ABALONES AGE PREDICTION

Sébastien Meyer, Ziru Niu and 2 other students
École polytechnique, France
{firstname.lastname}@polytechnique.edu

Introduction

Abalones are one type of reef-dwelling marine snails. It is difficult to tell the ages of abalones because their shellsizes not only depend on how old they are, but also depend on the availability of food. The study of age is usually by obtaining a stained sample of the shell and looking at the number of rings through a microscope. We are interested in using some of abalones physical measurements, especially the height measurement to predict their ages. Biologists believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages. In particular, that a larger height is associated with an older age.

The dataset and its description are available at `https://archive.ics.uci.edu/ml/datasets/Abalone`.

1 Description of the data

We first summarize the data by drawing the scatterplots and correlations with *sex* in each colour (see **Figure 1**). We can see that the data points that we have are evenly distributed among *sex* feature. Regarding the dependent variable *rings*, we observe that values range from 1 to 29, which correspond to ages from 2.5 to 30.5 years. Median and mean are relatively close, with a small skew for the dependent variables. We see that the number of rings generally has positive correlations with each variable, but a simple linear dependence does not seem clear. There are very high correlations between the different variables, which can make the explanation of our models more difficult. The separation between infants and adults is clear in almost all plots, but the distributions of variables for both male and female abalones are very similar. This indicates that the major difference is between infants and adults and might be a better variable for modeling. Thus, the number of rings cannot be described simply as a linear regression of the height, and we need to do some transformation of the features and the response variable.

2 Simple linear model

We first start with the simple linear model as in the assumption by the biologists

$$rings = \beta_0 + \beta_1 * height + \epsilon.$$

The assumptions to verify under this hypothesis are as follow:

1. Errors are centered: $\forall i = 1..n$, $\mathbb{E}_\beta[\epsilon_i] = 0$
2. Errors have homoscedastic variance: $\forall i = 1..n$, $\mathrm{Var}_\beta[\epsilon_i] = \sigma^2 > 0$
3. Errors are uncorrelated: $\forall i \neq j$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$
4. Errors are Gaussian: $\forall i = 1..n$, $\epsilon_i \hookrightarrow N(0, \sigma^2)$

We also remove some outliers by selecting the subset of data that has a Cook distance below 0.1 for the simple linear model. The simple linear model might not be the best model for prediction, as a lot of points in the test data set fall outside of the predicted 95% confidence interval. By drawing the autoplot (see **Figure 2b**) and computing the relevant tests, we see that none of the assumptions are satisfied. Indeed, our observations are as follows for the corresponding assumptions:

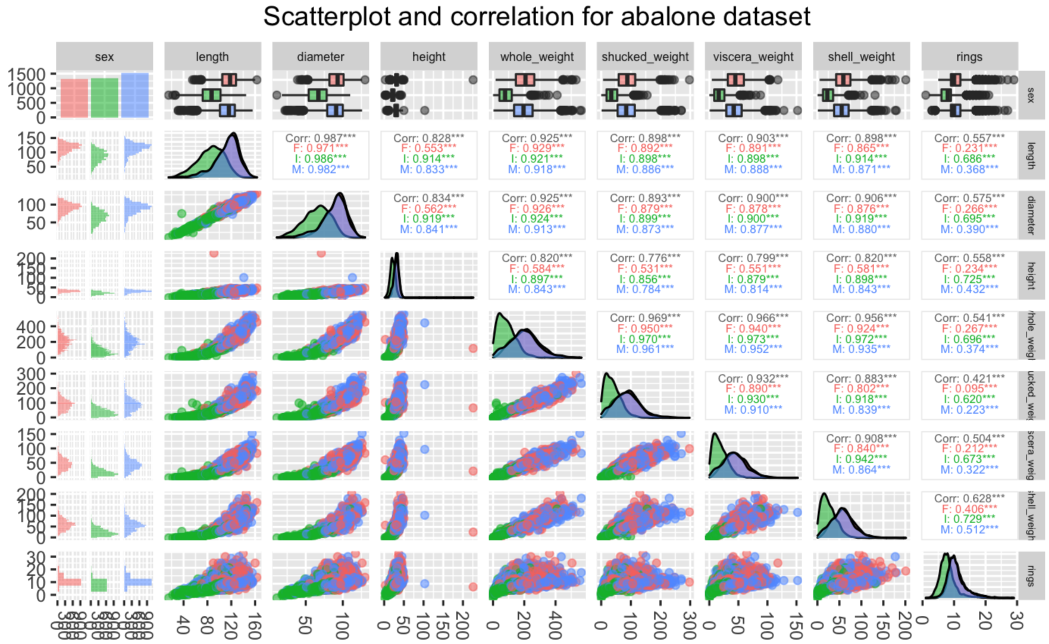


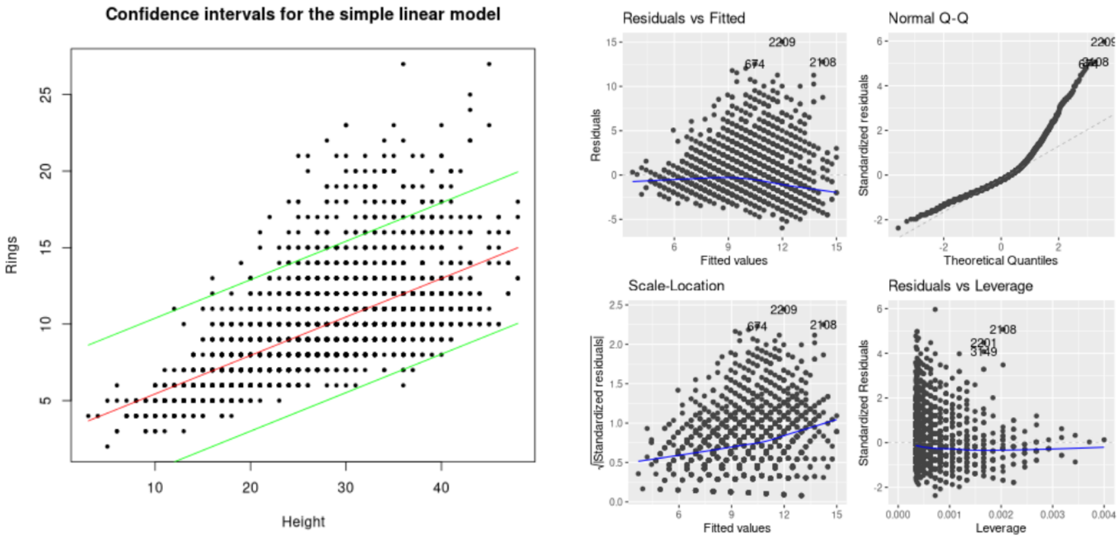
Figure 1: Initial scatterplot.

1. Errors are centered: Great decrease in *Residuals vs. Fitted* plot

2. Errors have homoscedastic variance: *Scale-Location* plot shows that residuals are not equally spread, and *Breush-Pagan* test's p-value is $\ll 0.05$

3. Errors are uncorrelated: The autocorrelation exists for all lags (plot not shown here) and *Durbin-Watson* test's p-value is $\ll 0.05$

4. Errors are Gaussian: *Q-Q* plot shows that there is a clear deviation for higher quantiles and *Shapiro-Wilk* test's p-value is $\ll 0.05$



(a) Predicted age

(b) Autoplot

Figure 2: Simple linear model autoplot and predictions.

We thus conclude that none of the assumptions are satisfied. A log transformation of our dependent variable might improve the results. We try with $\log(rings) = \beta_0+\beta_1*height+\epsilon$ and $\log(rings) = \beta_0+\beta_1*height+\beta_2*height^2+\epsilon$. Below, we show that this model allows to verify some of the assumptions (plots are not shown for clarity):

1. Errors are centered: Slight decrease in *Residuals vs. Fitted* plot
2. Errors have homoscedastic variance: *Scale-Location* plot shows that residuals are more evenly spread, and *Breush-Pagan* test's p-value is 0.1612
3. Errors are uncorrelated: The autocorrelation exists for all lags but is less important and *Durbin-Watson* test's p-value is $\ll 0.05$
4. Errors are Gaussian: *Q-Q* plot shows that there is a slight deviation for higher quantiles and *Shapiro-Wilk* test's p-value is $\ll 0.05$

Therefore, our simple linear model allows to verify at least assumptions 1 and 2. By anova test, we see that the term $height^2$ is significant. Thus, we take $\log(rings) = \beta_0 + \beta_1 * height + \beta_2 * height^2 + \epsilon$ as the simple linear model. The test Mean Square Error (MSE) of the model is approx. 6.902.

3 Multivariate linear model

We perform a logarithm transformation of all the other features, and we draw the scatter plot (see **Figure 3**). It appears that there is a positive linear relationship between $log(rings)$ and $\log$ of other dependent variables.

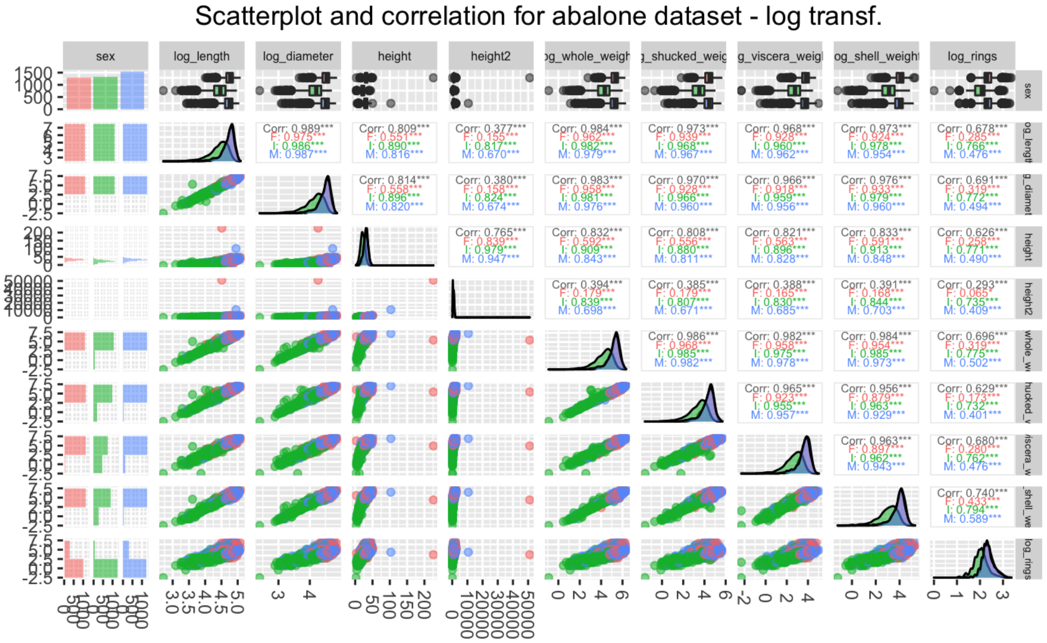


Figure 3: Scatterplot after transformation.

In addition, we observe that the separation between infants and adults is clear but the distributions of variables for both male and female abalones are similar, we take *adulthood* as 1 if $sex = M$ or F, and 0 if $sex = I$. After removing the outliers with Cook's distance, we have this first model:

$$\log(rings) = \beta_0 + \beta_1 * \log(length) + \beta_2 * height + \beta_3 * height^2 + \beta_4 * \log(shucked\_weight) \\ +\beta_5 * \log(whole\_weight) + \beta_6 * \log(shell\_weight) + \beta_7 * adulthood + \epsilon.$$

We find that all the coefficients are significant, and the final MSE on test data is reduced below 5.

We perform a forward feature selection to improve the model. We use a model consisting of all known features and interaction terms with *adulthood*, and we obtain the following model by forward feature selection.

$$\begin{aligned}\log(rings) = \beta_0 + \beta_1 * \log(shell\_weight) + \beta_2 * \log(shucked\_weight) + \beta_3 * \log(whole\_weight) \\ + \beta_4 * length + \beta_5 * adulthood + \beta_6 * height + \beta_7 * height^2 + \beta_8 * viscera\_weight \\ + \beta_9 * shucked\_weight + \beta_{10} * \log(diameter) + \beta_{11} * \log(length) + \beta_{12} shell\_weight \\ + \beta_{13} * \log(shucked\_weight) * adulthood + \beta_{14} * \log(whole\_weight) * adulthood \\ + \beta_{15} * viscera\_weight * adulthood + \beta_{16} * height * adulthood \\ + \beta_{17} * shell\_weight * adulthood + \beta_{18} shucked\_weight * adulthood + \epsilon. \qquad (1)\end{aligned}$$

The MSE of this model is 4.847. We manually find an improved model with a lower MSE by slight modification of the above model as

$$\begin{aligned}\log(rings) = \beta_0 + \beta_1 * \log(length) + \beta_2 * height + \beta_3 * \log(shell\_weight) + \beta_4 * viscera\_weight \\ + \beta_5 * \log(shucked\_weight) * adulthood + \beta_6 * height * adulthood \\ + \beta_7 * shucked\_weight * adulthood + \beta_8 * \log(whole\_weight) * adulthood \\ + \beta_9 * whole\_weight * adulthood + \beta_{10} * length * adulthood + \epsilon. \qquad (2)\end{aligned}$$

with a MSE of 4.759. Thus, we choose Model (2) as the final multivariate model. However, after drawing the autoplot, the multivariate model only satisfies assumption 1 but does not satisfy assumptions 2, 3, nor 4. We can however note that the autocorrelation of residuals is less important than for the other linear models, yet it remains above the confidence interval around zero.

4 Other models

As a final comparison, we decide to try other well-known machine learning models. We use Random Forest (randomForest), Lasso and ElasticNet (glmnet), XGBoost (xgbTree) and kernel estimators (QBAsyDist) methods. Kernel estimators are only based on one feature, which is *height*. For each method, we run the method with the following 3 to 6 models:

1. Initial model, *rings* against all other variables without transformation (rf, glmnet, xgbTree)
2. Full model, *rings* with all variables in Model 1 (rf, glmnet, xgbTree)
3. Full model, $\log(rings)$ with all variables in Model 1 (glmnet)
4. Multivariate model, *rings* with all variables in Model 2 (rf, glmnet, xgbTree)
5. Multivariate model, $\log(rings)$ with all variables in Model 2 (glmnet)
6. Multivariate model, *rings* with all variables in Model 2 and grid search (rf, xgbTree)

We keep the model with the lowest MSE for each method, and the selected model is shown in Table 1.

For kernel method, we show the result of the grid search corresponding to the kernel method in **Figure 4**. The optimal bandwidth is 4.52 and the MSE of corresponding kernel estimator is 14.100. This shows that kernel method is not good for predicting the age of abalones, at least with our selection of features and parameters.

It is worth mentioning that we have tried the kernel method based on the feature *height*. Here we should pay attention to the fact that a test point outside of the range of training values might return a NA value when doing the prediction (because the density is estimated 0 at the outlier, which means that our trained model thought the outlier could never happen). For the choice of kernel type, we used Gaussian kernel. Finally we did a cross-validation to choose the optimal bandwidth.

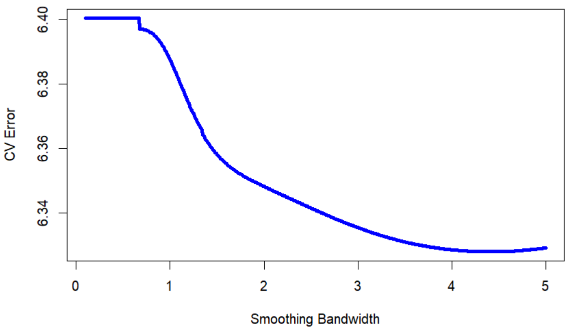


Figure 4: Finding the optimal bandwidth for kernel estimator by cross-validation.

Conclusion

The MSE for each method is shown in **Table 1** as well as the corresponding model from the 6 described in the previous section. In conclusion, we can see that the MSE of multivariate linear model is slightly less than what we can obtain using other machine learning models. This shows that multivariate linear model is a pretty good method to predict the abalones age after transformation of the variables. This result is not surprising as from the scatterplot, we observe that the linear model seems to be the most adapted model for the dependent variable. The other machine learning models have a limited impact in improving the results.

Method	MSE	Model
kernel	14.100	*Height* and grid search
Simple linear model	6.902	*Height* and $Height^2$
XGBoost	5.000	(6)
Random Forest	4.966	(6)
glmnet	4.824	(4)
Multivariate linear model	4.759	(5)

Table 1: MSE for different models