**STA 206 Final Project**

**The Prediction of the Age of Abalone**

**Yu Zhu 914480813 zoeyu@ucdavis.edu**

**Yu Su 915488050 yusu@ucdavis.edu**

**Abstract**

Out of the interest of predicting the age of an abalone, we try to find the relationship between Rings, which represent the age, and other measurement of properties and build a model to achieve the prediction. In our project, we use a dataset with 4177 cases and 8 predicator variables to fit the model. We fit a first-order model, a second-order model and a model with some quadratic and interaction terms as our candidate models. We abandon the first-order model by internal validation and pick the second-order model, which contains some quadratic terms and interaction terms of numeric variables. After comparing all aspects of three candidate models, we get our final model with good ability of prediction. And then we remove all the outliers and fit the final model on the rest.

**Introduction**

**1. The relevant background of the topic**

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, may be used to predict the age.

**2. Questions of interest**

Firstly, we are interested in relationship between age (response) and other variables (predictors). Next, we want to build a variance-bias trade-off model. Furthermore, we want to predict the age of an abalone.

**3. The motivation of the project**

Since it is not easy to measure the age of an abalone directly, getting the age by other measurable properties of abalone is reasonable. It is a pretty common phenomenon in real world that we need to use other relevant properties to infer the response which is not easy to be measured or time and money consuming. Naturally, for our topic, we want to find the relationship between the size, weight, sex and age of abalone.

**4. The basic description of the dataset**

There’re 9 variables in the dataset, and three of them are to describe the size of the abalone: ‘Length’, ‘Diameter’ and ‘Height’. And four of them are to describe the weight of abalone: ‘Whole\_weight’, ‘Shucked\_weight’, ’Viscera\_weight’ and ‘Shell\_weight’. The variable ‘Sex’ is a qualitative variable with three levels. Our response discrete variable is ‘Rings’ which has direct relation with the age of abalone: +1.5 gives the age of the abalone. These variables were measured on 4177 abalones.

**Methods and Results**

**1. Processing of the data**

At the beginning, we classify all the variables (table 1). ‘Sex’ is factor, ‘Rings’ is integer, and the rest are numeric. In conclusion, except ‘Sex’ is a qualitative variable, others are quantitative variables.

Then we summarize the variables (table 2) and combing the information from table 1 and 2, there is no missing value in our dataset, our data is complete so that we can analyze them directly.

**1.1 Predictor variables**

First of all, we attempt to explore the distributions of each predictor variable:

**1.1.1 Quantitative:**

From the histograms of each numeric predictor variable (figure 1), “Length” and “Diameter”, are left-skewed and others are right-skewed. But variable “Height” concentrates around 0.2.

From the boxplots of each numeric variable (figure 2) and summary of the data, there are some obvious outliers and most of the variables are approximately symmetric. Also, all the numeric variables have similar scales so that correlation transformation is not necessary.

**1.1.2 Qualitative:**

From the pie chart (figure 3) for the only qualitative variable ‘Sex’ and the summary of this variable, we find that this variable has three levels: infant, male and female. Meanwhile, all of the levels have similar frequency.

**1.2 Response variable**

Next, we explore the response variable:

From the histogram of the response variable ‘Rings‘ (figure 5), it is a little bit right- skewed. So we may take some transformation next.

**1.3 Relationships among variables**

Furthermore, we want to find out the relationships among these variables:

From pairwise scatter plot (figure 4), we find that many pairwise variables are linearly or nonlinearly correlated. For example, it is pretty obvious that ‘Length’ is linearly correlated with ‘Diameter’ and ‘Height’, and nonlinearly correlated with ‘Whole\_weight’, ‘Shucked\_weight’, ‘Viscera\_weight’, and ‘Shell\_weight’. Besides, the response variable ‘rings’ is positively correlated with other variables.

From their pairwise correlation matrix (table 3), the response variable is positively correlated with other variables. Besides, some predictor variables are highly correlated with each other.

From the boxplot of response variable ‘Rings’ and predictor ‘sex’ (figure 6), female and male have similar distribution and level “I” has lower frequency. Still, there are some obvious outliers.

**2. Preliminary model investigation**

**2.1 Initial fitting**

We initially fit a first-order model with all variables and use it find out whether transformations are needed.

Since the histogram shows that the response is slightly right-skewed, we use box-cox procedure to find a proper transformation. The box-cox plot (figure 7) suggests a logarithm transformation () for the response variable. The histogram after transformation is more symmetric (figure 8). Meanwhile, we compare the residual plots between two models using original response variable and transformed response variable (figure 9). We can find that logarithm transformation is able to remedy departures from model assumptions.

However, since there are some outliers influence the whole distribution, we should identify these outstanding outliers in our model and remove them later.

**2.2 Linearity, normality and error variance consistency inspection**

We can first assume the full model is the correct model (fit 1).

Fit 1:

From the summary of fit 1 (figure 11), the variables are all significant at 0.01 level except ‘Length’ and ‘SexM’, the MSE is 0.041 and the R square is 59.91%.

From the residuals vs fitted value plot (figure 9), it suggests no obvious nonlinearity. The residuals present a pattern and the points do not randomly lie around 0, the reason might be that the response variable is discrete. From the Q-Q plot (figure 9), it suggests heavy-tail, which means it has some deviations from normal. So we need to fit a better model.

**3. Model selection (find a variance-bias-trade-off model)**

**3.1 First-order model selection**

Under the observation of the full model, we can find that the information variables contain are basically about the length and weight, which might lead to large multicolinearity among X variables. From the table of VIF of fit 1(figure 12), the vif of ‘Length’, ’Diameter’ and all the weight related variables are larger than 10, we can take it as an indication of high multicolinearity. The vif of ‘Whole\_weight’ is the largest: 111.65.

Combing the information from added-variable plots (figure 13), the variable ‘Whole\_weight’ doesn’t show a largest coefficient of determination (‘Shucked\_weight’ has the largest coefficient of determination, others almost have the same coefficient of determination.) and provide the most information, so we can delete the variable ‘Whole\_weight’ in the first-order model. So the new model fit 2 is shown:

Fit 2:

We use this model as the new full model, and then perform stepwise regression procedure by criterion AIC(stepAIC() function).

Fit 3 is the null-model, after the AIC, the new model model 1 is shown after stepAIC(figure 14):

Model 1:

AIC equals to -13250.03.

From the summary of the model 1(figure 14), the variable ‘Viscera\_weight’ and ‘Length’ are not very significant. MSE equals to 0.04182, R square equals to 59.11%.

From the plots of model 1 (figure 15), the residuals plots perform better than the model without deleting the ‘whole\_weight’.

**3.2 Second- order model selection**

We try some high-order terms vs residuals and interaction terms vs residuals plots (figure 10), and find some high-order and interaction terms are needed. From the scatter plot before we find the response variable may have quadratic relationship with the variables related to weight. According to three plots of residuals vs quadratic terms of ‘shell\_weight’, ‘shucked\_weight’ and ‘viscera\_weight’, residuals seems to have linear relation with ‘shucked\_weight’ and ‘viscera\_weight’. So we can add these two quadratic terms in the X variables’ pool.

Besides, we randomly tried three interaction terms: ‘length’\*‘shell\_weight’, ‘diameter’ \*‘viscera\_weight’ and ‘height’\*‘shucked\_weight’. The plots show that the the residuals have linear relation with ‘diameter’\*‘viscera\_weight’ and ‘height’\*‘shucked\_weight’. So we can add these two interaction terms in the X variables’ pool.

To fit the model with quadratic term reducing the influence of multicolinearity, we decide to center the variables of ‘shucked\_weight’ and ‘viscera\_weight’.

Fit 3 is the null-model, after AIC, the new model 2 is shown after stepAIC():

Model 2:

AIC equals to -13584.1.

From the summary of the model Model 2 (figure 16), the variable ‘Length’ is not very significant. MSE equals to 0.03857, R square equals to 62.31%.

From the plots of model 2 (figure 17), the residuals vs fitted plot shows the residuals randomly and evenly disperse around zero.

**3.3 New interaction terms of quantitative and qualitative variables**

We draw the marginal boxplots for each quantitative variable (figure 18) and find that all kinds of weight and size have similar relationship with Sex, that female is slightly larger and heavier than male, and both of them are larger than infant with certainty. Therefore, we decide to add interaction terms ‘Sex\*Diameter ‘and ‘Sex\*Shell\_weight’ in variables’ pool.

Fit 3 is the null-model, after AIC, the new model 3 is shown after stepAIC():

Model 3:

AIC equals to -13632.3.

From the summary of the model Model 3 (figure 19), the variable ‘Length’ and ‘Shell\_weight\*SexM’ is not very significant. MSE equals to 0.029241, R square equals to 65.92%.

From the plots of model 3 (figure 20), the residuals vs fitted plot shows the residuals randomly and evenly disperse around zero.

**4. Model diagnostic and validation**

**4.1Residuals plots of three models**

The first-order model, model 1, is not sufficient since residuals vs fitted value plot has a concave pattern and Q-Q plot tells that residuals have right-skewed distribution. model 2 is much better since its residuals disperse around 0 randomly and there is no special pattern in residuals plot but it is still moderately right-skewed. Model 3 seems to be as good as model 2 but contains 2 more variables. Later we will take internal and external validation to pick the final model and check its ability of prediction.

**4.2Internal validation (Ra,p^2, SSEp, MSEp and Pressp)**

We calculate adjusted R square and SSEp, MSEp, and Pressp (figure 21) and model 1 is worse than the other 2 because it has the smallest adjusted R square and largest SSEp, MSEp and Pressp. Model 2 and model 3 have similar performance according to these criterions though SSE and MSE of model 2 is slightly smaller than that of model 3. Therefore, we take the external validation to pick the final model.

**4.3External validation (mean(MSPE), mean(SSE))**

We use 5-fold cross validation. We divide the dataset into 5 parts and each time use 4 of them as training data and the rest as testing data. Self-defined function CV (cross validation) returns average criterion SSE, MSE, and Press. According to the table of criterions (figure 22) and parsimony principle, model 2 is our final model because they have similar SSE and MSE, but model 3 have 4 more variables (corresponding to different levels of qualitative variable “Sex”), so model 2 is a better choice.

**4.4 Outliers**

In this session, our goal is to find and remove all outliers and then regress the final model. We compute Bonferroni’s threshold to identify outliers in response direction and leverage of each case to detect the outliers in predictors’ direction, and use Cook’s distance to find the cases which are influential.

In response direction, there are 2 outliers. However, in predictors’ direction, there are 270 outliers altogether. And by Cook’s distance, we remove 215 outliers from the dataset and then get the final model fit.final.

Comparing the model before deleting ourliers and the model after that, error sum of square (SSE) decrease a lot (from 160.691 to 115.5388).

Plots (figure 23), summary and ANOVA table of the final model show it is a good model.

Final model:

**Conclusions and Discussion**

Under the researches above, we can find that the response variable ‘Rings’ has linear and non-linear (quadratic and interaction terms) relationship with predictor variables. And from the external validation, our model has a good prediction ability.

we can use the final model to predict the age of an abalone. For example, if we have an observation of data including these predictor variables, we can achieve the fitted value of ‘Rings’. Then we calculate the exponential of the fitted value and plus 1.5, so that we can get the predicted value of the age.

During the analysis progress, we meet three limitations. First, out response is discrete, but we regress it as it is a continuous which leads to the strange residuals plot in which scatter lie in a certain pattern. Second, the model is a little bit complex since there are just 8 predictor variables in original dataset, but the final model contains 11 terms. Third, too many cases are removed as outliers which wastes some data. To solve these limitations, we may make some further transformation for the response variable so that it can be continuous and meet the requirements of the assumptions. And may be we can reduce the number of the outliers needed to be removed and we still can get a good prediction for the age of abalone without losing too much information.

**References**

* 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 Islands of Bass Strait”, Sea Fisheries Division, Technical Report No. 4.
* Lecture notes from STA 206.