Project 1

Gavin Gunawardena

Background

Airoldi_Flury_Salvioni_JTheorBiol_1995: Discrimination Between Two Species of Microtus using both Classified and Unclassified Observations.

> Microtus subterraneus and M. multiplex are now considered to be two distinct species (Niethammer, 1982; Krapp, 1982), contrary to the older view of Ellerman & Morrison-Scott (1951). The two species differ in the number of chromosomes: 2n=52 or 54 for M. subterraneus, and 2n=46 or 48 for M. multiplex. Hybrids from the laboratory have reduced fertility (Meylan, 1972), and hybrids from the field, whose karyotypes would be clearly recognizable, have never been found (Krapp, 1982).
>
> The geographic ranges of distribution of M. subterraneus and M. multiplex overlap to some extent in the Alps of southern Switzerland and northern Italy (Niethammer, 1982; Krapp, 1982). M. subterraneus is smaller than M. multiplex in most measurements, and occurs at elevations from 1000 m to over 2000 m, except in the western part of its range (for example, Belgium and Brittany), where it is found in lower elevations. M. multiplex is found at similar elevations, but also at altitudes from 200–300 m south of the Alps (Ticino, Toscana).
>
> The two chromosomal types of M. subterraneus can be crossed in the laboratory (Meylan, 1970, 1972), but no hybrids have so far been found in the field. In M. multiplex, the two chromosomal types show a distinct distribution range, but they are morphologically indistinguishable, and a hybrid has been found in the field (Storch & Winking, 1977).
>
> No reliable criteria based on cranial morphology have been found to distinguish the two species. Saint Girons (1971) pointed out a difference in the sutures of the posterior parts of the premaxillary and nasal bones compared to the frontal one, but this criterion does not work well in many cases. For both paleontological and biogeographical research it would be useful to have a good rule for discriminating between the two species, because much of the data available are in form of skull remains, either fossilized or from owl pellets.
>
> The present study was initiated by a data collection consisting of eight morphometric variables measured by one of the authors (Salvioni) using a Nikon measure-scope (accuracy 1/1000 mm) and dial calipers (accuracy 1/100 mm). The sample consists of 288 specimens collected mostly in Central Europe (Alps and Jura mountains) and in Toscana. One peculiar aspect of this data set is that

the chromosomes of 89 specimens were analyzed to identify the species. Only the morphometric characteristics are available for the remaining 199 specimens…"

Project

Background & Objective

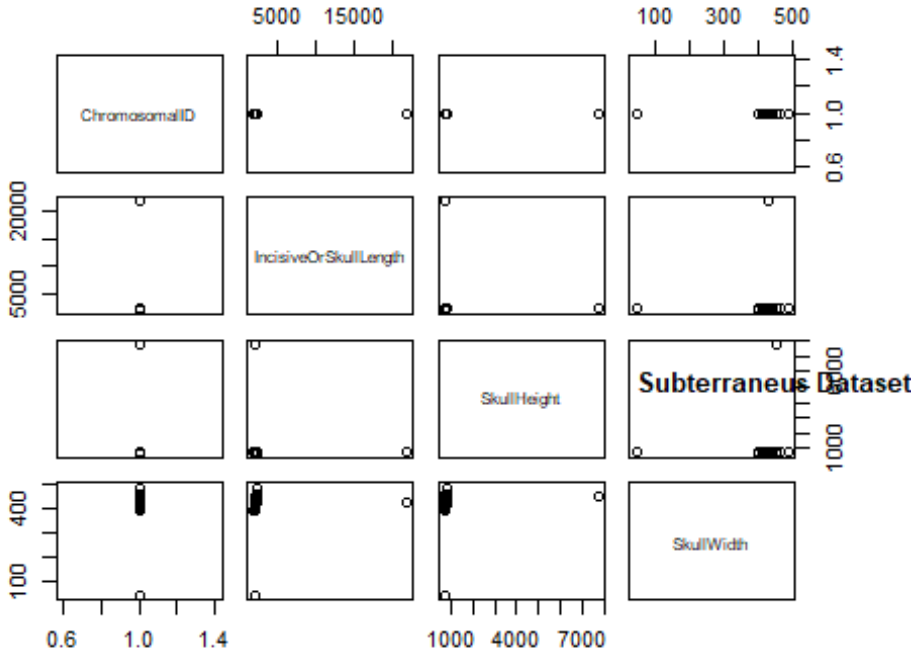
Since 1982, voles have been differentiated into two species, Microtus subterraneus and Microtus multiplex, mainly based on the number of chromosomes they each have. Hybrids of these two species are not common since they would have reduced fertility as discovered via lab testing. Other than their different habitats, the two species of voles are very difficult to reliably distinguish between. A study was conducted based on the collection of samples of fossilized vole skull remains in order to find a way to distinguish the two species of voles.

The objective of this project is to classify 199 unclassified vole skull samples via testing logistic regression models with cross-validation in order to find a model that will identify between the two species with a high percentage of accuracy. I have access to a dataset of 89 samples of voles that are close to evenly split between the two species, Microtus subterraneus and Microtus multiplex, classified via an analysis of their chromosomes. I also have access to 199 samples of voles that have not yet been classified. The variables included in the dataset that will be used for this classification include the condyle incisive length or skull length, the skull height above bullae, and the skull width across rostrum.

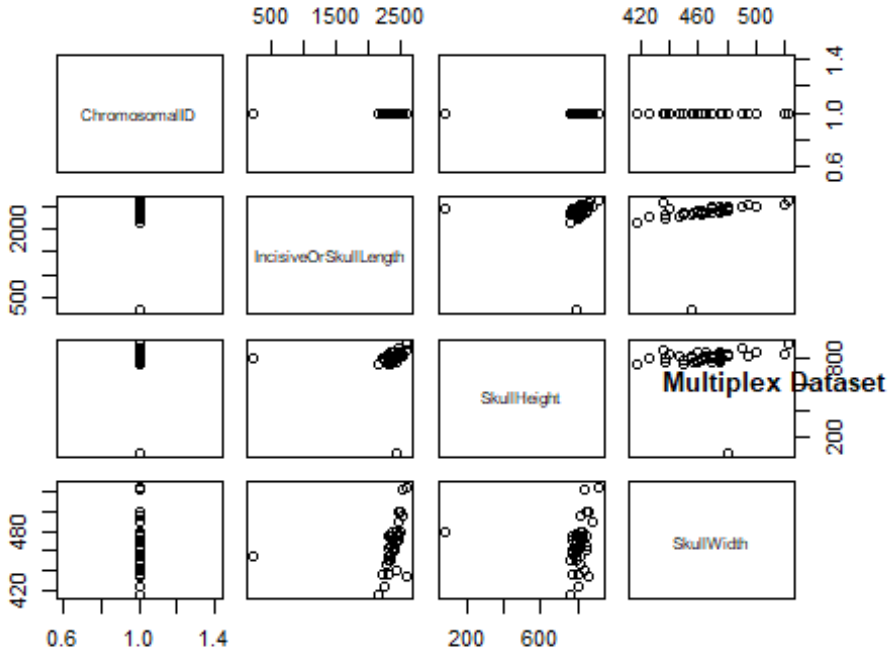
Original Dataset Analysis

After importing the data, I ended up with 3 different datasets, subterraneus, multiplex, and unknown, each with 3 independent variables and a target variable which was the classification ID. Via analysis, I found that most of the data was centered closely around the means of their respective classification ID and variable, except for a few outliers which were extremely large or small. After investigating these outliers, I concluded that the outliers are all at least 5 times larger or smaller than the mean of the samples, and the additional attributes of each individual sample that has an outlier, are not proportionally larger or smaller based on the outlier, but are instead close to the mean of the data. This has led me to believe that the outliers are data entry or measurement errors, worthy of being removed from the dataset, so I removed all samples(rows) that had a variable that was 5 times greater or 5 times less than the mean of the dataset. This ended up being 3 samples. In later tests of the model, I checked prediction accuracy through cross validation and found that removing the outliers improved accuracy by around 4 percent. I decided not to remove the outliers from the unclassified data, since the main purpose of removing the outliers is to improve the logistical model which I'll be using to classify the unclassified data. Here are some par plots, data summaries, and boxplots that show the distribution and outliers of the data:

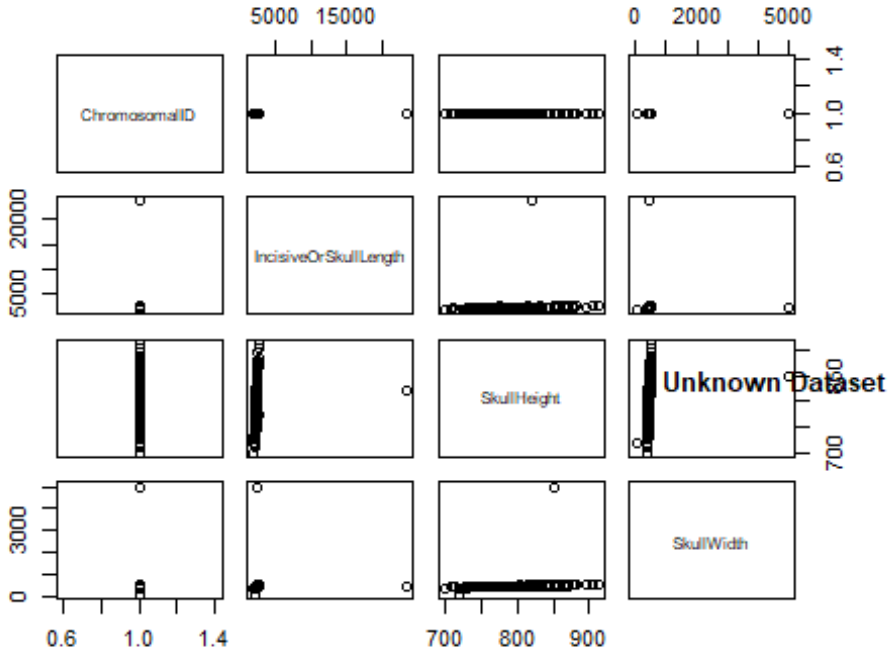
Data summaries and plots showing the discovery of major outliers



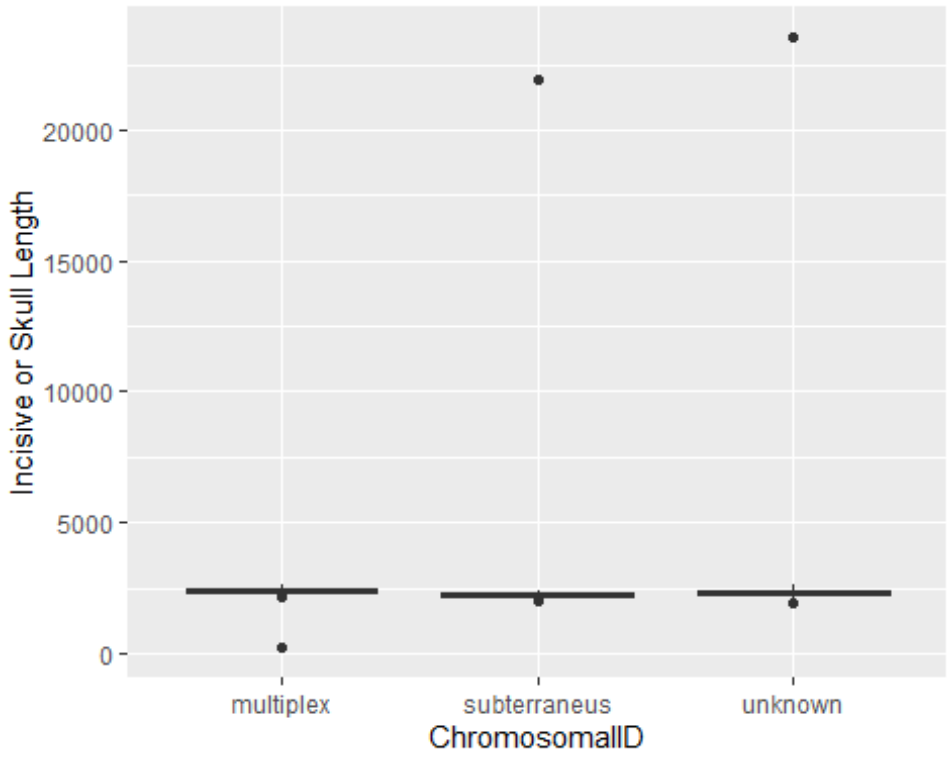

```
## [1] "Subterraneus Data Summary"

##  ChromosomalID      IncisiveOrSkullLength  SkullHeight       SkullWidth
##  Length:46          Min.   : 1965         Min.   : 715.0   Min.   : 42.0
##  Class :character   1st Qu.: 2176         1st Qu.: 741.2   1st Qu.:415.0
##  Mode  :character   Median : 2255         Median : 750.0   Median :425.0
##                     Mean   : 2655         Mean   : 909.2   Mean   :419.0
##                     3rd Qu.: 2290         3rd Qu.: 775.8   3rd Qu.:434.2
##                     Max.   :21899         Max.   :7722.0   Max.   :488.0
```

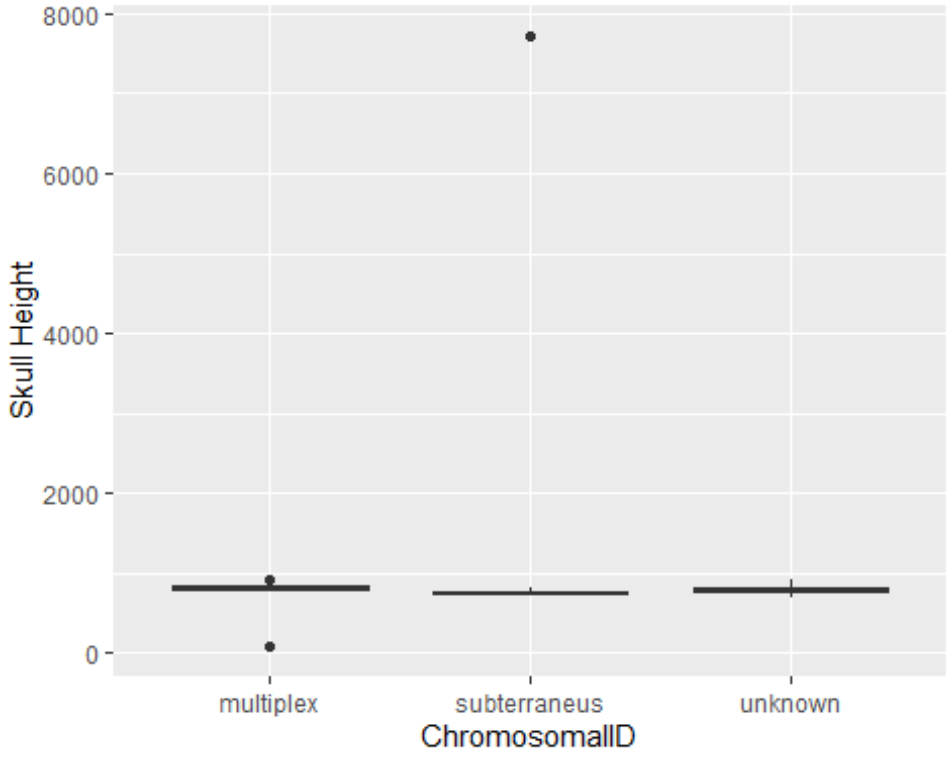


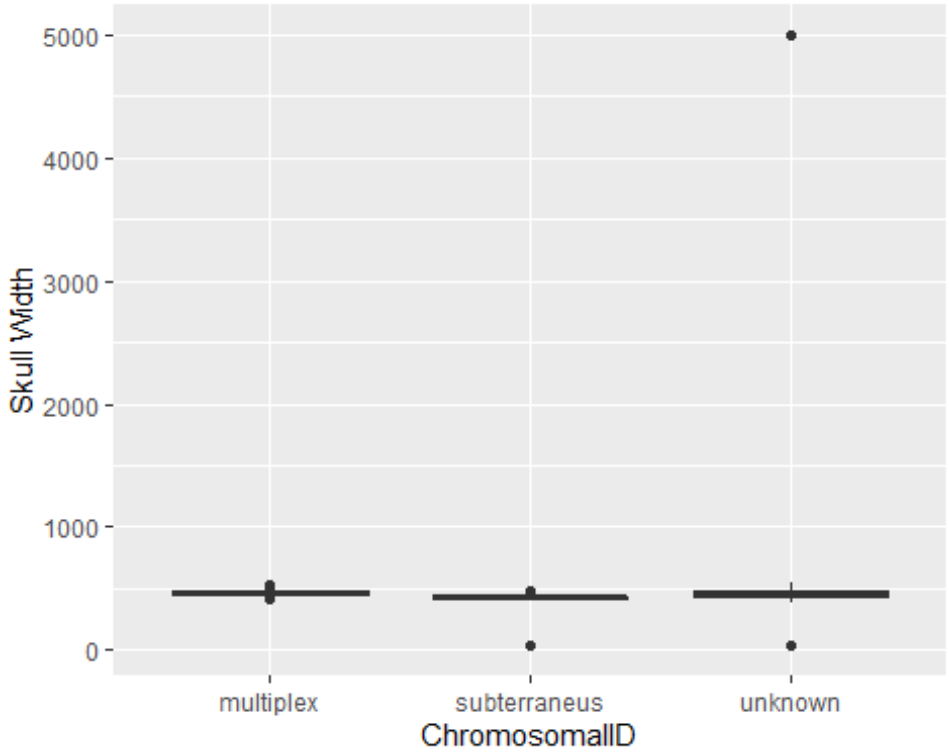

```
## [1] "Multiplex Data Summary"

##  ChromosomalID      IncisiveOrSkullLength  SkullHeight      SkullWidth
##  Length:44          Min.   : 234          Min.   : 84.0   Min.   :416.0
##  Class :character   1st Qu.:2330          1st Qu.:782.5   1st Qu.:455.0
##  Mode  :character   Median :2370          Median :804.5   Median :465.0
##                     Mean   :2337          Mean   :791.4   Mean   :466.1
##                     3rd Qu.:2453          3rd Qu.:828.5   3rd Qu.:475.0
##                     Max.   :2600          Max.   :910.0   Max.   :524.0
```




```
## [1] "Unknown Data Summary"

##  ChromosomalID      IncisiveOrSkullLength  SkullHeight      SkullWidth
##  Length:199         Min.   : 1908         Min.   :700.0   Min.   :  40.0
##  Class :character   1st Qu.: 2222         1st Qu.:760.0   1st Qu.: 428.0
##  Mode  :character   Median : 2320         Median :790.0   Median : 453.0
##                     Mean   : 2417         Mean   :794.5   Mean   : 473.7
##                     3rd Qu.: 2408         3rd Qu.:825.0   3rd Qu.: 475.0
##                     Max.   :23555         Max.   :912.0   Max.   :5000.0
```


Incisive or Skull Length
20000
15000
10000
5000
0
multiplex
subterraneus
unknown
ChromosomalID


Skull Height
8000
6000
4000
2000
0
multiplex
subterraneus
unknown
ChromosomalID



Standard logistical regression formula:

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \ldots \beta_q x_q$$

Replaced with the variables used in this scenario, this would be:

$$logit(ChromosomalID) = \log(ChromosomalID/(1 - ChromosomalID))$$
$$= 0 + \beta\_1\ (IncisiveOrSkullLength) + \beta\_2\ (SkullHeight) + \beta\_3\ (SkullWidth)$$

While testing the independent variables and combinations of the independent variables via the general linear model function in R, I found that none of them had a significant impact on the Chromosomal ID, as they each only had P values greater than .05. Furthermore, through running the general linear model function on the variables, squared versions of the variables, and with a log transformed incisive or skull length (this variable consistently had an extremely low intercept value), I found that leaving the variables untransformed, not squared, and utilizing only addition within the model seemed to give the best fit. The simplest model that included all of the independent variables had the lowest residual deviance, highest null deviance, and lowest AIC, indicating a good fit.

Null Deviance - how well the dependent variable can be predicted by a model via just the intercept term. Higher = Better

Residual Deviance - how well the dependent variable can be predicted by a model with predictor variables. Lower = Better

AIC(Akaike information criterion) - Used to determine the fit of a model. The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables. Lower = Better

source: https://www.statology.org/null-residual-deviance/

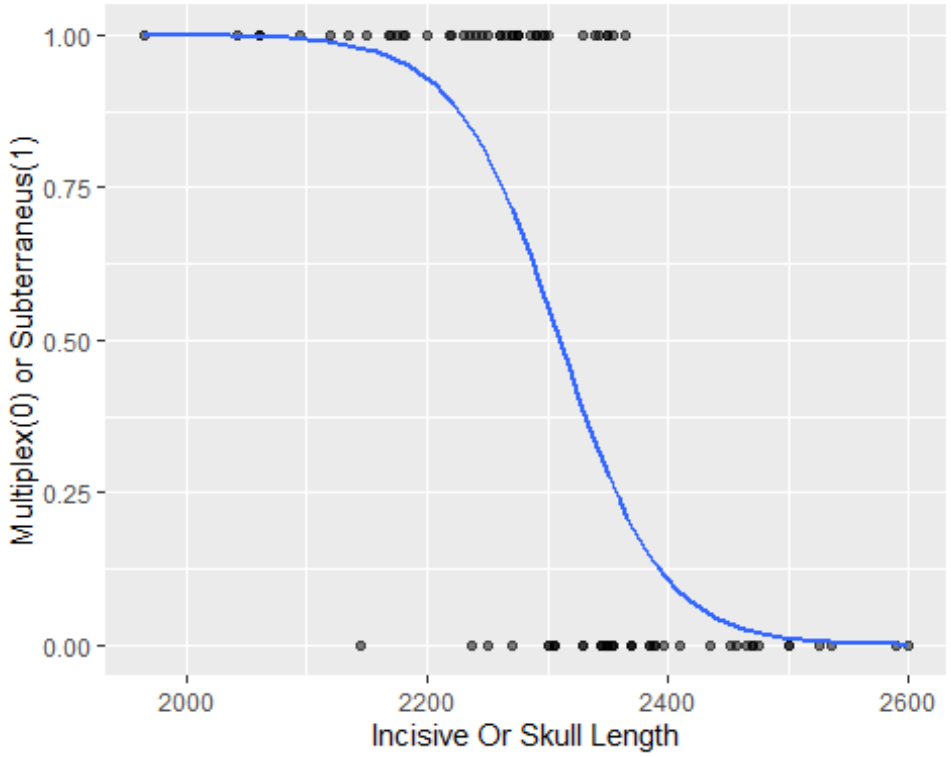
GLM Models and Fit Results

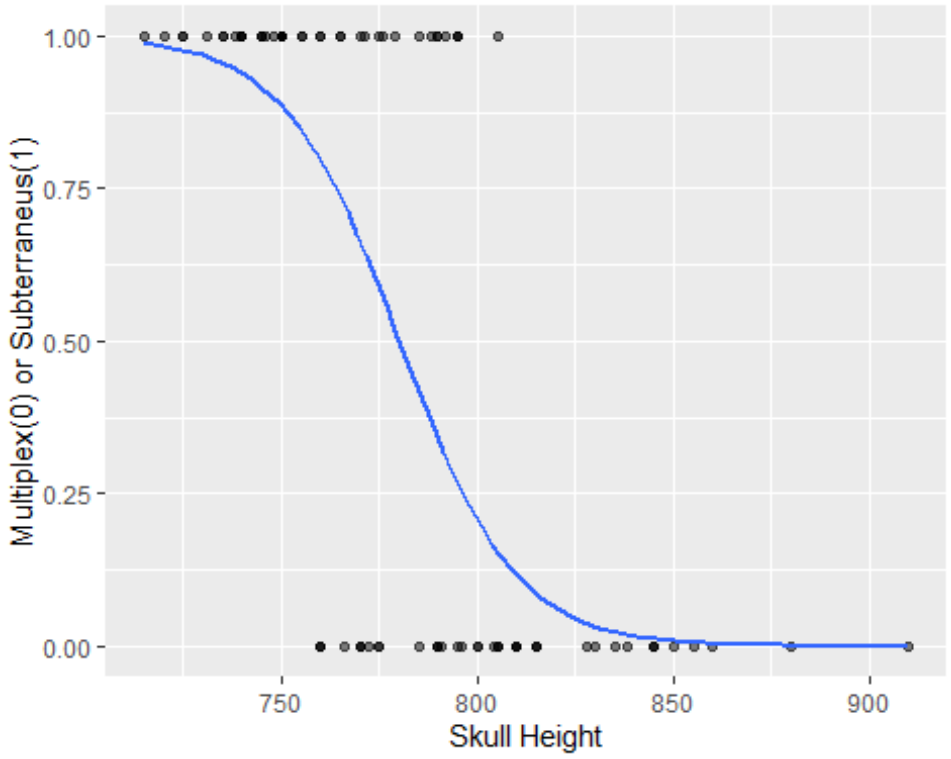
Formula	AIC	Residual_Deviance	Null_Deviance
ChromosomalIDBinary ~ IncisiveOrSkullLength * SkullHeight * SkullWidth	64.966	48.966	117.823
ChromosomalIDBinary ~ I(IncisiveOrSkullLength^2) * I(SkullHeight^2) * I(SkullWidth^2)	65.230	49.230	117.820
ChromosomalIDBinary ~ IncisiveOrSkullLength * SkullHeight * SkullWidth + 1	64.966	48.966	117.823
ChromosomalIDBinary ~ log(IncisiveOrSkullLength) * SkullHeight * SkullWidth	64.975	48.975	117.823
ChromosomalIDBinary ~ IncisiveOrSkullLength + SkullHeight + SkullWidth	57.965	49.965	117.823

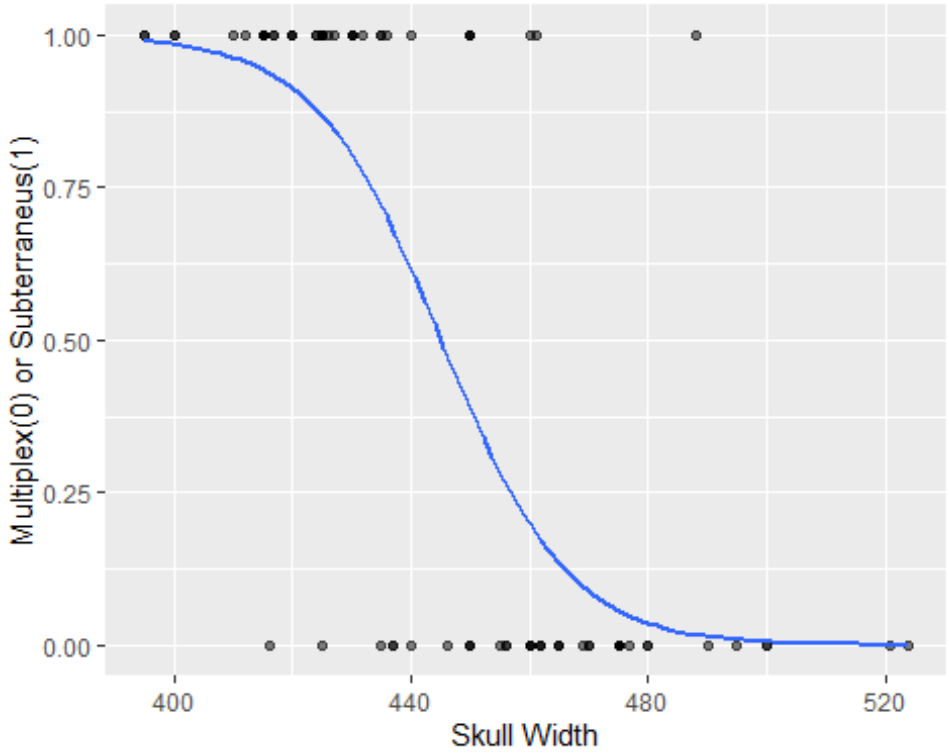
Model Selection and Analysis

```
Here, I ended up testing quite a few minor modifications to my chosen model o
f “ChromosomalIDBinary ~ IncisiveOrSkullLength + SkullHeight + SkullWidth” bu
t ended up sticking with the base form after finding that it was obtaining th
e most consistently high accuracy rate during multiple cross validation tests
. Furthermore, after looking through the advantages and disadvantages of the
different types of cross validation methods for testing a model, I ended up s
ettling on Repeated 10 times – Leave One Out Cross Validation. This method sp
lits the data into 10 evenly sized sections and then repeatedly trains the mo
del on 9 of those sections and tests with 1, changing which sections it train
s and tests with on each of 10 repeats. Its biggest disadvantage compared to
other versions of cross validation is that it’s very hardware intensive, but
I chose it since the dataset is small and thus it wouldn’t matter if the meth
od was hardware intensive.

## [1] "model 1.5 accuracy:"

## [1] 0.8901852
```

Logistic regression curve plot of the chosen model, showing the relationship between the independent variables and the chance that a vole skull is Multiplex or Subterraneus.:


1.00
0.75
0.50
0.25
0.00
Multiplex(0) or Subterraneus(1)
2000
2200
2400
2600
Incisive Or Skull Length


1.00
0.75
0.50
0.25
0.00
Multiplex(0) or Subterraneus(1)
750
800
850
900
Skull Height



Model Testing Results

After testing the beforementioned model with Repeated 10 times 10 fold – Leave One Out Cross Validation, I got an 88.18% accuracy rate. Also, with the model being simple, only utilizing addition of the independent variables, there is a low risk of overfitting. Furthermore, I tested the model on the unknown classification ID dataset and got results that seemed completely plausible. This R notebook outputs the results as a file called: "Vole Skulls Unknown Classified.csv".

Recommendations on Usefulness and Conclusion

I believe that with the high accuracy rate and the simplicity of the logistic regression model that I created for this project, it is fit for use with classifying vole skulls by species.

6. As a secondary component provide annotated code that replicates your analysis.

Source

Airoldi, J.-P., B. Flury, M. Salvioni (1996) “Discrimination between two species of Microtus using both classified and unclassified observations” Journal of Theoretical Biology 177:247-262

Vole Skull Excell Spread Sheet

Description

Microtus multiplex and M. subterraneus are difficult to distinguish morphologically. Here we have 43 known multiplex voles, 46 known subterraneus voles and a further 199 unidentified voles.

Data

288 Skulls from 2 species of voles 3 measurements on the vole skulls.

Group

a factor with levels -multiplex -subterraneus -unknown

Length

Condylo incisive length or skull length (0.01mm)

Height

Skull height above bullae (0.01mm)

Rostrum

Skull width across rostrum (0.01mm)

Details

89 species have been identified by means of chromosomal analysis.