STAT Project 2

Lauren O'Donnell (qsq6zz)

4/10/2022

Table of Contents

Abstract

max 2 pages here

Data Introduction

Data Description

The dataset selected is the `crabs` dataset imported from the `MASS` library. This dataset includes 200 observations and 7 variables. The dataset describes common features of the leptograpsus crab including sex, two color variations, and five measurements for the size of the crab.

The variables are: * *sp* : the species of the crab; O for orange crabs, B for blue crabs * *sex* : the gender of the crab: 1 for female, 2 for male * FL : frontal lobe size (mm) * RW : rear width (mm) * CL : carapace length (mm) * CW : carapase width (mm) * BD : body depth (mm)

Our group used both linear and logistic regression to find the optimal linear and logistic models that predicted the response variable CW for the linear model, and predict the *sp* for the logistic model. We used model selection techniques to achieve these goals. Our ultimate research questions were: * Is there a linear relationship between carapace width and other body measurements such as carapace length and body depth? * Do the two species of crab have significantly different carapace widths?

To determine if there is a linear relationship between carapce and the other bpdy measurements, we used linear regression to explore the relationships between the variables in the dataset, including relationships between potential predictor variables. This question, in context, helped explain whether these crabs grow proportionally, and whether certain body measurements could be predicted by others. If authorities wanted to set a minimum size regulation for the leptograpsus crab, this regression could help determine what measurement to use.

We used logistic regression to explore whether the two species of crab — blue and orange — have significantly different carapace widths. In context, a significant difference implies that any minimum size regulation should be different according to the species; otherwise, both species can be covered under the same regulation.

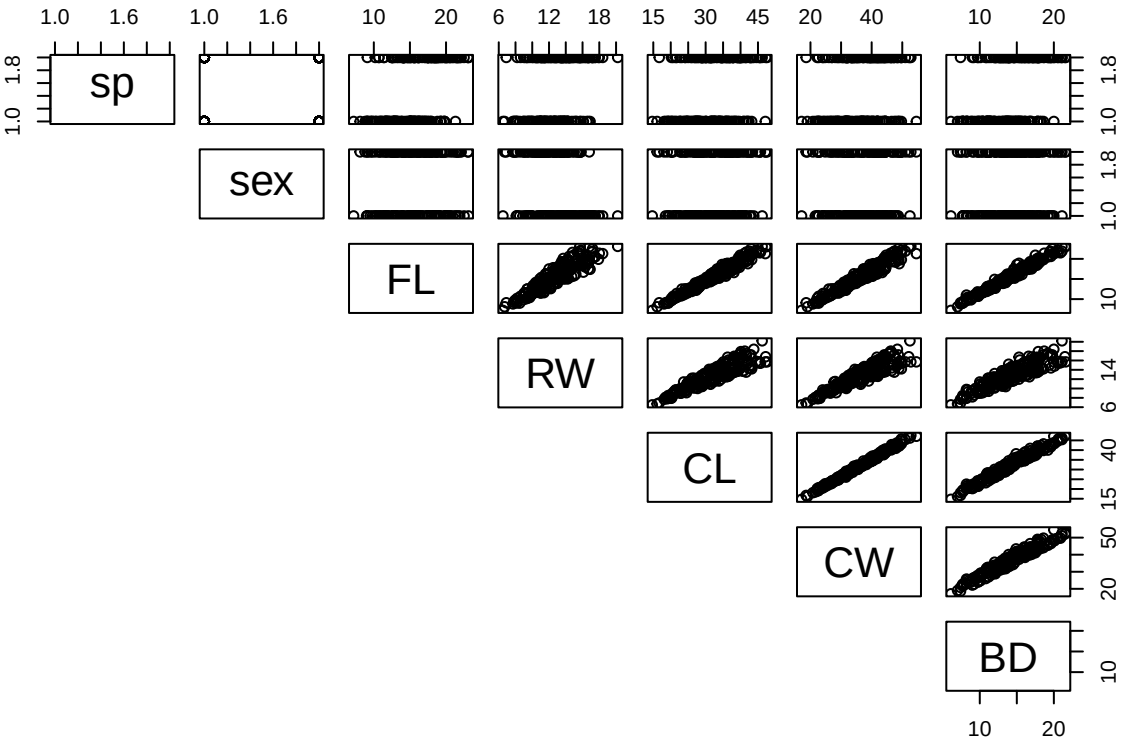
Data Cleaning

Data cleaning for the `crabs` dataset was minimal. Upon import, the dataset had two categorical variables, *sp* and *sex*, already recognized by R as factors. All observations had complete data.

Data Challenges

Upon analysis of the logistic regression model, we found that using `glm()` would not suffice for this model due to the complete separation of the variables. This lead to researching model selection techniques for logistic regression not previously explored in class. Research and reading led utilizing the `glmnet` package for model creating and selection.

Exploratory Data Analysis



Linear Regression

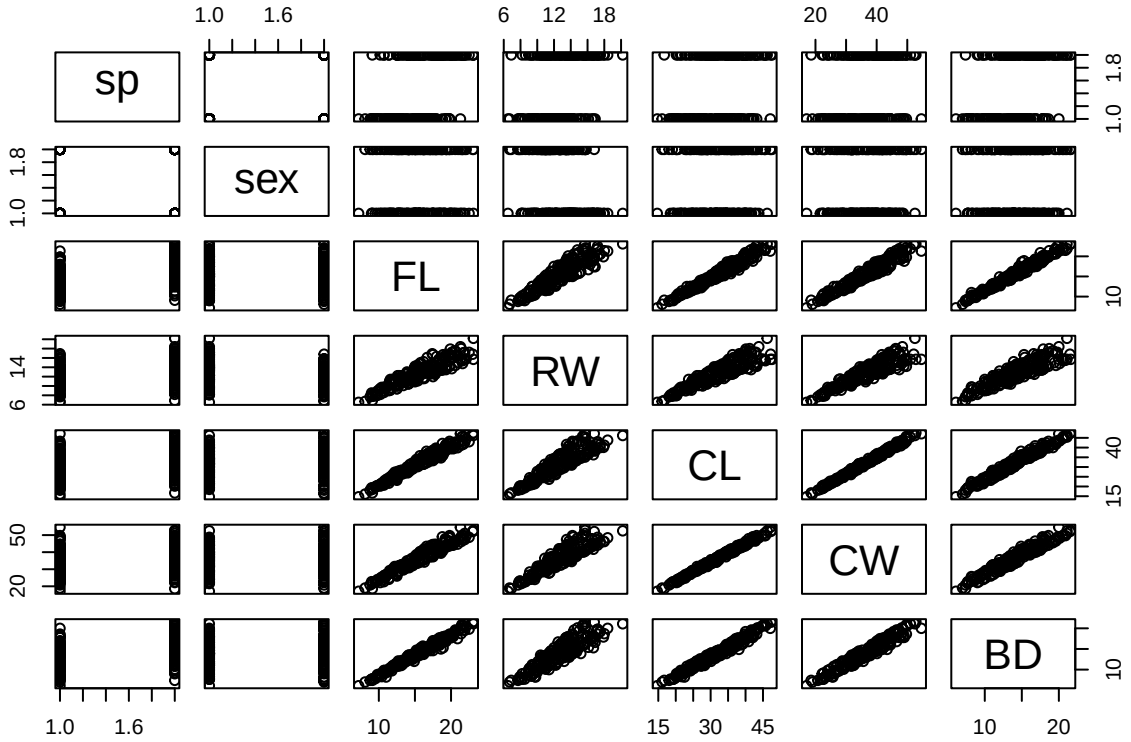
```
#GETTING BASIC INFO ABOUT DATA AND THE PREDICTORS
class(Data$sex)
```

```
## [1] "factor"
```

```
class(Data$sp)
```

```
## [1] "factor"
```

```
pairs(Data)
```




```
#all of them seem very linear
```

```
#SEEING WHICH MODEL IT SUGGESTS TO USE
allreg <- regsubsets(CW ~ ., data=Data, nbest = 2)
summary(allreg)
```

```
## Subset selection object
## Call: regsubsets.formula(CW ~ ., data = Data, nbest = 2)
## 6 Variables  (and intercept)
##      Forced in Forced out
## spO      FALSE      FALSE
## sexM     FALSE      FALSE
## FL       FALSE      FALSE
## RW       FALSE      FALSE
## CL       FALSE      FALSE
## BD       FALSE      FALSE
## 2 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          spO sexM FL  RW  CL  BD
## 1  ( 1 ) " " " "  " " " " "*" " "
## 1  ( 2 ) " " " "  " " " " " " "*"
## 2  ( 1 ) "*" " "  " " " " "*" " "
## 2  ( 2 ) " " " "  " " " " "*" "*"
## 3  ( 1 ) "*" " "  " " "*" "*" " "
## 3  ( 2 ) "*" "*"  " " " " "*" " "
## 4  ( 1 ) "*" " "  "*" "*" "*" " "
## 4  ( 2 ) "*" " "  " " "*" "*" "*"
## 5  ( 1 ) "*" "*"  "*" "*" "*" " "
## 5  ( 2 ) "*" " "  "*" "*" "*" "*"
## 6  ( 1 ) "*" "*"  "*" "*" "*" "*"
```

```r
which.max(summary(allreg)$adjr2) #7
```

```
## [1] 7
```

```r
which.min(summary(allreg)$cp) #7
```

```
## [1] 7
```

```r
which.min(summary(allreg)$bic) #7
```

```
## [1] 7
```

```r
regnull <- lm(CW ~ 1, data = Data)
regfull <- lm(CW ~ ., data = Data)

step(regfull, scope = list(lower=regnull, upper = regfull), direction = "backward")
```

```
## Start:  AIC=-323.31
## CW ~ sp + sex + FL + RW + CL + BD
##
##        Df Sum of Sq    RSS     AIC
## - BD    1     0.000 37.032 -325.31
## - sex   1     0.008 37.039 -325.27
## <none>              37.031 -323.31
## - FL    1     1.086 38.117 -319.53
## - RW    1     3.614 40.645 -306.69
## - sp    1    22.709 59.740 -229.66
## - CL    1    62.087 99.119 -128.40
##
## Step:  AIC=-325.31
## CW ~ sp + sex + FL + RW + CL
##
##        Df Sum of Sq     RSS     AIC
## - sex   1     0.008  37.039 -327.27
## <none>               37.032 -325.31
## - FL    1     1.096  38.127 -321.48
## - RW    1     3.617  40.649 -308.67
## - sp    1    33.948  70.979 -197.19
## - CL    1   112.745 149.777  -47.83
##
## Step:  AIC=-327.27
## CW ~ sp + FL + RW + CL
##
##        Df Sum of Sq     RSS     AIC
## <none>               37.039 -327.27
## - FL    1     1.097  38.136 -323.43
## - RW    1    10.964  48.003 -277.41
## - sp    1    35.675  72.714 -194.36
## - CL    1   160.458 197.498    5.48
```

```
##
## Call:
## lm(formula = CW ~ sp + FL + RW + CL, data = Data)
##
## Coefficients:
## (Intercept)          spO           FL           RW           CL
##      0.2971      -1.5079       0.1884       0.2250       0.9677
```

```
#all suggest lm(CW ~ sp + FL + RW + CL, data = Data)
```

```
#EVALUATING MODEL WITH TESTS/SEEING IF NEED TO TRANSFORM
aa <- lm(CW ~ sp + FL + RW + CL, data = Data)
summary(aa)
```

```
##
## Call:
## lm(formula = CW ~ sp + FL + RW + CL, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27048 -0.27146  0.00743  0.26648  1.17756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.29709    0.15728   1.889   0.0604 .
## spO         -1.50793    0.11003 -13.705  < 2e-16 ***
## FL           0.18839    0.07840   2.403   0.0172 *
## RW           0.22502    0.02962   7.597 1.23e-12 ***
## CL           0.96773    0.03330  29.065  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4358 on 195 degrees of freedom
## Multiple R-squared:  0.997,  Adjusted R-squared:  0.9969
## F-statistic: 1.618e+04 on 4 and 195 DF,  p-value: < 2.2e-16
```

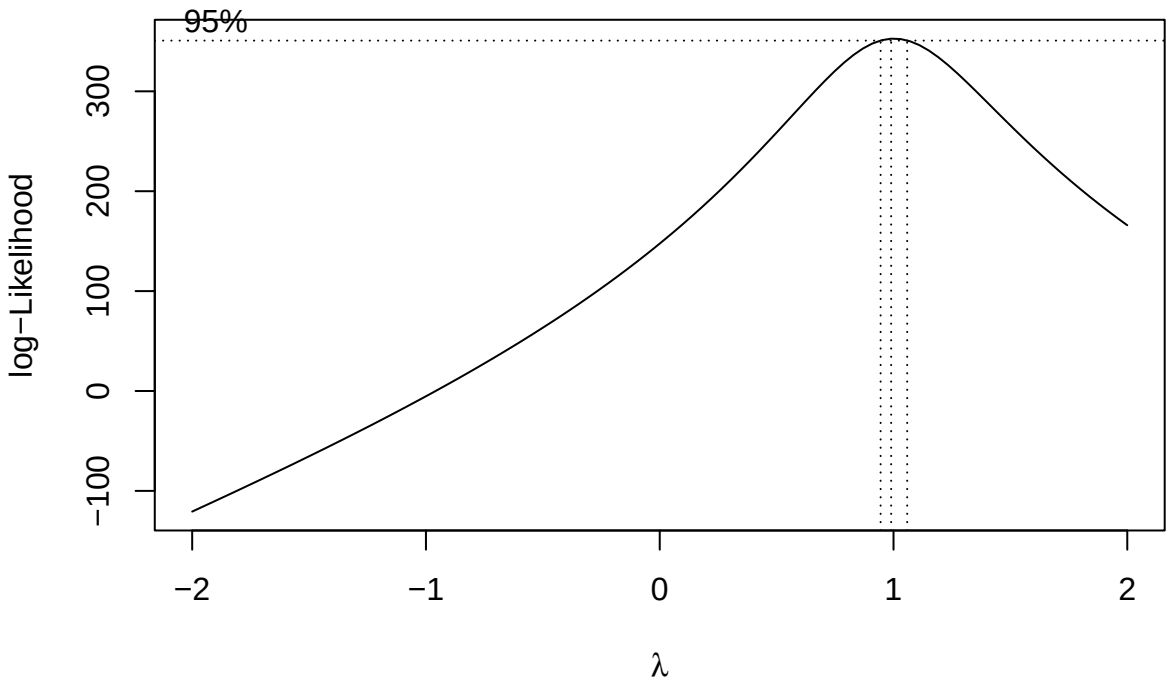
```
#F stat is high and p-value low so passes F test
#p-values for each predictor are low so significant (do not need to drop)


vif(aa)
```

```
##       spO        FL        RW        CL
##  3.186920 78.674089  6.085915 58.861592
```

```
#I think this suggests multicollinearity
```

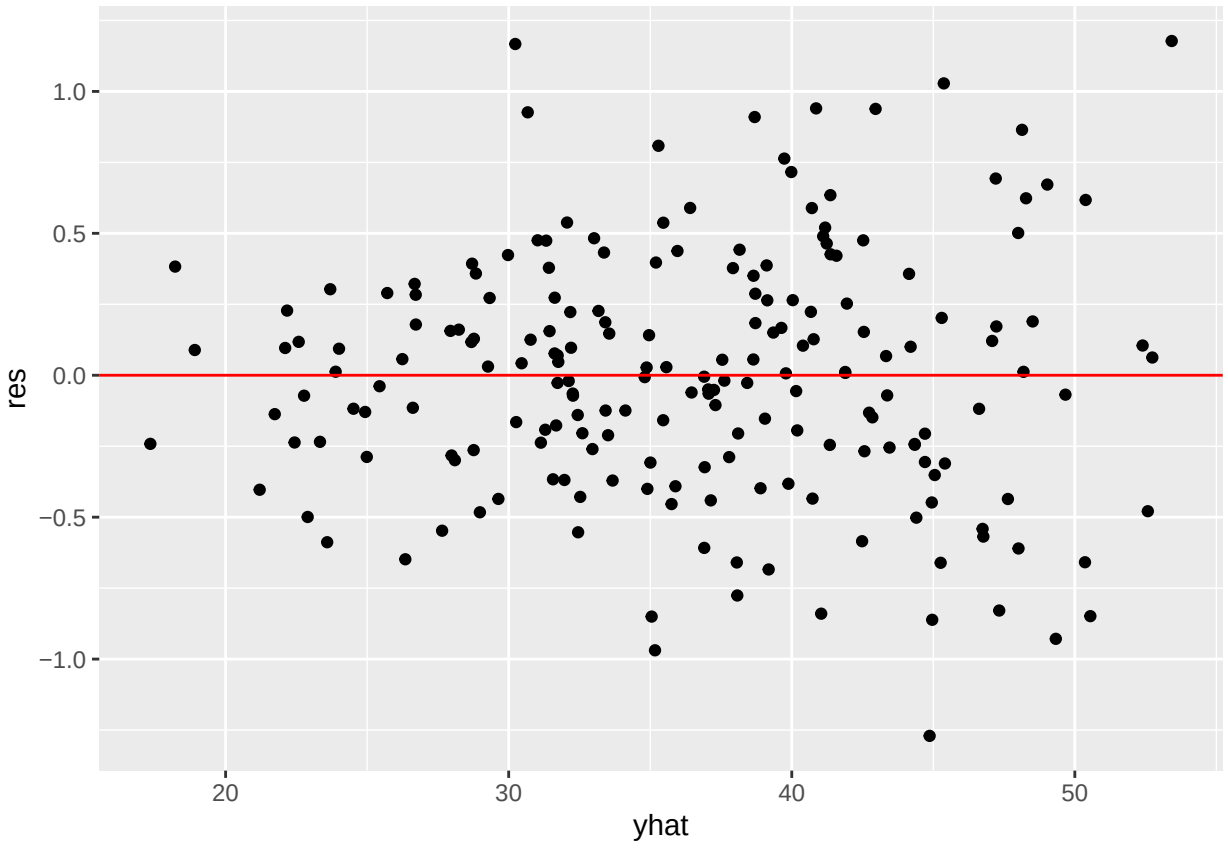
```
boxcox(aa)
```




```
#1 is in interval but we will plot residuals to make sure
```

```
yhat <- (aa$fitted.values)
res <- aa$residuals
Data <- data.frame(Data, yhat, res)

ggplot(Data, aes(x=yhat, y= res))+
  geom_point()+
  geom_hline(yintercept = 0, color = "red")
```




```
#residual plot looks really nice

#cannot seem to do qqnorm/acf plots???
```

```
#LEVERAGE AND OUTLIERS
n<-dim(data)[1]
p<-4
crit<-qt(1-0.05/(2*n), n-1-p)
ext.student.res<-rstudent(aa)
ext.student.res[abs(ext.student.res)>crit]
```

```
## named numeric(0)
```

```
#no outliers
lev<-lm.influence(aa)$hat
lev[lev>2*p/n]
```

```
## named numeric(0)
```

```
#none with leverage
```

```
DFFITS<-dffits(aa)
DFFITS[abs(DFFITS)>2*sqrt(p/n)]
```

```
## named numeric(0)
```

```
#50 and .674301???
COOKS<-cooks.distance(aa)
COOKS[COOKS>qf(0.5,p,n-p)]
```

```
## named numeric(0)
```

```
#none
```

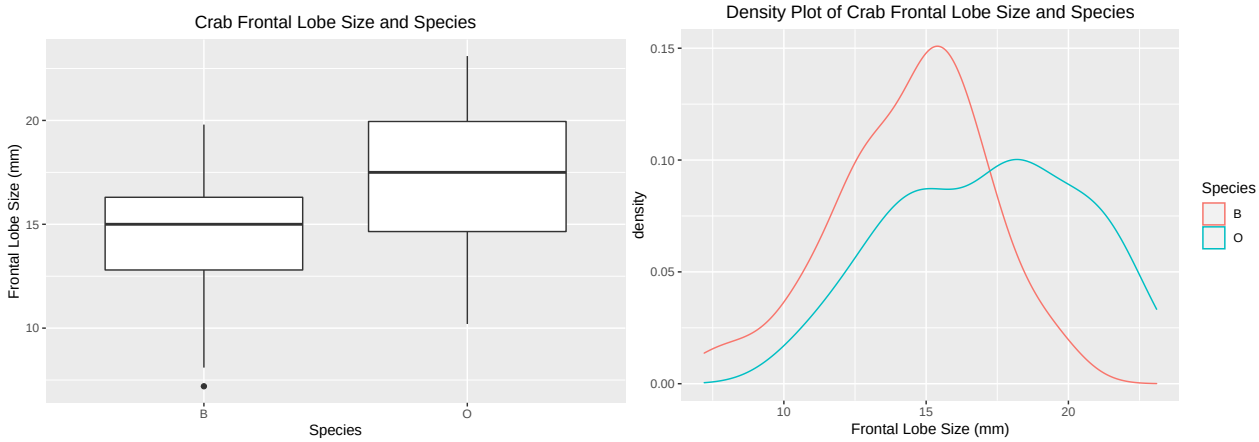
Logistic Regression

Splitting data into training and testing ses

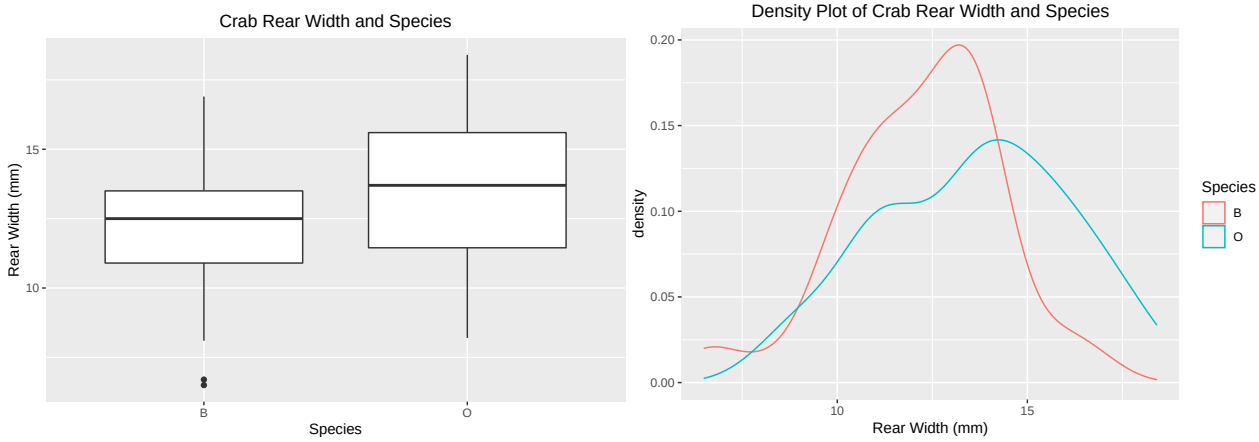
```
# splitting dataset into training and test sets
set.seed(1) # for reproducibility to get the same split
sample<-sample.int(nrow(Data), floor(.60*nrow(Data)), replace = F)
train <- Data[sample, ] # training data frame
test <- Data[-sample, ] # test data frame
```

Visualizations for Initial Analysis The y variable is species and we will control for sex in our logistic regression.

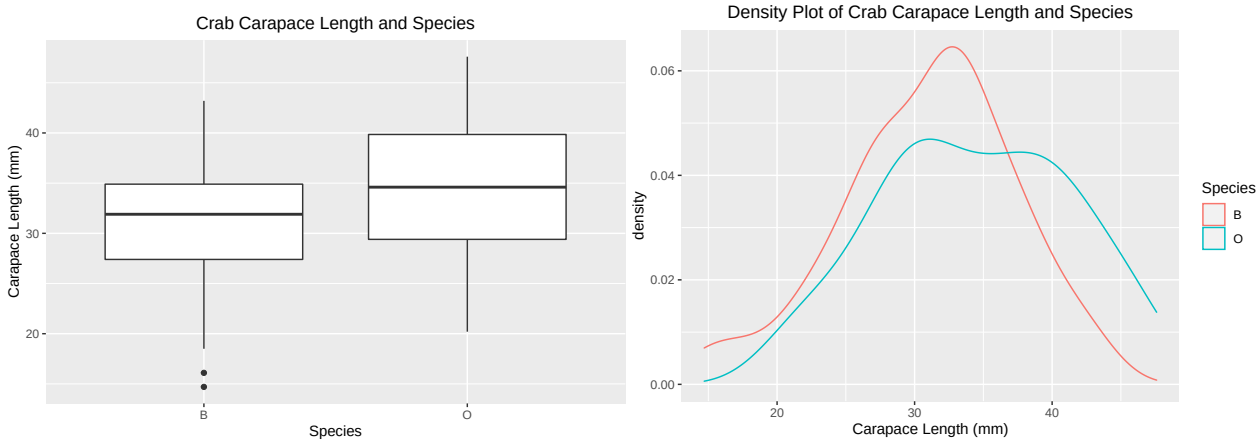
Frontal Lobe vs. Species



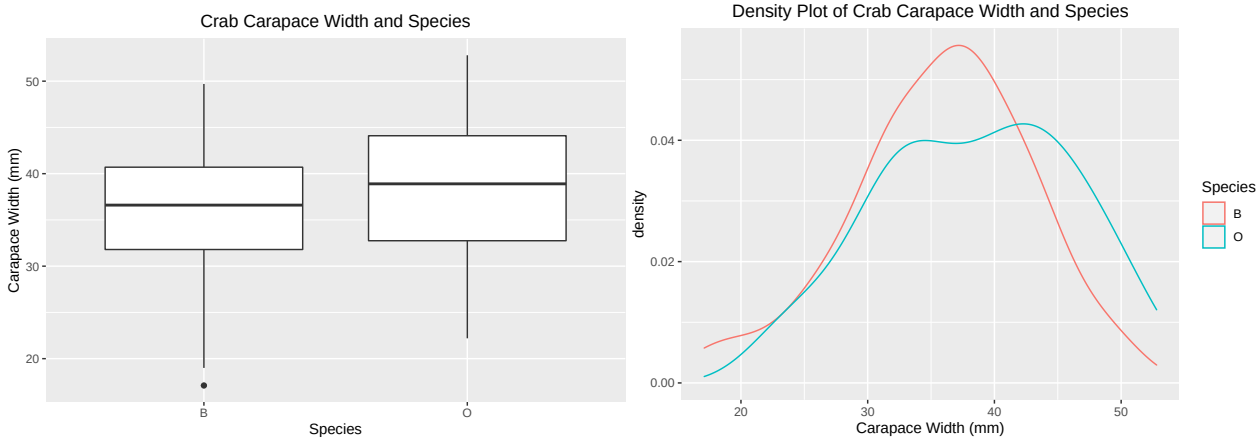
Rear Width vs. Species



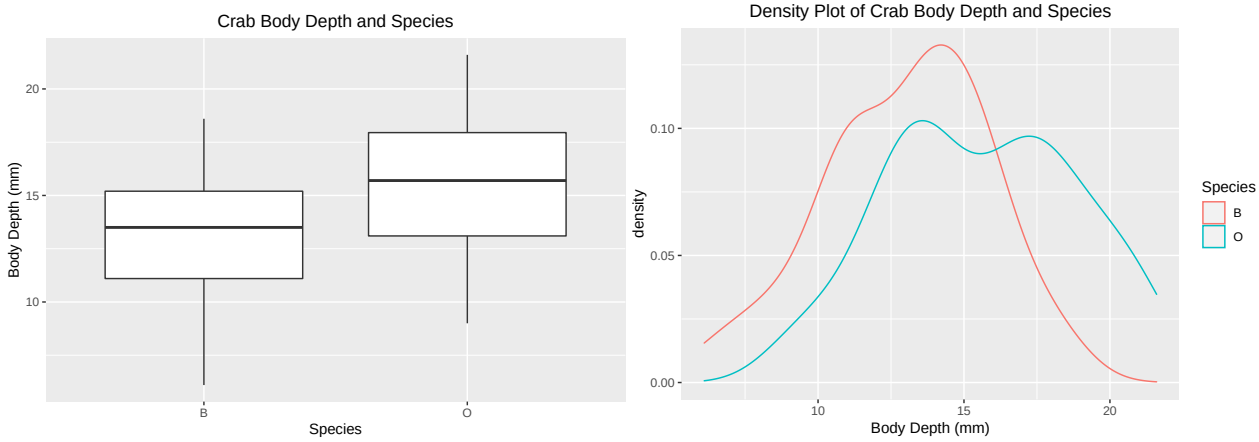
Carapace Length vs. Species



Carapace Width vs. Species



Carapace Width vs. Species



Regression Equation

```
result <- glm(sp ~sex +  FL + RW + CL + CW + BD, family = "binomial", data = train)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(result)
```

```
##
## Call:
## glm(formula = sp ~ sex + FL + RW + CL + CW + BD, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -3.003e-05  -2.100e-08   2.100e-08   2.100e-08   3.881e-05
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -59.51  673540.55       0        1
## sexM            20.79  198537.68       0        1
## FL              37.86  173321.94       0        1
## RW              13.86  126855.51       0        1
## CL              13.41  154407.14       0        1
## CW             -38.73   94869.94       0        1
## BD              19.30  152361.41       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.6472e+02  on 119  degrees of freedom
## Residual deviance: 3.5970e-09  on 113  degrees of freedom
## AIC: 14
##
## Number of Fisher Scoring iterations: 25
```

None of the variables as showing as significant and 2 error messages were received: Warning: glm.fit: algorithm did not converge Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

These errors are due to complete/perfect separation of the variables. Based on this information, we will instead focus our logistic model on only a few of the variables to predict the crab color.