Possum Analysis

Australian Possum Ear Conch Length Prediction Model

Spring
2022

Aussie Possum Enthusiasts

Timmy Diep, Ryan Sharp,
Hongjoon Kim, Angelo Randazzo

Photo of Trichosurus cunninghami (mountain brushtail possum) by Russell Best used under CC-BY 4.0 license,
Cropped by Hongjoon Kim

Table of Contents

Introduction

Discovered in the early 19th century, the Trichosurus caninus (mountain brushtail possums) are semi-arboreal, medium-sized, nocturnal marsupials that were thought to live throughout the wet sclerophyll forests of southeast Australia. In 2002, a study from Australian Journal of Zoology revealed the need to distinguish between two species: Trichosurus caninus, re-named as short-eared possum, and Trichosurus cunninghami, which retained the name, mountain brushtail possum (Lindenmeyer et al, p. 449-458). Short-eared possums live in a region stretching from southern Queensland to New South Wale ("Short-Eared Brushtail Possum Trichosurus caninus", Wires Northern River), while the other species live in a region extending from Victoria to central Queensland (McCreary, Animal Diversity Web). The purpose of this analysis is to identify an effective model that can predict the length of the ear conch of either of these possums with other physiological features, preferably with at least two continuous predictors. There are two motivations for predicting ear conch length: First, the ear conch can be difficult to measure due to its small size and geometry. Secondly, fragmentary physiological features left from decomposed remains of a possum might force one to measure few of the remaining features to predict its ear conch length with relative confidence; thus, we hope to aid one in such an extraneous circumstance.

The data set we used to achieve these goals was the *possum* data set from the Australia Journal of Zoology, accessed through the DAAG package (See References). We initially found it on Kaggle, but we loaded the data set using the DAAG library (See References). Consisting of 104 observations and 14 variables, the *possum* data set availed us with various physiological features of possums captured throughout select sites stretching from Victoria to central Queensland in 1995. (Note that the publishing date of the data set is before the mountain brushtail possum was split into two species in 2002.) We also utilized the *possumsite* data set, accessed through the DAAG package, and it supplemented the former data set with detailed information about the exact location of the sites, where the sample of possums were captured. After much exploration, we chose the following variables for our full model:

Response variable: Ear conch length (mm)

Predictor variables:

- Head length (in mm)
- Skull width (in mm)
- Tail length (in cm)
- Total Length (in cm)
- Foot Length (in mm)
- Eye (from medial canthus to lateral canthus) (in mm)
- Chest Girth (in cm)
- Belly Girth (in cm)
- Male (0 for female and 1 for male)
- Age
- Species (0 for T. caninus and 1 for T. cunninghami)

Finally, there are variables that were initially considered as possible predictors; However, as our team experimented with the data and progressively shaped the purpose of our analysis, a handful of variables were judged to be impertinent to our investigation. There was a variety of reasons for this. For instance, certain variable appears to be used for a specific purpose by the original compiler and his or her associate but ambiguous for others. Geographical variables were initially used to see if there is a geographical trend, but the small variety of observation values in these variables limited their usefulness. Certain categorical variables were swapped in favor of indicator variables for linear model fitting.

Variables deemed to be irrelevant for the purpose of the analysis:

- Site (Unique identification number for each trapping site)
- Case (Unique identification number for each possum being measured)
- Rowname (Unique name of the geographical area in which each trapping site was situated in)
    - This is a by-product of data merging.
    - The corresponding seven *indicator variables* are made from this variable.
- Longitude (of each trapping site)
- Latitude (of each trapping site)
- Altitude (of each trapping site)
- Sex ("m" for male and "f" for female)
- Pop ("Vic" for trapping sites in Victoria and "other" for trapping sites in New South Wales/ Queensland)

Data Cleaning

The original Possum dataset has 14 variables and 104 observations. We found the possum dataset on Kaggle, which informed us that this dataset was part of the DAAG R library. In the DAAG R library, our data set is called *possum*, and there is a supplementary dataset called *possumsites*. This supplementary set give use the names and geographical coordinates of the "site" variable used in *possum*. We then morphed and merged *possumsites* to give the site names and coordinates to *possum*.

Three of the possum dataset's 104 observation had null values, so we removed them.

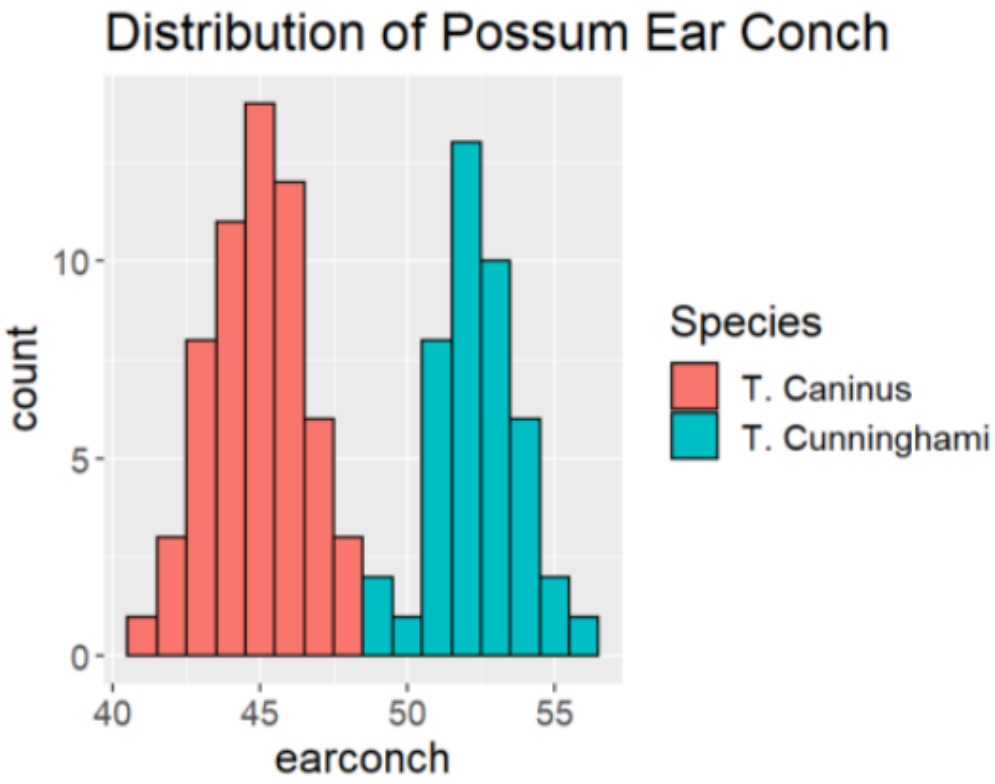
After more research, we found that in 2002, seven years after the data for our set was gathered, the scientific community learned that there are two separate species of possums within our data.

The possums from Southern Australia are generally larger than the possums from Northern Australia, so we used the location information and variable to create a new variable called species. We kept the location variables in possum, but we didn't use them since species covers what they would have told us.

	case	site	Pop	sex	age	hdlngth	skullw	totlngth	taill	footlgth	earconch	eye	chest	belly	rowname
1	1	1	Vic	m	8	94.1	60.4	89.0	36.0	74.5	54.5	15.2	28.0	36.0	Cambarville
2	2	1	Vic	f	6	92.5	57.6	91.5	36.5	72.5	51.2	16.0	28.5	33.0	Cambarville
3	3	1	Vic	f	6	94.0	60.0	95.5	39.0	75.4	51.9	15.5	30.0	34.0	Cambarville
4	4	1	Vic	f	6	93.2	57.1	92.0	38.0	76.1	52.2	15.2	28.0	34.0	Cambarville
5	5	1	Vic	f	2	91.5	56.3	85.5	36.0	71.0	53.2	15.1	28.5	33.0	Cambarville
6	6	1	Vic	f	1	93.1	54.8	90.5	35.5	73.2	53.6	14.2	30.0	32.0	Cambarville
7	7	1	Vic	m	2	95.3	58.2	89.5	36.0	71.5	52.0	14.2	30.0	34.5	Cambarville
8	8	1	Vic	f	6	94.8	57.6	91.0	37.0	72.7	53.9	14.5	29.0	34.0	Cambarville

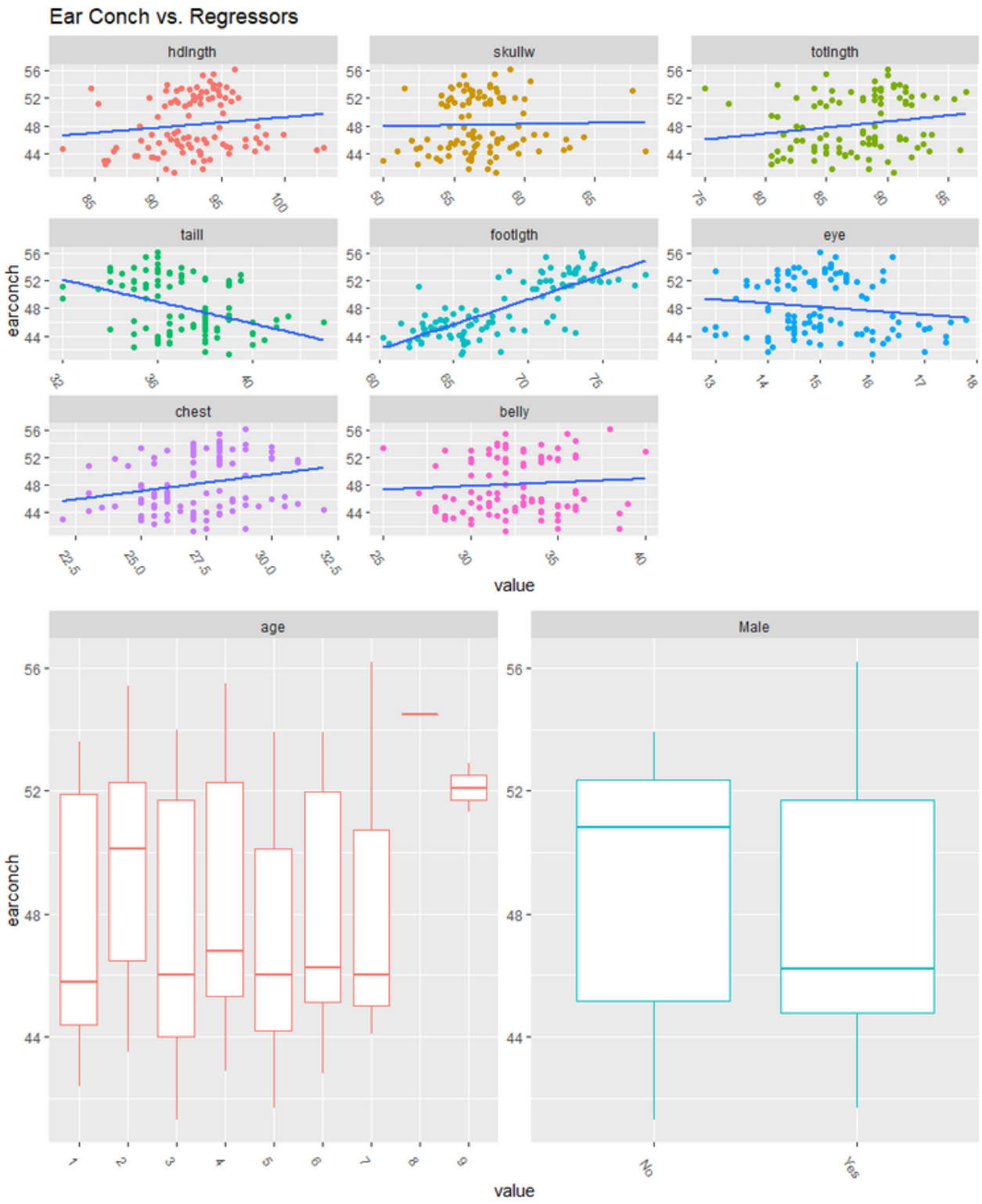
Longitude	Latitude	altitude	Cambarville	Bellbird	AllynRiver	WhianWhian	Byrangery	Conondale	Bulburin	Species	Male
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	1
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	0
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	0
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	0
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	0
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	0
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	1
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	0
145.8833	-37.55000	800	1	0	0	0	0	0	0	1	0

Exploratory Data Analysis

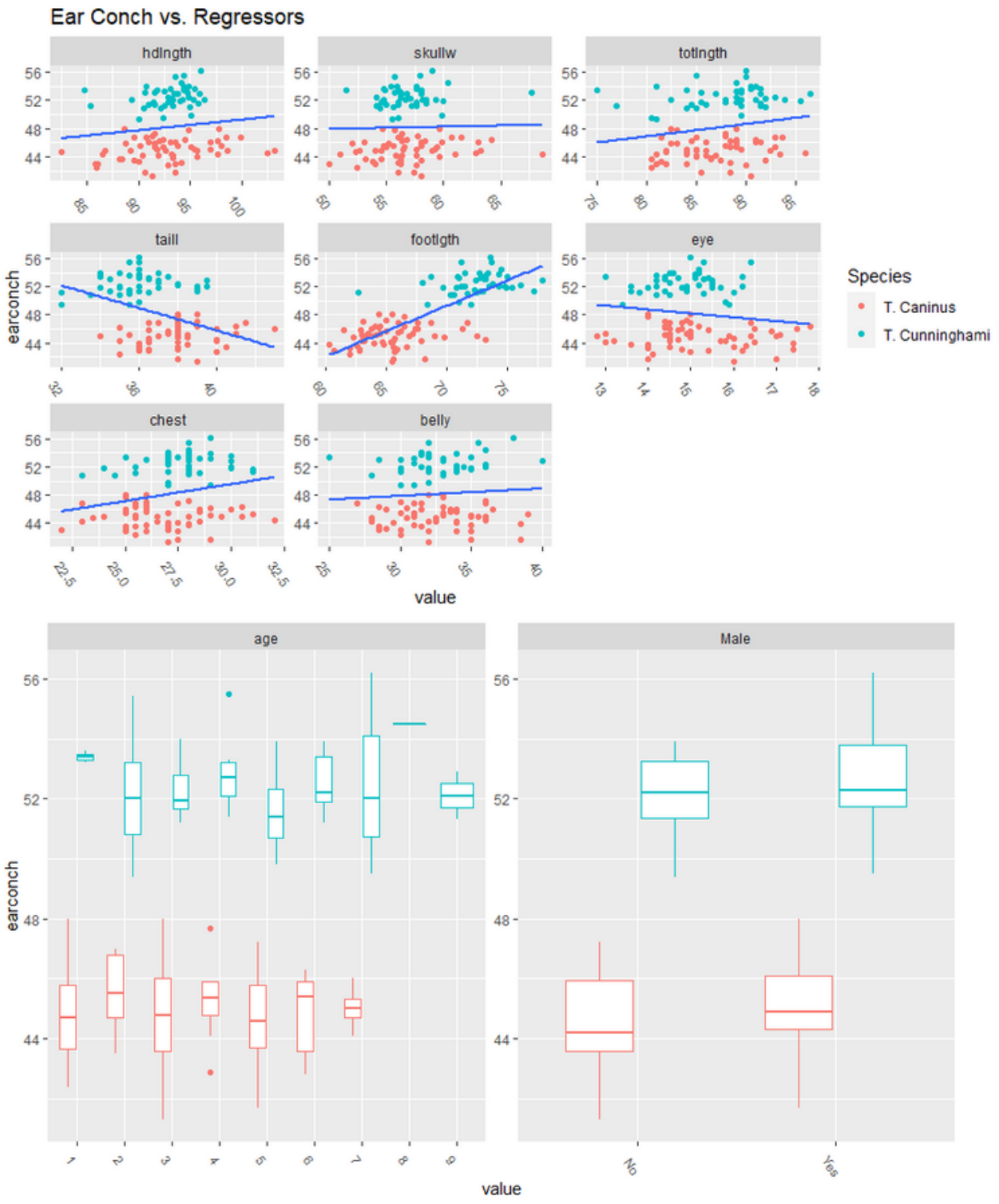


To begin our EDA, we examined the distribution of our response variable, the ear conch. In the histogram above, the ear conch's distribution is clearly bimodal. Fortunately, this is explained by the two species, T. Caninus and T. Cunninghami, having different mean ear conch size. Seeing that the range of ear conch size for each species is around 8 millimeters and the distribution is spread out within that range, this graph gives us hope that we should be able to build a useful model for predicting ear conch size.

Having considered the impact of the species variable on the distribution of the ear conch length, we then wanted to see if there were any other predictors that correlates with the ear conch length. We first approached this by constructing scatterplots (for continuous variables) and boxplots (for discrete variables) with ear conch length as the response:



We made some fascinating observations. First, there appears to be little to no linear relationship between each of the predictors and the response; the possible exceptions are tail length and foot length in which there is a negative trend in the former and a positive one in the latter. Second, we noticed that there is a horizontal gap cutting through each scatter plot. Knowing that the species in this data set was split into two different species in 2002, we distinguished between two species, and we found the following result:



Immediately, two aspects stood out to us. First, the lack of linear association between each of the predictors and the response became more apparent. Surprisingly, this seemed to impact the tail length and foot length plots, which initially appeared to have feasible linear association. As each predictor increase in value, there is neither general increase nor decrease in the response. We inferred that the species of these possums is a more significant predictor than all of the other predictors. Second, because a notable difference in the mean tail length and foot length between the two species, the two predictors appear to show some distinct characteristic of the two species.

Variable Selection

1. Looking at the full model, not many of our variables are significant. For this reason we decided to go with a forward approach starting with Species.

```
Call:
lm(formula = xEC ~ xAge + xHL + xSW + xTotL + xTailL + xFL +
    xEye + xChest + xBelly + xSpec + xSex)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0636 -0.9472 -0.1030  1.0304  3.4049

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.513271   5.041401   6.053 3.31e-08 ***
xAge        -0.052265   0.089404  -0.585    0.560
xHL          0.160607   0.080827   1.987    0.050 *
xSW          0.032873   0.075544   0.435    0.665
xTotL       -0.054227   0.072502  -0.748    0.456
xTailL       0.199059   0.136891   1.454    0.149
xFL          0.031075   0.082193   0.378    0.706
xEye        -0.221744   0.161261  -1.375    0.173
xChest      -0.142147   0.119903  -1.186    0.239
xBelly      -0.004283   0.076109  -0.056    0.955
xSpec1       7.918938   0.773086  10.243  < 2e-16 ***
xSexm        0.409309   0.340330   1.203    0.232
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.506 on 89 degrees of freedom
Multiple R-squared:  0.8775,    Adjusted R-squared:  0.8623
F-statistic: 57.95 on 11 and 89 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = xEC ~ xSpec + xHL)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4724 -1.0778 -0.0876  1.0329  3.5196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.89808    3.96753   8.796 4.94e-14 ***
xSpec1       7.53150    0.30299  24.857  < 2e-16 ***
xHL          0.10815    0.04279   2.528   0.0131 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.504 on 98 degrees of freedom
Multiple R-squared:  0.8655,    Adjusted R-squared:  0.8627
F-statistic: 315.2 on 2 and 98 DF,  p-value: < 2.2e-16

> anova(FAfit, fit)
Analysis of Variance Table

Model 1: xEC ~ xSpec + xHL
Model 2: xEC ~ xSpec
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     98 221.79
2     99 236.25 -1   -14.458 6.3885 0.01308 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

2. Length was then the next best predictor so we added it to the model and continued forward selection.

3. Chest girth is the third best predictor, and since we wanted our multilinear regression model to have multiple body measures as part of the predictors, we kept it with the 80% confidence it makes the model better. The inclusion of chest girth also made head length more significant.

```
Call:
lm(formula = xEC ~ xSpec + xHL + xChest)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3594 -0.8821 -0.0487  0.9359  3.4865

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.07592    3.98996   8.540 1.88e-13 ***
xSpec1       7.65378    0.31360  24.406  < 2e-16 ***
xHL          0.15705    0.05484   2.864  0.00513 **
xChest      -0.13909    0.09831  -1.415  0.16034
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.497 on 97 degrees of freedom
Multiple R-squared:  0.8682,    Adjusted R-squared:  0.8641
F-statistic:   213 on 3 and 97 DF,  p-value: < 2.2e-16

> anova(FAfit, fit)
Analysis of Variance Table

Model 1: xEC ~ xSpec + xHL + xChest
Model 2: xEC ~ xSpec + xHL
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     97 217.30
2     98 221.79 -1   -4.4839 2.0015 0.1603
>
```

4. The fourth round of forward approached show that any more additions made the model worse.

Comparing Models

Comparing to the full model to our final reduced model, the anova test has a P value of .5668 which is greater than .05 so we should accept the reduced model is better than the full.

```
Model 1: xEC ~ xSpec + xHL + xChest
Model 2: xEC ~ xAge + xHL + xSW + xTotL + xTailL + xFL + xEye + xChest +
    xBelly + xSpec + xSex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     97 217.30
2     89 201.98  8     15.32 0.8438 0.5668
> |
```

Variance Inflation Factor

From the output of the reduced model below, we see that the VIFs for our three predictor variables are less all than 2. Since a VIF value greater than 10 indicates multicollinearity, there is no sign of this in our predictor variables. Thus, there is no need to remove any of our three predictor variables. We also see below that the VIFs of the full model range from 1.25 to 6.5. Some of these VIFs are larger than the ones in the reduced model, but they are still less than 10, so they are not of concern.

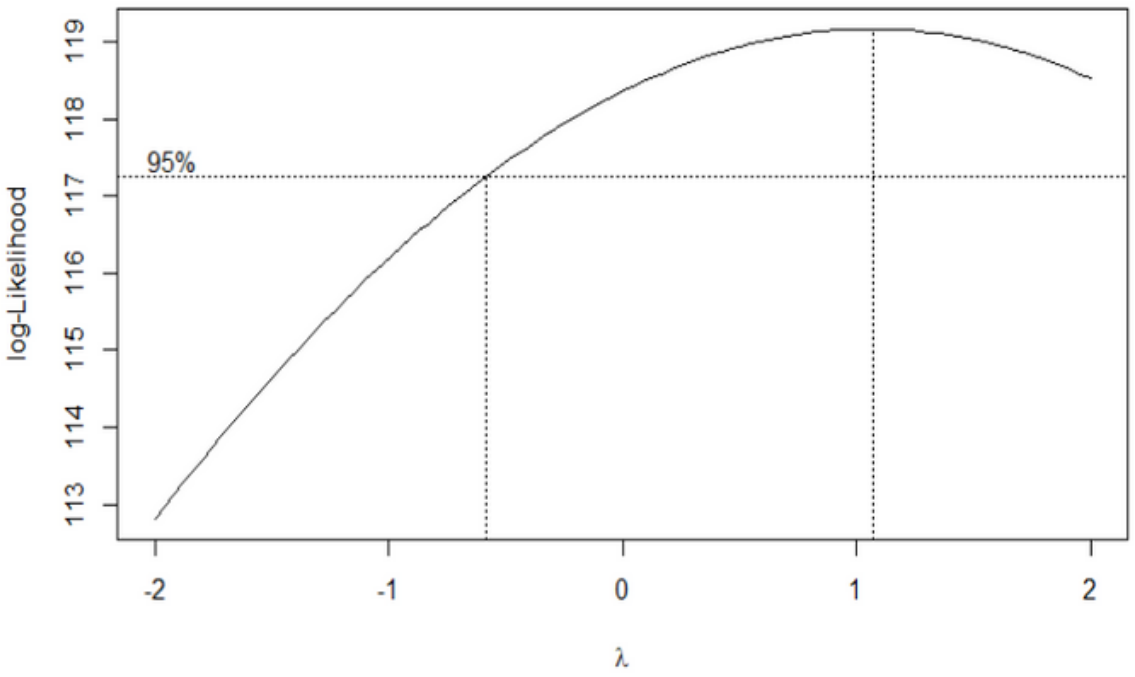
```
> #### VARIABLE INFLATION FACTORS
> vif(fit)
   xSpec       xHL    xChest
1.084029 1.661954 1.761754
> |
```

```
> vif(fullModel)
    xAge      xHL      xSW    xTotL   xTailL      xFL
1.291837 3.564145 2.420727 4.079586 3.209952 5.798500
> |
```

```
     xEye   xChest   xBelly    xSpec     xSex
1.284204 2.586710 1.899112 6.502869 1.252142
```

Box-Cox Method

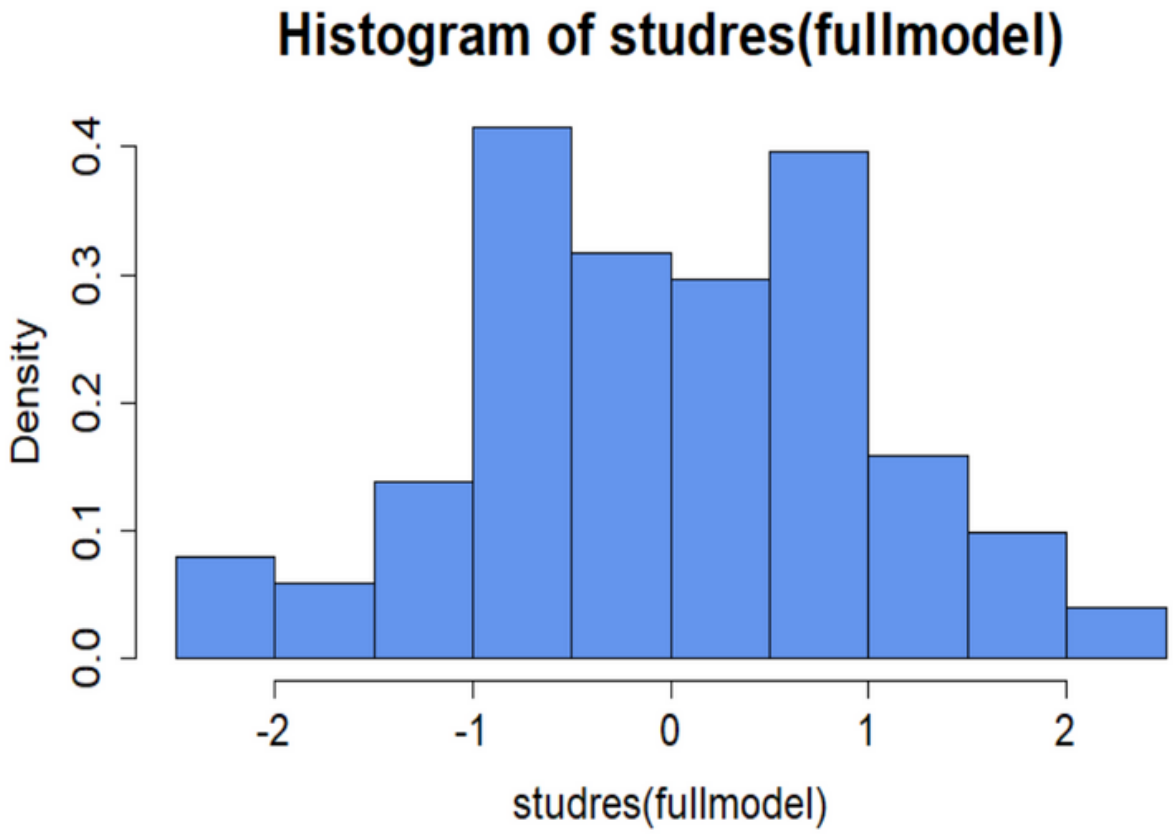
To test whether or not a transformation is necessary for our model, we take a look at the box-cox method. If a transformation was deemed necessary, then the Box-Cox method should provide possible suggestions on how the model can be transformed in order to normalize our residuals. In the case of our specific model, the graph shows a 95% confidence interval containing the lambda value of 1.0707. The reason the interval is quite large though is that the number of observations within our data set is fairly small so for a greater confidence of a smaller population, the interval needs to be quite large. Since the graph does show a lambda value of 1.0707, it is proven that it is unnecessary to perform any sort of transformation upon our data to achieve the normality assumption. To reinforce this idea, a Box-Cox transformation uses the lambda value and raises your target variable to that lambda power. In our case, this would mean taking the y^1.0707, which would lead to very little changes to our y-values (Ear Conch) and eventually our residuals.



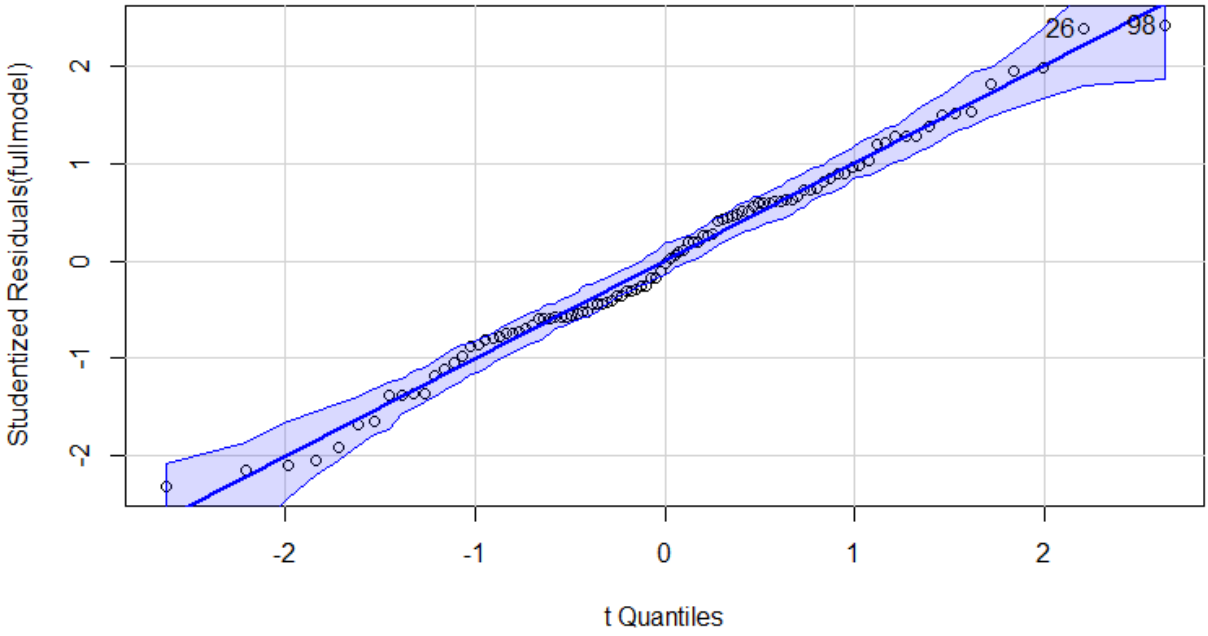
Now that there is no need for any sort of transformation, we now move on to residual analysis to further check if our residuals follow the normality assumption. If the errors of the model are not normally distributed, then the normality assumpution does not hold and the method of least squares cannot be applied to our model.

Residual Analysis

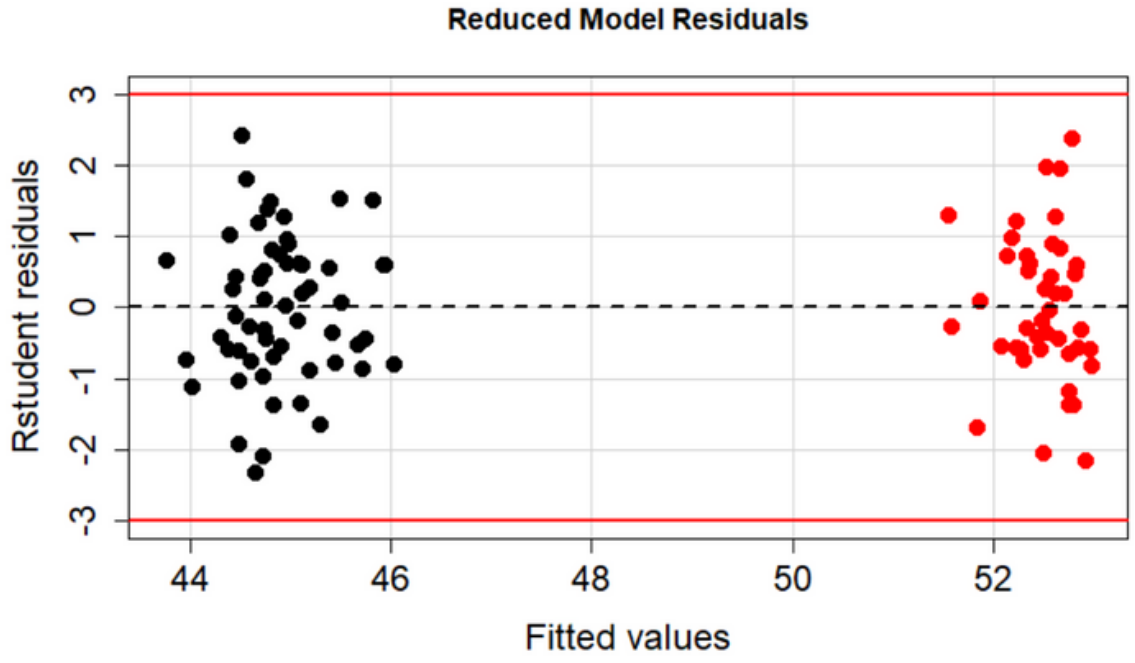
Histogram of studentized residuals: To analyze our residuals to prove that the normality assumption holds, it is necessary to look at the the histogram and QQ-plot of our models. Examining the histogram of the studentized residuals, we again see as a result of having two species that there is a bimodal graph. This is apparent as there are two separate peaks in the histogram which shows that the mean ear conch lengths vary between the two species. In further analysis, it is safe to assume that the our studentized residuals are still fairly normal. The histogram below shows a mixture distribution between two normal distributions with two different means. If the dataset we used had more observations, then it can be assumed that the residual histograms for individual species would be normally distributed.



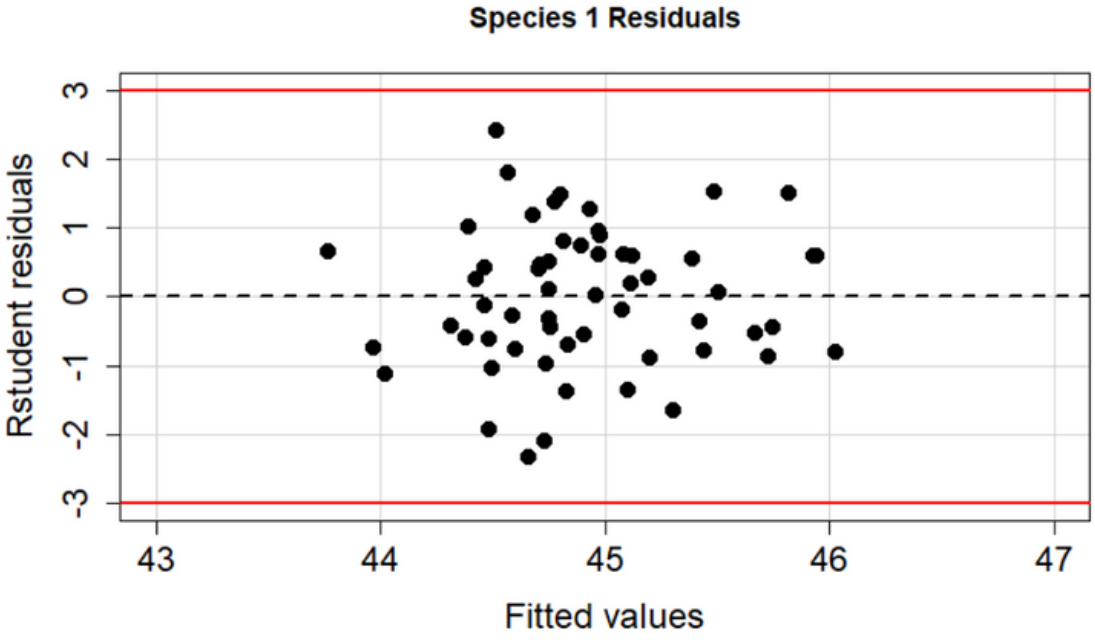
QQ-Plot: Taking a look now at the QQ-plot of the studentized residuals, we see that the observations all fit the straight line very well and all fall within the confidence bounds as well. The only observations that need to be taken into account are observations 26 and 98, which can be see as possible leverage points within our model that would require further analysis.



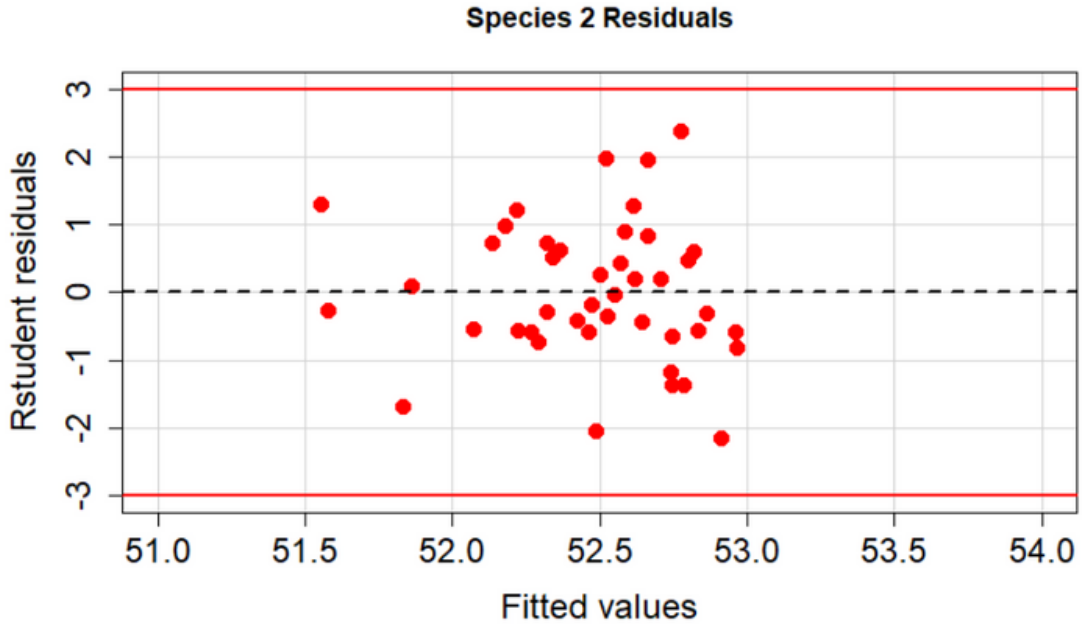
Fitted Values vs. Rstudent Residuals: Finally it is important to look at the fitted values plotted against the Rstudentized residuals to analyze the shapes of our residuals to further provide evidence that our residuals are following the normality assumption. At first glance, the residuals are fairly spread out but again due to the bimodality of our dataset, we would expect such a scatterplot to look this way. To further analyze the shapes of our residuals plotted against the fitted values, we take a zoomed in look on each set of clusters. Each set of clusters as well as the color of the points represents the species within our dataset; black being the T.Caninus and red being the T. Cunninghami. Regardless of the horizontal spread the current scatterplot has, we do recognize that none of the observations do exceed the designated cutoff values of 3 and -3. This idea supports that none of our observations are possible outliers in this case.



Taking a look at our T. Caninus species, we see that the residuals create a decent horizontal band which is the wanted outcome in examining the shape of our residuals. Again there is also no residual that exceeds our cutoff values so again we can say that there is no outliers for species 1.



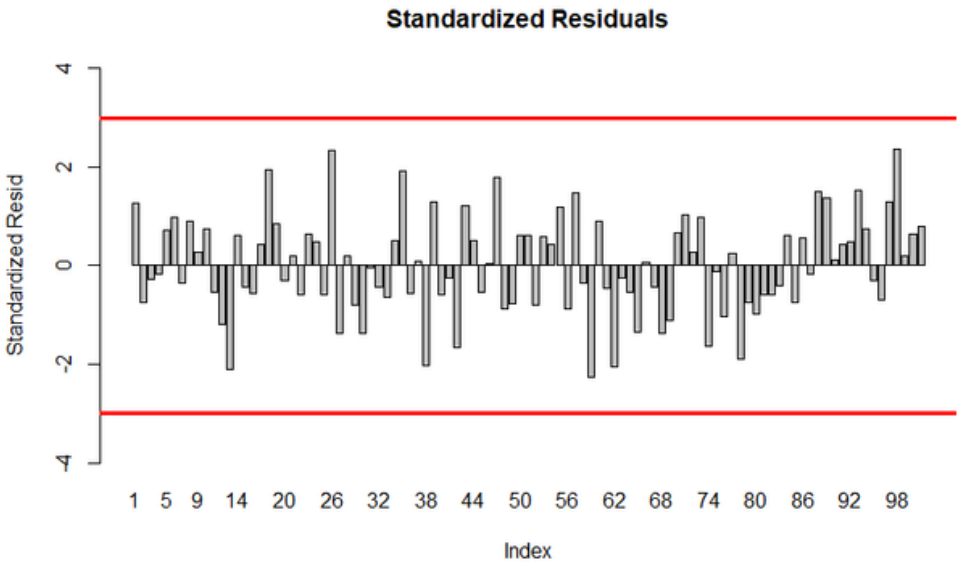
A similar approach is taken when looking at the T. Cunninghami species in which we conclude that there is no abnormal shape of our residuals. Species 2 does also vaguely show a horizontal band of residuals when plotted against the fitted values.



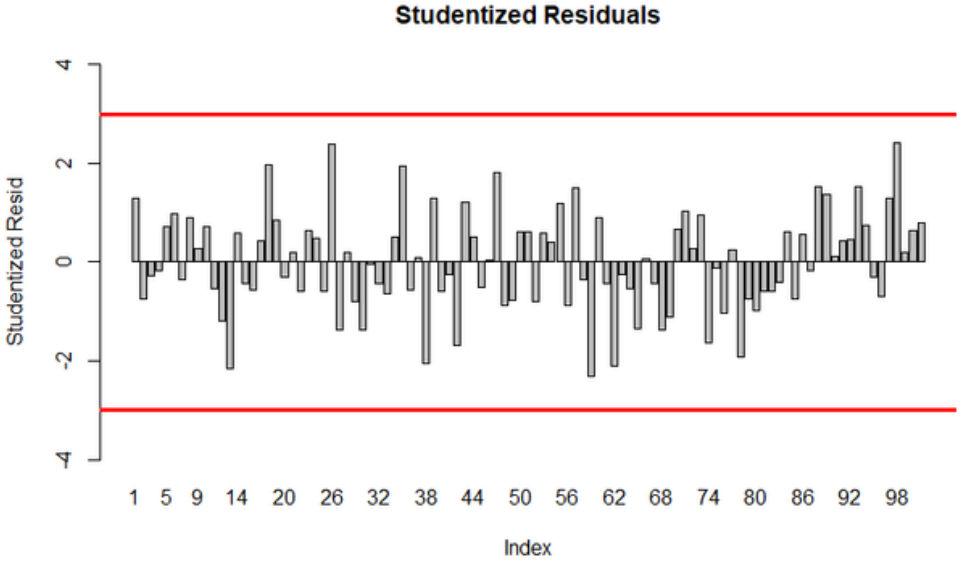
We are able to conclude that our residuals do follow the normality assumption and therefore the ability to perform the method of least squares is possible in determining our beta values for our model.

Influential Analysis

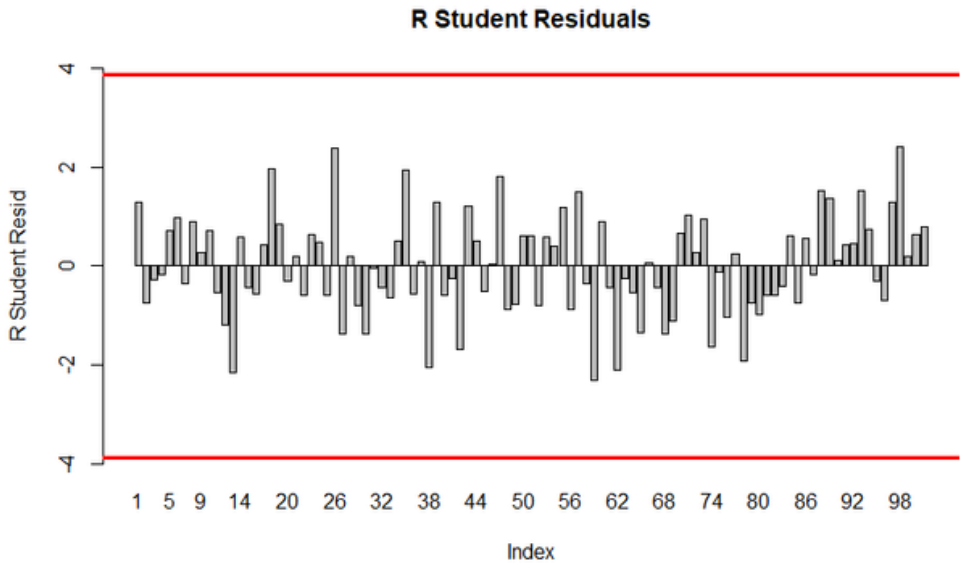
Standardized Residuals: To check for possible outlier points, we plot each method of forming our residuals and create cutoff values to determine if they of significance or not. The method of standardized residuals now takes the original residual value and then divides by the overall standard deviation of the original residuals. This in turn should allow for a more accurate assumption to check whether if there are outliers. Since our cutoff values are 3 and -3, we see that no observation can be identified as an outlier.



Studentized Residuals: We then look at the second method of residuals which is studentized. This method deletes observations one a time and looks to see how much the residuals have changed in deleted said observation. Looking at the studentized residuals, the cutoff values were set to 3 and -3 in which we are able to identify that none of the observations exceed the bounds and so we can conclude that with this method, there are still no outliers.



R Studentized Residuals: Lastly, we look at the R Student residuals method. We again see that after the method is performed, the shapes of the residuals stay fairly reminiscent of the previous two and that in calculating the cutoff values using 95% confidence, we see that there is no observation exceeding the bounds (cutoff value being .05 divided by 5 * 101.) To conclude, we see that there is no outliers present using these following methods.



Next, we want to consider any influential/leverage points for our model. To examine this, we utilized the "influence.measure" function in R. The following was the result, along with relevant plots to aid in understanding the result:

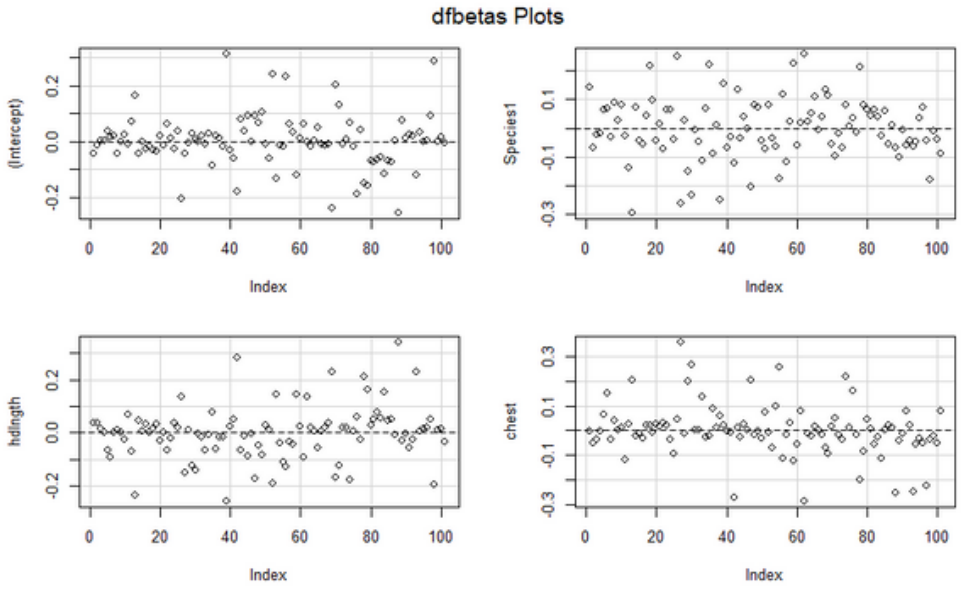
```
Potentially influential observations of
        lm(formula = earconch ~ Species + hdlngth + chest, data = poss2) :

   dfb.1_ dfb.Spc1 dfb.hdln dfb.chst dffit cov.r   cook.d hat
26 -0.20   0.25     0.14     0.04     0.43  0.86_*  0.04   0.03
52  0.24   0.08    -0.19    -0.01    -0.27  1.14_*  0.02   0.11
56  0.23   0.12    -0.13    -0.11    -0.31  1.14_*  0.02   0.11
59 -0.12   0.22     0.15    -0.13    -0.34  0.86_*  0.03   0.02
70  0.20  -0.05    -0.17     0.02     0.22  1.14_*  0.01   0.10
98  0.29  -0.18    -0.19    -0.04     0.43  0.85_*  0.04   0.03
```

DFBETA Test: The test yielded no evidence of an influential observation such that the absence of such observation significantly changes the estimation of any of the predictor coefficients. Our criterion for significance in this test is:

$$\text{Significant if } |DFBETAS_{j,i}| > 1$$
$$\text{for the ith observation with respect to } \hat{\beta}_j$$

We see that in the influential measure diagnostics and in the DFBETAS plot below that no observation has significant DFBETAS value for each predictor coefficient.



DFFIT Test: This test yielded no evidence of an influential observation such that the absence of such observation significantly changes the fitted value at the given observation. The criterion for significance in this test is:

$$\text{Significant if } |DFFITS_i| > 0.609 \approx 3\sqrt{\frac{4}{101-4}}$$
$$\text{for the ith observation}$$

We see that none of the values in the influence measure test has any significant DFFITS value.

COVRATIO Test: Using the following criterion, we found six potentially influential observation such that the inclusion of such observation impacts the stability of the estimation of the predictor coefficient matrix, β:

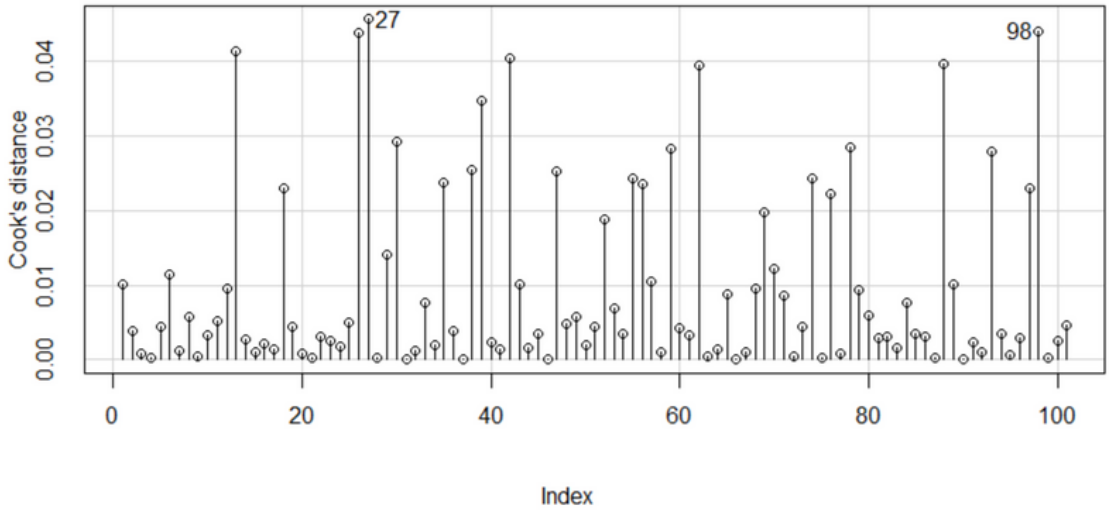
$$\text{Significant if } COVRATIO_i \epsilon (0, 1 - \frac{3*4}{101-4} \approx .876) \cup (1 + \frac{3*4}{101-4} \approx 1.124, \infty)$$
$$\text{for the ith observation}$$

Observations 26, 59, and 98 have cov.r < 1; therefore, the inclusion of these observation increases the generalized variance of the estimation of β, implying a less precise estimation of β. Observations 52, 56, and 70 have cov.r > 1; therefore, their inclusion decreases the generalized variance of the estimation of β, resulting in more precise estimation.

Cook's Distance Test: The test yielded no evidence of an influential observation such that the exclusion of such point implied a significant squared distance between the estimation of β based on all observation and that which excludes the ith observation. The criterion for significance was:

$$\text{Significant if } D_i > 0.845 \approx F_{0.5,4,101-4} \text{ for the ith observation}$$

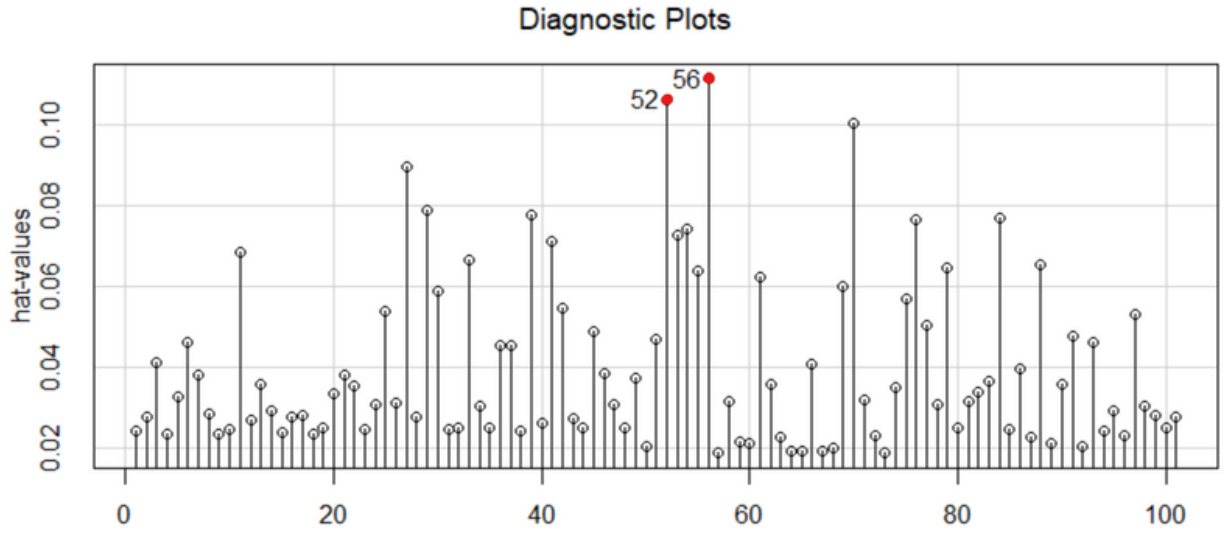
We saw that there is no observation with significant Cook's D value, as demonstrated by the influence measure test and the plot below:



Hat Matrix Test: The test did not indicate the existence of any points that exert significant leverage on the model across the X-space. The criterion for significance we utilized was:

$$\text{Significant if } h_{ii} > 0.119 \approx \frac{3*4}{101} \text{ for the ith observation}$$

While no observation in the data set had no significant hat matrix value, observation 52 and 56 come close in having significant hat matrix value that differ from the cut-off lower bound above by only about 0.09. Hence these observations do exert the most amount of leverage across the X-space of the model. The plot below confirms our result (red marks added):



The two points with the most amount of leverage across the X-space of the model (marked in red in the plot above) look like the following:

	Species	hdlngth	chest	earconch
52	0	103.1	30.5	44.9
56	0	102.5	32.0	44.5

Upon further investigation, we found that these belong to the species Trichosurus caninus (short-eared possum) but had an interesting set of characteristics. The head length for these possums were actually at least 98th percentile for the Trichosurus caninus. Their chest girth also at least 95th percentile for their own species. However, their ear sizes were below the 50th percentile for their own species. These might need to be investigated due to their extraordinarily large head length and chest girth. We noticed that these observations also had significant COVRATIO value greater than 1, as noted before, which somehow meant that, despite the leverage, these observations contributed to more precise estimation of β. Nevertheless, these observations do not leverage the model significantly enough to be considered for discarding.

Conclusion

From our analysis, we can conclude that the three best predictors of possum ear conch size are species, head length, and chest girth. Species is by far the best predictor of ear conch size, which was evident as soon as we began our exploratory data analysis. Although our dataset has many body measurements and other variables of the possums, the majority of them proved not to be useful in predicting ear conch size. With an adjusted R-squared value of 0.8623, our model performs pretty well at predicting the ear conch size. It is suspected that our model's performance is limited due to the dataset containing only about 101 observations. If our dataset had more observations, we could likely discover more significant relationships between the predictor variables and ear conch size.

Reflective Process

Reflecting back on the project, this small sample possum data set was more trouble than we could have ever thought. We started by trying to use the age as our response variable. These possums can live up to 15 or 19 years, but the oldest in our set was only 9. Also the majority of our ages were around 3 and 4, with the mean age being 3.8 and median being 3 After some linear modeling we weren't finding much linearity. Seeing that possums age to maturity around 3-4 years old, we tried to transform the data to make every age over 4 equal to 4, creating this matured group. But the models we were creating both before and after the transformation had weird fitted values plotted against the Rstudentized residuals graphs that were striped. After meeting with the professor, she explained that the response variable for multilinear regression needs to be continuous. The possum age only going from 1 to 9 at max made the variable basically discrete. The other response variables we were interested in testing at the beginning of the project were predicting site caught and sex, which are also both discrete. It was also around this time we learned our possum dataset had two different species of possum in it, split about 50/50. Trying models on one species or the other of possum left us with only 50 observations which was not nearly enough for accurate measures. Under guidance from the professor we picked one of the possum measures that was harder to measure and had a larger range of variability.

We used Ear Conch length as our response variable. Our fitted model was then the **species** of the possum, the **length** of the possums **head**; nose to neck; and the **girth** around the **chest** of the possum.

This dataset gave us exposure to working with messy real life data. There were not many observations and the linearity wasn't very prominent. This data set measuring one thing was really measuring and comparing two different things and we had to work around it. This dataset taught us to research and to dig and uncover patterns that can make or break a dataset.

References

Lindenmayer, D. B., Viggers, K. L., Cunningham, R. B., and Donnelly, C. F. 1995. Accessed Through www.kaggle.com/datasets/abrambeyer/openintro-possum.
Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala). Australian Journal of Zoology 43: 449-458

Kaggle Link: https://www.kaggle.com/datasets/abrambeyer/openintro-possum

John H. Maindonald and W. John Braun (2020). DAAG: Data Analysis and Graphics Data and Functions. R package version 1.24.
https://CRAN.R-project.org/package=DAAG

"Photo of Trichosurus Caninus (Short-Eared Possum) - Queensland Government,1977." Queensland Government, https://apps.des.qld.gov.au/species-search/details/?id=857#!lightbox-uid-3. Accessed 10 May 2022. Used under CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/legalcode). Cropped by Hongjoon Kim

Best, Russell. "Photo 75696171." INaturalist, https://www.inaturalist.org/photos/75696171. Accessed 10 May 2022. Used under CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/legalcode). Cropped by Hongjoon Kim

Lindenmayer D. B. , Dubach J. Viggers K. L. (2002) Geographic dimorphism in the mountain brushtail possum (Trichosurus caninus): the case for a new species . Australian Journal of Zoology 50, 369-393.

"Short-Eared Brushtail Possum." WIRES NORTHERN RIVERS, https://wiresnr.org/mountain-brushtail-possum/.

McCreary, Helen. "Trichosurus Cunninghami (Mountain Brushtail Possum)." Animal Diversity Web, https://animaldiversity.org/accounts/Trichosurus_cunninghami/.

Appendix

Member's Roles:

Hongjoon Kim - influential analysis and exploratory data analysis
Angelo Randazzo - model fitting and variable selection
Ryan Sharp - model fitting and exploratory data analysis
Timmy Diep - residual and transformation analysis

R Code:

```
#####################################################
# Group Aussie Possum Enthusiasts Report code
#
# Timmy Diep
# Angelo Randazzo
# Hongjoon Kim
# Ryan Sharp
#
#
#####################################################

library(dplyr) # For general data manipulation
library(tibble) # For data frame manipulation
library(ggplot2)
library(DAAG)
library(gridExtra)
library(car)
library(MASS)
library(reshape2)

###################### DATA CLEANING ############################
dirtypossumDF <- possum # LOAD Possum data set from DAAG library

possumDF <- na.omit(dirtypossumDF) # remove NA's

psDF <- possumsites # Load the possum site data set from the DAAG library

psDF <- rownames_to_column(psDF) #make possum site rownames into colums
site <- seq_len(nrow(psDF)) # Get the site number
psDF <- cbind(psDF, site) # Attach site number as column

poss1 <- right_join(possumDF, psDF, by = "site") # Combine by Site (dpylr)

# Split site names into separate columns and put a 1 if a possum was caught in that site
poss1$Cambarville = ifelse(poss1$rowname == "Cambarville", 1, 0)
poss1$Bellbird = ifelse(poss1$rowname == "Bellbird" , 1, 0)
poss1$AllynRiver = ifelse(poss1$rowname == "Allyn River" , 1, 0)
```

```
poss1$WhianWhian = ifelse(poss1$rowname == "Whian Whian" , 1, 0)
poss1$Byrangery = ifelse(poss1$rowname == "Byrangery" , 1, 0)
poss1$Conondale = ifelse(poss1$rowname == "Conondale" , 1, 0)
poss1$Bulburin = ifelse(poss1$rowname == "Bulburin" , 1, 0)

# If a possum was caught in Cambarville and Bellbird, aka the Southern possums have a 1
and northern possums have a zero
# Southern possums are Cunninghami species
poss1$Species = poss1$Cambarville + poss1$Bellbird
poss1$Species = as.factor(poss1$Species) # converting to categorical
poss1$Male = ifelse(poss1$sex == "m", 1, 0)
poss1$Male <- factor(poss1$Male, levels = c(0,1))



##################################################################
################
# LOOKING FOR A GOOD RESPONSE VARIABLE
# Ear Conch as a good range and since it is had to measure the ear conch,
# there is a reason to predict it's measurement

#                   Head Length: Var 12.38: Range = 20.6
var(poss1$hdlngth)
range(poss1$hdlngth)
103.1 - 82.5

#                   Skull Witdh: Var 9.62 : Range 18.6
var(poss1$skullw)
range(poss1$skullw)
68.6-50

#                   Total Length: Var 17.61 : Range 21.5
var(poss1$totlngth)
range(poss1$totlngth)
96.5 - 75

#                   Tail Length: Var  3.88: Range = 11
var(poss1$taill)
range(poss1$taill)
43-32

#                   Foot Length: Var 19.48 : Range 17.6
var(poss1$footlgth)
range(poss1$footlgth)
77.9 - 60.3
```

```
#                       Ear Conch: Var 16.49 : Range 14.9
var(poss1$earconch)
range(poss1$earconch)
56.2 - 41.3

#                       Eye Width: Var 1.12 : Range 5
var(poss1$eye)
range(poss1$eye)
17.8-12.8

#                       Chest Width: var 4.08: Range 10
var(poss1$chest)
range(poss1$chest)
32-22

#                       Belly Girth: var 7.44: Range 15
var(poss1$belly)
range(poss1$belly)
40 -25

###################### Exploratory Data Analysis ######################

# Distribution of Ear Conch
ggplot(poss1, aes(x=earconch, fill=Species)) +
 geom_histogram(color="black", binwidth=1) +
  ggtitle("Distribution of Possum Ear Conch") +
  theme(text = element_text(size = 15))

# Prepare Data Frame for dplyr::melt:
newposs2 <- poss1# Assign to new var for minor tweaking

newposs2$Pop <- ifelse(newposs2$Pop == "Vic", 1, 2) # Change Pop to 1 & 2
melt <- dplyr::select(newposs2,-case, -sex, -rowname,
                      -Cambarville, -Bellbird, -AllynRiver,
                      -WhianWhian, -Byrangery, -Conondale,
                      -Bulburin, -Longitude, -Latitude,
                      -altitude) # Assign new data frame

melt$Species <- as.numeric(melt$Species) # Convert Species to numeric

melt$Male <- as.numeric(melt$Male) # Convert Male to numeric

# Melt 1: Ear Conch Length vs. Continuous Predictors

meltcont <- dplyr::select(melt, -site,-Pop, -Male, -age) # Remove Discrete Var.

meltcont$Species <- ifelse(meltcont$Species == 2, "T. Cunninghami", "T. Caninus")
# Rename Species
```

```
earconchdf1 <- melt(data = meltcont, id = c("Species","earconch"))
# Melt w/ Species & earconch as variables

ggplot(data = earconchdf1, aes(x = value, y = earconch)) +
  geom_point(aes(color = variable)) +
  facet_wrap(~variable, scales = "free") +
  theme(axis.text.x = element_text(angle = -55, size = 8)) +
  ggtitle("Ear Conch vs. Regressors") +
  geom_smooth(method = "lm", se = FALSE) +
  theme(legend.position = "none")# Scatter Plots w/o Species Distinction

ggplot(data = earconchdf1, aes(x = value, y = earconch)) +
  geom_point(aes(color = Species)) +
  facet_wrap(~variable, scales = "free") +
  theme(axis.text.x = element_text(angle = -55, size = 8)) +
  ggtitle("Ear Conch vs. Regressors") +
  geom_smooth(method = "lm", se = FALSE) # Scatter Plots w/ Species Distinction


# Melt 2: Ear Conch Length vs. Discrete Predictors:

meltdisc <- dplyr::select(melt, Species, age, Male, earconch)
# Select Discrete Var.

meltdisc$Species <- ifelse(meltdisc$Species == 2, "T. Cunninghami", "T. Caninus")
# Rename Species

earconchdf2 <- melt(data = meltdisc, id = c("Species","earconch"))
# Melt w/ Species & earconch as variables

for (i in 1:length(earconchdf2$value)) {
  if (earconchdf2$variable[i] == "Male") {
    if(earconchdf2$value[i] == 2) {
      earconchdf2$value[i] <- "Yes"
    }
    else {
      earconchdf2$value[i] <- "No"
    }
  }
  else {
    next
  }
} # Re-assigned values corresponding to Male (Coercion into character vector)
```

```
ggplot(data = earconchdf2, aes(x = value, y = earconch, group = value)) +
  geom_boxplot(aes(color = variable))+
  facet_wrap(~variable, scales = "free") +
  theme(axis.text.x = element_text(angle = -55, size = 8)) +
  theme(legend.position = "none") +
  ggtitle("Ear Conch vs. Age/Sex") # Box Plots w/o Species Distinction

ggplot(data = earconchdf2, aes(x = value, y = earconch)) +
  geom_boxplot(aes(color = Species)) +
  facet_wrap(~variable, scales = "free") +
  theme(axis.text.x = element_text(angle = -55, size = 8)) +
  theme(legend.position = "none") +
  ggtitle("Ear Conch vs. Age/Sex") # Box Plots w/ Species Distinction

############## Look at full model and get fitted model ######################

xAge <- poss1$age
xHL <- poss1$hdlngth
xSW <- poss1$skullw
xTotL <- poss1$totlngth
xTailL <- poss1$taill
xFL <- poss1$footlgth
xEC <- poss1$earconch
xEye <- poss1$eye
xChest <- poss1$chest
xBelly <- poss1$belly
xSpec <- poss1$Species
xSex <- poss1$sex

fullModel <- lm(xEC ~ xAge + xHL + xSW + xTotL + xTailL + xFL + xEye + xChest + xBelly +
xSpec + xSex)

summary(fullModel)

vif(fullModel)
```

```
################## Forward Approach ################################

FAfit <- lm(xEC ~ xAge)
summary(FAfit)
FAfit <- lm(xEC ~ xHL)
summary(FAfit)
FAfit <- lm(xEC ~ xSW)
summary(FAfit)
FAfit <- lm(xEC ~ xTotL)
summary(FAfit)
FAfit <- lm(xEC ~ xTailL)
summary(FAfit)
FAfit <- lm(xEC ~ xFL)
summary(FAfit)
FAfit <- lm(xEC ~ xEye)
summary(FAfit)
FAfit <- lm(xEC ~ xChest)
summary(FAfit)
FAfit <- lm(xEC ~ xBelly)
summary(FAfit)
FAfit <- lm(xEC ~ xSex)
summary(FAfit)
################################## SPECIES IS BEST
FAfit <- lm(xEC ~ xSpec)
summary(FAfit)
###################################

# Round 2
fit <- lm(xEC ~ xSpec)

FAfit <- lm(xEC ~ xSpec + xAge)
summary(FAfit)
anova(FAfit, fit)
```

```
################################### Head Length is best
FAfit <- lm(xEC ~ xSpec + xHL)
summary(FAfit)
anova(FAfit, fit)
###################################

FAfit <- lm(xEC ~ xSpec + xSW)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xTotL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xTailL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xFL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xEye)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xChest)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xBelly)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xSex)
summary(FAfit)
anova(FAfit, fit)

# ROUND 3

fit <- lm(xEC ~ xSpec + xHL)

FAfit <- lm(xEC ~ xSpec + xHL + xChest)
summary(FAfit)
anova(FAfit, fit)
```

```
FAfit <- lm(xEC ~ xSpec + xHL + xAge)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xSW)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xTotL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xTailL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xFL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xEye)
summary(FAfit)
anova(FAfit, fit)

############################################## Chest is next best with
P of .16
FAfit <- lm(xEC ~ xSpec + xHL + xChest)
summary(FAfit)
anova(FAfit, fit)
##############################################

FAfit <- lm(xEC ~ xSpec + xHL + xBelly)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xSex)
summary(FAfit)
anova(FAfit, fit)

# ROUND 4
fit <- lm(xEC ~ xSpec + xHL + xChest)
```

```
################################ None of these make and model much better
FAfit <- lm(xEC ~ xSpec + xHL + xChest + xAge)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xChest + xSW)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xChest + xTotL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xChest + xTailL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xChest + xFL)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xChest + xEye)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xChest + xBelly)
summary(FAfit)
anova(FAfit, fit)

FAfit <- lm(xEC ~ xSpec + xHL + xChest + xSex)
summary(FAfit)
anova(FAfit, fit)

################################################
# FINAL MODEL
fit <- lm(xEC ~ xSpec + xHL + xChest)

anova(fit, fullModel)

############## VARIABLE INFLATION FACTORS ######################
vif(fit)

vif(fullModel)
```

```
################## Box - Cox  &  Residual Analysis ######################
### BOX-COX TRANSFORMATION
bc <- boxcox(fit, plotit = TRUE)
(bc.power <- bc$x[which.max(bc$y)])

# Studentized Residuals
studres(fit)
range(studres(fit))

barplot(height = studres(fit), names.arg = 1:101,
       main = "Studentized Residuals", xlab = "Index",
       ylab = "Studentized Resid", ylim=c(-4,4))
abline(h=3, col = 'Red', lwd = 3)
abline(h=-3, col = 'Red', lwd = 3)

cor.qt <- qt(cor.level, 95, lower.tail=F)
Rstudent > cor.qt
barplot(height = Rstudent, names.arg = 1:101,
       main = "R Student Residuals", xlab = "Index",
       ylab = "R Student Resid", ylim=c(-4,4))
abline(h=cor.qt , col = "Red", lwd=3)
abline(h=-cor.qt , col = "Red", lwd=3)

# Residuals vs. Fitted Values
par(mfrow=c(1,2))
hist(studres(fit), breaks=10, freq=F, col="cornflowerblue",
    cex.axis=1.5, cex.lab=1.5, cex.main=2, title = 'Histogram of studres')
qqPlot(fit)

residualPlot(fit, type="rstudent", quadratic=F, col = poss1$Species,
           pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5, main = 'Reduced Model Residuals'
           , ylim = c(-3, 3))
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)
```

```
# Species 1 Residuals vs. Fitted Values
residualPlot(fit, type="rstudent", quadratic=F, col = poss1$Species,
 pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5, main = 'Species 1 Residuals',
 xlim = c(43, 47), ylim = c(-3, 3))
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

# Species 2 Residuals vs. Fitted Values
residualPlot(fit, type="rstudent", quadratic=F, col = poss1$Species,
 pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5, main = 'Species 2 Residuals',
 xlim = c(51, 54), ylim = c(-3, 3))
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

################### Influential Analysis ##############################
# Rstudent Residuals
Rstudent <- rstudent(fit)
range(Rstudent)
cor.level <- 0.05/(5*101)

themod <- lm(earconch ~ Species + hdlngth + chest, data = poss1) # Final Model
summary(themod) # Summary of the Final Model

myInf <- influence.measures(themod) # Leverage/Influential Points Diagnostics
summary(myInf) # Summary of the Diagnostics

dfbetasPlots(themod,intercept=T) # DFBETA Plots

influenceIndexPlot(themod, vars=c("Cook", "hat")) # Cook's & Hat-value Plots
```