Prediction of the age of the abalones

Marcelo Colares da Silva
Department of Teleinformatics Engineering
Universidade Federal do Ceará
Fortaleza, Brasil
colaresmarcelo2018@gmail.com

***Abstract*—The object of the present study is the abalones, edible mollusks, which aims to make the prediction of their ages, which requires that a shell of the animal be cut for the counting of its dwarfs, which cannot be done with the live animal. So, the objective is to apply linear regression techniques, penalized models and PLS to generate a linear model with its physical characteristics as input, as well as to evaluate the accuracy of specific predictions.**

***Index Terms*—statistical learning, linear regression, penalized models, abalones, PLS.**

I. Introduction

As the first regression technique, linear least squares regression is used to predict the age of the abalone based on its physical attributes. Subsequently, penalized models are applied to check if there is any difference in accuracy compared to the first model, finally, PLS is used to reduce the dimensionality of the predictors and generate a model with an input that is a combination of the originals.

II. Methods

The data set has 9 predictors with 4177 observations. Variables include type, length, diameter, height, total weight, bark weight, viscera weight, bark weight and number of rings, which will be used to predict age. Since the age of the abalone is equal to the number of rings added to 1.5, then this value is added to all the rows in the Rings column (number of rings) in Table I, it will be called Age (age) during manipulation of the data, and it will be the output of the model. The Typoe attribute is the gender of the abalone, represented as a string M (1527 observations), F (1307 observations) and I (1342 observations) for men, women and children, respectively.

TABLE I
Data characteristics.

Nome	Data type	Unit	Observações
Type	categorical		4177
LongestShell	continuous	mm	4177
Diameter	continuous	mm	4177
Height	continuous	mm	4177
WholeWeight	continuous	gramas	4177
ShuckedWeight	continuous	grams	4177
VisceraWeight	contínuo	grams	4177
ShellWeight	continuous	grams	4177
Rings	continuous		4177

Figure 1 contains the scatter plots of all predictors in relation to age:

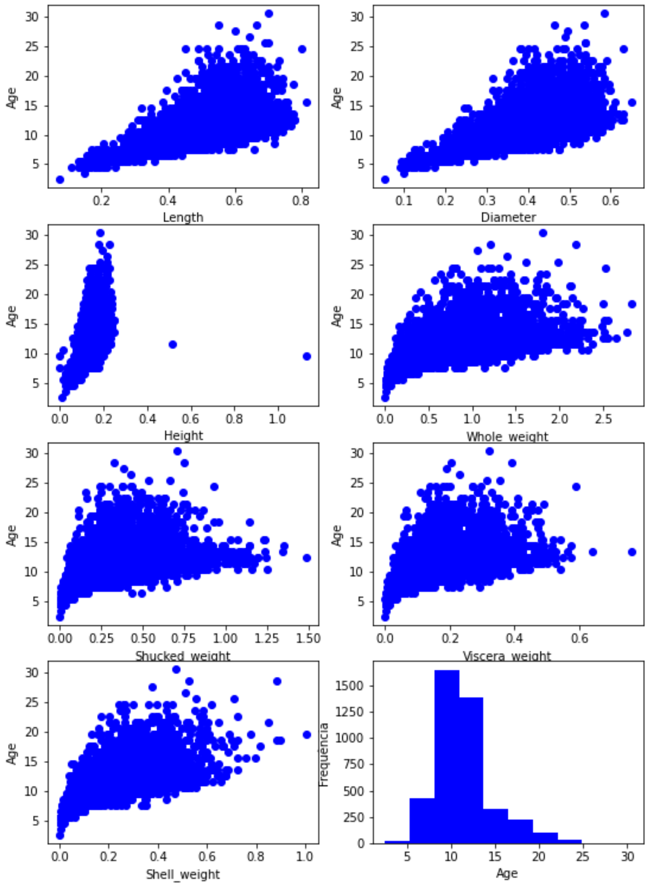

Fig. 1. Scatter-plot de todos os preditores em relação a idade.

As the predictors have different scales, it is necessary to perform a pre-processing on the data in order to correct asymmetry problems and possible outliers and approximate their respective distributions to normal distributions of zero mean and unit variance. For this, the Yeo-Johnson transformation is used. Where y_i corresponds to the vector of data to be transformed. The transformation is given by:

$$y_i^{\lambda} = \begin{cases} \left((y_i+1)^{\lambda} - 1 \right) / \lambda & if \ \lambda \neq 0, y \geq 0 \\ log\left(y_i + 1\right) & if \ \lambda = 0, y \geq 0 \\ -\left[(-yi+1)^{(2-\lambda)} - 1 \right] / (2-\lambda) & if \ \lambda \neq 2, y < 0 \\ -log\left(-y_i + 1\right) & if \ \lambda = 2, y < 0 \end{cases} \tag{1}$$

Following is the scatter plot of the predictors in relation to age after pre-processing.

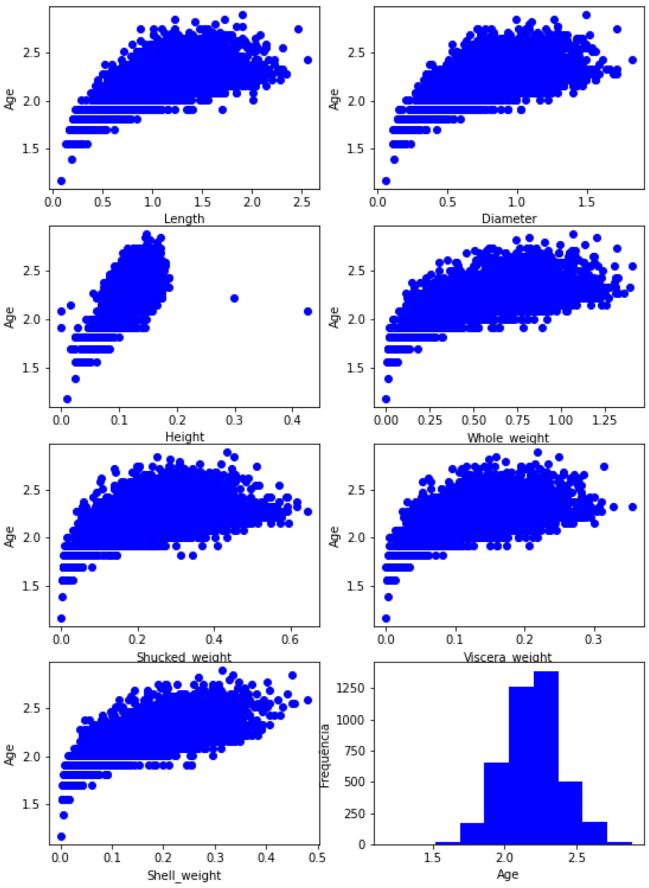

Fig. 2. scatter-plot of all age-related predictors processed.

After treating the data, it is necessary to verify the correlation of the predictors with age for this, using the correlation matrix, which has its content composed of the correlation coefficients ρ with $\rho(i,j) = \rho_{di,dj}$ and each pair, where i and j represent the rows and columns of the matrix.

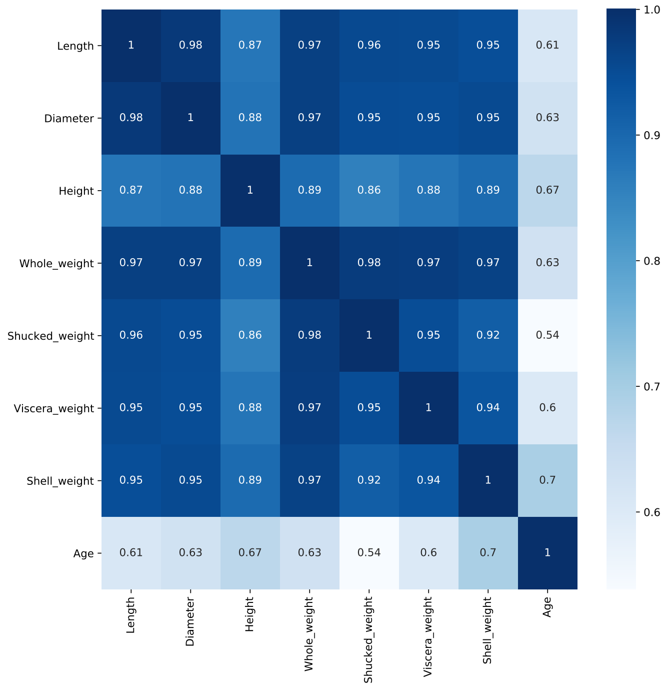

Fig. 3. Correlation matrix

It appears that the line corresponding to age (Age) has high values of correlation with the other predictors, allowing the application of linear models.

A. *Linear least squares regression*

Linear regression is a simple approach to supervised learning to predict a quantitative response from an output Y at the based on a single input predictor variable X. it is necessary that there is an approximately linear relationship between X and Y, for a relationship to be established as follows between them:

$$Y \approx \beta_0 + \beta_1 X \tag{2}$$

Where β_0 and β_1 are two unknown constants to be determined in the process, they represent the intercept and slope terms in the linear model, respectively. The data are used to estimate the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, which will later return the estimation for the output $\hat{y}$. In this process there is the presence of an irreducible error.

$$e_i = y_i - \hat{y}_i \tag{3}$$

The objective is to minimize the sum of the square errors:

$$RSS = e_1^2 + e_1^2 + ...e_N^2 \tag{4}$$

As $\hat{y} = b\hat{e}ta_0 + b\hat{e}ta_1 x$, we define the cost function that will be minimized:

$$J\left(\beta_0,\beta_1\right)=\sum_{i=1}^{N} e_i^2=\sum_{i=1}^{N}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2 \tag{5}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the points where (5) has its minimum value, that is, points where its derivative in relation to the respective $beta$ is null :

$$\frac{\partial J}{\partial \beta_0}=\frac{1}{2}\sum_{i=1}^{N}-2\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0 \tag{6}$$

$$\frac{\partial J}{\partial \beta_1}=\frac{1}{2}\sum_{i=1}^{N}-2x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0 \tag{7}$$

Two equations are obtained with two unknowns, solving the system of equations we have that:

$$\beta_1=\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum\left(x_i-\bar{x}\right)^2}=\frac{cov(x,y)}{var(x)} \tag{8}$$

$$\beta_0=\bar{y}-\beta 1\bar{x} \tag{9}$$

B. Penalty-Ridge regression models

The coefficients produced by the regression of ordinary least squares are unbiased having a lower variance. Since the MSE (which will be defined below) consists of a combination of variance and bias, it is possible to generate models with smaller MSEs, however, this makes the estimates of the parameters obtained to be biased. It is normal for a small increase in bias to lead to a considerable drop in variance, producing a MSE less than the least squares regression coefficients. A consequence of the large correlations between the variances of the predictor is that the variance can become very large. A possible solution would be to penalize the sum of the quadratic errors. In the present study, emph Ridge regression was used, which adds a penalty to the sum of the square regression parameter:

$$RSS_{L2}=\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2+\lambda\sum_{j=1}^{N}\beta_j^2 \tag{10}$$

The sub-index $L2$ means that the model has a square order penalty on parameter estimates, which means that parameter estimates are only allowed to become large if there is a proportional reduction in RSS. This method reduces the estimates to 0 as the λ penalty becomes large. When penalizing the model, a compensation is made between the variance and the model bias.

C. Dimensionality reduction-PLS

As a way of reducing the dimensionality of the data, PLS *(Partial least squares)* is used, which consists of a method that first identifies a new set $Z_1, Z_2, ..., Z_N$ that are combinations lines of the original predictors, somewhat similar to the PCA, in which later ones are adjusted to a linear model using least squares using these N new features. As a requirement, the predictors must be zero mean and unit variance, so the importance of the pre-processing previously applied arises. PLS calculates the first feature Z_1 by setting each $\phi_j 1$ equal to the coefficient of the simple linear regression. Then the first direction is calculated by:

$$Z_1=\sum_{j=1}^{N}\phi_{j1}X_1 \tag{11}$$

PLS gives greater weight to predictors that are more strongly linked to the model's output. In order to determine the second direction of the PLS, first it is necessary to set each variable to Z_1, regressing each variable to Z_1 and taking the residuals. Such residues can be interpreted as data that was not explained by the first direction Z_1 of the PLS. . Subsequently, Z_2 is calculated using these orthogonalized data in exactly the same way as Z_1 was calculated based on the original data. This procedure is repeated for N times to calculate the components components $Z_1, Z_2, ..., Z_N$. At the end of the process, the least squares are used to fit a linear model to the Y output.

D. Evaluation metrics

It is necessary to develop some measure of precision to evaluate the effectiveness of the model. There are different ways to measure this accuracy. A qualitative metric for performance evaluation is the mean square error (MSE) is given by the sum of the squared errors divided by the total of samples:

$$MSE=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2 \tag{12}$$

The obtained value can be interpreted as the average distance of zero residues or also as the average distance between the observed values and those generated by the model. Taking the square root of MSE we define $RMSE$:

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2} \tag{13}$$

The $RMSE$ will then be the model's performance metric, in terms of explaining the data, R^2 is adopted, which represents the proportion of variance for a dependent variable that is explained by an independent variable in a regression model, which is defined as follows:

$$R^2=1-\frac{RSS}{TSS} \tag{14}$$

where $RSS=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$, which can be interpreted as a dispersion measure of the data generated by in relation to the originals and $TSS=\sum_{i=1}^{N}\left(y_i-\bar{y}\right)$, which measures the variance from the output.

E. Validation

To evaluate the effectiveness of the model using the $RMSE$ it is necessary to use resampling techniques. That consists of separating a subset of samples to train the model (training sequence) and another part of the remaining samples are used to estimate the effectiveness of the model (test sequence). As an example, we have k-fold cross-validation, in which the samples are randomly divided into sets of sets of approximately equal sizes. Retained samples are predicted by this model and used to assess performance. The first subset consists of the training set and this is repeated with the second subset and so on. The R^2 is calculated from each set of samples retained. The choice of k is usually 5 or 10 , although there is no standard. The present study adopted $k = 5$

III. Results of

A. *Linear least squares regression:*

Applying the method using the equations (8) and (9), the following table was obtained, which has the averages of $RMSE$ and R^2 resulting from the model test:

TABLE II
$RMSE$ and R^2 averages obtained during the model test.

Nome	RMSE-Test	R^2-Test
LongestShell	0.1504	0.40
Diameter	0.1479	0.42
Height	0.1337	0.52
WholeWeight	0.1478	0.42
ShuckedWeight	0.161	0.31
VisceraWeight	0.1372	0.50
ShellWeight	0.1342	0.42

As it is possible to observe, the linear regression produced estimates that were really divergent from the median R^2 values, that is, the models generated did not perform well. The following are the scatter plots with the training data (blue), test (red) along with the line generated by the model. Even several models were generated, now all the predictors are gathered in order to generate a multivariate model It is possible to observe that a model that considers all the characteristics of the abalones can be more efficient both in training and testing, as can be seen in the previous table.

TABLE III
$RMSE$ and R^2 averages obtained during the model test.

	Train	Test
RMSE	0.0117	0.0119
R2	0.94	0.94

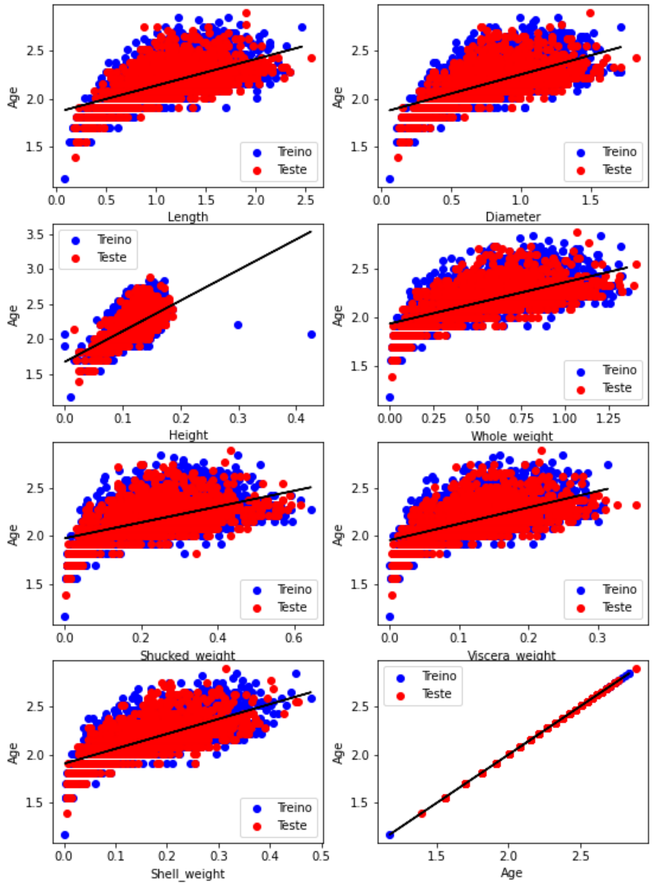

Fig. 4. Linear regression obtained for the predictors.

B. *Ridge Regression Penalized Models*

This step aims to make use of a penalized model, *Ridge regression*. However, it is necessary to determine a value of λ, which in relation to $RMSE$, presents a performance better or equal to that obtained in the previous item, the values will be tested vary $n = 1, 2, .., N$, so that:

$$\lambda^* = arg\ min\ RMSE_{\lambda n} \tag{15}$$

In the present study, $N = 10$ was adopted and 10 values of λ between 0.001 and 0.1 were tested for each predictor using the training sequence, then using the optimal values of λ, the one that produced the lowest $RMSE$, the respective $RMSE$ and R^2 of each predictor with the test sequence were determined. The following table has the average values of $RMSE$ for training and testing as well as the R^2 of testing.

TABLE IV
RMSE AND R^2 OBTAINED FOR THE OPTIMAL λ, WITH λ BETWEEN 0.001 AND 0.1.

Predictor	$RMSE$-**Train**	$RMSE$-**Test**	R^2	λ
LongestShell	0.1505	0.1533	0.37	0.1
Diameter	0.1477	0.1506	0.39	0.056
Height	0.144	0.1372	0.5	0.012
WholeWeight	0.144	0.1502	0.4	0.012
ShuckedWeight	0.144	0.1629	0.29	0.012
VisceraWeight	0.144	0.1533	0.37	0.012
ShellWeight	0.1362	0.1403	0.47	0.001

The following figure shows the $RMSE$ obitdo for the LongestShell predictor, ranging from from 0.001 to 0.15, it is possible that increasing λ improves the accuracy of the data causing $RMSE$ to fall until the optimum value of λ, then $RMSE$ starts to fall, that is, the precision model.

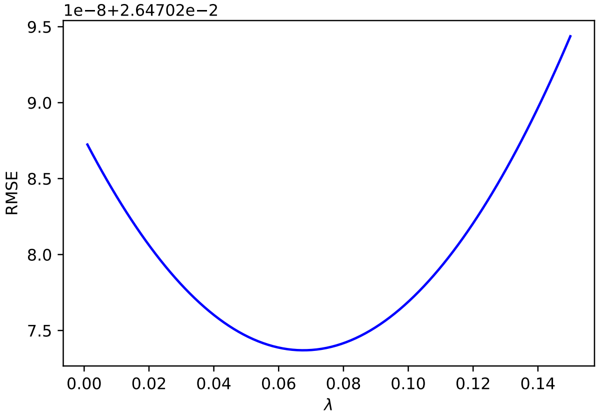

Fig. 5. RMSE according to λ

The compromise between bias and variance is clear, as increasing the model's complexity after a certain point generates an increase in error since the variance increases as the bias falls.

C. *Dimensionality reduction-PLS*

The following figure illustrates the value generated for $RMSE$ using 5-fold cross-validation, varying the number of components. It appears that 2 components result in a relatively low $RMSE$, indicating that the model was able to learn well.

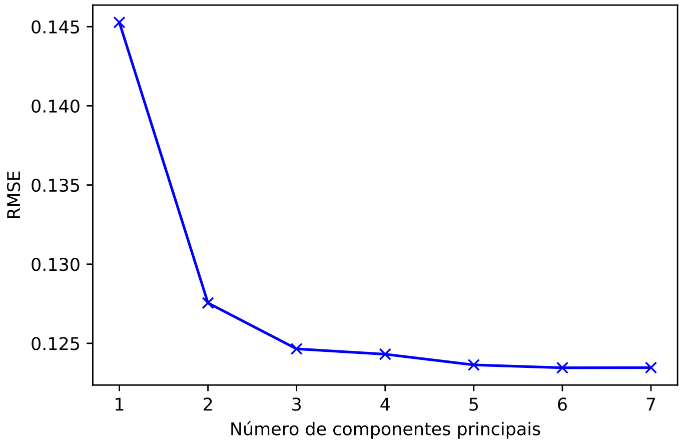

Fig. 6. RMSE em função do número de componentes principais.

Below is the scatter plot of the two components as a function of age, it is observed that the component pc_1, in blue, appears to be better correlated with the age of the abalone, thus explaining a greater percentage of the variance of the data, as in the PCA (*main component analysis*), while the second component does not show an apparent correlation.

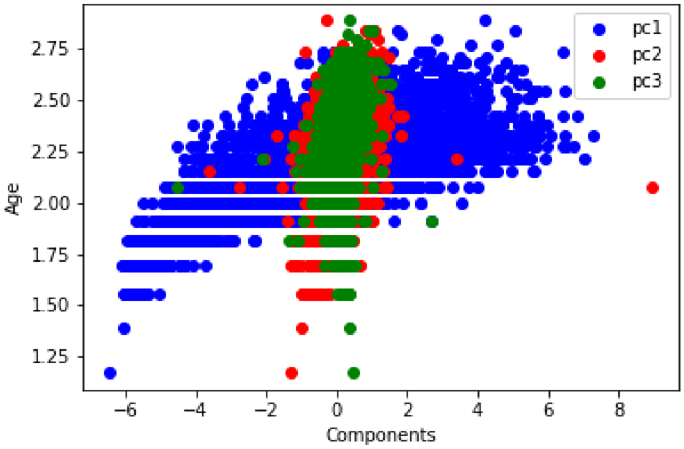

Fig. 7. Main components x Age.

TABLE V
RMSE AND R^2 EARNED DURING TRAINING AND TESTING

	Treino	**Teste**
RMSE	0.0155	0.014
R^2	0.9	0.89

In view of the data in Table V, it is concluded that, as in the cross-validation, the model loses some effectiveness during the test, however, its performance is still good, producing a low $RMSE$.

References

[1] Yeo and R.A. Johnson, "A New Family of Power Transformations to Improve Normality or Symmetry", Biometrika 87.4 (2000).

[2] Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani, An Introduction to Statistical Learning: with Applications in R. Springer Publishing Company, 2014.

[3] McCullagh, Peter; Nelder, John (1989). Generalized Linear Models, Second Edition. Boca Raton: Chapman and Hall/CRC. ISBN 0-412-31760-5.

[4] "Regularization Path For Generalized linear Models by Coordinate Descent", Friedman, Hastie Tibshirani, J Stat Softw, 2010 (Paper).

[5] Jacob A. Wegelin. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle, 2000.

[6] Han, Yizhen. (2019). Machine Learning Project - Predict the Age of Abalone. 10.13140/RG.2.2.21738.88009.