Stat601 Final Problem 2: Microtus

James Young

December 16, 2019

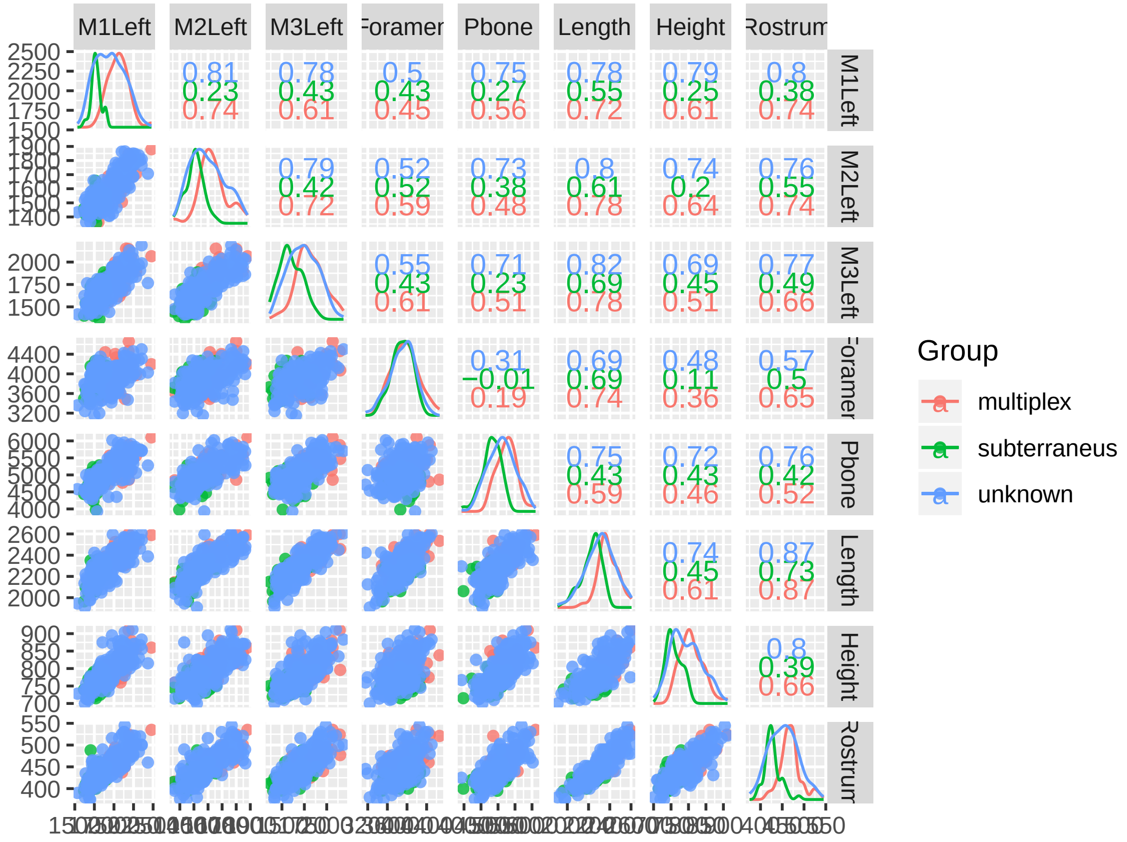
The Problem

Given a dataset, which will be further characterized and analyzed below, I have been asked to create a model that can distinguish between two species of microtus. The two species are multiplex and subterraneus. I will create a general linear model for a binary response, a logistic regression, selecting for variables that create a model with the best predictive power at a tradeoff for AIC.

The Data

The data contains the response factor *Group* and 8 predictor variables which are integers on a continuom including *M1Left*, *M2Left*, *M3Left*, *Foramen*, *Pbone*, *Length*, *Height*, and *Rostrum*. We can see the *Group* variable has 43 known multiplex species, 46 known subterraneus, and 199 unknown species. The input variables pertain to measurements of the skull, which may help discern the species the skull belonged to. Looking further below, we can see that some variables, such as *M1Left* and *Rostrum* have some visible separation between multiplex and subterraneus species and other variables such as *Foramen* that seem visually identical between the species. The distribution of the unknown species measurements hit that it is probably a fairly balanced mix between subterraneus and multiplex.

Group	M1Left	M2Left	M3Left	Foramen	Pbone	Length	H
multiplex : 43	Min. :1534	Min. :1355	Min. :1361	Min. :3155	Min. :3928	Min. :1908	Min
subterraneus: 46	1st Qu.:1783	1st Qu.:1503	1st Qu.:1595	1st Qu.:3751	1st Qu.:4815	1st Qu.:2227	1st Q
unknown :199	Median :1923	Median :1570	Median :1724	Median :3932	Median :5079	Median :2312	Medi
NA	Mean :1935	Mean :1589	Mean :1727	Mean :3913	Mean :5082	Mean :2309	Mea
NA	3rd Qu.:2074	3rd Qu.:1660	3rd Qu.:1856	3rd Qu.:4080	3rd Qu.:5328	3rd Qu.:2388	3rd Q
NA	Max. :2479	Max. :1880	Max. :2187	Max. :4662	Max. :6104	Max. :2605	Max



The Model

To build the model, first I must remove all of the unknown samples. We are left with samples previously determined to be subterraneus or multiplex by genomic testing. Then I ran a logistic regression with all 8 variables as input variables and removed variables to get to the lowest AIC using the step function. This formula was "glm(Group~M1Left+M3Left+Foramen+Length+Height, binomial)". I ran this through 5-fold cross-validation to determine the error rate.

```
##
## Call:
## glm(formula = Group ~ M1Left + M3Left + Foramen + Length + Height,
##     family = binomial, data = microtus2)
##
## Deviance Residuals:
##      Min        1Q    Median       3Q      Max
## -2.26335  -0.00138   0.00013   0.05223  1.14144
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 187.830585 101.914533   1.843   0.0653 .
## M1Left       -0.058382   0.026760  -2.182   0.0291 *
## M3Left        0.024869   0.016656   1.493   0.1354
## Foramen       0.011898   0.007164   1.661   0.0968 .
## Length       -0.041467   0.029516  -1.405   0.1600
## Height       -0.092972   0.071107  -1.307   0.1910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.279  on 88  degrees of freedom
## Residual deviance:  15.703  on 83  degrees of freedom
## AIC: 27.703
##
## Number of Fisher Scoring iterations: 10
```

Taking Significant Variables

I will take M1Left as it is the only statistically significant variable at alpha = 0.05 and I will also take Foramen as it is close 0.05. I will use these two variables to make another logistic regression model, run it through 5-fold cross-validation, and compare the results for model selection. This new model will be called Step Significant and be of the form "glm(Group ~M1Left+Foramen, binomial)"

Model Selection

The Step model and the Step Significant model both had similar AIC, with the Step model being slightly lower. However, the step significant model had a lower error rate so I am choosing that for my predictive model. The seeds are set in the Rmd file so it is reproducible.

Model	AIC	Error1	Error2
Step Model	27.7026442847978	0.101123595505618	0.0964524681227118
Step Model Significant	28.0490448973609	0.0561797752808989	0.0493624542355763

Predictions

When binarized, microplex becomes "0" and subterraneus becomes "1". My predictive model, when split at a probability threshold of 0.5, predicts there are 121 multiplex and 78 subterraneus samples. Based on the error rate, this should be a balanced prediction and we can expect to be wrong about 7% of the time. The predictions have also been submitted as a CSV file.

```
##   0   1
## 121  78
```

Conclusion

The model made here, using *M1Left* and *Foramen* as input variables to predict wheter a microtus was multiplex or subterraneus resulted in predictive ability with an error rate of approximately 0.07 or 7% using 5-fold cross validation. This suggests the model will generalize fairly well as k-fold cross validation is generally considered a robust predictor of the generalized ability of a model.