# Analysis of Microtus data

Yamuna Dhungana

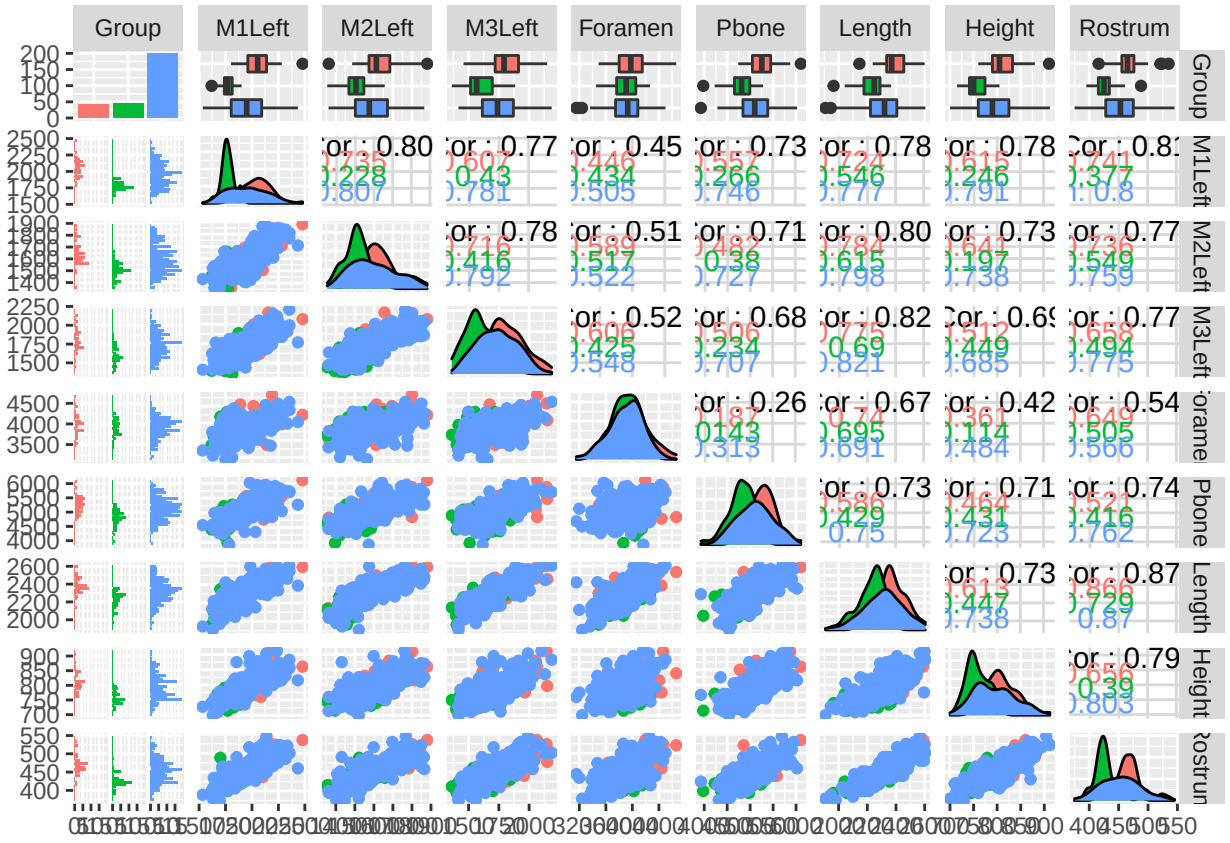
The data was taken from "Discrimination between two species of Microtus using both classified and unclassified observations journal of Theoretical Biology" by Airoldi, J.-P., B. Flury, M. Salvioni (1996). Our dataset, Microtus used from the library "Flury". The data consist of 288 no of samples of a vole, collected mostly in Europe (the Alps and Jura mountains) and in Toscana. The dataset was initiated by a data collection consisting of eight morphometric variables measured by one of the authors (Salvioni) using a Nikon measurescore(accuracy1/1000mm). 89 of the total data set is classified by its species with a detailed explanation of the chromosome data. Whereas the remaining 199 sets of data are yet to identify its species. Hence, the main objective of our finalproject is to predict the species of the unknown group. The data set consist of 288 observations with a factor indicating the species and observations on a further eight variables: Group: Species of the data with multiple, subterraneous, and unknown.M1Left: Width of the upper left molar 1 (0.001mm),M2Left: Width of the upper left molar 2 (0.001mm),M3Left: Width of the 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), and Rostrum: Skull width across rostrum (0.01mm)

In order to work on our problem, we must load our dataset Microtus and some of the libraries. Here I have loaded data and libraries. To begin with, I have visualized our data via pairwise plot to view the relation between the variables of the dataset.

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




The pairwise plot shows the relationship between the variables. The upper right of the graph shows the correlation values between the variables. The lower left of the pairwise plot shows the scatter plot of the data. Likewise, the bar and the boxplot also can be seen in the plot. The graph at the diagonal is the density plot of the data. Color red representing the multiple, green as subterraneous, and blue as the unknown species. The coefficient for the variables in the upper-right of the plot shows most of the variables were either moderately or highly correlated with each other.

The data is then divided into known and the unknown data for the further analysis. Then, we fit the logistic model to our data.

## Logistic model

```
##
## Call:
## glm(formula = Group ~ M1Left + Height + Foramen, family = binomial,
##     data = Known_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.20430  -0.01969   0.01008   0.10191   1.24456
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 80.578764  31.313670   2.573  0.01007 *
## M1Left      -0.042164   0.014111  -2.988  0.00281 **
## Height      -0.026826   0.028312  -0.948  0.34336
## Foramen      0.004990   0.003317   1.504  0.13253
```

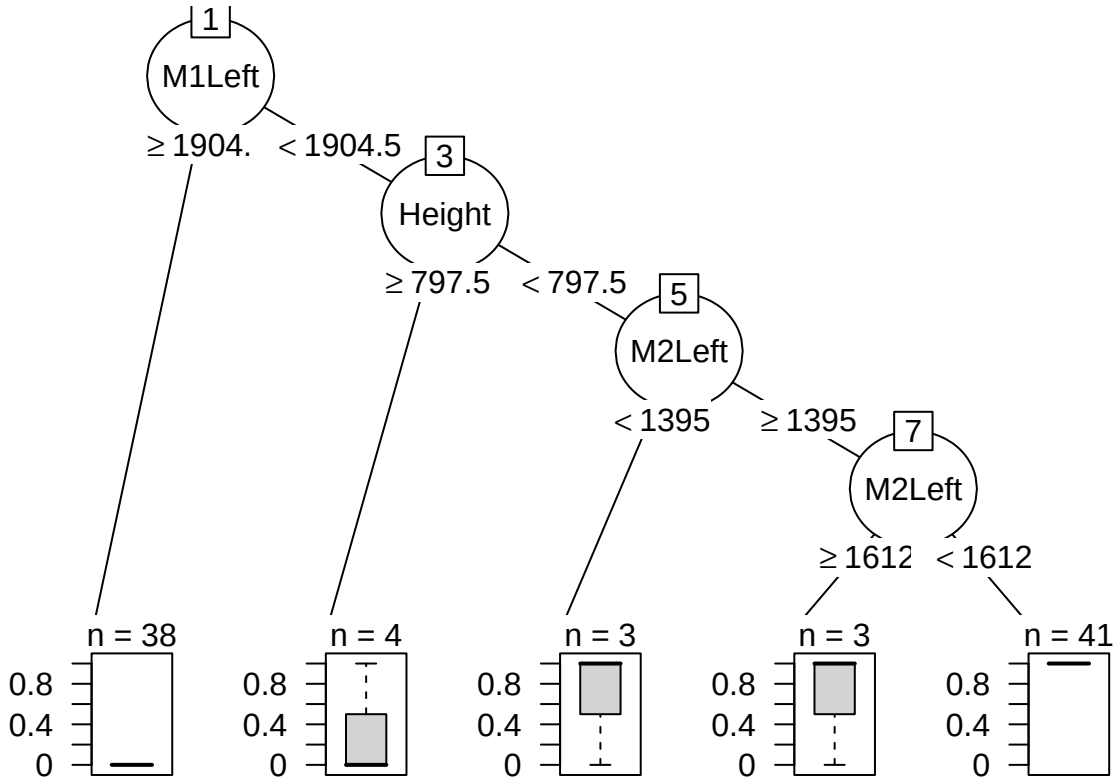
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.28  on 88  degrees of freedom
## Residual deviance:  21.10  on 85  degrees of freedom
## AIC: 29.1
##
## Number of Fisher Scoring iterations: 8

## [1] 29.09982

## [1] 0.03622752

## Cross validation with 10 fold:

## [1] 0.05180155
```

The logistic model (GLM) with the formula `Group~M1Left+Height+Foramen`, where the group is the response and the other M1Left, Height, and foramen as the predictor variables. The deviance residuals from the summary show that the data are negatively skewed. The variable L1left looks statistically significant at 0.05. The null deviance is 123.28 that shows how well our response variable has predicted the outfitted model, including only the intercept (grand mean) whereas, residual deviance with 21.10 inclusion of independent variables. Deviance is a measure of the goodness of fit of a model. The lowest numbers always indicate a good fit. The AIC of my first model, the logistic model, is 29.1. The Mean square error of my fitted model is 0.0362. Whereas cross-validation with 10 fold is 0.0475, and it is slightly more than the mean square error.

## Decision tree

Now, I also construct a regression tree selecting all the variables.

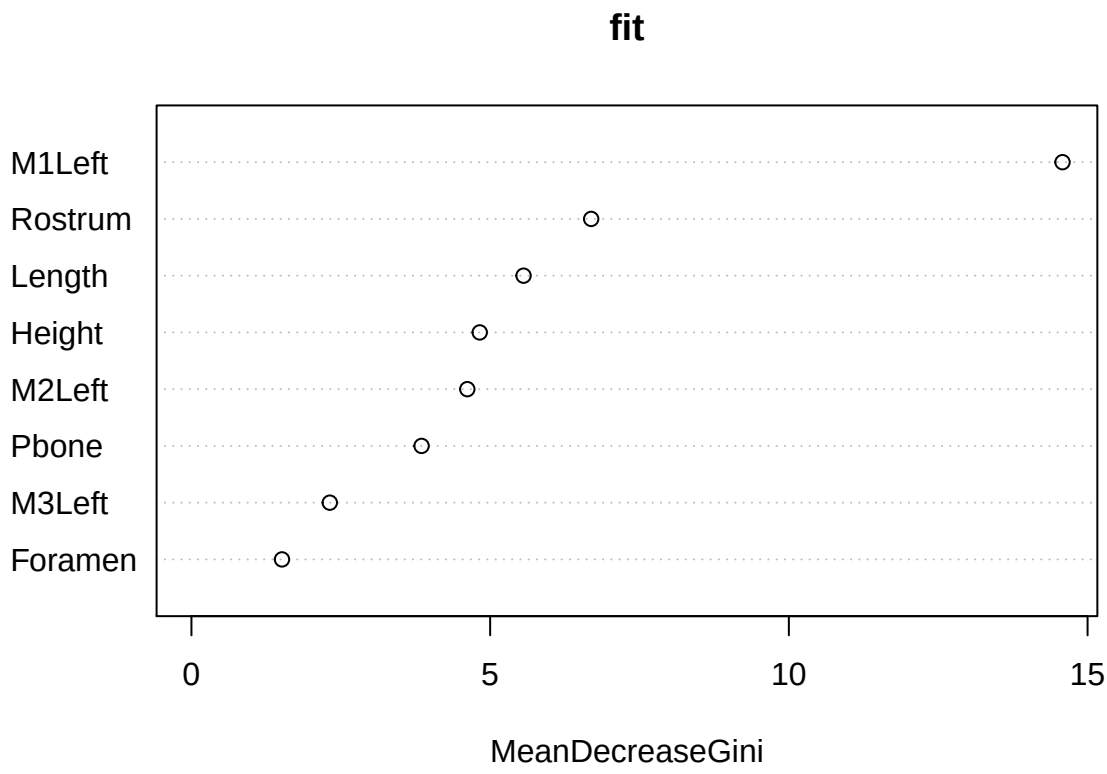
```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: libcoin

##
## Attaching package: 'partykit'

## The following objects are masked from 'package:party':
##
##     cforest, ctree, ctree_control, edge_simple, mob, mob_control,
##     node_barplot, node_bivplot, node_boxplot, node_inner, node_surv,
##     node_terminal, varimp
```




Now, I also construct a regression tree selecting all the variables to see which variables were chosen. In the regression tree, the main root of the tree is M1Left with the height and M2Left. These variables are the same as the model I chose above. These results seemed to follow the summary results from the model selected below with step regression, which is why the step model seems better than my GLM above.

Here, M2Left has replicated therefore, I plot the Mean DecreaseinGini plot. This plot shows the average (mean) of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. A higher Mean Decrease in Gini indicates higher variable importance. I would replace the duplicates with the other variable by viewing the importance of the variables from the mean DecreaseinGini plot.

```
## Loading required package: randomForest

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```




The mean Decrease in Gini shows that the Foramen have the least importance whereas, M1Left and the Rostrum variables have the highest importance.

### Stepwise selection

In order to see which model performed better with the data, I used the stepwise selection method.

```
## Start:  AIC=32.96
## Group ~ M1Left + M2Left + M3Left + Foramen + Pbone + Length +
##     Height + Rostrum
##
##           Df Deviance    AIC
## - M2Left   1   14.965 30.965
## - Pbone    1   15.288 31.288
## - Rostrum  1   15.627 31.627
## <none>         14.962 32.962
## - Length   1   17.330 33.330
## - Height   1   18.744 34.744
## - Foramen  1   19.434 35.434
## - M3Left   1   20.654 36.654
## - M1Left   1   40.753 56.753
##
## Step:  AIC=30.97
## Group ~ M1Left + M3Left + Foramen + Pbone + Length + Height +
##     Rostrum
##
##           Df Deviance    AIC
```

```
## - Pbone    1   15.306 29.306
## - Rostrum  1   15.627 29.627
## <none>         14.965 30.965
## - Length   1   18.268 32.268
## - Height   1   18.945 32.945
## + M2Left   1   14.962 32.962
## - Foramen  1   19.965 33.965
## - M3Left   1   20.763 34.763
## - M1Left   1   42.436 56.436
##
## Step:  AIC=29.31
## Group ~ M1Left + M3Left + Foramen + Length + Height + Rostrum
##
##           Df Deviance    AIC
## - Rostrum  1   15.703 27.703
## <none>         15.306 29.306
## - Length   1   18.625 30.625
## - Height   1   18.951 30.951
## + Pbone    1   14.965 30.965
## + M2Left   1   15.288 31.288
## - M3Left   1   20.855 32.855
## - Foramen  1   21.418 33.418
## - M1Left   1   42.970 54.970
##
## Step:  AIC=27.7
## Group ~ M1Left + M3Left + Foramen + Length + Height
##
##           Df Deviance    AIC
## <none>         15.703 27.703
## - Length   1   18.960 28.960
## - Height   1   19.019 29.019
## + Rostrum  1   15.306 29.306
## + Pbone    1   15.627 29.627
## + M2Left   1   15.694 29.694
## - M3Left   1   21.039 31.039
## - Foramen  1   21.463 31.463
## - M1Left   1   46.843 56.843

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

## [1] 27.70264

##
## Call:
## glm(formula = Group ~ M1Left + M3Left + Foramen + Length + Height,
##     family = binomial, data = Known_data)
##
## 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

## [1] 0.02740134

## Cross validation with 10 fold:

## [1] 0.05186471
```

The third model that fitted is stepwise selection in the hope of obtaining a better model than the previous model. I hoped to improve the selection of variable in my model.First, I used all the variables in both directions. AIC of the stepwise selection varied from 30 to 27.7. The lowest of the AIC that we got is 27.7 with the group as the response variable and M1left + M3Left + Foramen+ Length+ Height and the predictor variables. The previous models suggest that the height and the foramen were not statistically significant, whereas this model suggests that using these variables gives us the lowest AIC. Therefore, this model indicates that the omission of Foramen and Length is not a good idea. With all the variables as the predictor variable, The AIC of the model is the highest whereas, eliminating two variables M2Left and Rostrum, decreased AIC to the lowest was interesting to view how the elimination can change the performance of the model. The MSE of the model is 0.027 is slightly smaller than the Logistic model. The CV of the model is 0.060 is slightly more than the above model.

Hence, eventually, the best model that I chose from the step selection will use this model for predicting the rest of the unknown data (species). The AIC for this model was 27.70 is smaller (thus better) than our previous models. Also, the MSE for this model was only 0.027 is smaller (thus better) than our previous models. Also, the cross-validation error was 060 is a bit higher than our MSE, but still better than the CV of our model. Anyway, I will be selecting this model with AIC 0.27 from obtained step regression for the prediction purpose.

The unknown data is predicted and the 5 of them is also printed below.

```
##
## Call:
## glm(formula = Final_formula, family = binomial, data = Known_data)
##
## 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 .
## Height       -0.092972   0.071107  -1.307   0.1910
```

```
## Length       -0.041467   0.029516  -1.405   0.1600
## ---
## 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
```

```
## This is the first 5 rows of the predictions for unclassified observations
```

```
##            Group M1Left M2Left M3Left Foramen Pbone Length Height Rostrum
## 90     multiplex   1841   1562   1585    3750  5024   2232    821     430
## 91 subterraneous   1770   1459   1542    3856  4542   2140    755     405
## 92 subterraneous   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
```

Now, I want to count the total no of predicted values. We found out that 129 of the total unknown data is classified as the multiplex species and 70 of the total unknown data is classified as the subterraneous species.

```
## Count of predicted class:
```

```
##
##     multiplex subterraneous
##           129            70
```