Module 2A

Andrew Estes

8/22/2021

### Pre-Analysis Set-Up

##### Installing Packages

library(ggpubr)  
library(car)  
library(caret)  
library(tidyverse)  
library(ggplot2)  
library(dplyr)  
library(MASS)  
library(sensemakr)

##### Importing and View Data

abalone.url <-  
 "https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data"  
  
abalone.data <- read.csv(abalone.url, header=FALSE) #data does not have variable names  
  
head(abalone.data)

## V1 V2 V3 V4 V5 V6 V7 V8 V9  
## 1 M 0.455 0.365 0.095 0.5140 0.2245 0.1010 0.150 15  
## 2 M 0.350 0.265 0.090 0.2255 0.0995 0.0485 0.070 7  
## 3 F 0.530 0.420 0.135 0.6770 0.2565 0.1415 0.210 9  
## 4 M 0.440 0.365 0.125 0.5160 0.2155 0.1140 0.155 10  
## 5 I 0.330 0.255 0.080 0.2050 0.0895 0.0395 0.055 7  
## 6 I 0.425 0.300 0.095 0.3515 0.1410 0.0775 0.120 8

### 1) What is the class of the data abalone.data? (Use class() function)

class(abalone.data)

## [1] "data.frame"

##### Data Clean Up

*Add Column Names to Data Frame*

colnames(abalone.data)[1] <- "sex"  
colnames(abalone.data)[2] <- "length"  
colnames(abalone.data)[3] <- "diameter"  
colnames(abalone.data)[4] <- "height"  
colnames(abalone.data)[5] <- "whole\_weight"  
colnames(abalone.data)[6] <- "shucked\_weight"  
colnames(abalone.data)[7] <- "viscera\_weight"  
colnames(abalone.data)[8] <- "shell\_weight"  
colnames(abalone.data)[9] <- "rings"

*New Output*

head(abalone.data)

## sex length diameter height whole\_weight shucked\_weight viscera\_weight  
## 1 M 0.455 0.365 0.095 0.5140 0.2245 0.1010  
## 2 M 0.350 0.265 0.090 0.2255 0.0995 0.0485  
## 3 F 0.530 0.420 0.135 0.6770 0.2565 0.1415  
## 4 M 0.440 0.365 0.125 0.5160 0.2155 0.1140  
## 5 I 0.330 0.255 0.080 0.2050 0.0895 0.0395  
## 6 I 0.425 0.300 0.095 0.3515 0.1410 0.0775  
## shell\_weight rings  
## 1 0.150 15  
## 2 0.070 7  
## 3 0.210 9  
## 4 0.155 10  
## 5 0.055 7  
## 6 0.120 8

### 2) What is the datatype of variable diameter? (Use typeof() function)

typeof(abalone.data$diameter)

## [1] "double"

### 3) Use the function summary() to find basic description of diameter and age (rings) of abalone.

summary(abalone.data$diameter)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0550 0.3500 0.4250 0.4079 0.4800 0.6500

summary(abalone.data$rings)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 8.000 9.000 9.934 11.000 29.000

### 4) Use the function mean() to find the mean of diameter of female abalone.

female.abalone <- dplyr::filter(abalone.data,sex=='F')  
mean(female.abalone$diameter)

## [1] 0.4547322

<https://rstudio-pubs-static.s3.amazonaws.com/408480_f4bdfd9620c84a9598e512f1a59e66f5.html>

### 5) Assume Y: diameter and X: rings. What is〖(X^’ X)〗^(-1) X^’ Y ?

x <- abalone.data$rings  
y <- abalone.data$diameter  
  
transposed <- t(x) %\*% x  
inversed <- solve(transposed)  
part2 <- t(x) %\*% y  
answer <- inversed %\*%part2  
answer

## [,1]  
## [1,] 0.03883339

### 6) Can you think of an interesting question about Abalone?

I had no real goals for Abalone, but instead chose to take this opportunity to explore the different packages and capabilities RStudio provides. I ran histograms, side-by-side-by-side box plots, linear regression models, scatter plots of those models, sensitivity analysis of a model, and statistical tests of covariance, correlation, collinearity, and fit of analysis for the regression model.

The biggest improvement I would make to this section is the output. Instead of having 4 histograms all follwing each other taking up a half a page or even more, I would manipulate the size of the output so all 4 histograms would be on the same page in a 2x2 table.

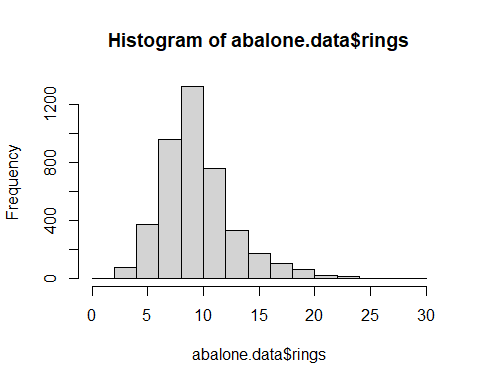
*I need to make the data and graphs more readable.*

The first step in any analysis was to continue the separation of the abalone by sex. Since the females were previously partitioned, only the males and infants are required

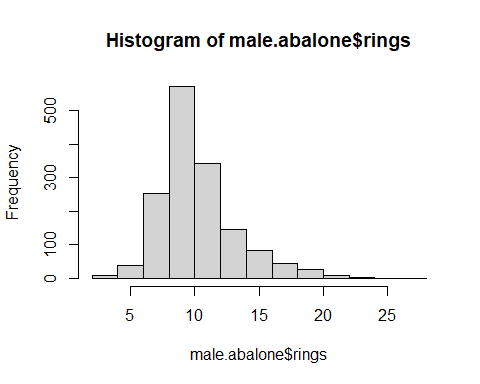
male.abalone=dplyr::filter(abalone.data,sex=='M')  
indet.abalone=dplyr::filter(abalone.data,sex=='I')

Now we can run a plots and tests for the data as a whole. Below is a histogram of the entire data with a focus on the rings, length, weight, and diameter by total data set, male, female, and indeterminate genders.

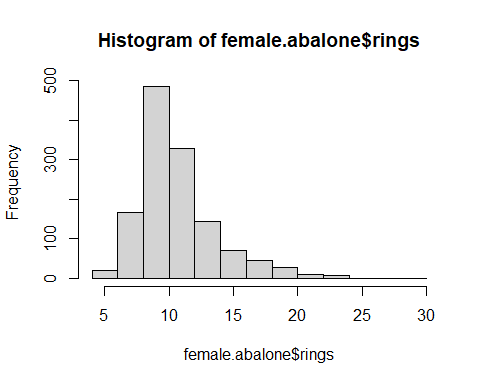
hist(abalone.data$rings)



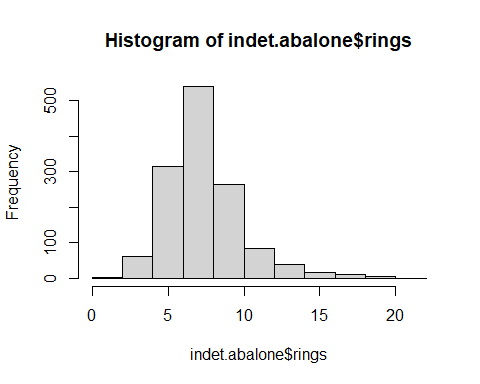
hist(male.abalone$rings)



hist(female.abalone$rings)

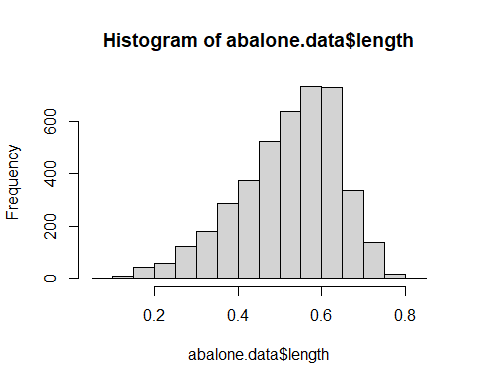


hist(indet.abalone$rings)

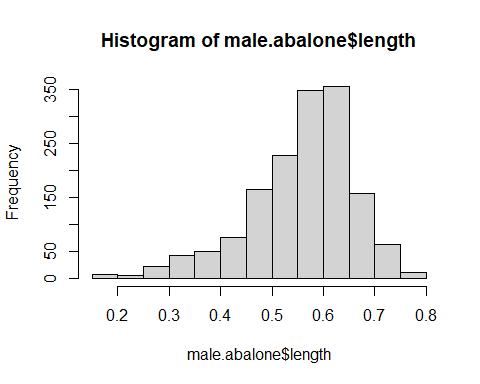


Here are the histograms of the data focused on length by gender

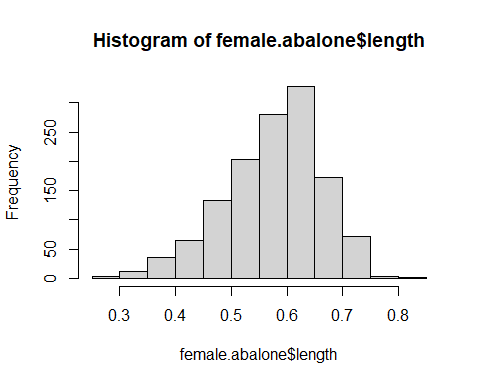
hist(abalone.data$length)



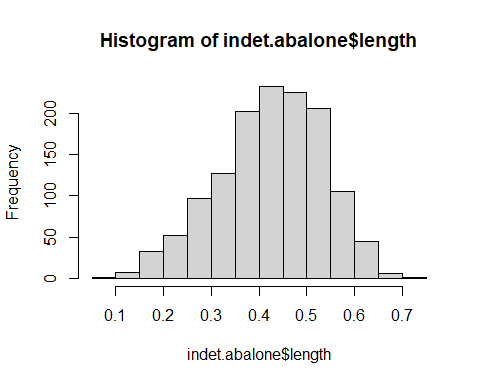
hist(male.abalone$length)



hist(female.abalone$length)

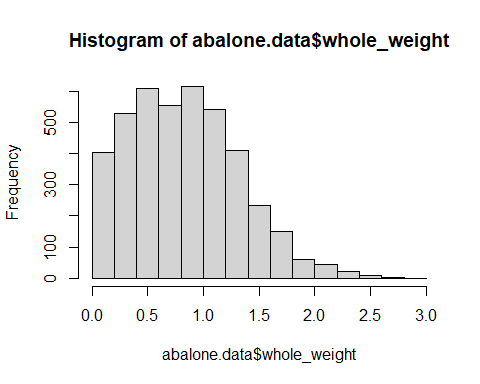


hist(indet.abalone$length)

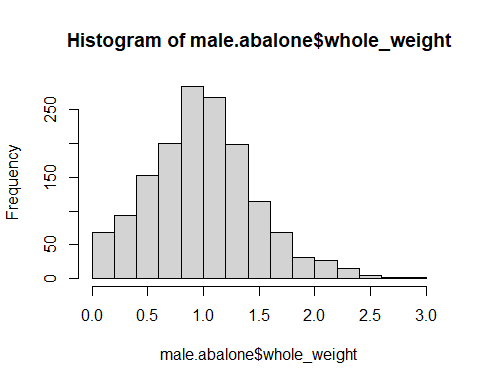


Here are the histograms of the data focused on whole weight by gender

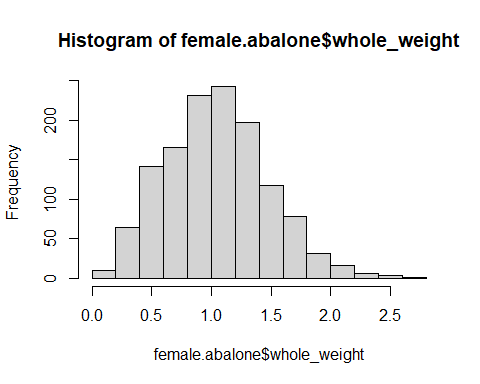
hist(abalone.data$whole\_weight)



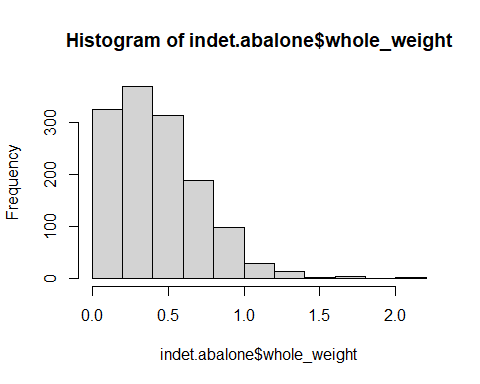
hist(male.abalone$whole\_weight)



hist(female.abalone$whole\_weight)

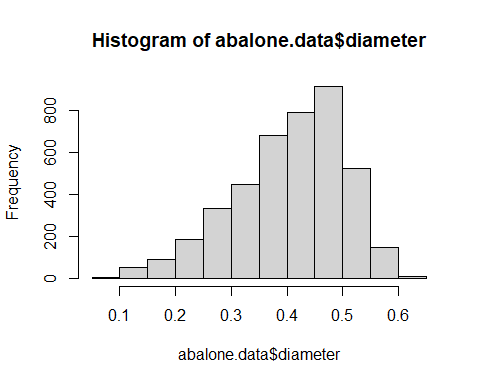


hist(indet.abalone$whole\_weight)

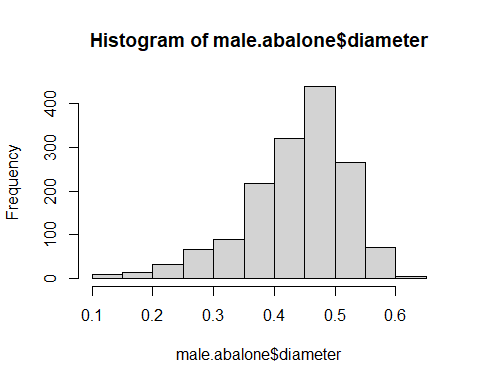


Here are the histograms of the data focused on diameter by gender

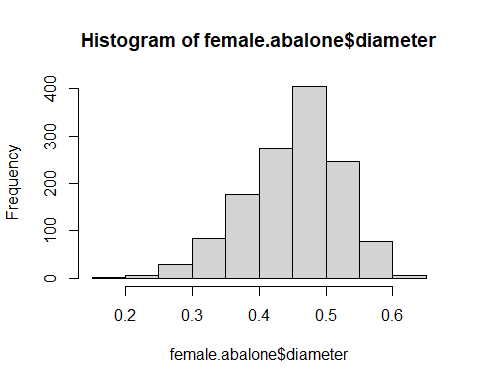
hist(abalone.data$diameter)



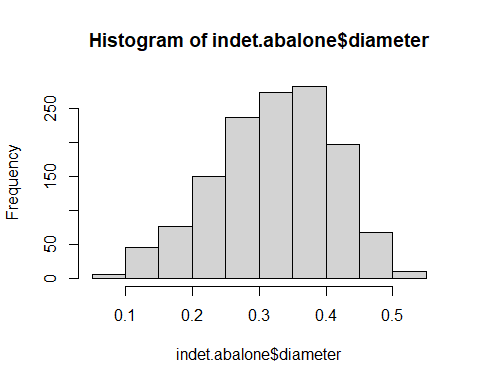
hist(male.abalone$diameter)



hist(female.abalone$diameter)

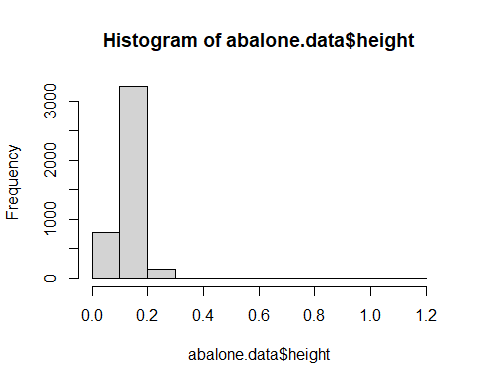


hist(indet.abalone$diameter)

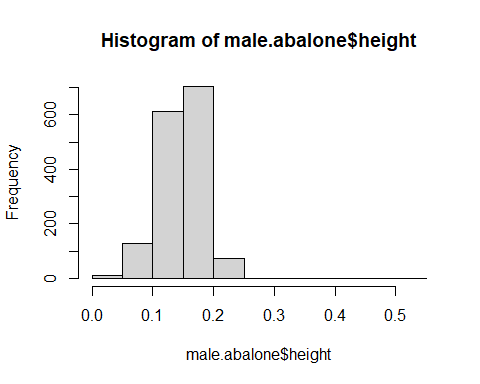


Here are the histograms of the data focused on height by gender

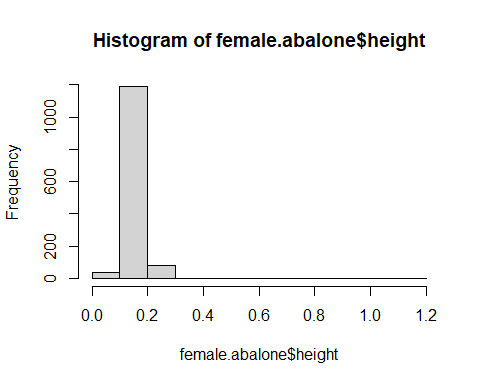
hist(abalone.data$height)



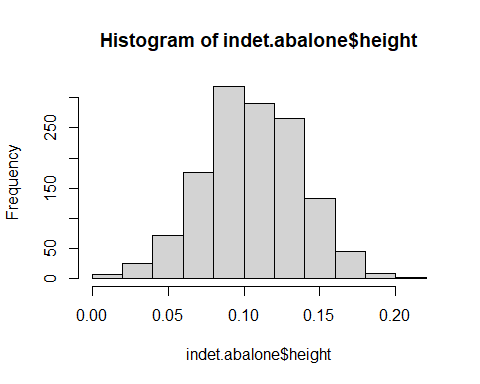
hist(male.abalone$height)



hist(female.abalone$height)

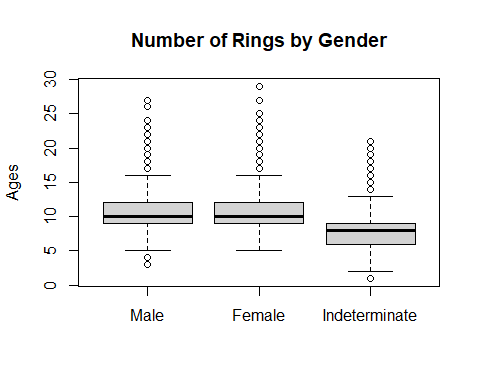


hist(indet.abalone$height)

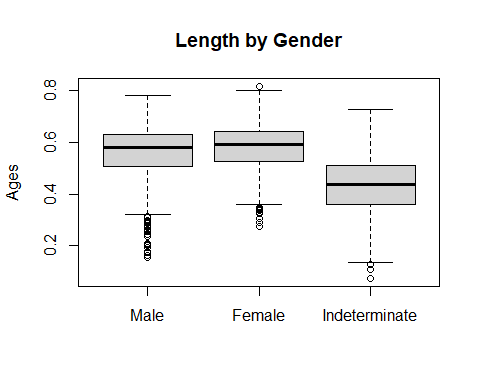


After running through some new statistical tests, I went back to last weeks assignment and created a side-by-side boxplot. As you can see, there is no real difference in ages based upon gender. This is to be expected as Abalone do not have a natural cannibalistic relationship like the praying mantis or black widow. One thing I would like to expound upon further is the utilization of ggplot2 to increase the color schema for readability.

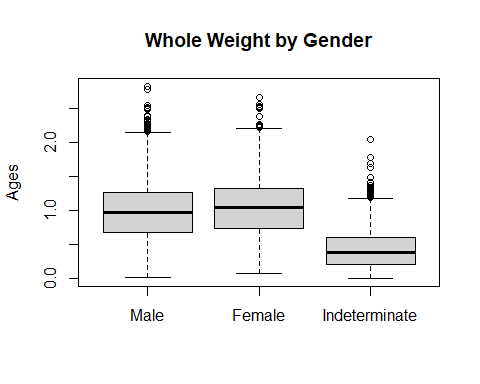
Male.Rings = abalone.data$rings[which(abalone.data$sex=='M')]   
Female.Rings = abalone.data$rings[which(abalone.data$sex=='F')]  
Indet.Rings = abalone.data$rings[which(abalone.data$sex=='I')]  
my.bp <- boxplot(Male.Rings, Female.Rings, Indet.Rings,  
 main = "Number of Rings by Gender",  
 ylab = "Ages",  
 names = c("Male", "Female", "Indeterminate"))



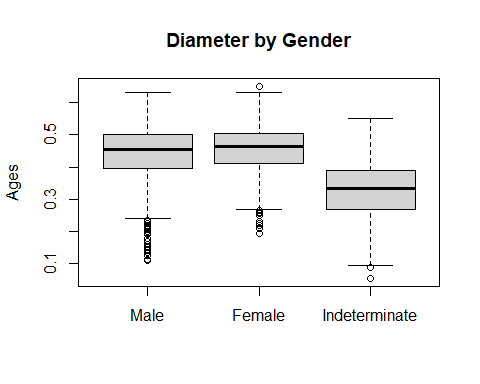
Male.Length = abalone.data$length[which(abalone.data$sex=='M')]   
Female.Length = abalone.data$length[which(abalone.data$sex=='F')]  
Indet.Length = abalone.data$length[which(abalone.data$sex=='I')]  
my.bp <- boxplot(Male.Length, Female.Length, Indet.Length,  
 main = "Length by Gender",  
 ylab = "Ages",  
 names = c("Male", "Female", "Indeterminate"))



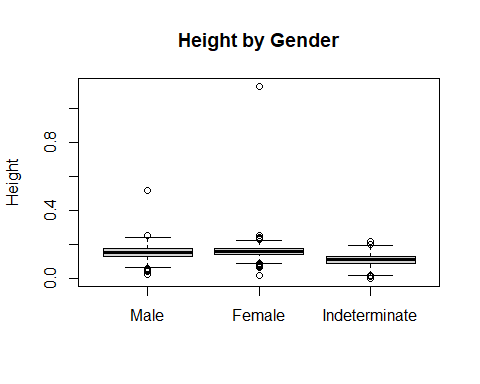
Male.WWeight = abalone.data$whole\_weight[which(abalone.data$sex=='M')]   
Female.WWeight = abalone.data$whole\_weight[which(abalone.data$sex=='F')]  
Indet.WWeight = abalone.data$whole\_weight[which(abalone.data$sex=='I')]  
my.bp <- boxplot(Male.WWeight, Female.WWeight, Indet.WWeight,  
 main = "Whole Weight by Gender",  
 ylab = "Ages",  
 names = c("Male", "Female", "Indeterminate"))



Male.Diameter = male.abalone$diameter  
Female.Diameter = female.abalone$diameter  
Indet.Diameter = indet.abalone$diameter  
my.bp <- boxplot(Male.Diameter, Female.Diameter, Indet.Diameter,  
 main = "Diameter by Gender",  
 ylab = "Ages",  
 names = c("Male", "Female", "Indeterminate"))



Male.Height = abalone.data$height[which(abalone.data$sex=='M')]   
Female.Height = abalone.data$height[which(abalone.data$sex=='F')]  
Indet.Height = abalone.data$height[which(abalone.data$sex=='I')]  
my.bp <- boxplot(Male.Height, Female.Height, Indet.Height,  
 main = "Height by Gender",  
 ylab = "Height",  
 names = c("Male", "Female", "Indeterminate"))



Some basic correlation, covariance, and standar deviation calculations

corLengthAge <- cor.test(abalone.data$length, abalone.data$rings)  
corWeightAge <- cor.test(abalone.data$whole\_weight, abalone.data$rings)   
corHeightAge <- cor.test(abalone.data$height, abalone.data$rings)  
  
covLengthAge <- cov(abalone.data$length, abalone.data$rings)  
covWeighthAge <- cov(abalone.data$whole\_weight, abalone.data$rings)  
covHeighthAge <- cov(abalone.data$height, abalone.data$rings)  
  
corLengthAge

##   
## Pearson's product-moment correlation  
##   
## data: abalone.data$length and abalone.data$rings  
## t = 43.303, df = 4175, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5354322 0.5773001  
## sample estimates:  
## cor   
## 0.5567196

corWeightAge

##   
## Pearson's product-moment correlation  
##   
## data: abalone.data$whole\_weight and abalone.data$rings  
## t = 41.498, df = 4175, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5185606 0.5615148  
## sample estimates:  
## cor   
## 0.5403897

corHeightAge

##   
## Pearson's product-moment correlation  
##   
## data: abalone.data$height and abalone.data$rings  
## t = 43.388, df = 4175, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5362051 0.5780225  
## sample estimates:  
## cor   
## 0.5574673

covLengthAge

## [1] 0.2155617

covWeighthAge

## [1] 0.8544085

covHeighthAge

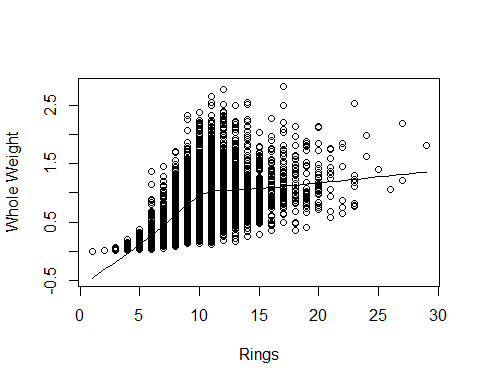
## [1] 0.07517865

sd(abalone.data$rings)

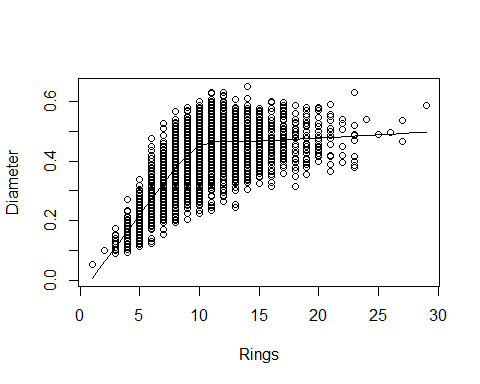
## [1] 3.224169

Smooth scatter plot for rings vs whole weight, diameter. Smooth scatter plot for whole weight and length vs diameter. Smooth scatter plot for length vs height

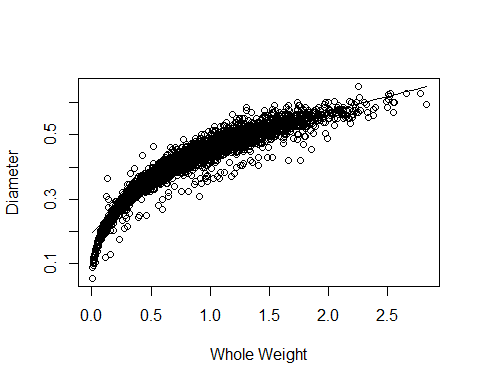
scatter.smooth(x=abalone.data$rings, y=abalone.data$whole\_weight,  
 xlab = "Rings",  
 ylab = "Whole Weight"  
)



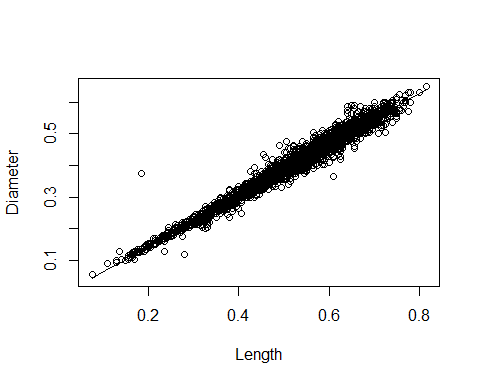
scatter.smooth(x=abalone.data$rings, y=abalone.data$diameter,  
 xlab = "Rings",  
 ylab = "Diameter"  
)



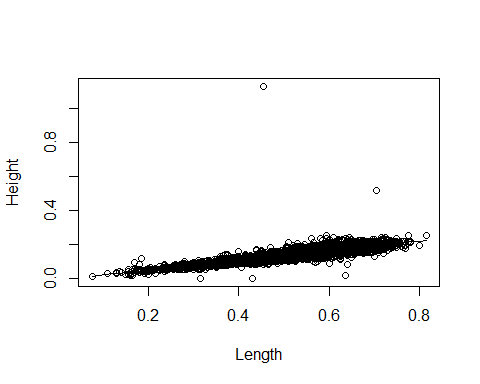
scatter.smooth(x=abalone.data$whole\_weight, y=abalone.data$diameter,  
 xlab = "Whole Weight",  
 ylab = "Diameter"  
)



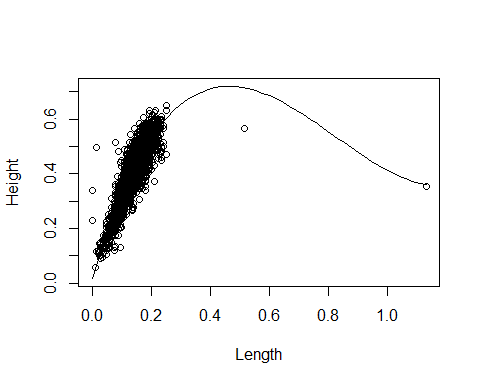
scatter.smooth(x=abalone.data$length, y=abalone.data$diameter,  
 xlab = "Length",  
 ylab = "Diameter"  
)



scatter.smooth(x=abalone.data$length, y=abalone.data$height,  
 xlab = "Length",  
 ylab = "Height"  
)



scatter.smooth(x=abalone.data$height, y=abalone.data$diameter,  
 xlab = "Length",  
 ylab = "Height"  
)



Additional least squares estimation via matrix math. Entirely focused on whole weight vs shucked weight, viscera weight, length, diameter, and height.

ww.matrix <- abalone.data$whole\_weight  
sw.matrix <- abalone.data$shucked\_weight  
v.matrix <- abalone.data$viscera\_weight  
l.matrix <- abalone.data$length  
d.matrix <- abalone.data$diameter  
h.matrix <- abalone.data$height  
head(abalone.data)

## sex length diameter height whole\_weight shucked\_weight viscera\_weight  
## 1 M 0.455 0.365 0.095 0.5140 0.2245 0.1010  
## 2 M 0.350 0.265 0.090 0.2255 0.0995 0.0485  
## 3 F 0.530 0.420 0.135 0.6770 0.2565 0.1415  
## 4 M 0.440 0.365 0.125 0.5160 0.2155 0.1140  
## 5 I 0.330 0.255 0.080 0.2050 0.0895 0.0395  
## 6 I 0.425 0.300 0.095 0.3515 0.1410 0.0775  
## shell\_weight rings  
## 1 0.150 15  
## 2 0.070 7  
## 3 0.210 9  
## 4 0.155 10  
## 5 0.055 7  
## 6 0.120 8

t.ww <- t(ww.matrix) %\*% ww.matrix  
inversed.ww <- ginv(ww.matrix)  
part2.sw <- t(ww.matrix) %\*% sw.matrix  
answer.sw <- inversed %\*%part2.sw  
answer.sw

## [,1]  
## [1,] 0.003697741

part2.v <- t(ww.matrix) %\*% v.matrix  
answer.v <- inversed %\*%part2.v  
answer.v

## [,1]  
## [1,] 0.001848333

part2.l <- t(ww.matrix) %\*% l.matrix  
answer.l <- inversed %\*%part2.l  
answer.l

## [,1]  
## [1,] 0.004480867

part2.d <- t(ww.matrix) %\*% d.matrix  
answer.d <- inversed %\*%part2.d  
answer.d

## [,1]  
## [1,] 0.00351199

part2.h <- t(ww.matrix) %\*% h.matrix  
answer.h <- inversed %\*%part2.h  
answer.h

## [,1]  
## [1,] 0.001214097

Linear Regression models Whole weight vs Rings and Diameter Viscera Weight vs Whole Weight and Shell Weight Whole Weight vs Diameter and Height and Length

lm.ring <- lm(abalone.data$rings ~ abalone.data$diameter)  
lm.ring

##   
## Call:  
## lm(formula = abalone.data$rings ~ abalone.data$diameter)  
##   
## Coefficients:  
## (Intercept) abalone.data$diameter   
## 2.319 18.670

lm.vweight <- lm(abalone.data$viscera\_weight ~ abalone.data$shell\_weight + abalone.data$whole\_weight)  
lm.vweight

##   
## Call:  
## lm(formula = abalone.data$viscera\_weight ~ abalone.data$shell\_weight +   
## abalone.data$whole\_weight)  
##   
## Coefficients:  
## (Intercept) abalone.data$shell\_weight   
## 0.003557 -0.140496   
## abalone.data$whole\_weight   
## 0.254110

lm.wweight <- lm(abalone.data$whole\_weight ~   
 abalone.data$diameter +   
 abalone.data$height +   
 abalone.data$length)  
lm.wweight

##   
## Call:  
## lm(formula = abalone.data$whole\_weight ~ abalone.data$diameter +   
## abalone.data$height + abalone.data$length)  
##   
## Coefficients:  
## (Intercept) abalone.data$diameter abalone.data$height   
## -1.095 1.857 1.753   
## abalone.data$length   
## 1.759

Testing collinearity using VIF function

car::vif(lm.wweight)

## abalone.data$diameter abalone.data$height abalone.data$length   
## 39.555221 3.288735 38.277084

AOV tests regression fit

aov(lm.ring)

## Call:  
## aov(formula = lm.ring)  
##   
## Terms:  
## abalone.data$diameter Residuals  
## Sum of Squares 14335.66 29074.97  
## Deg. of Freedom 1 4175  
##   
## Residual standard error: 2.638951  
## Estimated effects may be unbalanced

aov(lm.vweight)

## Call:  
## aov(formula = lm.vweight)  
##   
## Terms:  
## abalone.data$shell\_weight abalone.data$whole\_weight Residuals  
## Sum of Squares 41.33685 5.66082 3.17815  
## Deg. of Freedom 1 1 4174  
##   
## Residual standard error: 0.02759376  
## Estimated effects may be unbalanced

aov(lm.wweight)

## Call:  
## aov(formula = lm.wweight)  
##   
## Terms:  
## abalone.data$diameter abalone.data$height abalone.data$length  
## Sum of Squares 860.1018 7.4881 4.8682  
## Deg. of Freedom 1 1 1  
## Residuals  
## Sum of Squares 131.7922  
## Deg. of Freedom 4173  
##   
## Residual standard error: 0.1777136  
## Estimated effects may be unbalanced

Summaries of the three regression tests

summary(lm.ring)

##   
## Call:  
## lm(formula = abalone.data$rings ~ abalone.data$diameter)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1868 -1.6932 -0.7200 0.9066 15.9999   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.3186 0.1727 13.42 <2e-16 \*\*\*  
## abalone.data$diameter 18.6699 0.4115 45.37 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.639 on 4175 degrees of freedom  
## Multiple R-squared: 0.3302, Adjusted R-squared: 0.3301   
## F-statistic: 2059 on 1 and 4175 DF, p-value: < 2.2e-16

summary(lm.vweight)

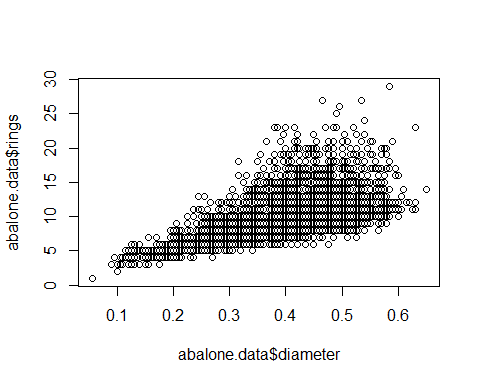
##   
## Call:  
## lm(formula = abalone.data$viscera\_weight ~ abalone.data$shell\_weight +   
## abalone.data$whole\_weight)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.176520 -0.012757 -0.001288 0.012442 0.221178   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0035567 0.0008511 4.179 2.99e-05 \*\*\*  
## abalone.data$shell\_weight -0.1404959 0.0103821 -13.532 < 2e-16 \*\*\*  
## abalone.data$whole\_weight 0.2541100 0.0029471 86.224 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02759 on 4174 degrees of freedom  
## Multiple R-squared: 0.9367, Adjusted R-squared: 0.9366   
## F-statistic: 3.086e+04 on 2 and 4174 DF, p-value: < 2.2e-16

summary(lm.wweight)

##   
## Call:  
## lm(formula = abalone.data$whole\_weight ~ abalone.data$diameter +   
## abalone.data$height + abalone.data$length)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.75120 -0.11499 -0.04166 0.07183 1.12838   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.09479 0.01277 -85.72 <2e-16 \*\*\*  
## abalone.data$diameter 1.85674 0.17428 10.65 <2e-16 \*\*\*  
## abalone.data$height 1.75268 0.11923 14.70 <2e-16 \*\*\*  
## abalone.data$length 1.75895 0.14167 12.41 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1777 on 4173 degrees of freedom  
## Multiple R-squared: 0.8688, Adjusted R-squared: 0.8687   
## F-statistic: 9208 on 3 and 4173 DF, p-value: < 2.2e-16

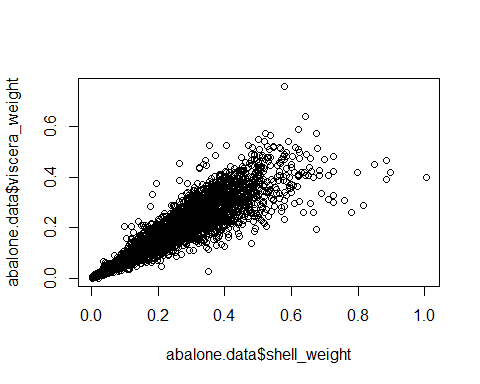
Plot of Rings vs Diameter

plot1 <- plot(abalone.data$rings ~ abalone.data$diameter, data = abalone.data)



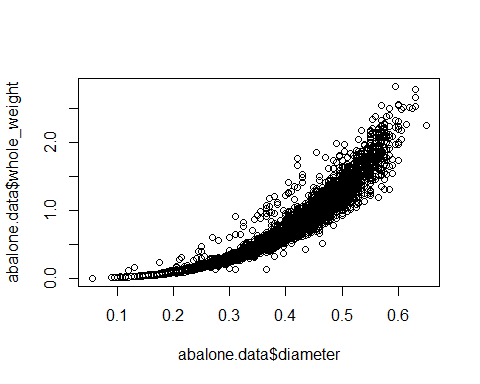
Plot of Viscera Weight vs Shell Weight

plot(abalone.data$viscera\_weight ~ abalone.data$shell\_weight, data = abalone.data)



Plot of Whole Weight vs Diameter

plot(abalone.data$whole\_weight ~ abalone.data$diameter, data = abalone.data)



Sensitivity analysis

wweight.sensitivity <- sensemakr(model = lm.wweight,   
 treatment = "abalone.data$height",  
 benchmark\_covariates = "abalone.data$diameter",  
 kd = 1:3,  
 ky = 1:3,   
 q = 1,  
 alpha = 0.05,   
 reduce = TRUE)  
wweight.sensitivity

## Sensitivity Analysis to Unobserved Confounding  
##   
## Model Formula: abalone.data$whole\_weight ~ abalone.data$diameter + abalone.data$height +   
## abalone.data$length  
##   
## Null hypothesis: q = 1 and reduce = TRUE   
##   
## Unadjusted Estimates of ' abalone.data$height ':  
## Coef. estimate: 1.75268   
## Standard Error: 0.11923   
## t-value: 14.69957   
##   
## Sensitivity Statistics:  
## Partial R2 of treatment with outcome: 0.04923   
## Robustness Value, q = 1 : 0.20313   
## Robustness Value, q = 1 alpha = 0.05 : 0.17871   
##   
## For more information, check summary.

<https://cran.r-project.org/web/packages/sensemakr/vignettes/sensemakr.html>

<http://faculty.cas.usf.edu/mbrannick/regression/regma.htm>

<https://escholarship.org/content/qt2k89z4z7/qt2k89z4z7_noSplash_79a83593875d17076c19b260e6c60fd4.pdf>

<http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r#install-and-load-required-r-packages>

<https://rstudio-pubs-static.s3.amazonaws.com/408480_f4bdfd9620c84a9598e512f1a59e66f5.html>

<http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/>

<https://github.com/nishitpatel01/predicting-age-of-abalone-using-regression>