Question 1

Amin Baabol

11/21/2020

Introduction

"Subterraneus" and "Multiplex" are by enlarge considered two distinct species by biologists. However, it has not been easy to distinguish between the two species. Microtus, also known as Voles, are small roden-like burrow animals that are geographically spread out across western Asian, Europe and North America. Our interest lies in establishing a "best-fit" statistical and machine learning model that will help biologists' identity or distinguish between two Microtus species. The data consist of eight morphometric variables collected from fossilized bird pellets, using Nikon-scope with an accuracy of 1/1000 mm and a dial caliper with an accuracy of 1/100 mm. There are 299 specimens of which only 89 specimens whose specie has been identified. Furthermore, research indicates that there are no reliable criteria based on cranial morphology that can distinguish the two species. Our analysis process will use the 89 identified samples to construct a model that will classify the rest of the unclassified 199 specimens. It is important to note that while we strive to refine our analysis the small sample size provided can potentially reduced the likelihood of detecting a statistically significant result.

Methodology

will be imported from the *Flurry* library as *microtus*. As stated in the introduction, only 89 samples have their specie classified as either "multiplex" or "subterranean" and the remaining 199 samples are unknown. It is apparent that this is a classification problem with a binary outcome. There are various machine learning models we can employ for the given classification problem; however, the problem statement specifically requires us to develop a classification model using a generalized-linear model family. Hence, we are going to develop a logistic regression model using the *glm* function in base R. Also, it is a good practice to employ various, competing models to refining our model and variables selection process. Any figures referred to will be attached an an appendix at the end of report.

An important concept to keep in regarding the coefficient estimates of logistic regression is the concept of log of "odds ratio".Unlike regular linear regression models,logistic regression model predicts the probability of observing specie "multiplex",conveniently coded as 1 and the probability of observing specie "subterraneous" coded as 0.The probability of observing 1 over 0 also known as odds ratio,is calculated as $P/1-P$.The logit link function will then take the logarithm of the odds ratio and will increase with unrestricted range as the P increases from 0 to 1.

Assumptions

The following assumptions are made in the process of a building binomial logistic regression model: 1. Predicted outcome is binary or discrete 2. The continuous explanatory variables follow normal Gaussian distribution 3. A linear relationship exists between the independent explanatory variables and the logit output 4. No outliers that exert undue influence on the model 5. No troublesome multicollinearity.

Data Visualization

The aim of the data visualization is to assess the univariate frequency distribution of the variables to ensure the normal distribution assumption isn't violated. By observing the box-plots

of the all the variables separated by specie interesting differences in distributions emerge. First,the histograms plots shown in Figure 1a-h indicate relatively normal distribution of the variables for each specie with varying degrees of skewness.In particular,Figure 1a and Figure 1b indicate that there are very little overlap of the distributions of upper left molar 1 width (M1Left) and uper left molar 2 (M2Left) of the two species. M1Left of subterraneus is narrower distribution centered around 1700mm, while Multiplex's M1Left is more flatten with a mean of 2054mm. Figure 1c-1f shows the two species fairly overlap in the distributions of M3Left, Foramen,Pbone,and Length.Lastly, Figures 1g-1h show a distinction in Height and Rostrum distributions of the two species.By observing the box-plots of the all the variables separated by specie interesting differences and outliers emerged. Moreover, the density plots show that the probability of observing subterraneus left upper molar1 is significantly higher at around 1600mm-1800mm,whereas left upper molar1 for multiplex is highest around 1900mm-2200mm.The other variables show slight differences but nothing as different as "M1Left". This means we expect to see significant differences in M1Left mean for the two species.

Model Selection

The initial model fitted is a logistic regression model to be used as a reference to compare to the subsequent fitted models. This baseline model contained only the constant intercept. A total of seven models were fitted with increasing parameters. Subsequently, a log-likelihood ratio test was performed to compare model fitness. This test calculates the probability of observing parameters that optimize the coefficient estimates of the two compared models. In other words, it analyzes the log-likelihood of the two models compared and see if their difference is statistically significant. If the difference is indeed significant, the more complex model is chosen. On the other hand, if the p-value is not significant at the 0.05 level then the simpler model is selected. Hence, Model0 which only contained the intercept was compared with Model1 which has one parameter. The resulting p-value is 2e-16, so Model1 was selected. Next, Model1 was compared to Model2 which has two parameters. The resulting p-value is 0.01098, which led us to select Model2 over Model1.Next, Model2 against Model3 and the p-value is 0.5769. The subsequent comparisons tests failed to reject the null hypothesis that the difference between the log-likelihood of the compared models is not significant. Therefore, Model 2 was selected to move forward. Also, we could have compared the means of the square residuals of the models and picked one with the lowest MSE.

Model 2

Model2 which is fitted with only two predictors (M1Left, Foremen) shows in the model summary that the intercept, M1Left and Foremen are all highly significant at the 0.05 level. The Null deviance is 123.279 with 88 degrees of freedom, while the residual deviance is 22.049 with 86 degrees of freedom. The residual deviance indicates how well Model2 predicts with the included parameters. The AIC which penalizes for having more variables is very low at 29. 738.The mean square of the model residuals is 3.679744, which very low. Finding the MSE of the model uses the

$$MSE = (1/n) * \sum_{i=1}^{n} (Observered - Predicted)^2$$

Additionally, Cook's distance was computed to detect the presence of any highly influential outliers. The threshold or the cutoff line for cook's distance is 0.02 The observations [21,], [24,] were detected as outliers with mild influence having only passed the conservative threshold of 0.02. Observation [3,] is seen as an outlier with extreme influence having passed the both 0.02 and 1 thresholds. However, it is not omitted from the training dataset due to the small sample size we have.

```
## [1] 0.021085
```

Conclusion

We began our analysis by assessing the descriptive statistics of the microtus data. The distribution of the boxplots of the predictor variables suggested that several of the eight explanatory variables contain outliers and the correlation plot also indicated many of these variables are highly correlated. We fitted seven models and performed log-likelihood ratio tests using anova. This method was intended to isolate the important explanatory variables and reduce the model complexity without compromising its performance. The second model with only two predictors was deemed "best" because we failed to reject the null hypothesis that the less complex model is a better fit than the more complex model. This model reduced the residual deviance from 28.5 to 22. The mean square error for the selected model was 3.68%.To verify that no outliers were exerting undue influence on the model's performance Cook's distance,however, no observation was ultimately removed.Lastly,a 10 folk cross-validation was used to ensure the model wasn't simply too overly-optimistic.The cross validation mean square error was slightly higher at 4.71%, it was none the less, a within 5% error margin.While every effort was made to ensure the quality of this analysis,however, I recommend collecting more samples because the coefficient estimates are small. This means the log-odds of correctly classifying the default specie (1) is lower.

Appendix

Figure 1a: Width of upper left molar 1

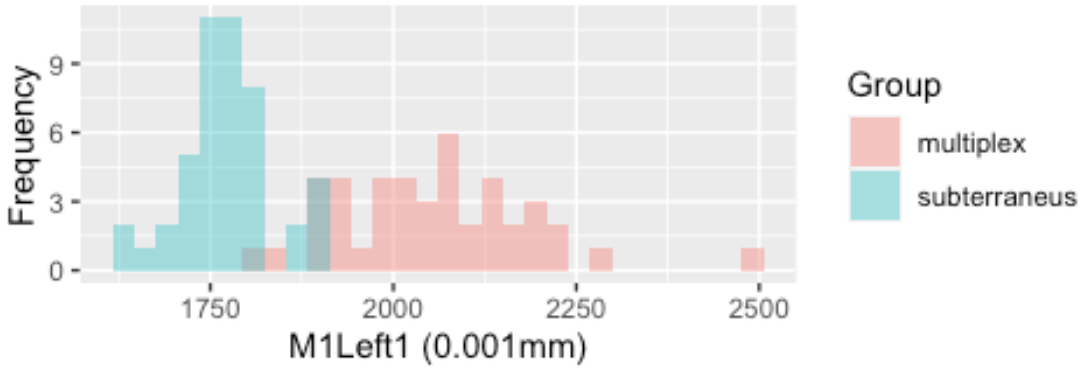


Figure 1b: Width of upper left molar 2

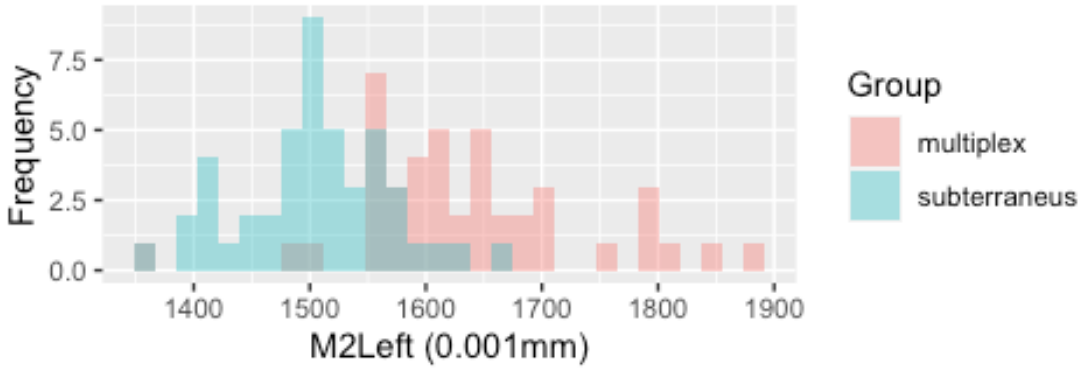


Figure 1c: Width of upper left molar 3

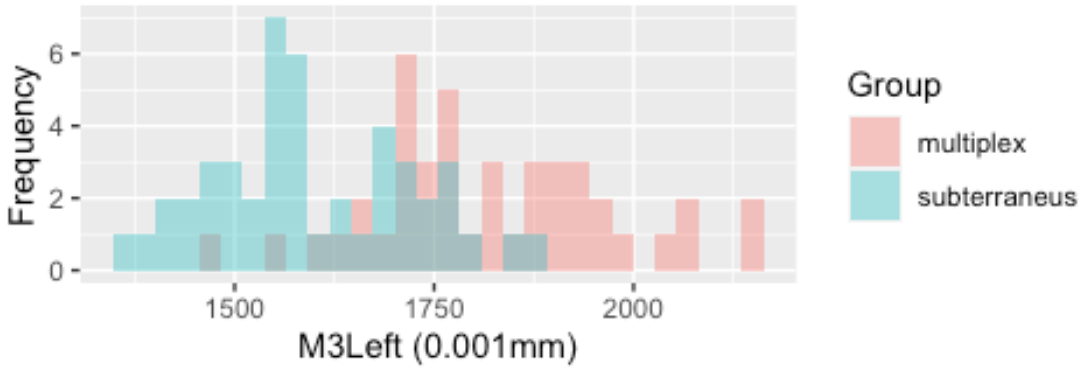


Figure 1d: Length of incisive foramen

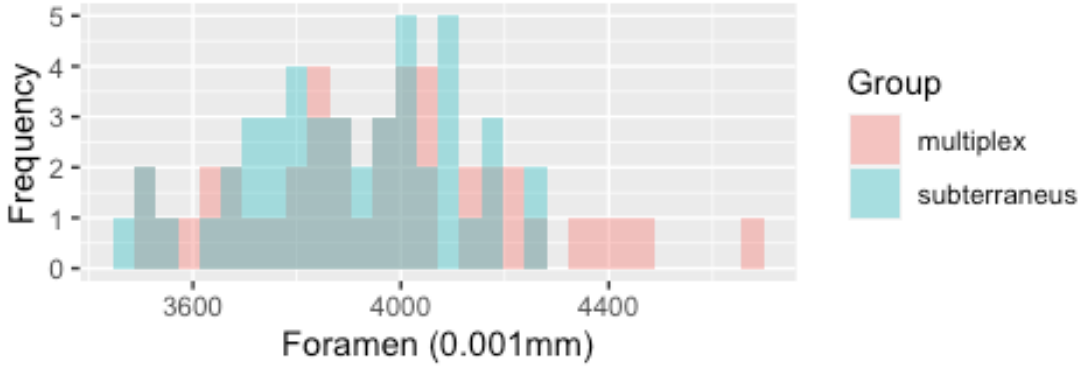


Figure 1e: Length of palatal bone

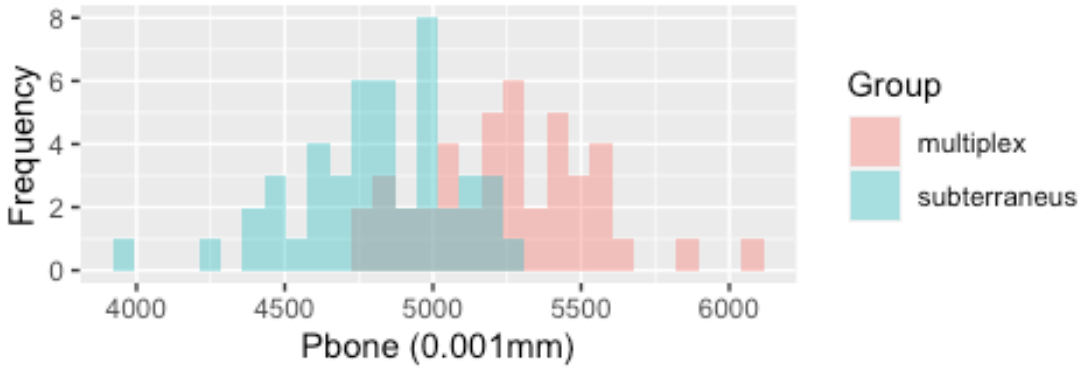


Figure 1f: Condylo incisive length or skull length

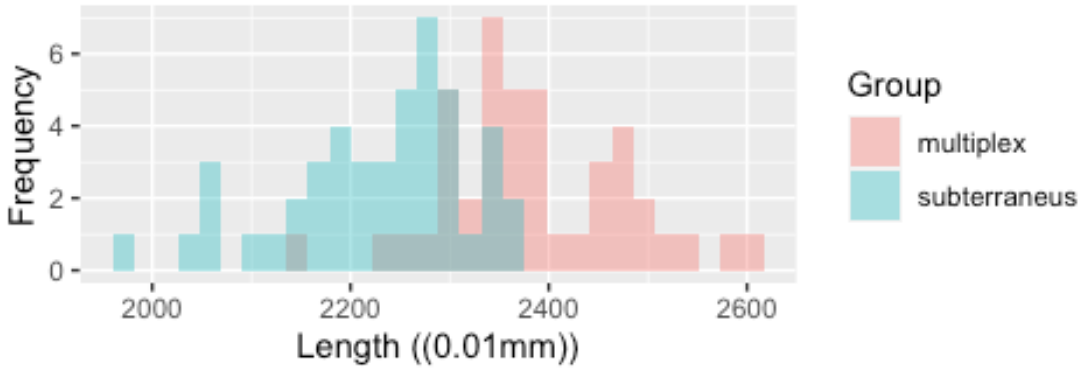


Figure 1g: Skull height above bullae

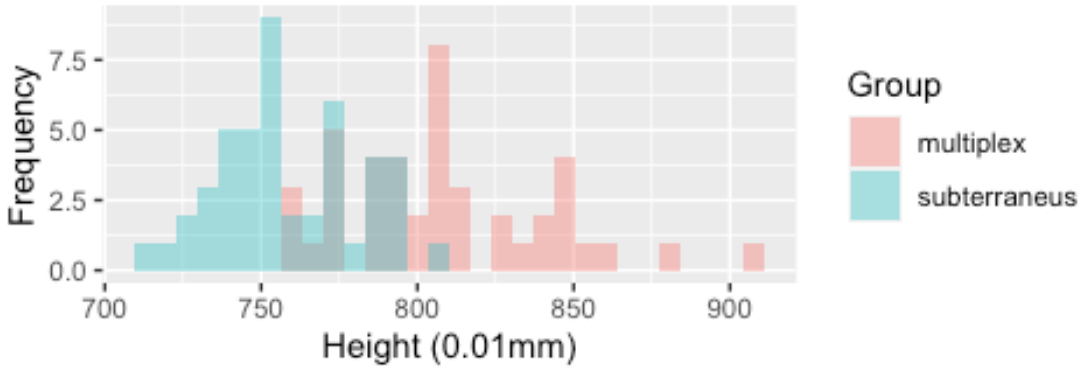


Figure 1h: Skull width across rostrum

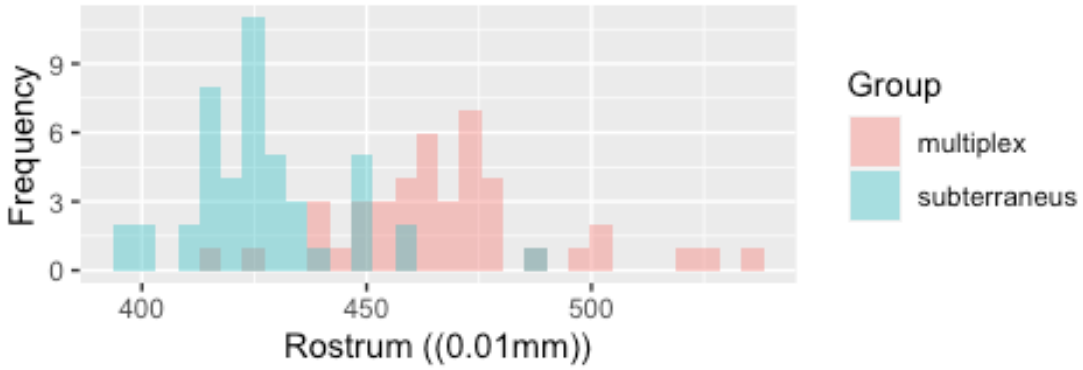


Figure 2a: Width of upper left molar 1

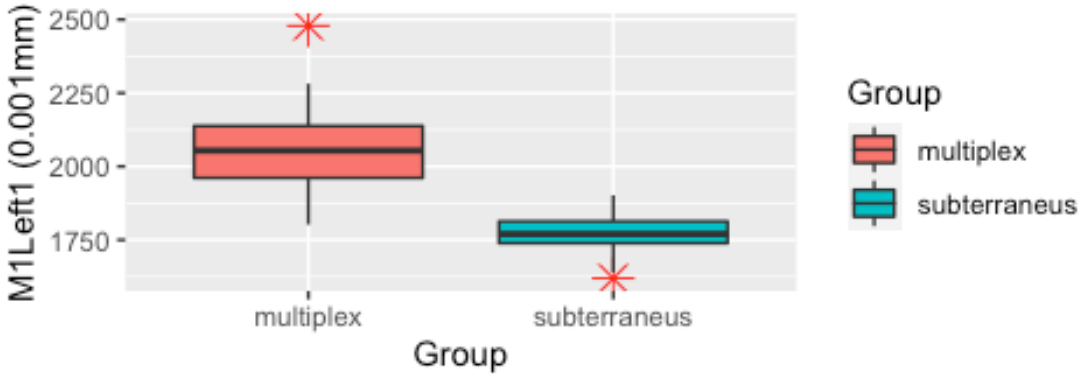


Figure 2b: Width of upper left molar 2

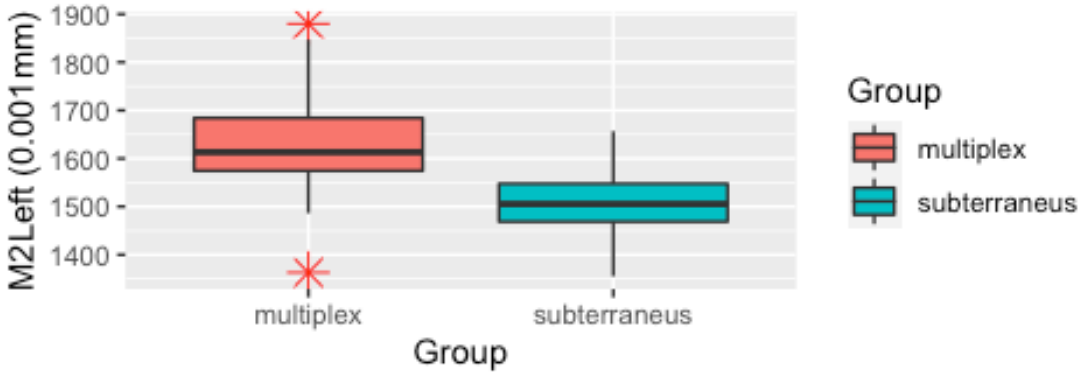


Figure 2c: Width of upper left molar 3

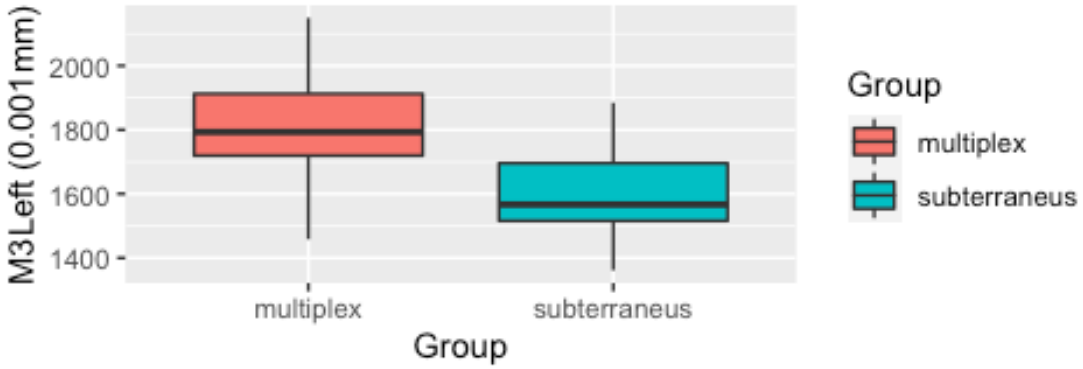


Figure 2d: Length of incisive foramen

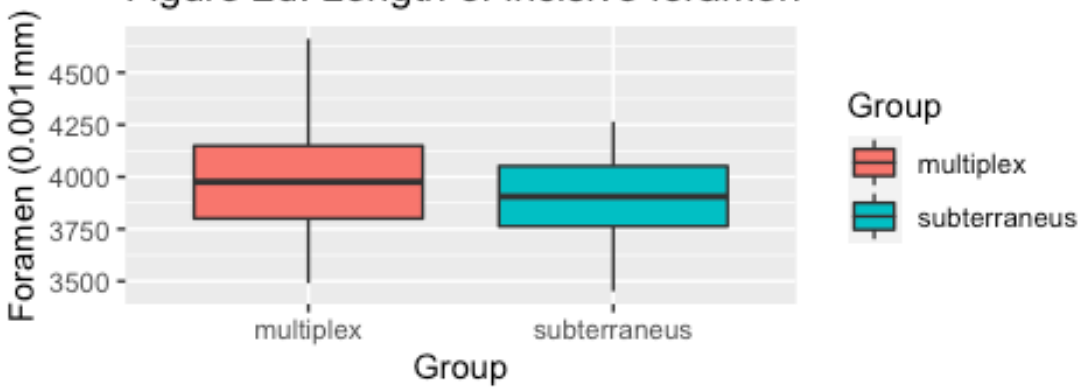


Figure 2e: Length of palatal bone

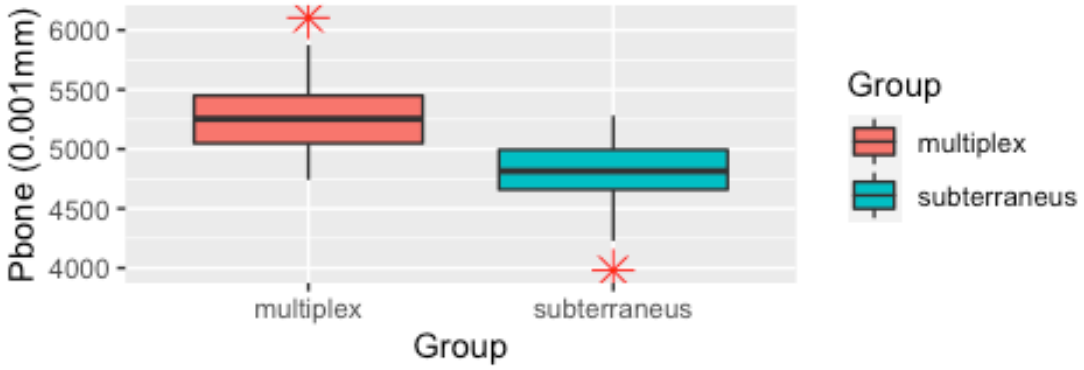


Figure 2f: Condylo incisive length or skull length

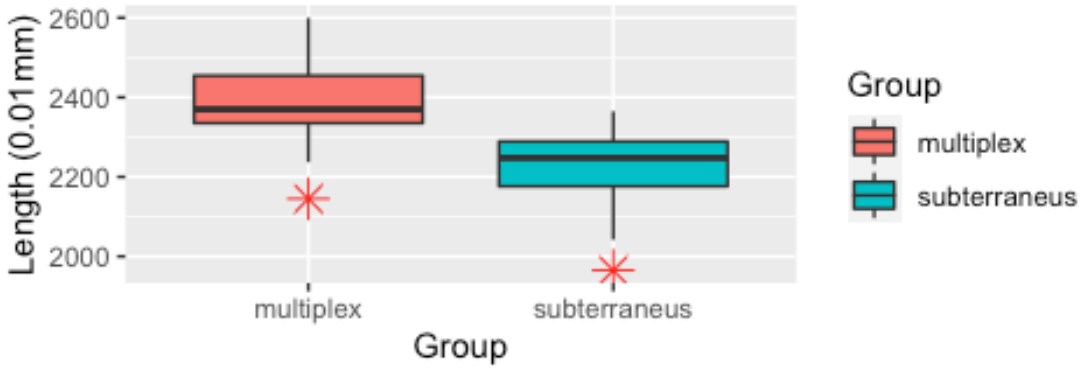


Figure 2g: Skull height above bullae

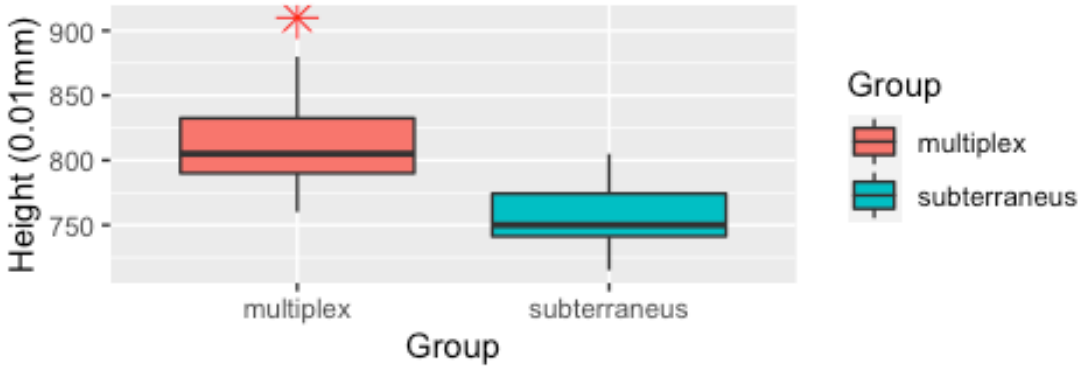
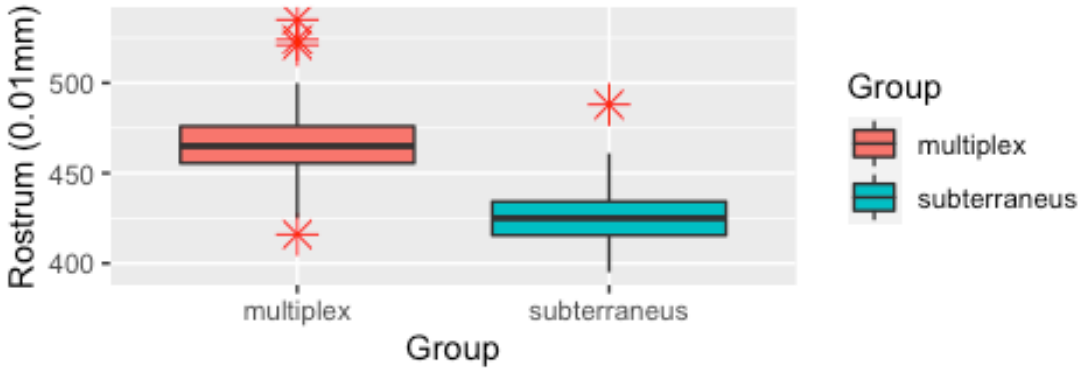
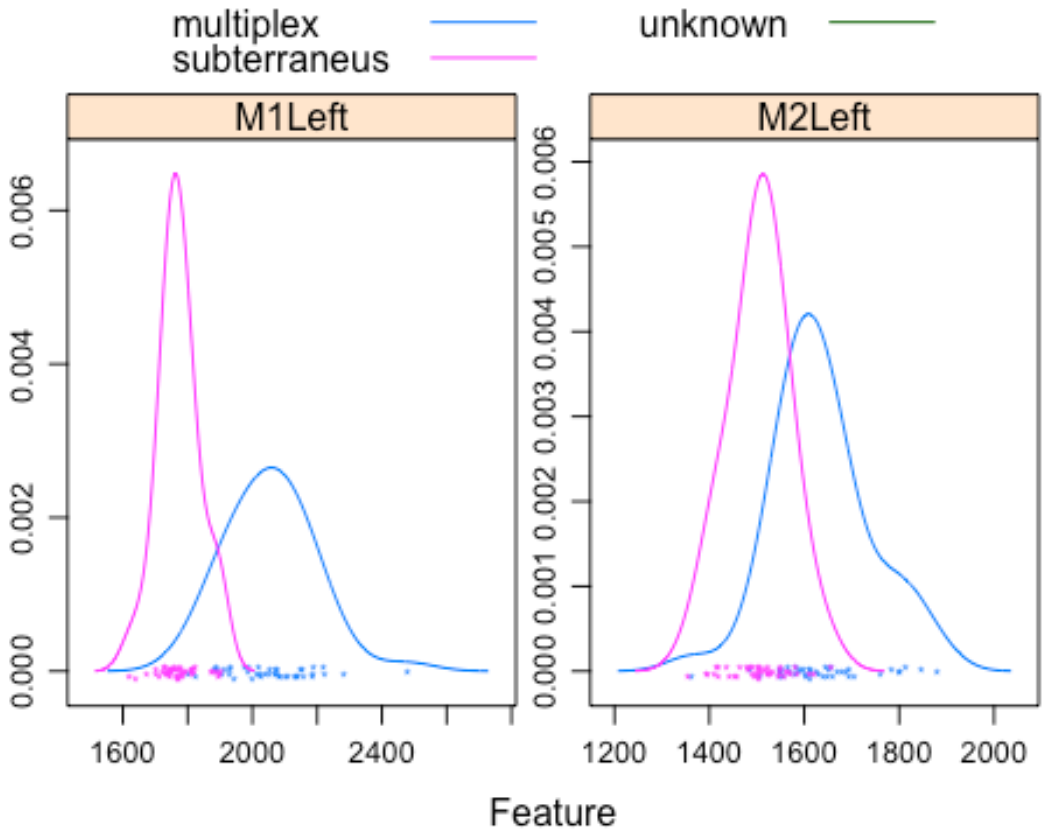
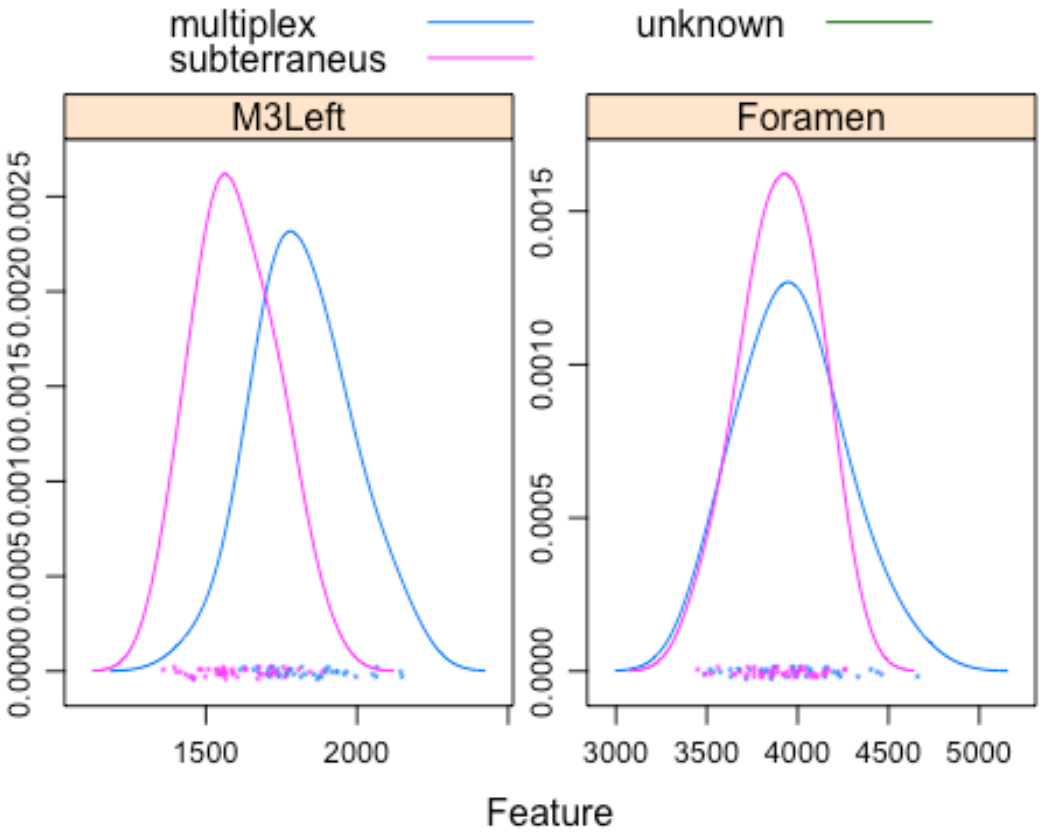
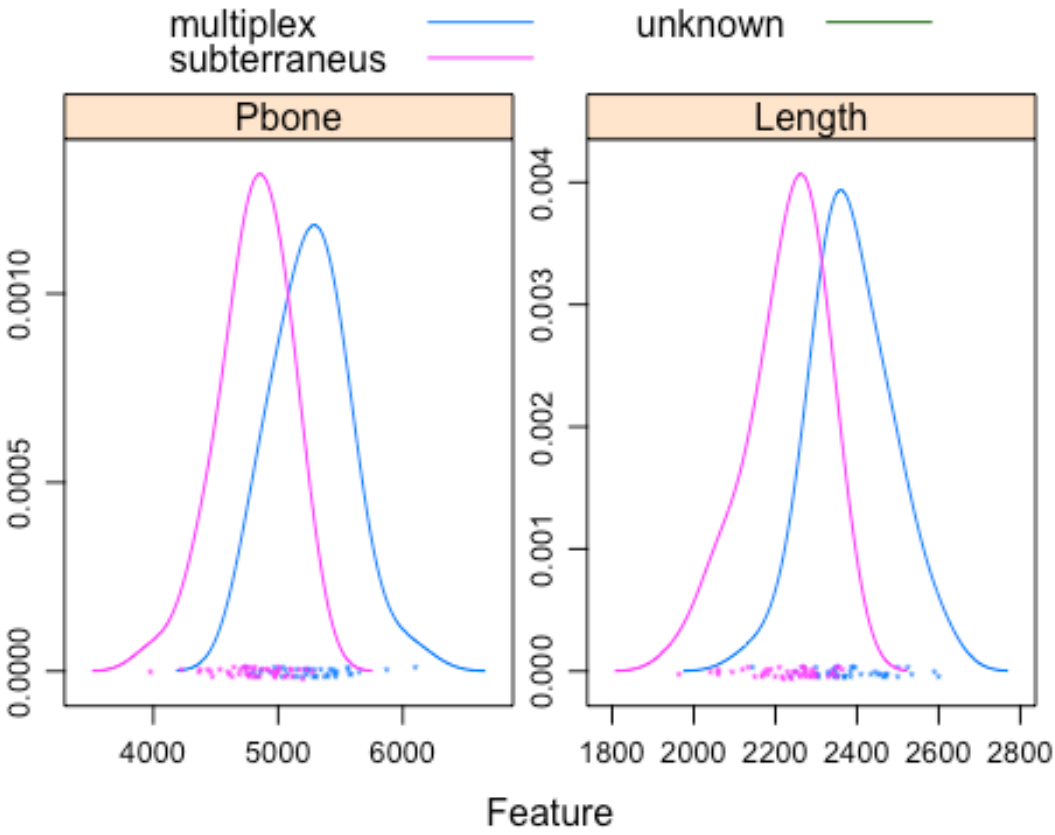


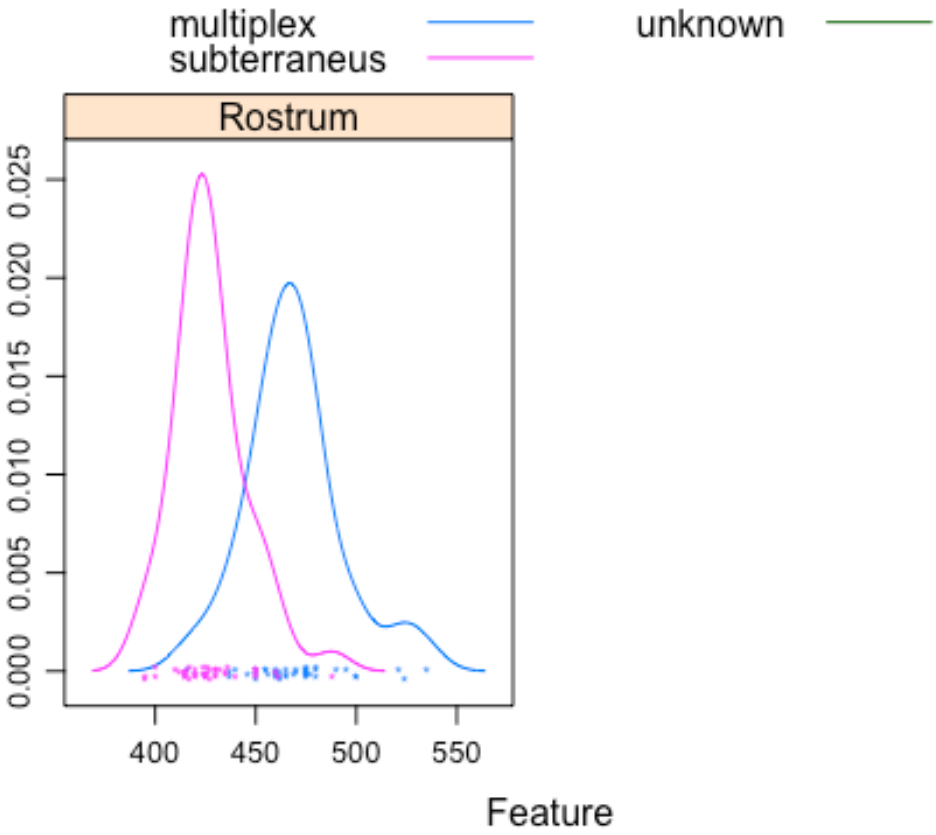
Figure 2h: Skull width across rostrum




multiplex
subterraneus
unknown
M1Left
M2Left
0.000
0.002
0.004
0.006
0.000
0.001
0.002
0.003
0.004
0.005
0.006
1600
2000
2400
1200
1400
1600
1800
2000
Feature


multiplex
subterraneus
unknown
M3Left
Foramen
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
1500
2000
3000
3500
4000
4500
5000
Feature


multiplex
subterraneus
unknown
Pbone
Length
0.0000
0.0005
0.0010
0.000
0.001
0.002
0.003
0.004
4000
5000
6000
1800
2000
2200
2400
2600
2800
Feature


multiplex
subterraneus
unknown
Rostrum
0.000
0.005
0.010
0.015
0.020
0.025
400
450
500
550
Feature

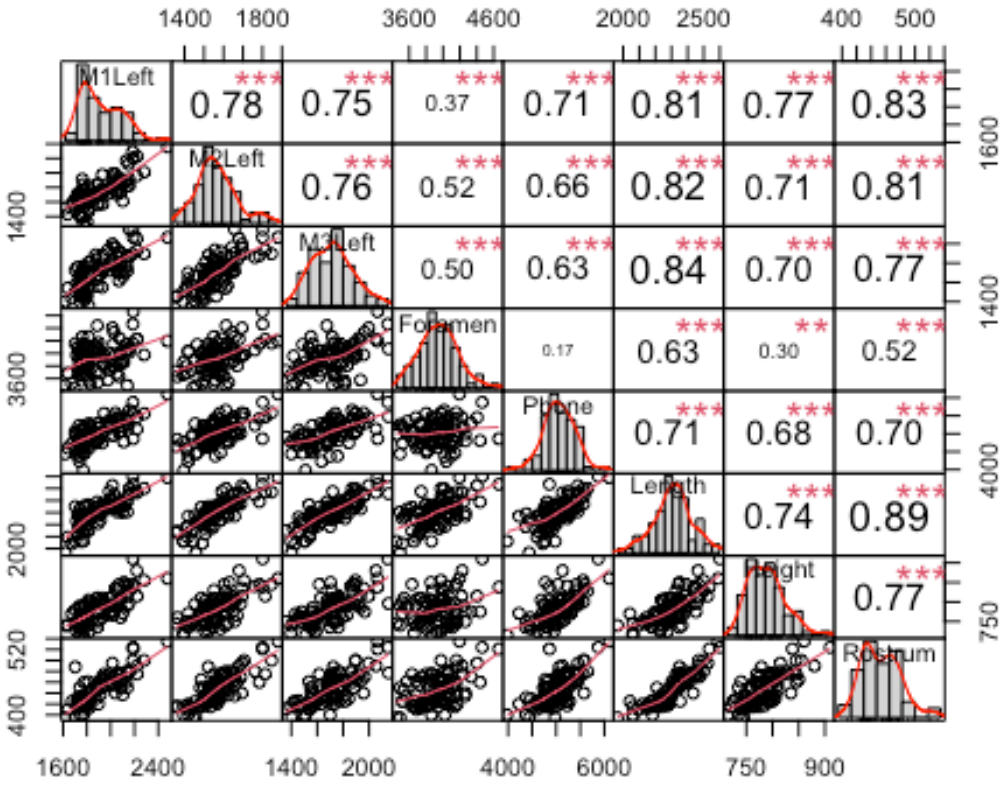

M1Left
0.78
0.75
0.37
0.71
0.81
0.77
0.83
0.76
0.52
0.66
0.82
0.71
0.81
0.50
0.63
0.84
0.70
0.77
0.17
0.63
0.30
0.52
0.71
0.68
0.70
0.74
0.89
0.77
1400 1800
3600 4600
2000 2500
400 500
1600 2400
1400 2000
4000 6000
750 900
400 520
2000
3600
1400
1600
1400
4000
750

Figure 5: Influencial Outliers