A New Method of Measuring the Age of Abalone Based on Data Visualization Analysis

Runze Guo, Junmin Luo and Wuqi Gao*
School of Computer Science and Engineering, Xi’an Technological University, Xi’an, China

*Corresponding author e-mail: gaowuqi@xatu.edu.cn

Abstract. This project uses a new way to count the abalone age, which use abalone’s physical characteristics to predict by multiple linear regression. After the model is trained, when we catch a new abalone, we can use a computer to replace the labor to a certain extent, saving costs to the enterprise. Results are given in a visualization of the data.

Keywords: Abalone, Multi-Linear Regression, Data Visualization

1. Introduction

Predicting values of numeric or continuous attributes is known as regression in statistics. Predicting real values is an important challenge for machine learning because many problems encountered in real life involve regression [1]. For a long time, people who want to know the age of abalone have to be achieved through physical measurement. Judging the age of abalone is a boring and time-consuming task. It needs to cut the abalone shell through the cone, then staining it, and then observe the number of its rings through a microscope to determine it [2]. However, there may be more than one way to know the age of abalone. We can predict the age of abalone through some more easily obtained measurements. For example: weather patterns and location. The purpose of this analysis is to use abalone multi-dimension data to predict the age of abalone. These estimates can be used to create a new way to predict to help frontline workers or factory and let them work on computers instead of manual methods, which can reduce production costs. From our own prior knowledge, it is more suitable to use the regression method. Thus, we want to build muti-linear regression models to do some explore.

Figure 1. Abalone
Table 1. Advantages and disadvantages of multiple linear regression

| Advantages                                                                 | Disadvantages                                      |
|---------------------------------------------------------------------------|----------------------------------------------------|
| Up to now, It is the most commonly used method for numerical data modeling | Made strong assumptions about the data.            |
| Applicable to almost all data.                                            | The form of the model must be specified by the user in advance. |
| Provides an estimate of the strength and size of the relationship between features (variables) and results. | Does not handle missing data well. |
|                                                                           | Only numerical features can be processed, so classification data requires additional processing. |

2. Dataset

Data comes from an original (non-machine-learning) study: "The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288) and donated by Sam Waugh (Department of Computer Science, University of Tasmania)[3]. We download the dataset from UCL website.

Table 2. Data characteristics description table

| Name         | Data Type | Measurement 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                               |
| Rings        | integer   | --               | +1.5 gives the age in years                      |

3. Methodology

3.1. Collect/Observe Data

For analysis purposes, this article uses a data set from a web page. The data set contains the real attribute values of multiple abalone, including Sex, Length, Diameter, Height, Whole. Weight, Shucked. Weight, Viscera. Weight, Shell. Weight, Rings and other properties [4]. Therefore, it can roughly reflect the situation of abalone in the real world.

3.2. Explore and Prepare Data
The first step is to prepare data. We need to deal with missing values. Generally, there will be some problems with the dataset we got for the first time [5]. Thus, we use the ‘sum’ function to determine if there are missing values and the completeness of the data. The results are shown below:

```r
> sum(is.na(abalone))
[1] 0
> sum(complete.cases(abalone))
[1] 4177
```

**Figure 2.** Data prepared results

As the result shown, this is a complete data set!

After the previous step, the appropriate data set has been collected in this article, which needs to be initially processed in this article. In this section, in general, one may need to check whether values are available, convert values, detect/eliminate outliers, or some other section. However, in this data set, after testing, the data set basically meets the conditions for further processing [6]. Therefore, the relationship between features (correlation coefficient matrix) and the relationship between visual features (scatter diagram matrix) are mainly explored in this part. In doing so, this article calls the functions of cor (), pairs (), and pairs. Panels () to complete the determination. The first two functions can be called directly from Rstudio, while the last requires the psych package to be pre-installed in Rstudion. The final results are as follows (due to space constraints, only visualizations are listed here. To view additional results, see the source code):

**Figure 3.** Scatter diagram matrix

In view of the above results, all variables except Height are positively correlated with abalone age. Therefore, in the initial variable selection, this article will take all variables except the Height variable as the dependent variables.

### 3.3. Train Model Based On Data

This paper adopts multiple linear regression model for modeling. The function lm () that needs to call is included by default in Rstudio software, so people can call it directly. After modeling the six variables selected above, the results are as follows (sex variable will not be considered because it is a factor value and make the process more complex):
3.4. Evaluate the Performance of the Model

After getting the model of training, this paper needs to evaluate the performance of the model. In doing so, the functions of summary() and plot are called to perform significance and residual analysis of the function [7].

The final result is as follows:

```
call:  
ln(formula = rings ~ length + diameter + whole.weight + shucked.weight +  
viscera.weight + shell.weight, data = abalone)

coefficients:  
(Intercept)     length    diameter   whole.weight  shucked.weight   viscera.weight  shell.weight  
            3.307       -1.194      15.345         9.362          -20.437          -9.340         9.485

Residuals:  
Min      1Q    Median      3Q      Max  
-7.8722 -1.3770  0.9027  13.9450

Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept)    3.3076    0.2676  12.358  < 2e-16 ***  
Length         5.3198    1.4366   3.705  0.0001 **  
Diameter       2.2717    6.8695   0.341  0.7348    
Whole.weight   -0.8246    2.4780  -0.336  0.7382    
Shucked.weight -0.2460    2.4780  -0.100  0.9202    
Viscera.weight -0.7531    1.1414  -0.665  0.5098    
Shell.weight   -0.0132    1.1415  -0.011  0.9912    
```

It is not difficult to see from the result graph that the significance of the Length variable is not high enough. In addition, there is a faint contour line in the last figure (bottom right) of the residual analysis diagram, which means that there may be outliers in the model.

3.5. Improve the Performance of Model

After the performance evaluation, it is not difficult to find some problems in the current model. Since Length is not significant for dependent variables, the Length variable is removed later in this article for modeling. This article will take a step further after the performance assessment of the new model is completed. However, in the actual operation process, the results obtained by repeating steps 3 and 4 in this paper are found to be completely in line with the expectation. The details will be detailed in the next section.

4. Results

According to the above methods, the age prediction model of abalone adopted in this paper is as follows:

```
> shb.model15

call:  
ln(formula = rings ~ diameter + whole.weight + shucked.weight +  
viscera.weight + shell.weight, data = abalone)

Coefficients:  
(Intercept)  diameter   whole.weight  shucked.weight  viscera.weight  shell.weight  
            3.238       14.028       9.369          -20.480         -9.418         9.506
```

It is not difficult to see from the result graph that the significance of the Length variable is not high enough. In addition, there is a faint contour line in the last figure (bottom right) of the residual analysis diagram, which means that there may be outliers in the model.
Figure 7. Significance analysis of the final model and Residuals analysis of the final model

From the significance analysis results of the model, it can be seen that:

The Intercept of the model and the significance detection of other factors are all 0.01. Meanwhile, looking down, p-value is also less than 2.2e-16. Therefore, in terms of single independent variable and the whole independent variable, both of them passed the significance test for the dependent variable of abalone age. The value of f-statistic also indicates that there is a strong relationship between the abalone age variable being tested and its multiple independent variables. In addition, we can see the two values of Multiple r-squared and Adjusted r-squared, namely “goodness of fit” and “goodness of fit of correction”. These two values refer to the degree to which the regression equation fits the sample. Here we can see that the modified goodness of fit is 0.5204, which means that the fitting degree is only about 50%. For real-world regression models, it is not uncommon for r-squared values to be quite low, so the r-squared value of 0.5204 is not actually very bad. Given the nature of medical costs, some of these errors are noteworthy, but not surprising [8].

From the residual analysis results of the model, it can be seen that:

A. Residuals and fitting values (top left).
The data points appearing between the residual and the fitted value are roughly evenly distributed on both sides of the red line, showing a random distribution. The red line has no obvious shape characteristics and represents a smooth curve.

B. Residual q-q (top right).
The data points in the figure are arranged in a diagonal line, approximately tending to a straight line. Intuitively it's a normal distribution.

C. Standardized residual square root and fitting value (bottom left).
On both sides of the red line, the data points in the figure are distributed, and show a random distribution. Furthermore, without visible form features, the red line provides a smooth curve.

D. Standardized residuals and leverage (bottom right).
In the figure, the absence of the red contour line means that there are no irregular points in the data that influence the regression results in particular. Residuals of Conclusion comply with the hypothesis.

To summarize, the residuals meet the assumptions.

5. Discussion
In the above prediction model of abalone age, this paper realizes the modeling of multiple regression model for the prediction of abalone age by analyzing the parameters of Diameter, Whole, Weight and Shell [9]. Weight of abalone. The model can predict the age of abalone which is difficult to measure by using the physical data of abalone which is relatively easier to get. At the same time, the result analysis also proves that the significance analysis and residual analysis of the model also meet certain standards. Therefore, this model is a successful modeling analysis case [10].

In the process of experiment, this paper also tries other models. For example, all the attributes in the dataset are modeled as dependent variables. However, the results show that the Length variable has a low significance and its value is greater than 0.1. At the same time, according to the analysis of the
standardized residuals and the leverage value, there are abnormal points in the model. The details are shown in the following figure:

![Figure 8. Significance analysis and Residuals analysis of the compared model 1](image)

In addition, this article tried to Length, Diameter, Whole. The weight, Shucked. Weight, Viscera. Weight, Shell. Weight. The weight as the dependent variable to modeling. Compared with the previous model, the model omitted the height variable because it was the only one in the interaction diagram that decreased as the rings increased. The results show that, although the outliers are improved in the residual analysis, the Length variable still fails to meet the condition of 0.1 in the significance analysis. The results are as follows:

![Figure 9. Significance analysis and Residuals analysis of the compared model 2](image)

Through the comparison between the above experimental results and the modeling method adopted in this paper, it is not difficult to see that the modeling method implemented in this paper is the best. But there are also disadvantages. For example, in the significance analysis, the value of r-squared is not very large; In residual analysis, the data points only show a certain degree of uniform distribution, but not a perfect normal distribution. Finally, the modeling method adopted in this paper is relatively simple, only to study the performance of different linear models with different independent variables. In the following process, this paper may further use nonlinear regression, curve regression, lasso regression and other models to conduct in-depth research on this problem.

References

[1] Allhan Uysal and H. Altay Guvenir. Instance-Based Regression by Partitioning Feature Projections. Applied. 2004.

[2] UCL (1995) Abalone Data Set. Available: https://archive.ics.uci.edu/ml/datasets/Abalone. Last accessed 19th Mar 2020. W. Strunk Jr., E.B. White, The Elements of Style, third ed., Macmillan, New York, 1979.
[3] Sam Waugh (1995) "Extending and benchmarking Cascade-Correlation", PhD thesis, Computer Science Department, University of Tasmania.

[4] David Clark, Zoltan Schreter, Anthony Adams "A Quantitative Comparison of Dystal and Backpropagation", submitted to the Australian Conference on Neural Networks (ACNN'96).

[5] Michie, D., Spiegelhalter, D.J. and Taylor, C.C., 1994. Machine learning. Neural and Statistical Classification, 13(1994), pp.1-298.

[6] Lin, A., & Meyers, M. A. (2005). Growth and structure in abalone shell. Materials Science and Engineering: A, 390(1-2), 27-41.

[7] Chatterjee, S., & Hadi, A. S. (2015). Regression analysis by example. John Wiley & Sons.

[8] Daoud, J. I. (2017, December). Multicollinearity and regression analysis. In Journal of Physics: Conference Series (Vol. 949, No. 1, p. 012009). IOP Publishing.

[9] Vorobeychik, Y., & Kantarcioglu, M. (2018). Adversarial machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 12(3), 1-169.

[10] Raschka, S., & Mirjalili, V. (2017). Python machine learning. Packt Publishing Ltd.