poly(shellweight, 2)1    109.63537   12.35103   8.877  < 2e-16 ***
poly(shellweight, 2)2     -9.73013    5.34988  -1.819   0.0690 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.411689)

    Null deviance: 34584  on 3340  degrees of freedom
Residual deviance: 14664  on 3324  degrees of freedom
AIC: 14459

Number of Fisher Scoring iterations: 2
```

Fig. 9 Initial polynomial regression (Photo credit: Original)

According to the result in Fig. 9, the mean square error of this basic polynomial model is 4.472435, which is already lower than the best version of linear model, thus polynomial model definitely is

better than linear model in this regression problem. Just like the linear model, the p-value of the SexM is still large in the polynomial model, thus replace all male with female.

```
[1] 4.470623

Call:
glm(formula = Age ~ Sex + poly(Length, 2) + poly(Diameter, 2) +
    poly(Height, 2) + poly(wholeweight, 2) + poly(shuckedweight,
    2) + poly(visceraweight, 2) + poly(shellweight, 2), data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.3048  -1.2668  -0.2903   0.8873  15.7309

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.63654    0.04883 238.322  < 2e-16 ***
SexI                      -0.65705    0.09996  -6.573 5.70e-11 ***
poly(Length, 2)1         -36.53444   15.62130  -2.339  0.01941 *
poly(Length, 2)2         -20.34316    8.86730  -2.294  0.02184 *
poly(Diameter, 2)1        23.74473   15.42654   1.539  0.12385
poly(Diameter, 2)2        -6.03339    8.99316  -0.671  0.50234
poly(Height, 2)1          29.97289    6.04296   4.960 7.40e-07 ***
poly(Height, 2)2          -6.61079    2.75762  -2.397  0.01657 *
poly(wholeweight, 2)1    315.42362   26.16480  12.055  < 2e-16 ***
poly(wholeweight, 2)2    -64.79913   11.73530  -5.522 3.61e-08 ***
poly(shuckedweight, 2)1 -279.01571   12.77313 -21.844  < 2e-16 ***
poly(shuckedweight, 2)2   61.41213    6.66680   9.212  < 2e-16 ***
poly(visceraweight, 2)1  -64.01542   10.48885  -6.103 1.16e-09 ***
poly(visceraweight, 2)2   16.02462    5.81722   2.755  0.00591 **
poly(shellweight, 2)1    109.49840   12.35071   8.866  < 2e-16 ***
poly(shellweight, 2)2     -9.63414    5.34929  -1.801  0.07179 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 4.411921)

    Null deviance: 34584  on 3340  degrees of freedom
Residual deviance: 14670  on 3325  degrees of freedom
AIC: 14458

Number of Fisher Scoring iterations: 2
```

Fig. 10 Result after replacing male with female (Photo credit: Original)

As shown in the Fig. 10, after replacing male with female, the mean square also decrease slightly like the linear model. Then check the covariance table, and adjust the power of each variables, remove those with large p-value, stop until the mean square error reaches the lowest point. Considering in the polynomial model, there will be much more combination between variables, thus we construct new variable according to the covariance table one at a time in order to make the model less complicated.

```
set.seed(1)
cv.error.10=rep(0,10)
for (i in 1:10){
poly.fit=glm(Age~Sex+poly(Length,2)+poly(Diameter,1)+poly(Height,2)+poly(wholeweight,2)+poly(shuckedweight,4)+poly(visceraweight,2)+poly(shellweight,2)+poly(wholeweight*shuc
kedweight,1),data=train_data)
cv.error.10[i]=cv.glm(train_data, poly.fit, K=10)$delta[1]
}
mean(cv.error.10)
summary(poly.fit)
```

```
[1] 4.35839

Call:
glm(formula = Age ~ Sex + poly(Length, 2) + poly(Diameter, 1) +
    poly(Height, 2) + poly(wholeweight, 2) + poly(shuckedweight,
    4) + poly(visceraweight, 2) + poly(shellweight, 2) + poly(wholeweight *
    shuckedweight, 1), data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2336  -1.2305  -0.2745   0.8801  15.6119

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         11.63747    0.04828 241.050  < 2e-16 ***
SexI                                -0.65988    0.09890  -6.672 2.94e-11 ***
poly(Length, 2)1                   -33.95297   14.34759  -2.366 0.018016 *
poly(Length, 2)2                   -19.65262    4.66830  -4.210 2.62e-05 ***
poly(Diameter, 1)                   23.67555   14.05902   1.826 0.067899 .
poly(Height, 2)1                    28.75734    5.97289   4.815 1.54e-06 ***
poly(Height, 2)2                    -6.77615    2.71068  -2.500 0.012474 *
poly(wholeweight, 2)1              749.56478   56.01761  13.381  < 2e-16 ***
poly(wholeweight, 2)2               79.57658   21.18907   3.756 0.000176 ***
poly(shuckedweight, 4)1            187.48014   55.61724   3.371 0.000758 ***
poly(shuckedweight, 4)2            228.15871   20.94653  10.892  < 2e-16 ***
poly(shuckedweight, 4)3             -2.62825    3.35429  -0.784 0.433360
poly(shuckedweight, 4)4             -8.28237    2.66007  -3.114 0.001864 **
poly(visceraweight, 2)1            -60.09921   10.39047  -5.784 7.97e-09 ***
poly(visceraweight, 2)2             17.75552    5.75550   3.085 0.002052 **
poly(shellweight, 2)1              102.56329   12.23713   8.381  < 2e-16 ***
poly(shellweight, 2)2              -16.47109    5.40100  -3.050 0.002309 **
poly(wholeweight * shuckedweight, 1) -942.57410  109.65099  -8.596  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 4.309025)

    Null deviance: 34584  on 3340  degrees of freedom
Residual deviance: 14319  on 3323  degrees of freedom
AIC: 14382

Number of Fisher Scoring iterations: 2
```

Fig. 11 Model after adjusting (Photo credit: Original)

Final version is shown in the Fig. 11, the mean square error is 4.35839, and when trying to construct another new variables, no matter what variable is being constructed, the mean square error can not be decreased. Therefore, this would be the best version of polynomial model. Using the function lm(),

```
poly=lm(Age~Sex+poly(Length,2)+poly(Diameter,1)+poly(Height,2)+poly(wholeweight,2)+poly(shuckedweight,4)+poly(viscerweight,2)+poly(shellweight,2)+poly(wholeweight*shuckedweight,1),data=train_data)
summary(poly)
```

```
Call:
lm(formula = Age ~ Sex + poly(Length, 2) + poly(Diameter, 1) +
    poly(Height, 2) + poly(wholeweight, 2) + poly(shuckedweight,
    4) + poly(viscerweight, 2) + poly(shellweight, 2) + poly(wholeweight *
    shuckedweight, 1), data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2336 -1.2305 -0.2745  0.8801 15.6119

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       11.63747    0.04828 241.050  < 2e-16 ***
SexI                              -0.65988    0.09890  -6.672 2.94e-11 ***
poly(Length, 2)1                 -33.95297   14.34759  -2.366 0.018016 *
poly(Length, 2)2                 -19.65262    4.66830  -4.210 2.62e-05 ***
poly(Diameter, 1)                 25.67555   14.05902   1.826 0.067899 .
poly(Height, 2)1                  28.75734    5.97289   4.815 1.54e-06 ***
poly(Height, 2)2                  -6.77615    2.71068  -2.500 0.012474 *
poly(wholeweight, 2)1            749.56478   56.01761  13.381  < 2e-16 ***
poly(wholeweight, 2)2             79.57658   21.18907   3.756 0.000176 ***
poly(shuckedweight, 4)1          187.48014   55.61724   3.371 0.000758 ***
poly(shuckedweight, 4)2          228.15871   20.94653  10.892  < 2e-16 ***
poly(shuckedweight, 4)3           -2.62825    3.35429  -0.784 0.433360
poly(shuckedweight, 4)4           -8.28237    2.66007  -3.114 0.001864 **
poly(viscerweight, 2)1           -60.09921   10.39047  -5.784 7.97e-09 ***
poly(viscerweight, 2)2            17.75552    5.75550   3.085 0.002052 **
poly(shellweight, 2)1            102.56329   12.23713   8.381  < 2e-16 ***
poly(shellweight, 2)2            -16.47109    5.40100  -3.050 0.002309 **
poly(wholeweight * shuckedweight, 1) -942.57410  109.65099  -8.596  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.076 on 3323 degrees of freedom
Multiple R-squared:  0.586,     Adjusted R-squared:  0.5838
F-statistic: 276.6 on 17 and 3323 DF,  p-value: < 2.2e-16
```

Fig. 12 result of lm() function (Photo credit: Original)

In the Fig. 12, the multiple R square of the polynomial model is shown, which is 0.586, and it's larger than the linear regression model.

In conclusion, the best version of polynomial model has mean square error 4.35839 and multiple R square 0.586, which is more accurate than the linear model. The polynomial model is

Age=11.64 - 0.66*Sex(equals to 1 if it's Infant) - 33.95*Length - 19.65*Length^2 + 25.68*Diameter + 28.76*Height - 6.78*Height^2 + 749.56*whole weight + 79.58*whole weight^2 + 187.48*shucked weight + 228.16*shucked weight^2 - 2.63*shucked weight^3 - 8.28*shucked weight^2 - 60.10*viscera weight + 17.76*viscera weight^2 + 102.56*shell weight - 16.47*shell weight^2 - 942.57*whole weight *shell weight .

3.3.3 Random Forest

The last model considered is random forest, with ntree equals to 500 and mtry equals to 3. Using the randomforest package, fit the model.

```
library(randomForest)
set.seed(1)
cv.error.10=rep(0,10)

for (i in 1:10){
  rfm=randomForest(Age~.,data=train_data,
                   ntree=500,
                   mtry=3,
                   proximity=TRUE,
                   importance=TRUE
                  )
cv.error.10[i]=cv.glm(train_data, rfm, K=10)$delta[1]
}
summary(rfm)
mean(cv.error.10)
```

```
                Length   Class  Mode
call                 7 -none- call
type                 1 -none- character
predicted         3341 -none- numeric
mse                500 -none- numeric
rsq                500 -none- numeric
oob.times         3341 -none- numeric
importance          16 -none- numeric
importanceSD         8 -none- numeric
localImportance      0 -none- NULL
proximity     11162281 -none- numeric
ntree                1 -none- numeric
mtry                 1 -none- numeric
forest              11 -none- list
coefs                0 -none- NULL
y                 3341 -none- numeric
test                 0 -none- NULL
inbag                0 -none- NULL
terms                3 terms  call
[1] 4.475051
```

Fig. 13 Initial random forest model (Photo credit: Original)

The mean square error in Fig. 13 shows it's not a bad result, and then check the importance plot. Replace all male with female just like the first two models

```
train_data[train_data$Sex=="M",1]<-"F"
set.seed(1)
cv.error.10=rep(0,10)
for (i in 1:10){
rfm=randomForest(Age~.,data=train_data,
                 ntree=500,
                 mtry=3,
                 proximity=TRUE,
                 importance=TRUE
                )
cv.error.10[i]=cv.glm(train_data, rfm, K=10)$delta[1]
}
summary(rfm)
mean(cv.error.10)
```

```
                Length   Class  Mode
call                 7 -none- call
type                 1 -none- character
predicted         3341 -none- numeric
mse                500 -none- numeric
rsq                500 -none- numeric
oob.times         3341 -none- numeric
importance          16 -none- numeric
importanceSD         8 -none- numeric
localImportance      0 -none- NULL
proximity     11162281 -none- numeric
ntree                1 -none- numeric
mtry                 1 -none- numeric
forest              11 -none- list
coefs                0 -none- NULL
y                 3341 -none- numeric
test                 0 -none- NULL
inbag                0 -none- NULL
terms                3 terms  call
[1] 4.466598
```

Fig. 14 The importance plot (Photo credit: Original)

As shown in Fig. 14, after the replacement, the mean square error do decrease slightly, then when trying to remove variable with large p-value or add new variables constructed by correlated variable, the mean square error do not decrease anymore.

In conclusion, the final version of random forest has mean square error 4.466598, which is lower than linear model and larger than polynomial model. The data set has some limitations that may affect the research. First, the data is collected in 27 years ago, as the change of environment, the data may not be the best option to fit the model today. What's more, there are different kinds of abalone in the whole world, and the dataset didn't only contains 4177 instances, which may be not enough to cover all kinds of abalone, thus may affect the accuracy of the model.

3.4. Limitation

The data set has some limitations that may affect the research. First, the data is collected in 27 years ago, as the development of environment, the data may not be the best option to fit the model today. What's more, there are different varieties of abalone in the whole world, and different kind of abalone has different physical characters. However, the data only contains 4177 instances and does not categorized abalone according to its varieties, thus the model may be not accurate for some other species of abalone.

As for the model training process, limitations also exist. For the polynomial model and the random forest model, there may be better combination of independent variables, however, the number is too large and we can't try all of them. Moreover, running the random forest model in a regular computer is really a time-consuming process, which makes it difficult to adjust the model and try new possibilities. The correlation problem may be able to solve in a better way with different combinations of independent variables for all three models.

Finally, the criteria for model evaluation are relatively simple, more complete evaluation of the models are possible.

4. Conclusion

In conclusion, studying about abalone is a meaningful and beneficial work for many people and countries whose lives and economy rely on it, and constructing a model to predict its age can save researchers a lot of time and work. In this research, three models are considered to predict the age of abalone, the linear regression model, polynomial regression model and the random forest, R-squares and mean square error are used to evaluate the model. Moreover, 10-fold cross validation is used to calculate the mean square error in order to decrease the randomness. After processing the data, analyzing the variables, constructing new variables for the correlation problem and variables selection, polynomial model appears to be the best model among three, it has the lowest 10-fold mean square error and largest multiple R-square. Though random forest has a higher mean square error, it has shown great potential in making the predictions. However, because of the limitation of computer, further development in random forest is difficult, but it would not be a surprise if an updated random forest model has lower mean square error. This research provide a new model to predict the age of abalone according to its physical measurements and a direction to do further research, it is making the research about abalone more convenient and thus making contribution to the development of abalone industry.

References

[1] 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 the Islands of Bass Strait. Sea Fisheries Division. Marine Research Laboratories - Taroona, Department of Primary Industry and Fisheries, Tasmania

[2] Cook, P. (2014) The Worldwide Abalone Industry. Modern Economy, 5, 1181-1186. doi: 10.4236/me.2014.513110.

[3] Caihuan Ke. (2013). Current situation and prospect of abalone aquaculture industry in China. Chinese Fisheries (1), 4.

[4] Taohua Liu, & Muzi Hou. (2016). Discriminant analysis and cluster analysis in the abalone age classification. Journal of Shaoyang University: Natural Science Edition, 13 (1), 5.

[5] Hang XING. (2007). Establishment and reliability test of multiple linear regression model. Journal of Vocational University (4), 3.

[6] Linghui Fu, & Wang Huiwen. (2004). A comparative study of modeling methods for polynomial regression. Journal of Mathematical Statistics and Management, 23(1), 5.

[7] Xinhai LI. (2013). Application of random forest model in classification and regression analysis. Chinese Journal of Applied Entomology, 50(4), 8.

[8] Miles, J. (2005). R-Squared, Adjusted R-Squared. In Encyclopedia of Statistics in Behavioral Science (eds B.S. Everitt and D.C. Howell). https://doi.org/10.1002/0470013192.bsa526

[9] David M. Allen (1971) Mean Square Error of Prediction as a Criterion for Selecting Variables, Technometrics, 13:3, 469-475, DOI: 10.1080/00401706.1971.10488811

[10] Berrar, D. (2019). Cross-Validation.