Predicting Microtus Species Between Subterraneus and Multplex

Determining the best logistic regression model for determining 199 unknown species of microtus based on 8 different mouth, skull, and bone measurements. Based on a stepwise regression feature selection and different fit tests, it was determined that the best predictors for determining the type of species was the upper left molar (M1Left) width and the length of the incisive foramen (Foramen). However, for predictive purposes, the suggestions of the stepwise regression was used. Finally, it was determined that a logistic regression may not be the best model to determine species type, and different classification models, such as KNN and decision tree should be tested.

Author: John Herbert

Institution: Dakota State University

Abstract

The aim of this study is to determine the best logistic regression model to predict unknown Microtus species types. This will be done by using a stepwise regression to determine which variables have the best fit based on the Akaike Information Criterion (AIC). In addition, multicollinearity was tested using a Variance Inflation Factor (VIF) function and variables were removed and tested against the stepwise regression model. Also, a model was tested based on the p values of the cofficients for each variable in the model, and 2 were chosen with the lowest p values and compared against the other models. The models were chosen on AIC, Mean Squared Error (MSE), and the error rate from a 10 fold cross validation model.

External Libraries

Packages and tools used for this analysis:

- **Flury** package for the *microtus* dataset
- **dplyr** and **tidyr** package sued for data manipulation
- **knitr** package used for *kable* function used to format tables
- **htmltools** package used for formatting pdf document
- **ggplot2** package for graphing
- **gridExtra** package for to output plots side by side
- **boot** package for logistic regression function
- **moments** for skewness and kurtosis calculations
- **gggally** package for pair plot comparison graph
- **stats** package used for stepwise regression
- **boot** package used for cv.glm (cross validation calcuation)
- **car** package for multicollinearity testing

Methodology

Data

This study was conducted by Airoldi, J.P. and Flury, M. Salvioni in 1995. This study's goal was to determine a visual method of classifying a Microtus between two species types: Multiplex and Subterraneus. These species can be determined based on chromosome count, however the goal is to see if there is an easy way to determine the difference.

The data consists of 3 target variables: multiplex, subterraneus, and unknown. There are also 8 input variables to determine the classification:

- **Group**: factor with levels multiplex subterraneus unknown
- **M1Left**: Width of upper left molar 1 (0.001mm)
- **M2Left**: Width of upper left molar 2 (0.001mm)
- **M3Left**: Width of upper left molar 3 (0.001mm)
- **Foramen**: Length of incisive foramen (0.001mm)
- **Pbone**: Length of palatal bone (0.001mm)
- **Length**: Condylo incisive length or skull length (0.01mm)
- **Height**: Skull height above bullae (0.01mm)
- **Rostrum**: Skull width across rostrum (0.01mm)

In addition, there are a total of 199 records: 100 in the unknown group, 43 in multiplex, and 46 in suberraneus.

Data Manipulation and Exploration

Since the unknown Group subset is the variable that needs to be predicted, all rows containing the unknown Group classifications will be seperated into a test dataset, and the other classifications will be subset into a training dataset to form a model for prediction. This is being seperated and only the training dataset will be used for model fitting because I do not want the test data to be influenced by the model fitting in any way.

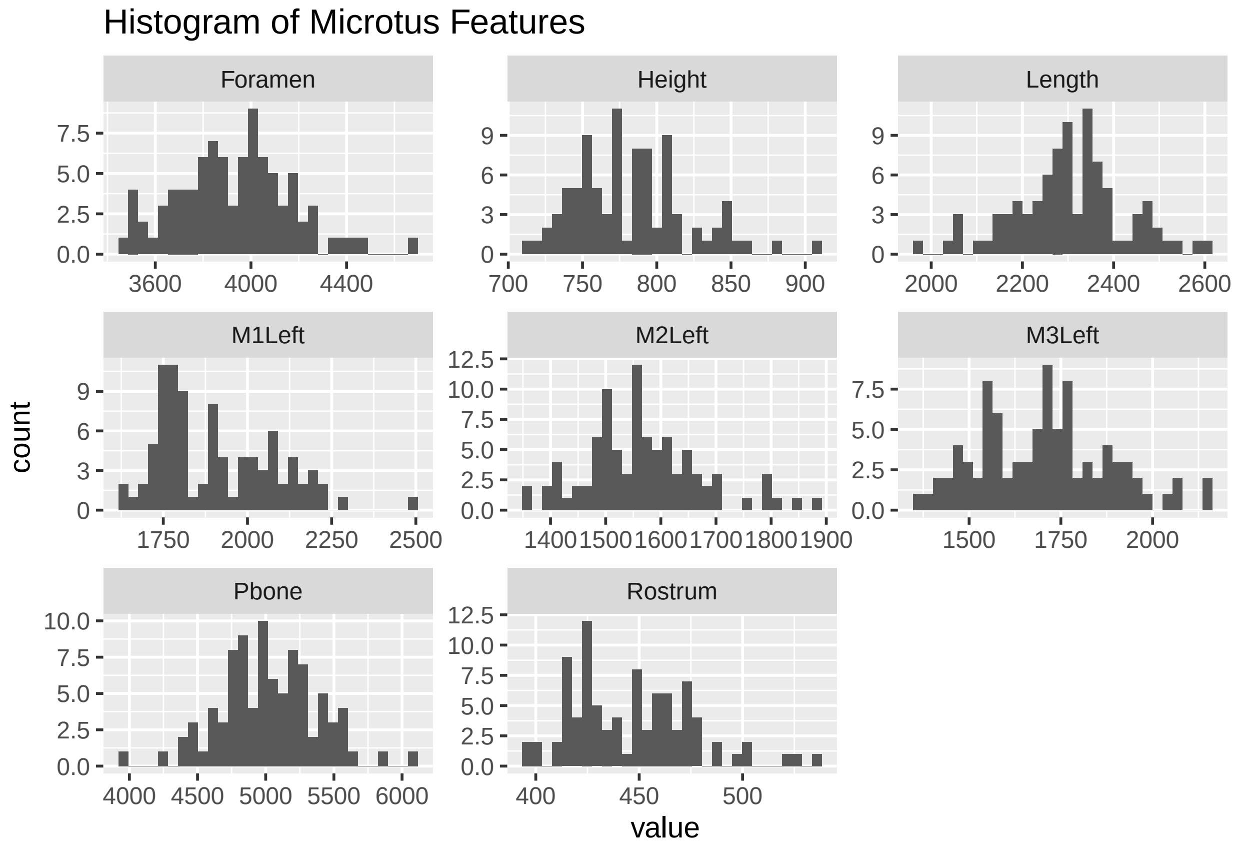
Below is a summary of descriptive statistics for each of the variables in the training dataset. We can determine that all values appear to be scaled appropriately (in mm), there are no null values, and appears to be no large outliers or unreasonable/false values.

Table 1: Summary of Training Subset

Group	M1Left	M2Left	M3Left	Foramen	Pbone	Length	Height	Rostrum
multiplex :43	Min. :1619	Min. :1355	Min. :1361	Min. :3451	Min. :3980	Min. :1965	Min. :715.0	Min. :395.0
subterraneus:46	1st Qu.:1770	1st Qu.:1504	1st Qu.:1561	1st Qu.:3764	1st Qu.:4773	1st Qu.:2237	1st Qu.:750.0	1st Qu.:425.0
unknown : 0	Median :1885	Median :1551	Median :1712	Median :3941	Median :5004	Median :2300	Median :776.0	Median :450.0
NA	Mean :1909	Mean :1568	Mean :1705	Mean :3932	Mean :5025	Mean :2304	Mean :782.9	Mean :447.2
NA	3rd Qu.:2052	3rd Qu.:1621	3rd Qu.:1815	3rd Qu.:4078	3rd Qu.:5254	3rd Qu.:2370	3rd Qu.:805.0	3rd Qu.:465.0
NA	Max. :2479	Max. :1880	Max. :2150	Max. :4662	Max. :6104	Max. :2600	Max. :910.0	Max. :535.0

In addition, a historam of each of the input variables in the dataset was graphed to determine normality of the data and determine if there are any outliers that need manipulation.. Based on the below graph, the data appears to be relatively normal, except the *Height* variable appears to be right skewed. There are a few outliers in a few of the variables, however when run on a stepwise regression, they do not appear to skew the results, so they are kept in.

This was done in **ggplot2** with the *gather* and *geom_histogram* functions. In addition, *facet_wrap* was used to format each histoplot into one visual.



There are a few variables that appear right skewed, therefore skewness was calculated to determine if any variables need log transformation to normalize. Based on the table below, each variables is below absolute 1, but some are above 0.5 which mean there is some skewness in some of the variables. However, log transformation did not affect the scoring, fit, or variable selection of the model, so it was not be transformed.

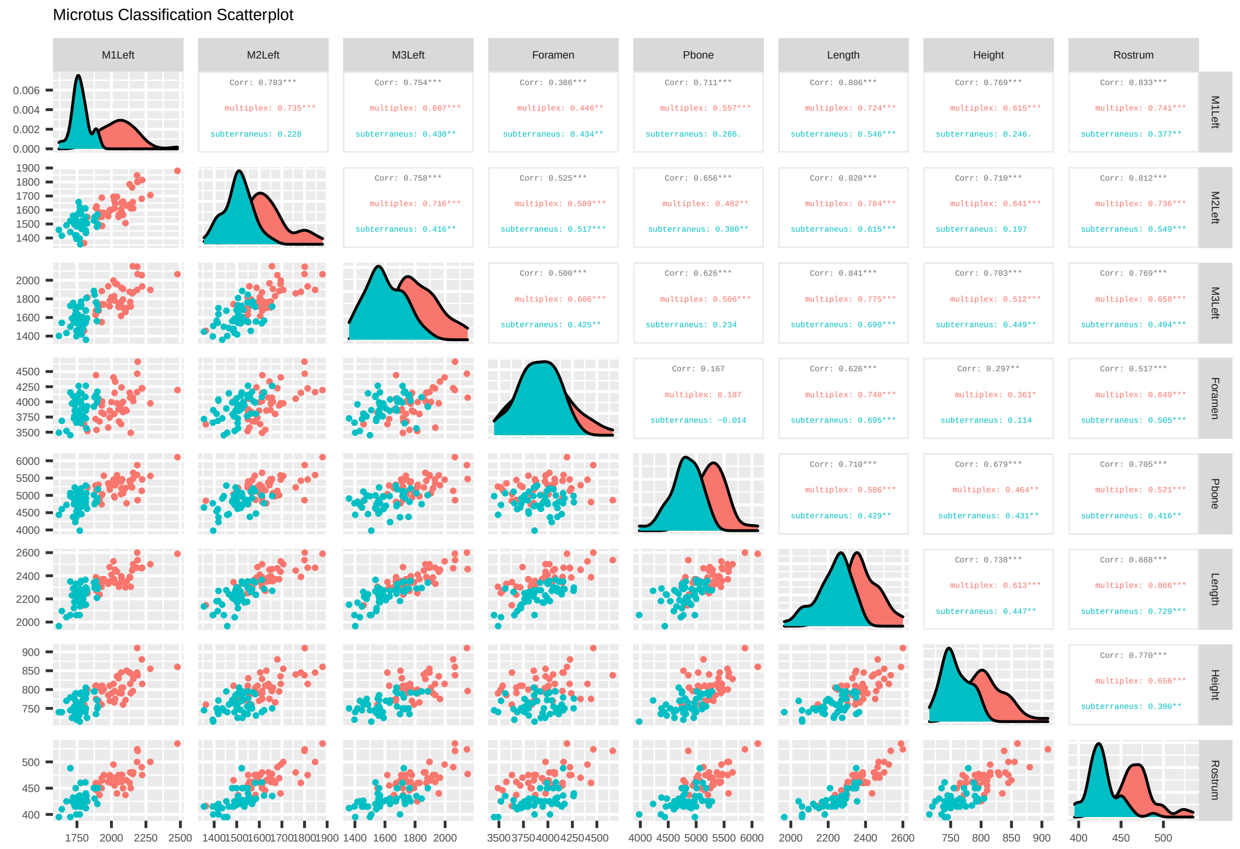
the *skewness* function in the **moments** package was us for the skewness test.

Table 2: Skewness of Microtus Features

Feature	Skewness
M1Left	0.6850374
M2Left	0.6364370
M3Left	0.3320895
Foramen	0.2769374
Pbone	0.0348765
Length	-0.1249235
Height	0.6941826
Rostrum	0.5825419

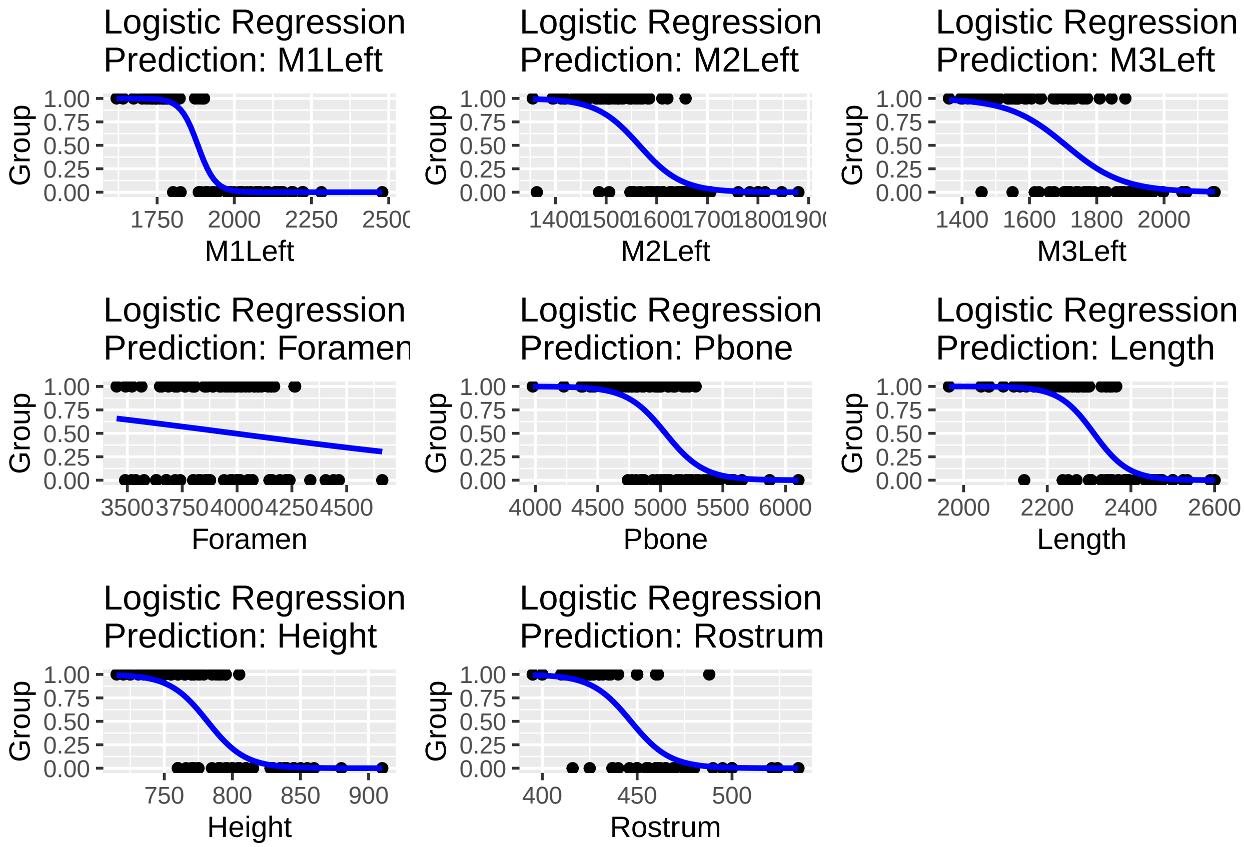
A pairs plot was used to determine the density of each group in relation to each variable. There does appear to be a linear relationship in the data, with a clear seperation of classes in the *M1Left*, *M2Left*, and *Rostrum* variables.There appears to be a strong seperation for th *M1Left* variable specifically.

This was made using the **ggpairs** function in the **GGally** package in conjunction with other **ggplot2** functions.



A scatterplot and logistic regression line were plotted for each of the variables vs. the target. Based on this, it appears that multiple inputs are fairly good at predicting which species the microtus are at high and low measurements, but there is a lot of cross over in the middle ranges. Specifically, *M1Left* has the least cross over in measurements, while the *Foramen* variables has the most. This is also shown by a steep sigmoid function line for *M1Left*, while the *Foramen* function appears to be linear and has a fair amount of cross over in measurments.

This graph was made using **ggplot2* in conjunction with *geom_poimt*, and *stat_smooth* functions for the scatterplot and linear regression lines.



Feature Selection

In order to determine which variables should be kept in the model, a stepwise regression (forwards and backwards) was used based on AIC as a scoring metric. First, the data was fit using a binomial logistic regression with Group as the target, and all the other variables as the inputs.

The model was made fitting the model with the *glm* function and then using the *step* function in the **stats** package and setting the direction to 'both' for the forwards and backwards stepwise. A seed was set in order to produce the same results.

The result of the stepwise regression were 5 remaining variables from the original 8 with an AIC of 27. *M1Left* is the only variable with a p value below the 0.05 signifiance level and *Foramen* is below the 0.10 signifiance.

Table 3: Model 1: Stepwise Feature Selection

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	187.8305847	101.9145325	1.843021	0.0653260
M1Left	-0.0583819	0.0267599	-2.181696	0.0291320
M3Left	0.0248694	0.0166559	1.493129	0.1354034
Foramen	0.0118977	0.0071640	1.660755	0.0967627
Length	-0.0414671	0.0295157	-1.404918	0.1600455
Height	-0.0929716	0.0711074	-1.307481	0.1910493

##

```
## AIC for Stepwise Regression is 27.70264
```

When fitting the variables based on the lowest AIC of different combinations of independent variables. The results are an AIC of 27.7 and a mean squared error of 3.18. These appear fairly high, however it should be compared with other models to determine if it is truely the best fitted logistic regression model.

These metrics were calculated using the *aic* call in the *glm* function and the residuals call on the model for the MSE.

Table 4: Model 1 Measurements

Measure	Metric
AIC	27.702644
MSE	3.186422

When examining the variables, especially the molar length, there appears to be some multicolinearity in the variables. While this normally does not affect the accuracy of a model's predictions, if the goal of the study is to determine a simple way to determine the difference in species, a VIF test should be run, and the variables with a score above 10 should be removed and compared to the stepwise regression.

Starting with the step wise regression variables as a base, we can see from the test below that *M3Left* and *Length* have high multicollinearity and will be removed for Model 2 to see if a simplier model can improve or produce the same scores as the stepwise. If it does, this would be the model of choice for our predictions.

THe multicollinearity test was conducted using the *vif* function in the **car** package on the glm model with the stepwise regression chosen variables.

Table 5: Multicollinearity Test on Model 1

	x
M1Left	3.929237
M3Left	15.933123
Foramen	7.815833
Length	10.744455
Height	4.319241

Model 2: Variance Inflation Adjusted

The logistic function for Model 2 will include *M1Left*, *Foramen*, and *Height* as the input variables.

BaSed on the summaries below, *M1Left* has a signifiance below 0.05, while the other 2 variables are not significant according to the p tests. In addition, AIC is 29 and MSE is 4, both worse than the stepwise regression model.

Table 6: Coefficients of Model 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	80.5787640	31.3136704	2.5732775	0.0100740
M1Left	-0.0421641	0.0141114	-2.9879394	0.0028087
Foramen	0.0049898	0.0033173	1.5041839	0.1325340
Height	-0.0268263	0.0283115	-0.9475399	0.3433637

Table 7: Model 2 Measurements

Measure	Score
AIC	29.099816
MSE	4.077587

However, if we run a VIF function, we can see that we have removed the multicolinearity from the variables.

Table 8: Multicollinearilty Test on Model 2

	x
M1Left	1.505918
Foramen	1.739465
Height	1.198528

Model 3: Trimmed

As mentioned above, a simplier model would be better, therefore we will only include the 2 variables with the most significant p values: *M1Left* and *Foramen*.

According to the logistic regession summary below, Model 3 has a AIC of 28 and a MSE of 7. While AIC improved for this model, MSE got worse. This makes sense since AIC measures how well the model explains the greatest amount of variation using the fewest possible independent variables. Since we reduced the number of variables *Foramen* now has signifiance below the 0.05 p value, and since the variable coefficients are fairly low, the AIC improved. However, there may have been some useful information in the *Height* variable (most likely at the upper and lower values) that improved MSE in Model 2.

Table 9: Coefficients of Model 3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	62.8044523	20.6610801	3.039747	0.0023678
M1Left	-0.0472459	0.0140909	-3.352927	0.0007996
Foramen	0.0066369	0.0031917	2.079406	0.0375801

Table 10: Model 3 Measurements

Measure	Model_3
AIC	28.04904
MSE	7.47215

Model Selection

In order to determine which model to use in predicting the test data (unknown species group), I will run a 10 fold cross validation on all 3 models. Since there are no huge improvements or differences between the models, the MSE scores can vary significantly depending on the seed I use. Therefore, I created a loop of 1,000 random tests and took an average of each score for each model to determine which one actually performs the best. Below are the average MSE scores for each model based on the cross validation test.

Based on the results below, Model 3 has the loweest MSE of all 3 models, however, Model 1 and model 3 appear to be very close, thereore I will run a chi squared test to determine if there is statistical signifiance between the 2 models.

This table was made by creating a for loop of 1,000 iterations. Within the for loop, I set the seed to randomly select a number between 1 and 1,000,000 each iteration. The results of each iterations were put into a data frame, and the mean of the error rates for each model were recorded and shown below.

Table 11: 10 Fold CV Error of 3 Models

Model	MSE
Model 1	0.0670337
Model 2	0.0833371
Model 3	0.0610449

From the chi squared test below, there is no statistical signifiance at the 0.05 level, but there is signifiance at the 0.10. This test is between Model 1 and 3, since the scores were so similar.

The test was created using the *anova* function and setting teset equal to 'Chisq'.

Table 12: Chi Squared Test Model 1 vs. Model 3

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
83	15.70264	NA	NA	NA
86	22.04904	-3	-6.346401	0.0959209

Summary

Based on the scores below, Model 1 has the best fit. It has the lowest model AIC and MSE of the 3, and while it does not have the lowest error rate from the cross validation, the difference between that and the Model 3 is not not significant at the 0.05 level. Therefore, Model 1 will be used for the predictions of the unknown species group.

Table 13: Model Summary Results

Models	AIC	MSE	CV
Model 1	27.70264	3.186422	0.0670337
Model 2	29.09982	4.077587	0.0833371
Model 3	28.04904	7.472150	0.0610449

Predictions

The predictions of the unknown species groups is attached in the *microtus_pred.csv* file. This was done by converting the predictions to a binomial class (0,1). If the prediction was greater than or equal to 0.5, it was assigned to the *subterraneus* class, otherwise it was assigned to the *multiplex* class.

Below is a head of the data export to confirm everything was coded correctly.

Table 14: Head of Teset Prediction Export

	Pred	M1Left	M2Left	M3Left	Foramen	Pbone	Length	Height	Rostrum
90	multiplex	1841	1562	1585	3750	5024	2232	821	430
91	subterraneus	1770	1459	1542	3856	4542	2140	755	405
92	subterraneus	1785	1573	1616	4165	3928	2295	767	425
93	multiplex	2095	1660	1870	3937	5218	2355	842	490
94	multiplex	1976	1666	1704	4058	5235	2335	814	481
95	multiplex	1980	1643	1950	3569	6020	2355	815	460

Conclusion

In conclusion, while the model with the features chosen from the stepwise regression model were used for predictions, a binomial logistic regression may not be the best model for this problem. Based on the fact that the error rates were still fairly, high, improvements to the model were not that significant, and the coefficients were small. Further analysis would be need and different methods, such as K-Nearest Neighbor and decision tree should be teseted to see if results improve.

Bibliography

microtus: Microtus classification (more vole data)'

How do I generate a histogram for each column of my table?

R: plot histogram of all columns in a data.frame

Scatterplot matrices (pair plots) with cdata and ggplot2

How to change correlation text size in ggpairs()

Change Font Size of ggplot2 Plot in R (5 Examples) | Axis Text, Main Title & Legend

Stepwise Regression Essentials in R

Binary classifier evaluation metrics: error rate, KS statistic, AUROC, lift, gains table

An introduction to the Akaike information criterion

Model Selection Approaches

Generalized Linear Models in R, Part 1: Calculating Predicted Probability in Binary Logistic Regression

Logistic regression