Abalone Age Prediction
Assignment 4 – Model Creation: Regression Using R

Rachel Armington
University of Maryland University College
DATA 610 – Decision Management Systems
Fall 2015, Section 9040
Professor Charles Knode
November 15, 2015

Introduction

This assignment uses the data set abalone.data from the UCI Machine Learning Repository. The data set includes eight numeric attributes: length (of shell), diameter, height (with meat in shell), whole weight, shucked weight (weight of meat), viscera weight (gut weight after bleeding), shell weight (after drying), and age; one factor attribute: sex; and one integer attribute: rings (see Figure 1 for a screenshot of the data set). The data represents information for 4,177 abalone snails.

There were no observed mistakes or missing variables in the data set. However, for purposes of regression analysis, I converted the factor attribute (sex) to integer by coding it (0 = Infant, 1 = Male, and 2 = Female). This made it easier to run regression models on the data.

According to Mayukh (2010), age determination of abalone snails "is a cumbersome process: It involves cutting a sample of the shell, staining it, and counting the number of rings through a microscope." An individual can then determine the age of an abalone by adding 1.5 to the number of counted rings. Prior estimation studies on abalones, such as that conducted by Hiran Mayukh, have already determined that relationships exist between the physical measurements (e.g. sex, length, diameter, etc.) and age. Thus, the purpose of this assignment is to create a linear regression model to estimate the age of an abalone using more easily obtainable physical measurements.

Regression Model and Results

Since the goal of this model is to predict age using easily obtainable physical measurements, I included all variables except Rings (a measurement obtained through the long, tedious process described in the previous section). Age is also directly related to (a derivative of) Rings – thus making it inappropriate to include Rings in this model. Based on correlation tests run in RStudio, I hypothesized that all eight of the other physical measurements would relate to Age, as well as be needed in order to more accurately predict Age. More specifically, I thought that Shell Weight (cor = 0.6276) and Diameter (cor = 0.5747) would have the strongest relation to Age, whereas Sex (cor = 0.4014) and Shucked Weight (cor = 0.4209) would have the weakest relation to Age.

Based on the results from RStudio, the multiple linear regression model is: $y = 4.609 + 0.364x_{Sex} - 0.842x_{Length} + 11.709x_{Diameter} + 11.024x_{Height} + 9.064x_{Whole.Weight} - 19.729x_{Shucked.Weight} - 10.527x_{Viscera.Weight} + 8.669x_{Shell.Weight}$, where the dependent variable (y) is Age and the independent variables (x_{Sex}, x_{Length}, $x_{Diameter}$, x_{Height}, $x_{Whole.Weight}$, $x_{Shucked.Weight}$, $x_{Viscera.Weight}$, and $x_{Shell.Weight}$) are Sex, Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight, respectively. The y-intercept (4.609) is the estimated age of an abalone with Sex = Infant (coded 0), Length = 0mm, Diameter = 0mm, Height = 0mm, Whole Weight = 0gr, Shucked Weight = 0gr, Viscera Weight = 0gr, and Shell Weight = 0gr.

In addition, the estimated coefficients of the eight independent variables vary in the direction and magnitude of the impact on the dependent variable, Age. More specifically, the following independent variables have a positive (+) impact on the Age of abalones, from greatest to least: Diameter (11.709), Height (11.024), Whole Weight (9.064), Shell Weight (8.669), and Sex (0.364). On the other hand, the following independent variables have a negative (-) impact on Age, from greatest negative impact to least negative impact: Shucked Weight (-19.729), Viscera Weight (-10.527), and Length (-0.842). Therefore, Diameter has the greatest positive impact on the dependent variable, Age, while Shucked Weight has the greatest negative impact. For example, based on these estimated coefficients, an abalone with a Diameter of 1mm can be

associated with an age increase of 11.709 years, controlling for the other seven independent variables. Contrastingly, an abalone with a Shucked Weight of 1gr can be associated with an age decrease of 19.729 years, controlling for the other seven independent variables.

Viewing a summary of the model in RStudio yields a regression table (reproduced in Table 1 of Appendix B). The asterisks in the right-hand column of the regression table signify the "level of the statistical significance of a regression coefficient" (Miller, 2014). In this case, seven of the eight variables have three asterisks (***), meaning $p < 0.001$. These seven variables are statistically significant at the 5% significance level (0.05) because $0.001 < 0.05$. Therefore, the only variable that is not statistically significant at the 5% significance level is Length (no asterisks).

Creating the regression model in RStudio also shows us that Multiple R-Squared equals 0.5334, meaning that approximately 53.34% of variation in Age can be explained by the X-variables Sex, Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight. The F-statistic (595.5 on 8 and 4168 DF) and the *p*-value (< 2.2e-16) indicate the overall significance of the model. The model tests the null hypothesis that all coefficients except the intercept are equal to 0 (see the full null hypothesis in Appendix B). This suggests that the model is significant with a *p*-value of less than 2.2e-16. We can reject the null hypothesis that $\beta = 0$ at the 0.05 significance level. Finally, the residual standard error indicates how far observed Age values are from the predicted Age values.

Comparing the RStudio regression model with the Predixion R Linear Regression model shows some similarities, but more differences. First, in terms of independent variables, the regression formula developed in Predixion contains two different variables for Sex (Sex: 0 and Sex: 1), instead of just one like in the model developed in RStudio. The remaining seven variables are the same. Second, in terms of the direction of the coefficients, the coefficient for Sex is positive (0.3645) in the RStudio-developed model, whereas both Sex coefficients are negative (Sex: 0 = -0.9704, Sex: 1 = -0.0326) in the Predixion-developed model. The direction of the remaining seven variables is consistent between the two models. Furthermore, in terms of the magnitude of the coefficients, Diameter and Shucked Weight have the largest positive and negative impact, respectively, on the dependent variable in both models. However, the coefficients differ slightly between the models (e.g. Diameter: 11.7094 RStudio vs. 12.6911 Predixion). The y-intercept also differs (4.6090 RStudio vs. 5.5069 Predixion). These differences may be attributed to the number of observations (1253) that Predixion used to develop the model, as opposed to the full 4177 observations.

Activity 6: Alternate Models

I used the same abalone.data.csv data set to develop two alternate decision tree models – one using the tree() package and one using the rpart() package. However, for this model, I did not code the Sex independent variable in the data set; I left it in factor format. Similar to the multiple linear regression model, I chose to include all of the independent variables, except for Rings. Again, the dependent variable, Age, is a derivative of Rings, so Rings is not appropriate to include in the model. The purpose of creating these decision trees models is to identify and explain patterns in the data set.

In order to read the plots, one should start at the root (top) node and follow the branches down to the terminal (bottom) nodes. In both decision tree models, Shell Weight is the most significant variable because that is the root node of the trees.

However, the decision tree models also differ in several ways (other than in looks). The tree() package model (Figure 3 of Appendix B) has 11 terminal nodes and it contains the independent variables Shell Weight, Viscera Weight, and Shucked Weight. This is interesting because Viscera Weight and Shucked Weight had the two most negative impacts on the dependent variable, Age, as seen in the linear regression model.

On the other hand, the rpart() package model (Figure 4 of Appendix B) has five terminal nodes and it includes the independent variables Shell Weight and Diameter. Unlike the tree() model, this one contains the variable that had the largest positive impact on the dependent variable, Age. In addition, unlike the tree() model, the rpart() model includes the proportion of the population that resides in every node. For example, 100% of the population resides in the root node. Then, only 28% of the population resides in the second node (have a Shell Weight of less than 0.14), while 72% of the population resides in the third node (do not have a Shell Weight of less than 0.14). The rpart() package uses the CART (Classification & Regression Trees) decision tree algorithm (Stephens, 2014). Due to its easier-to-read design and higher level of detail, the rpart() plot seems like the more effective decision tree plot to incorporate into an organization.

Activity 7: Data Visualization

Figure 5 of Appendix B contains four residual plots of the Abalone data set. The Residuals vs. Fitted plot shows heteroscedasticity and reflects a megaphone shape, which indicates that the linear model is a better fit for smaller x-values, but not necessarily for larger x-values. In other words, this plot reveals that variance increases with x. It also tells us that variance is increasing. According to Mike Marin (2013b), this megaphone shape suggests that "larger predicted values are associated with larger errors or residuals." Furthermore, the red line is somewhat flat, which tells us that the linearity assumption is met.

Furthermore, the QQ, or Quantile-Quantile, plot indicates that the y-values, or standardized residuals, are somewhat normally distributed. The Scale-Location plot also shows heteroscedasticity and increasing variance, judging by the upward sloping red line. Finally, the Residuals vs. Leverage plot shows a point (2052) with high leverage and a large negative standardized residual (gung, 2013). Future studies should try to fix the heteroscedasticity of the model by performing a data transformation, such as log transformation.

Figure 6 of Appendix B is an interactive plot of the linear regression model variables: Age and Diameter. Since Sex is categorical (Infant, Male, Female), I was able to color code the plot by this independent variable. This plot reflects some interesting results. In all cases (Infant, Male, and Female), it appears that as the variable Diameter increases, so does the variable Age. In other words, there is a positive correlation between the independent variables Diameter and Age – across Infant, Male, and Female abalones. In addition, this plot makes it easier to visualize the pattern between Diameter and Sex. The average Diameter is greatest in female abalone shells (0.4547mm). Not surprisingly, the average Diameter of male abalone shells (0.4393mm) is greater than the average Diameter of infant abalone shells (0.3265mm). Finally, this plot shows that the maximum Age is highest for Females (30.5 years), followed by Males (28.5 years), then Infants (22.5 years). With that said, the following question may arise: are Infant abalones classified as Infants until they reach a certain age, or are these abalones considered Infants throughout the course of the study? Future studies should review the methodology and data collection procedures of the original study.

References

gung. (2013, July 29). Interpreting plot.lm(). Retrieved from http://stats.stackexchange.com/questions/58141/interpreting-plot-lm

Lichman, M. (2013). UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from https://archive.ics.uci.edu/ml/datasets/Abalone

MarinStatsLectures. (2013, November 13). Checking linear regression assumptions in R (R tutorial 5.2) [video file]. Retrieved from https://www.youtube.com/watch?v=eTZ4VUZHzxw&list=SPqzoL9-eJTNBDdKgJgJzaQcY6OXmsXAHU&feature=iv&src_vid=q1RD5ECsSB0&annotation_id=annotation_468341255

MarinStatsLectures. (2013, November 22). Multiple linear regression in R (R tutorial 5.3) [video file]. Retrieved from https://www.youtube.com/watch?v=q1RD5ECsSB0

MarinStatsLectures. (2015, July 6). Multiple linear regression with interaction in R (R tutorial 5.9) [video file]. Retrieved from https://www.youtube.com/watch?v=8YuuIsoYqsg

Mayukh, H. (2010). Age of abalones using physical characteristics: A classification problem. Retrieved from http://homepages.cae.wisc.edu/~ece539/fall10/project/Mayukh_rpt.pdf

Miller, S. V. (2014, August 13). Reading a regression table: A guide for students. Retrieved from http://svmiller.com/blog/2014/08/reading-a-regression-table-a-guide-for-students/

Sharif, A. A. (2013, November 5). DSO 530: Decision trees in R (regression) [video file]. Retrieved from https://www.youtube.com/watch?v=LziT4fJDB4I

StatsDirect Limited. (2015). P values. Retrieved from http://www.statsdirect.com/help/default.htm#basics/p_values.htm

Stephens, T. (2014, October 1). Titanic: Getting started with R – part 3: Decision trees. Retrieved from http://trevorstephens.com/post/72923766261/titanic-getting-started-with-r-part-3-decision

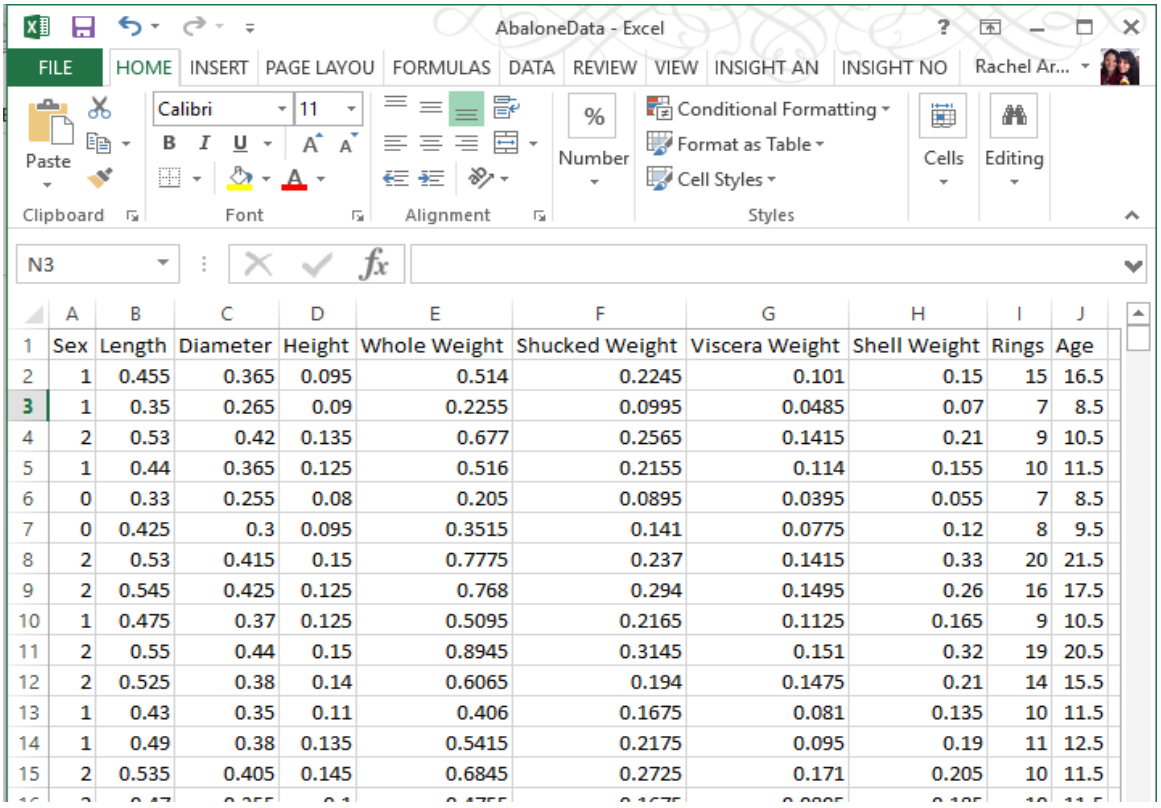
Appendix A

```
###Abalone R Source Code for Assignment 4###

# Set the working directory
projectWD <- "/Users/Rachel/Google Drive/UMUC/DATA 610/Assignment 4"
setwd(projectWD)
getwd()

# Read in the data, save it as AbaloneData, and attach the data
AbaloneData <- read.csv(file.choose(), header = T)
attach(AbaloneData)

# View a frame of data set
str(AbaloneData)

# Find the correlation between the X-variables ("Sex", "Length",
"Diameter", "Height", "Whole.Weight", "Shucked.Weight", "Viscera.Weight",
and "Shell.Weight") and the Y-variable ("Age").
cor(Sex, Age)
cor(Length, Age)
cor(Diameter, Age)
cor(Height, Age)
cor(Whole.Weight, Age)
cor(Shucked.Weight, Age)
cor(Viscera.Weight, Age)
cor(Shell.Weight, Age)

# Fit a model using Length, Diameter, Height, Whole Weight, Shucked
Weight, Viscera Weight, and Shell Weight as X-Variables, and Age as the Y-
variable
model1 <-lm(Age ~ Sex + Length + Diameter + Height + Whole.Weight +
Shucked.Weight + Viscera.Weight + Shell.Weight)

# Get a summary of the model.
summary(model1)

# Look at the coefficients for the model
model1$coefficients

# Make all four plots appear at the same time. Check the regression
diagnostic plots and add the regression link to the scatterplot.
par(mfrow=c(2,2))
plot(model1)

# Calculate Pearson's correlation between Diameter and Shell Weight.
Diameter and Shell Weight are somewhat bounded together due to the high
correlation.
cor(Diameter, Shell.Weight, method="pearson")

# Ask for confidence intervals for the model coefficients
confint(model1, conf.level=0.95)


## Decision Trees (Regression)
```

```
# Split the data set in two for training and testing
library(tree)
install.packages('rattle')
install.packages('rpart.plot')
install.packages('RColorBrewer')
library(rattle)
library(rpart.plot)
library(RColorBrewer)
set.seed(1)
train = sample(1:nrow(AbaloneData),nrow(AbaloneData)/2)
test = -train
training_data = AbaloneData[train,]
test_data = AbaloneData[test,]
test_Age = Age[test]

# fit a tree model based on training data set
tree <- tree(Age ~ Sex + Length + Diameter + Height + Whole.Weight +
Shucked.Weight + Viscera.Weight + Shell.Weight, training_data)
tree
plot(tree)
text(tree, pretty=0)

# check the accuracy of the model using the test data set
tree_pred = predict(tree, test_data)
mean((tree_pred - test_Age)^2) #5.581717

# cross validation for pruning the tree
cv_tree = cv.tree(tree)
names(cv_tree)
plot(cv_tree$size, cv_tree$dev, type = "b", xlab = "Tree Size", ylab =
"MSE")
which.min(cv_tree$dev)
cv_tree$size[1]

### Data Visualization ###

# Fancy Decision Tree
library(rpart)
install.packages('rattle')
install.packages('rpart.plot')
install.packages('RColorBrewer')
library(rattle)
library(rpart.plot)
library(RColorBrewer)
Tree2 <- rpart(Age ~ Sex + Length + Diameter + Height + Whole.Weight +
Shucked.Weight + Viscera.Weight + Shell.Weight, data=AbaloneData,
method="class")
Tree2
prp(Tree2)
text(Tree2)
fancyRpartPlot(Tree2)

# Linear Regression Visuals
```

```
TAB = table(AbaloneData$Sex)
barplot(TAB, legend=T)
pie(TAB)
chisq.test(TAB, correct=T)

# Age vs. Diameter, Sex
plot(Diameter[Sex=="M"], Age[Sex=="M"], col="blue", ylim=c(0,31),
xlim=c(0,0.7), xlab="Diameter", ylab="Age", main="Age vs. Diameter, Sex")
points(Diameter[Sex=="I"], Age[Sex=="I"], col="green",pch=16)
points(Diameter[Sex=="F"], Age[Sex=="F"], col="red",pch=16)
legend(0.05,30,legend=c("Male","Infant","Female"),col=c("blue","green","re
d"), pch=c(1,16),bty="n")
```

Appendix B



	A	B	C	D	E	F	G	H	I	J
1	Sex	Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings	Age
2	1	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	16.5
3	1	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	8.5
4	2	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	10.5
5	1	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	11.5
6	0	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	8.5
7	0	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	9.5
8	2	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	21.5
9	2	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	17.5
10	1	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	10.5
11	2	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	20.5
12	2	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14	15.5
13	1	0.43	0.35	0.11	0.406	0.1675	0.081	0.135	10	11.5
14	1	0.49	0.38	0.135	0.5415	0.2175	0.095	0.19	11	12.5
15	2	0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	10	11.5

Figure 1. Screenshot of "abalone.data" data set.

Null Hypothesis:

$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

$H_0 = \beta_{Sex} = \beta_{Length} = \beta_{Diameter} = \beta_{Height} = \beta_{Whole.Weight} = \beta_{Shucked.Weight} = \beta_{Viscera.Weight} = \beta_{Shell.Weight} = 0$

	Estimate	Std. Error	t value	Pr(>|t|)	Signif
(Intercept)	4.60903	0.26808	17.193	< 2e-16	***
Sex	0.36446	0.05088	7.163	9.30E-13	***
Length	-0.84207	1.81671	-0.464	0.643	
Diameter	11.7094	2.23562	5.238	1.71E-07	***
Height	11.02363	1.54294	7.145	1.06E-12	***
Whole.Weight	9.06357	0.72872	12.438	< 2e-16	***
Shucked.Weight	-19.7292	0.82118	-24.025	< 2e-16	***
Viscera.Weight	-10.52745	1.29986	-8.099	7.22E-16	***
Shell.Weight	8.66863	1.13001	7.671	2.11E-14	***

Table 1. Regression Table. Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Full Regression Formula - Age

Type	Regression Formula
General	5.5068704908753938 - 0.97036276557158019 * [Sex:0] - 0.03256818100284168 * [Sex:1] - 1.136109726942506 * [Length] + 12.69110562147698 * [Diameter] + 8.3201511940715456 * [Height] + 10.467881018219829 * [Whole Weight] - 20.71116878871673 * [Shucked Weight] - 12.30753994805962 * [Viscera Weight] + 6.2763559378380513 * [Shell Weight]

Coefficients for Age

Column	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.50687049087539	0.345999879848917	15.9158161941559	9.3528688709575E-55
Sex:0	-0.97036276557158	0.122609890144561	-7.91422914111981	3.503722482352O6E-15
Sex:1	-0.0325681810028417	0.100528749630666	-0.323968826057166	0.745984885930165
Length	-1.13610972694251	2.20719711660497	-0.514729617212454	0.606781012030457
Diameter	12.691105621477	2.72602837023278	4.65552954622892	3.37599353542463E-06
Height	8.32015119407155	1.69751757244624	4.90136380861238	1.00382339848324E-06
Whole Weight	10.4678810182198	0.909294947907614	11.5120853165494	5.06175344391055E-30
Shucked Weight	-20.7111687887167	1.02145333428397	-20.2761771816381	1.28791650654934E-85
Viscera Weight	-12.3075399480596	1.56409107517326	-7.86881284818806	5.00377931624513E-15
Shell Weight	6.27635593783805	1.37174182424799	4.57546444009523	4.95020997864493E-06

Figure 2. Predixion Regression Formula and Regression Table.

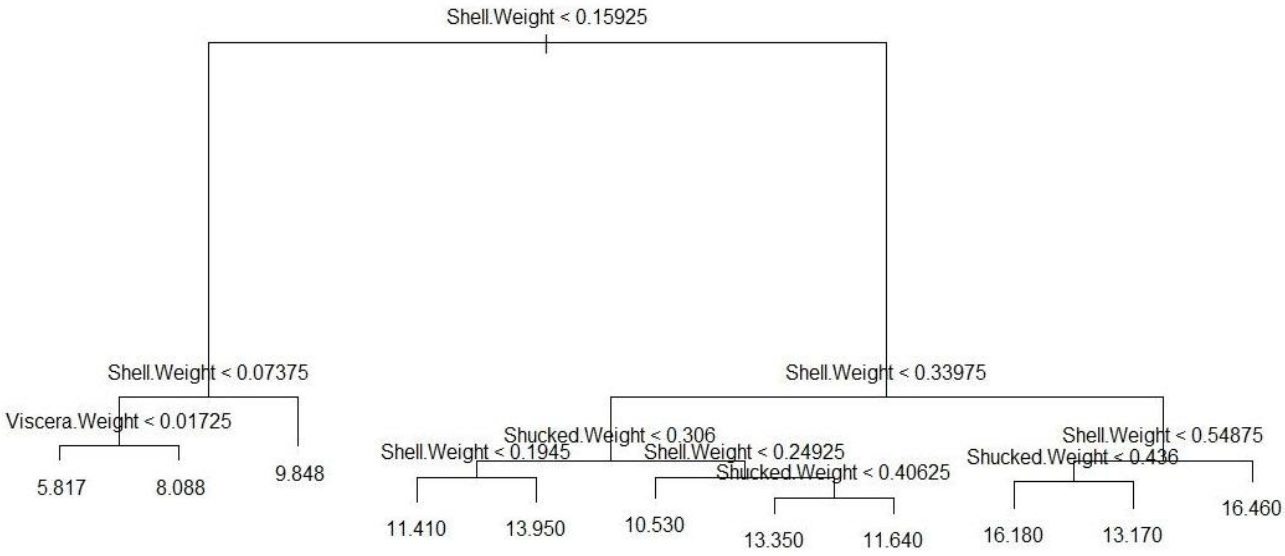


Figure 3. Decision Tree, tree() Package

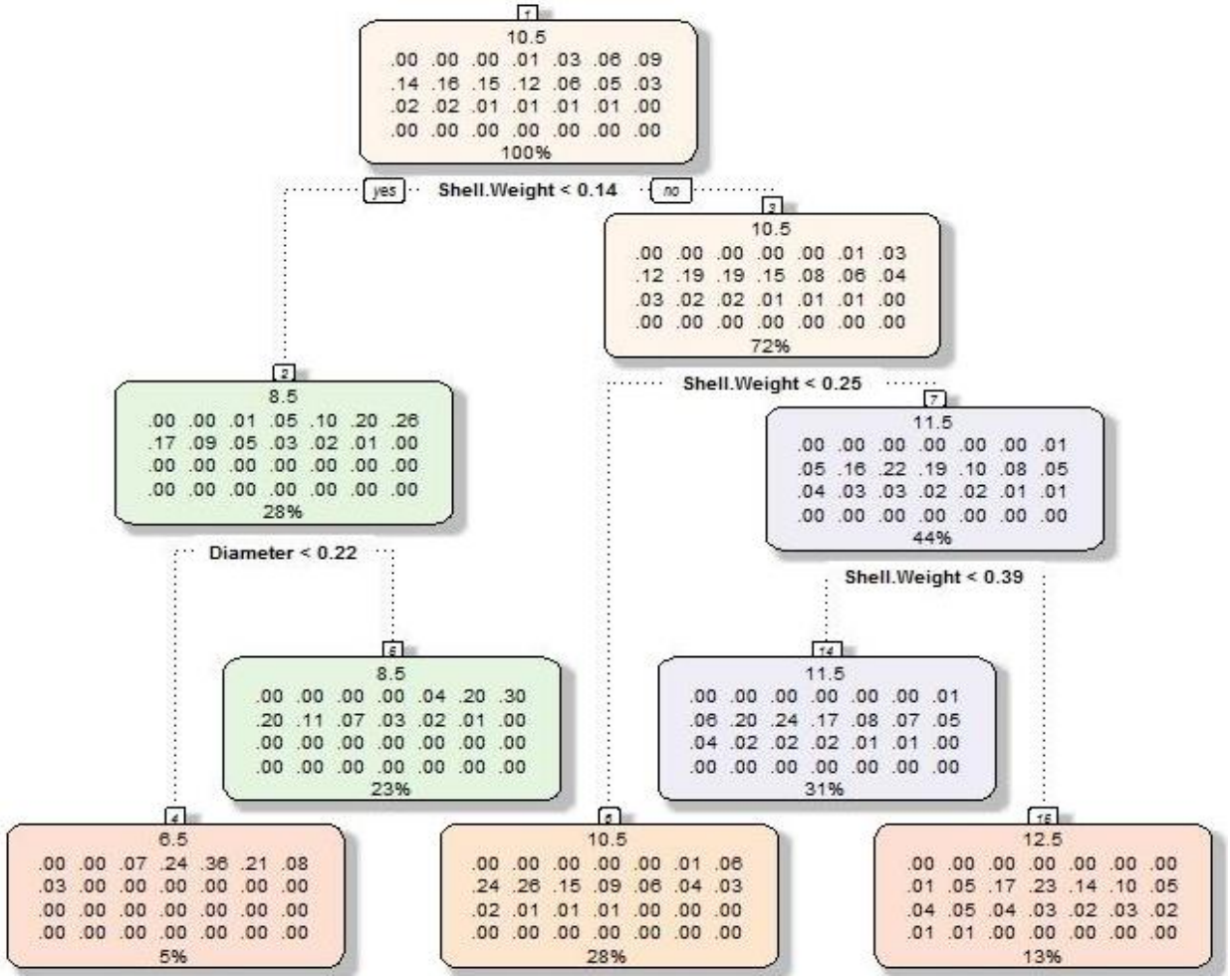


Figure 4. Decision Tree, rpart() Package

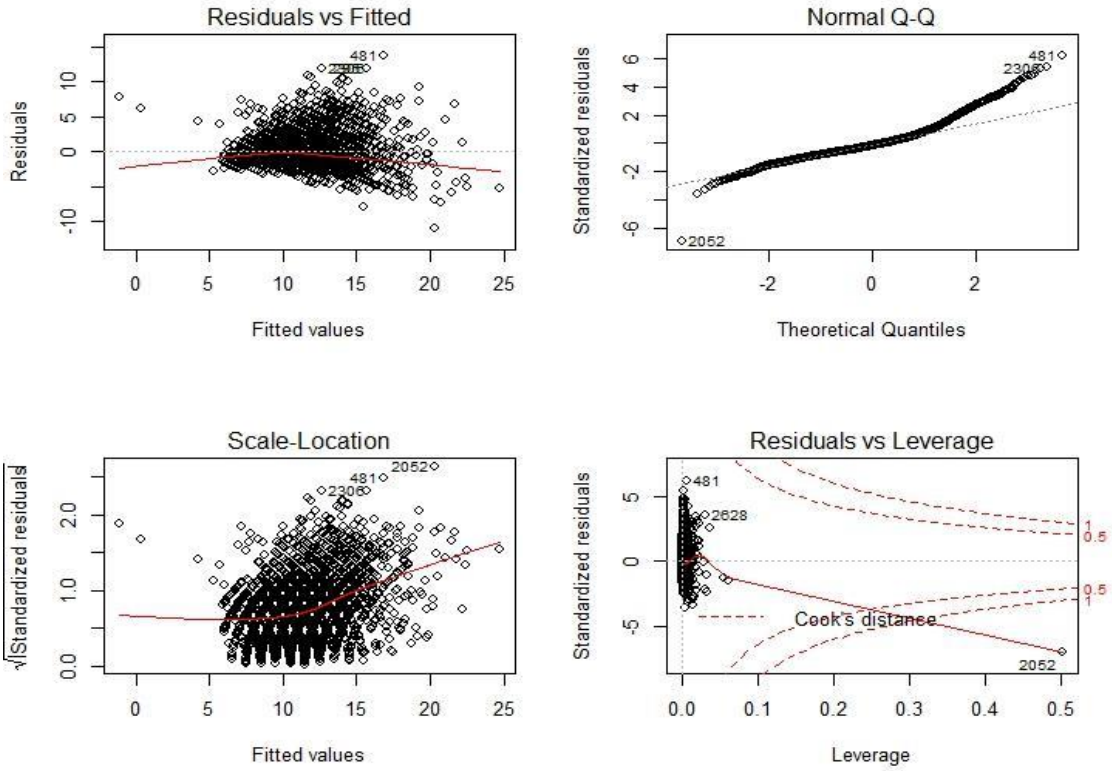


Figure 5. R Plots.

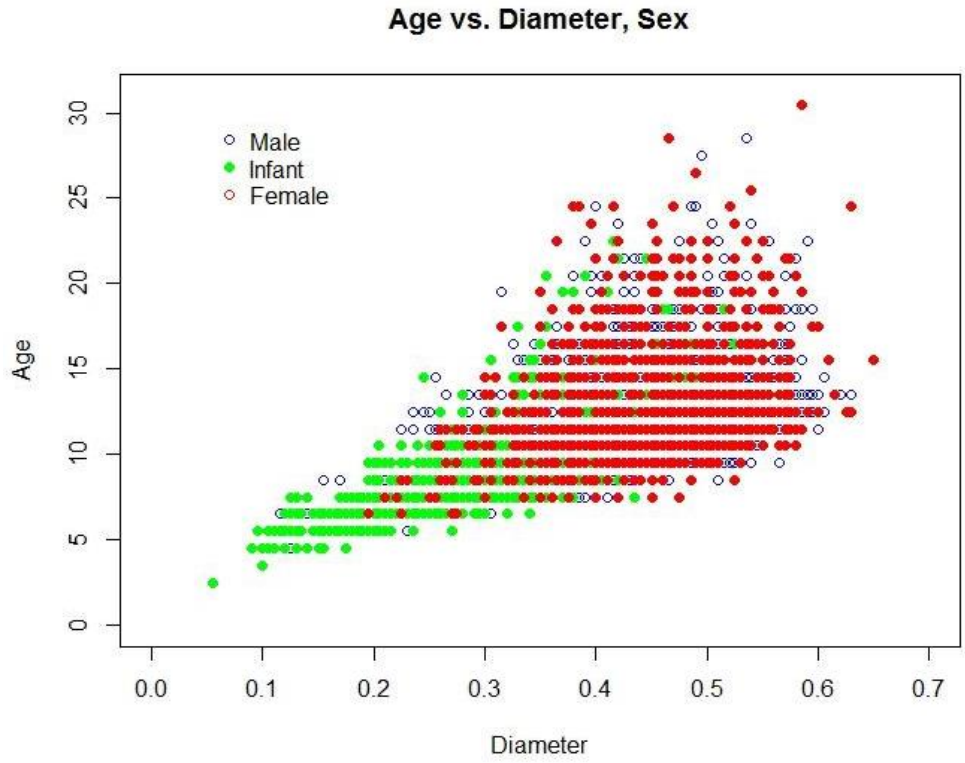


Figure 6. Age vs. Diameter Interaction, by Sex.