Milla Nielsen, Alexis Adzich, Rohan Saklani, Michael Gureghian

Introduction

Our research question is: In this study, a predictive model of the age of crabs was constructed to study how certain crab physical attribute variables affect the age of the crab. We will attempt to answer this question using a dataset from Kaggle called, “Crab Age Prediction”. This data set measured 3,893 crabs in the Boston area, recording their Age, Sex which is either Male, Female or Indeterminate, Length in feet, Diameter in feet, Height in feet, Weight in ounces, Viscera Weight in ounces which is the weight of the abdominal organs, Shucked Weight in ounces which is weight without shell, and Shell Weight in ounces. This dataset was collected to help crab farmers increase their harvests.

The method that we chose to model the data was a multivariate linear regression model. This model seems to be the obvious choice for multiple reasons. Firstly, we can guess that Age may have a positive linear relationship with variables such as weight, length and size. All animals increase in size as they get older. Additionally we can look at the scatterplot matrix. The matrix will be further elaborated on below when we discuss the distribution of each variable and the relationship between the variables, however based on this matrix we can conclude that a linear regression model is appropriate. This is because all the response variables have a positive linear relationship with the Age variable.

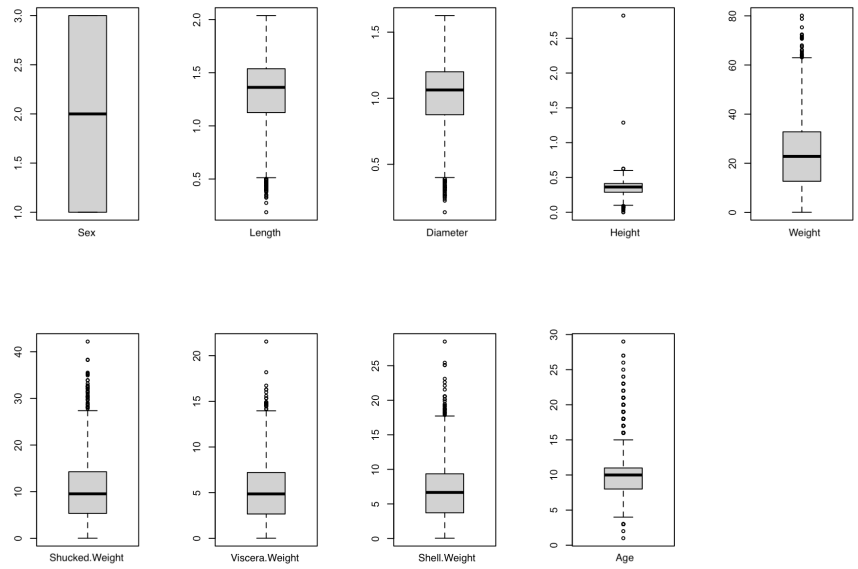
The paper’s structure will be as follows. It begins with an introduction which will introduce our project, our data, and the model type. Next the paper will cover the variables in our data set, describing their summary statistics and relationships with other variables. Then the paper will go through the model selection beginning with the full model, transformations, and then reduced versions. At each step of the way we will be using diagnostic tools to determine which models best represent the data. At the end of the paper we will discuss the project, its limitations and real-world implications.

Data Description

The summary statistics for each of our variables are as follows:

Feature	Mean	Standard Deviation
Sex	2.054	0.82
Length	1.31	0.3
Diameter	1.02	0.25
Height	0.35	0.1
Weight	23.6	13.9
Shucked.Weight	10.2	6.27
Viscera.Weight	5.14	3.1
Shell.Weight	6.8	3.94
Age	9.95	3.22

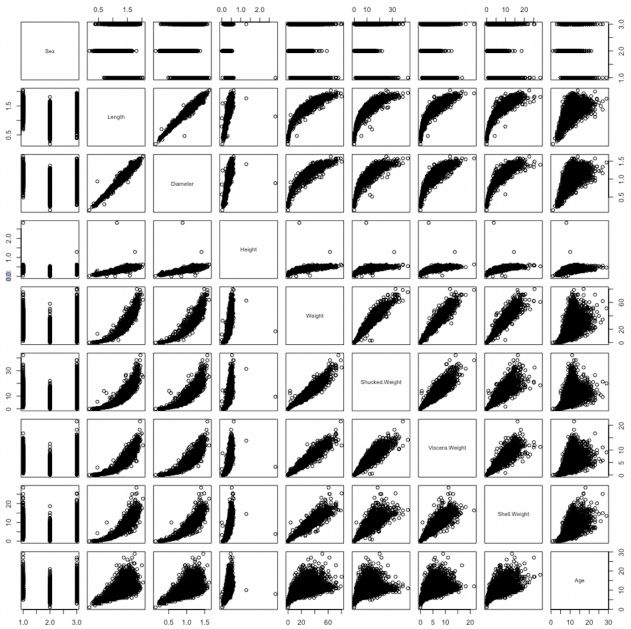
The boxplots of these variables are as follows:



The correlations of the variables are as follows, using the cor() function in R:

	Sex	Length	Diameter	Height	Weight	Shucked.Weight	Viscera.Weight	Shell.Weight	Age
Sex	1.000000000	-0.03497498	-0.03839399	-0.04162915	-0.0217729	-0.002352987	-0.0321105	-0.03621639	-0.03369951
Length	-0.034974977	1.00000000	0.98665319	0.82308098	0.9253735	0.898180660	0.9032528	0.89773627	0.55497328
Diameter	-0.038393987	0.98665319	1.00000000	0.82953154	0.9257697	0.893625712	0.8998103	0.90556111	0.57384432
Height	-0.041629151	0.82308098	0.82953154	1.00000000	0.8144051	0.770961137	0.7932717	0.81229046	0.55195636
Weight	-0.021772898	0.92537352	0.92576970	0.81440514	1.0000000	0.969077278	0.9655833	0.95526899	0.53881944
Shucked.Weight	-0.002352987	0.89818066	0.89362571	0.77096114	0.9690773	1.000000000	0.9312796	0.88240627	0.41875977
Viscera.Weight	-0.032110499	0.90325277	0.89981031	0.79327173	0.9655833	0.931279556	1.0000000	0.90610470	0.50132777
Shell.Weight	-0.036216389	0.89773627	0.90556111	0.81229046	0.9552690	0.882406265	0.9061047	1.00000000	0.62519499
Age	-0.033699510	0.55497328	0.57384432	0.55195636	0.5388194	0.418759773	0.5013278	0.62519499	1.00000000

As you can see above many of the variables are strongly correlated. Specifically all of the variables that have “Weight” in the name have correlations that are above 0.8. This is likely because they are variations of the same measurement. Our response variable age has around a 0.5 correlation with most of the other predictor variables.



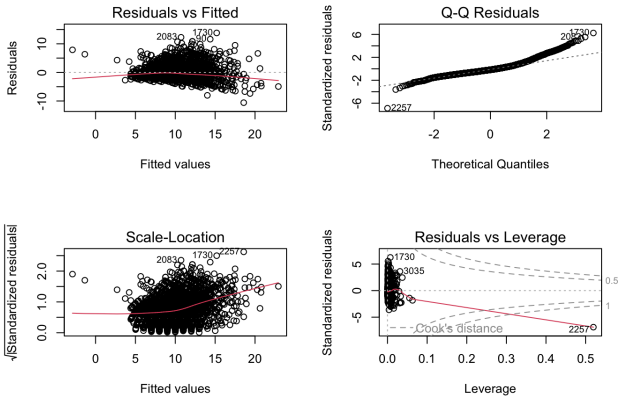
As you can see from the Scatterplot of the variables above, Age has a positive linear relationship with all of the predictor variables besides Sex which is difficult to interpret because of the categorical nature of the variable.

Transformations

i. Original Full Model

From these variables, we produce a full model from the relationship between Age and the sum of the predictor variables. The summary and diagnostic plots of this model are as shown:

```
##
## Call:
## lm(formula = Age ~ crab_data$Sex + Length + Diameter + Height +
##     Weight + Viscera.Weight + Shucked.Weight + Shell.Weight +
##     Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5818  -1.3683  -0.3821   0.9041  13.8069
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.82986    0.29655   9.543  < 2e-16 ***
## crab_data$Sex   0.06522    0.04334   1.505    0.132
## Length         -0.89036    0.75180  -1.184    0.236
## Diameter        5.84403    0.92293   6.332 2.70e-10 ***
## Height          4.52490    0.63051   7.177 8.52e-13 ***
## Weight          0.32655    0.02644  12.349  < 2e-16 ***
## Viscera.Weight -0.33782    0.04728  -7.146 1.06e-12 ***
## Shucked.Weight -0.71869    0.02985 -24.077  < 2e-16 ***
## Shell.Weight    0.29235    0.04137   7.067 1.87e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.219 on 3884 degrees of freedom
## Multiple R-squared:  0.5262, Adjusted R-squared:  0.5253
## F-statistic: 539.3 on 8 and 3884 DF,  p-value: < 2.2e-16
```



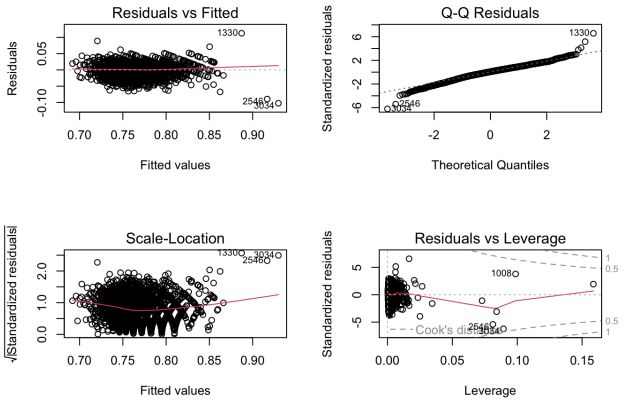
The Residuals vs Fitted plot shows that the error term is below zero, and has a slight curvature which implies some nonlinearity in the model. The Q-Q Residuals plot is heavily tailed towards both ends, showing a violation of normality. Likewise, the Scale-Location plot is not random and sees a clear upward trend, indicating that our error term does not have constant variance. The Residuals vs Leverage plot shows that there are many outliers, as well as one bad leverage point in observation 2257. Upon closer inspection, this observation describes a crab with a height drastically higher than any other observed height, with its other observations also not being well-proportioned towards this observation. Thus, we can conclude that this bad leverage point represents either a rare outlier or measurement error, but we still avoid removing it for the sake of the model's completeness.

The original full model sees the violation of virtually all its model assumptions, so we should consider a transformation.

ii. Box-Cox Transformation

We obtain one of our strongest models by taking the Box-Cox of the response and predictor variables simultaneously, which leads to the following summary and diagnostic output:

```
## Call:
## lm(formula = tAge ~ tSex + tLength + tDiameter + tHeight + tWeight +
##     tShucked.Weight + tViscera.Weight + tShell.Weight)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.102576 -0.009712  0.001275  0.011503  0.112742
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.0302766  0.0060563 170.117  < 2e-16 ***
## tSex            -0.0006558  0.0005873  -1.117 0.264211
## tLength          0.0479772  0.0098458   4.873 1.14e-06 ***
## tDiameter       -0.0366538  0.0122652  -2.988 0.002822 **
## tHeight         -0.0796964  0.0135606  -5.877 4.53e-09 ***
## tWeight         -0.2092879  0.0174626 -11.985  < 2e-16 ***
## tShucked.Weight  0.1757531  0.0077128  22.787  < 2e-16 ***
## tViscera.Weight  0.0346402  0.0089849   3.855 0.000117 ***
## tShell.Weight   -0.1421732  0.0115167 -12.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0173 on 3882 degrees of freedom
## Multiple R-squared:  0.646,  Adjusted R-squared:  0.6452
## F-statistic: 885.4 on 8 and 3882 DF,  p-value: < 2.2e-16
```



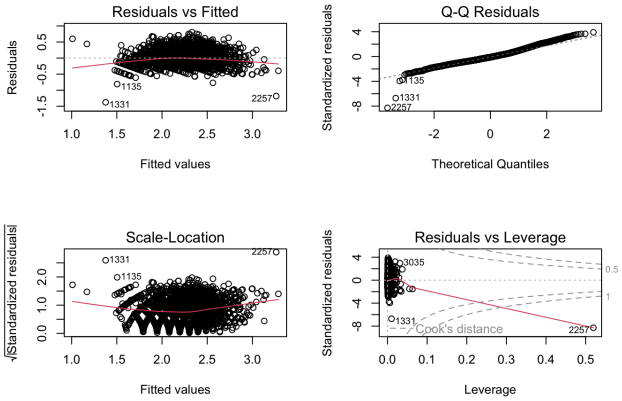
The Residuals vs Fitted plot shows stronger linearity and an error term which is statistically zero. Additionally, we have near-normality of the error term, although the Q-Q residuals plot still shows some tailing on both ends. Similarly, the Scale-Location plot again shows some upward trend in variance, although this non-constancy is significantly less drastic than in the full, unaltered model.

This model is valid. With the exception of our categorical variable, Sex, all of our variables are significant according to the summary output, and we have a usable R-squared of 0.646. If we did not find a better model, this would be perfectly reasonable to use in our study.

iii. Log(Age)

We produce a new model by taking the natural log of the response variable alone, as shown in the following summary and diagnostic output:

```
##
## Call:
## lm(formula = log(Age) ~ Sex + Length + Diameter + Height + Weight +
##     Viscera.Weight + Shucked.Weight + Shell.Weight, data = crab_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37648 -0.13832 -0.02302  0.11087  0.80433
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.221838   0.027572  44.315  < 2e-16 ***
## Sex             0.008621   0.004022   2.143   0.0322 *
## Length          0.140105   0.069777   2.008   0.0447 *
## Diameter        0.715068   0.085692   8.345  < 2e-16 ***
## Height          0.509831   0.058868   8.661  < 2e-16 ***
## Weight          0.022352   0.002475   9.033  < 2e-16 ***
## Viscera.Weight -0.025508   0.004395  -5.804 7.00e-09 ***
## Shucked.Weight -0.060632   0.002785 -21.775  < 2e-16 ***
## Shell.Weight    0.020561   0.003896   5.277 1.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.206 on 3882 degrees of freedom
## Multiple R-squared:  0.5842, Adjusted R-squared:  0.5833
## F-statistic: 681.8 on 8 and 3882 DF,  p-value: < 2.2e-16
```



Here we again see a near-zero error term with a slight curvature implying imperfect linearity. Likewise, the Q-Q Residuals plot again shows a mostly normal error term, although we now have quite heavy tailing towards the very bottom of the distribution because of some outliers. The Scale-Location plot tells a similar story as in our Box-Cox model, with near but not perfectly constant variance. Our Residuals vs Leverage plot still shows several outliers, and our bad leverage point in observation 2257 rears its head again.

Although the Box-Cox and Log(Age) models are both strong, we ultimately decide to proceed with the latter into variable selection due to its easier interpretability through the nature of log transform, as well as its marginally better diagnostic plots. However, we make a tradeoff here with this model's slightly lower R-squared of 0.5842 compared to the Box-Cox model's 0.646.

Variable Selection

After choosing our log(Age) transformation, our next step was to explore feature selection and if it may improve our model or decrease overfitting risks. To first explore this, we calculated the VIF for the independent variables, a metric for detecting multicollinearity between our predictors.

```
          Sex         Length       Diameter         Height
     1.010051      40.279353      41.472091       3.483633
       Weight Viscera.Weight Shucked.Weight   Shell.Weight
  108.327589      17.063282      27.994704      21.648861
```

While Sex and Height don't show evidence of multicollinearity, we do see some significantly high numbers supporting multicollinearity between similar variables such as Length and Diameter, or Weight, Viscera Weight, Shucked Weight, and Shell Weight. This makes sense as we'd intuitively expect these variables to be rather correlated for the same crab. Multicollinearity can cause our model to make less accurate predictions and also have a more difficult time distinguishing variable significance if many of them are correlated with each other.

This suggests there might be evidence for a model eliminating some of our features if they are dependent on another. To see if there's a more optimal model on less variables to prevent multicollinearity and concerns of overfitting, we performed stepwise regression. As outlined in our code, we used both forward and backward stepwise regression, evaluating on AIC score. Both models supported the full model with a final AIC score of -12286.83, however the forward stepwise regression also reported a very close AIC for a 7-variable subset model, dropping our Length feature. Due to our high multicollinearity, we wanted to explore this subset model further, first by evaluating its updated VIF scores.

```
          Sex       Diameter         Height         Weight
     1.010010       8.279845       3.481680     108.295668
Shucked.Weight Viscera.Weight   Shell.Weight
     27.809996      16.900522      21.632051
```

While still many of our variables show evidence for multicollinearity with values over 5, we see a significant decrease in many of the values, especially in the Diameter variable as it shrinks by a factor of 5 in the subset model's VIF. To compare the reduced model to the full model, we performed an ANOVA test to see if there's evidence that the reduced model is an improvement over our full model.

```
Analysis of Variance Table

Model 1: log(Age) ~ Sex + Diameter + Height + Weight + Shucked.Weight +
    Viscera.Weight + Shell.Weight
Model 2: log(Age) ~ Sex + Length + Diameter + Height + Weight + Viscera.Weight +
    Shucked.Weight + Shell.Weight
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   3883 164.86
2   3882 164.69  1   0.17103 4.0316 0.04473 *
```

Our ANOVA test resulted in a p value of .04473, which is just under the 5% significance level. Thus, we have sufficient evidence to reject the null hypothesis, which gives support for the full model and to reject the reduced model. So in conclusion, we can support that our final model is our full, 8-variable model with a log(Y) transformation on the Age response variable, as evaluated in our previous model selection section, with an Adjusted R-Squared value of 0.5833 and F statistic p-value of near 0. Our final model with intercept and coefficients for the predicted log(Age) is as follows:

$$\widehat{\log(\text{Age})} = 1.222 + 0.009 * \text{Sex} + 0.140 * \text{Length} + 0.715 * \text{Diameter} + 0.510 * \text{Height} + 0.022 * \text{Weight} - 0.026 * \text{Viscera.Weight} - 0.061 * \text{Shucked.Weight} + 0.021 * \text{Shell.Weight}$$

We can interpret the slopes in such a way that a unit increase in the Length variable reflects an increase in our response variable Age by a factor of $e^{0.140}$. This means that positive coefficients will result in an increase in Y for every unit increase of the predictor variable, and for negative coefficients, a unit increase in the predictor variable will result in a factor decrease in Y.

Discussion

In our project, our goal was to figure out a predictive model for estimating the age of crabs based on numerous physical attributes. From a dataset called "Crab Age Prediction" from Kaggle, we used a multivariate linear regression model to see the relationship between a crab's age with various predictors, like sex, length, diameter, height, weight, shucked.weight, viscera.weight, and shell.weight. We did this to help crab farmers because the age of a crab is a crucial indicator of prime harvest time.

Our final model selected was our full 8-variable model with a log(Y) transformation on the Age response variable. This model can be applied in a real world situation. From our model, as a crab ages, there is an increase in diameter, height, and weight of a crab, which aligns with real life, as crabs tend to become larger as they grow older. An article that can support this claim is https://link.springer.com/article/10.1007/s11250-009-9515-4.

Some limitations included a low R-squared and multicollinearity. The low R-squared might have occurred due to there being other factors affecting the crab age that is not included in the model. A way to improve this in the future would be to explore and include more predictors. For multicollinearity, there was very high correlation with weight and the predictors. Even though this can be fixed through transformations, multicollinearity can still alter the significance of predictors. So, we could remove the violating predictors which would reduce multicollinearity. Also, there was an odd pattern in our diagnostic plots, which shows the model cannot grasp the complexity of the relationship between the response variable and predictors. A way to improve this could be to use polynomial regression or add an interaction term. Additionally, we had trouble interpreting our categorical variable, Sex. This was likely due to Sex being categorized as Male, Female, and Indeterminate. It would've been easier if Sex was just a dummy variable with values of 0 and 1 instead of three categories.