STAT 601 Final Project: Microtus Data Set

Joseph McDonald

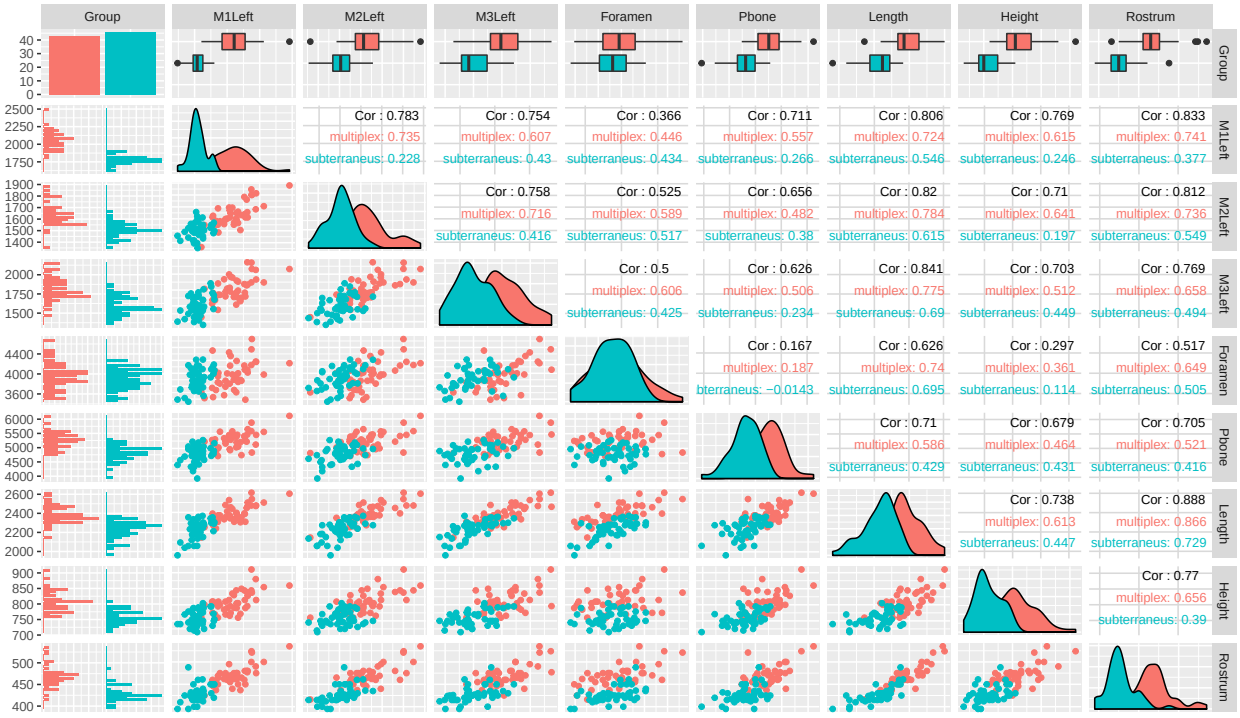
Introduction

The problem at hand is to classify the Microtus subterraneus and Microtus Multiplex based upon morphological data alone. We will use the **Microtus** dataset from the **Flury** package which includes 89 specimens whose chromosomes were analyzed to determine the correct species and 199 specimens with unknown species. We will attempt to fit and find the best generalized linear model using the data with given species with the goal of accurately predicting the species of the 199 unknown specimens.

Exploration

We begin by briefly exploring the data set. This data set contains eight predictor variables and the response variable, **Group**. There are no missing values and a brief look at the summary does not reveal any large concerns. After this examination, we split the data into two sets: a training set *m.train* which contains all specimens for which the species is known, a test set *m.test* containing the 199 specimen for which we do not know the species. We will now set aside our test set until the best model is chosen and use the training set for further analysis.

We examine a pairs plot of the training data for an overview of the relationship of the predictor variables to each other and the response variable. The first thing that jumps out is the scatter plots for the variable **M1Left**. We notice that there is a relatively high linear correlation between between **M1Left** and all of the other predictor variables accept for **Foramen**. We also note in both the box plot and scatter plots for **M1Left** that there seems to be a pretty clear distinction between the two classes of the response variable. Because of this we might expect **M1Left** to be a good predictor variable for our generlized linear model. Even though there are a lot of high correlations between predictor variables, none are high enough to warrant removing any predictor variables at this time.



Model Training and Selection

Next we will fit several GLM models and use the step function in R for variable selection using AIC as the discrimant criterion. We will try all three feature selection methods and compare the results to help determine our best model. The "backward", "forward", and "both" directions of feature selection return the following GLM formulas respectively:

```
## Group ~ M1Left + M3Left + Foramen + Length + Height
```

```
## Group ~ M1Left + Foramen
```

```
## Group ~ M1Left + Foramen
```

We will now fit two new GLM models using the predictor variables selected by the step function. An ANOVA test for the two models returns the p-value shown below which indicates that there is not a statistically significant difference between the simple and more complex models. Therefore, we chose the simple model, with only **M1Left** and **Foramen** as predictor variables. The selection of this model as the best model is also backed up by our earlier observations that **M1Left** appeared to be a strong predictor variable and was highly correlated with all of the other predictor variables except **Foramen**. While the correlations were not strong enough for us to eliminate any variables early on, intuitively, it makes sense that the best model would need only these two variables.

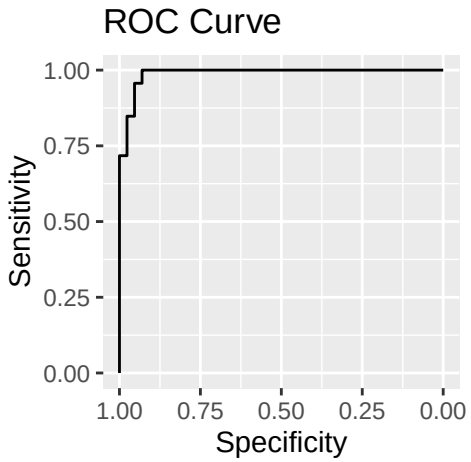
```
## [1] "p-value = 0.0959209421392917"
```

Model Evaluation

Lastly we move on to evaluate our model. Because our training set is very small we will use 10-fold cross validation to determine the error rate. We also examine the ROC curve and return the AUC.

```
## [1] "10-fold Cross Validation Error Rate = 0.0561797752808989"
```

```
## Area under the curve: 0.9889
```



The shape of the ROC curve and the AUC of .9889 indicate that our model is a very good fit of the training data. Any concerns of potential overfitting are mitigated by the low cross-validation error rate which is just under 6%.

Conclusions and Recommendations

Based upon the results of our model evaluation we conclude that our model ought to preform very well on new data. Through cross-validation we determined that the model performs well on "new" data with approximately 94% accuracy. Additionally, The ROC curve indicates that that this model can achieve a balance of both high sensitivity and specificity indicating that the error in the model should not be skewed toward either class. That is to say, that misclassifying either species as the other is roughly equally likely and in both cases this likelihood is relativley low. This model would be useful for determining the correct species of microtus for the specimens in situations in which high precision is not necessary. With this model we would expect that about 1 in 20 specimens would be missclassified. This could be very useful in some cases, such as identifying populations of microtus or summary investigations. On the other hand, for more precise scientific study, such as further study of the differences bewteen the two species, this model would not be accurate enough to provide reliable results.

Citations

Citation 1: https://stackoverflow.com/questions/8599685/how-to-change-correlation-text-size-in-ggpairs Changing font size for correlation text in ggpairs.
Citation 2: https://stackoverflow.com/questions/41577362/suppress-ggpairs-messages-when-generating-plot Suppressing plot progress for ggpairs.
Citation 3: https://rdrr.io/cran/pROC/man/ggroc.html assistance plotting roc in ggplot