STAT 601: Final project

Investigating the use of Microtus subterraneus and M. multiplex which are considered to be two distinct species

Kenneth Annan

12/04/2020

Introduction

This study seeks to investigate the use of Microtus *subterraneus* and M. multiplex which are 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. No reliable criteria based on cranial morphology have been found to distinguish the two species. That is, the objective is to analyze this effect using 1) fitting a *glm* model using the 89 specimens; 2) to predict the group membership of the remaining 199 specimen's using the best-fitted model, and 3) to explain the analysis of the dataset and draw recommendations on the usefulness of the predictions.

Data and methodology

The study employs the *Microtus* data located in the *Flury* library in R.The original data set contains 9 variables and 288 observations. However, the present study was initiated by a data collection consisting of eight morphometric variables (M1Left, M2Left, M3Left, Foramen, Pbone, Length, Height, Rostrum) measured by one of the authors (Salvioni) using a Nikon measure-scope (accuracy 1/1000 mm) and dial calipers (accuracy 1/100 mm). The dataset has the chromosomes of 89 specimens analyzed to identify the species. The remaining 199 of the specimens are unknown but their characteristics are available.

In order to develop a model from the 89 specimens that I can use to predict the group membership of the remaining 199 specimens, I subset the *Microtus* dataset for the 89 specimens(known data, and name it *data1*). The *data1* dataset is made up of 43 known Microtus multiplex and 46 known Microtus subterraneus. I then subset the data for the remaining 199 specimens (unknown data, and name it *data2*). The goal for this step is to develop a model that suits the known dataset and use it to predict the unknown chromosomes of the 199 specimens.

Also, I explored the data by finding the correlation of the known dataset(*data1*) by using *cor* and *ggpairs* command in R.I also used the *aggregate function* in R to find the mean of the known dataset(the multiplex and the subterraneus) and the unknown dataset.

In other to fit a glm model, I created a binary response variable *Group* that takes a value of 1 and 0 for M.multiplex and M.subterraneus, respectively. The intuition is that for any glm model(logistic regression model), the dependent variable should be binary.

Following *HSAUR3: A Handbook of Statistical Analyses Using R*, I fit two glm models. I used the glm command in R to fit the models. I used the regsubsets command in R that I learned in my time series class to check on the best variables to include in the model. The summary of the regsubsets command shows that the variables that are highly needed in the model are M1Left, Foramen, and Rostrum. I also made a boxplot of the three variables (M1Left, Foramen, and Rostrum)that were selected for using the regsubsets command.

I first fit a glm_model_1(model 1) for the predictors(M1Left, Foramen, and Rostrum), a second model named glm_model_2(model 2) for the predictors(M1Left, Foramen) using the results from the regsubsets. I also used analysis of variance (ANOVA) to check on my results. To make a recommendation on the best model, I compared the 10 fold cross-validation with a seed set at 100, Error rates, and the accuracy rate from both models. With the fitted models, I predicted the known chromosomes to compare to what was initially collected by Salvioni. I used the best model (model2) to predict the rest of the remaining 199 species. Again, I used the model best-fitted model to predict the chromosomes of the rest of the 199 specimens. I used the head command in R to show the first 8 rows of the predicted chromosomes.

Discussion of results

Table 1 shows the results of the correlation coefficients of the known dataset. Table 1 also shows the characteristics of the specimens. It can be seen that some of the variables are highly correlated with a correlation coefficient greater than 0.8. Also, Figure 1 shows the visual display of the results in Table 1. The graph looks at the correlation of the variables in the Microtus multiplex and the the Microtus subterraneus. There is a high correlation for Condylo incisive length or skull length and Skull width across rostrum, Width of upper left molar and Skull width across rostrum, and Width of upper left molar 3 and Condylo incisive length of skull length. The highest correlation is 0.88, I see it to be not reasonable to include both skull length and Rostrum in the same model since their correlation coefficient is approximately 0.90. Table 2 shows the means of the measurement of their cranial morphology. The results show that the means of the Microtus subterraneus are smaller in all measurements than the Microtus multiplex. Also, the boxplot of the the three variables shows that multiplex tends to have greater values for morphometric measurements than subterraneus for the variables *M1Left and Rostrum*. However, for the variable *Foramen*, both the morphometric measurements (multiplex) and subterraneus species show similar values.

Table 3 shows the variable selection method for using the regsubsets command in R.Table 3 serves as the benchmark for using models 1 and 2. The summary for models 1 and 2 are shown in Tables 4a and 4b.

Table 5 shows the model selection method using the K fold cross-validation at the seed of 100, accuracy rate, AIC, and the error rate. All the four measures of the model selection methods show that model 2 is the best model-see Table 5. Thus, using the Cross-Validation to predict the

accuracy of your model, model 2 has approximately 95.5% accuracy while model 1 has approximately 94.4% accuracy as shown in Table 5. Again, Table 5 shows that model 2 is the best model because model 2 has the lowest error rate. This shows that model 2 is preferable in predicting the unknown chromosomes of the dataset. Following the ongoing discussion, I estimated the best model as

$$\widehat{Group} = -62.8044523 + 0.0472459M1Left - 0.0066369Foramen \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The summary of the best model (model 2) shows that all the variables are statistically significant at the 5% predefined alpha level. It can also be seen that the M1Left can positively tell the difference in the two species while the Foramen negatively tell the difference of the group-see Table 4b and equation (1). Also, the ANOVA table also shows that M1left and Foramen are statistically significant at the 5% predefined alpha level.

Conclusion

The best-fitted model that is model 2 indicates that M1Left(Foramen) positively (negatively) affects the difference in the two species significantly. The prediction I made is significant enough to be used for future study. Thus, the accuracy rate of the best model fitted is high enough to measure this impact. Also, finding the M1Left and the foramen is enough to predict the chromosome of the species. The only concern about the study is that the sample size of the known dataset(*data1*) is too small for predicting the unknown dataset(*data2*). I predicted 121 M. multiples and 78 Microtus subterraneus as shown in Table 7.

Appendix

Table 1:correlation of the known dataset

	M1Left	M2Left	M3Left	Foramen	Pbone	Length	Height	Rostrum
M1Left	1.0000	0.7832	0.7537	0.3657	0.7113	0.8062	0.7691	0.8331
M2Left	0.7832	1.0000	0.7583	0.5249	0.6560	0.8198	0.7098	0.8117
M3Left	0.7537	0.7583	1.0000	0.5005	0.6261	0.8414	0.7025	0.7694
Foramen	0.3657	0.5249	0.5005	1.0000	0.1666	0.6263	0.2966	0.5175
Pbone	0.7113	0.6560	0.6261	0.1666	1.0000	0.7095	0.6794	0.7049
Length	0.8062	0.8198	0.8414	0.6263	0.7095	1.0000	0.7383	0.8882
Height	0.7691	0.7098	0.7025	0.2966	0.6794	0.7383	1.0000	0.7700

Rostrum	0.8331	0.8117	0.7694	0.5175	0.7049	0.8882	0.7700	1.0000

Table 2:Means of the groups

Group.1	M1Left	M2Left	M3Left	Foramen	Pbone	Length	Height	Rostrum
multiplex	2054.5	1636.5	1819.9	3966.5	5260.2	2386.0	809.4	468.7
subterraneus	1773.3	1504.6	1597.8	3899.0	4805.2	2226.7	758.1	427.2
unknown	1947.3	1597.8	1736.6	3904.5	5108.2	2311.4	794.4	452.9

Figure 1: correlation of the known dataset

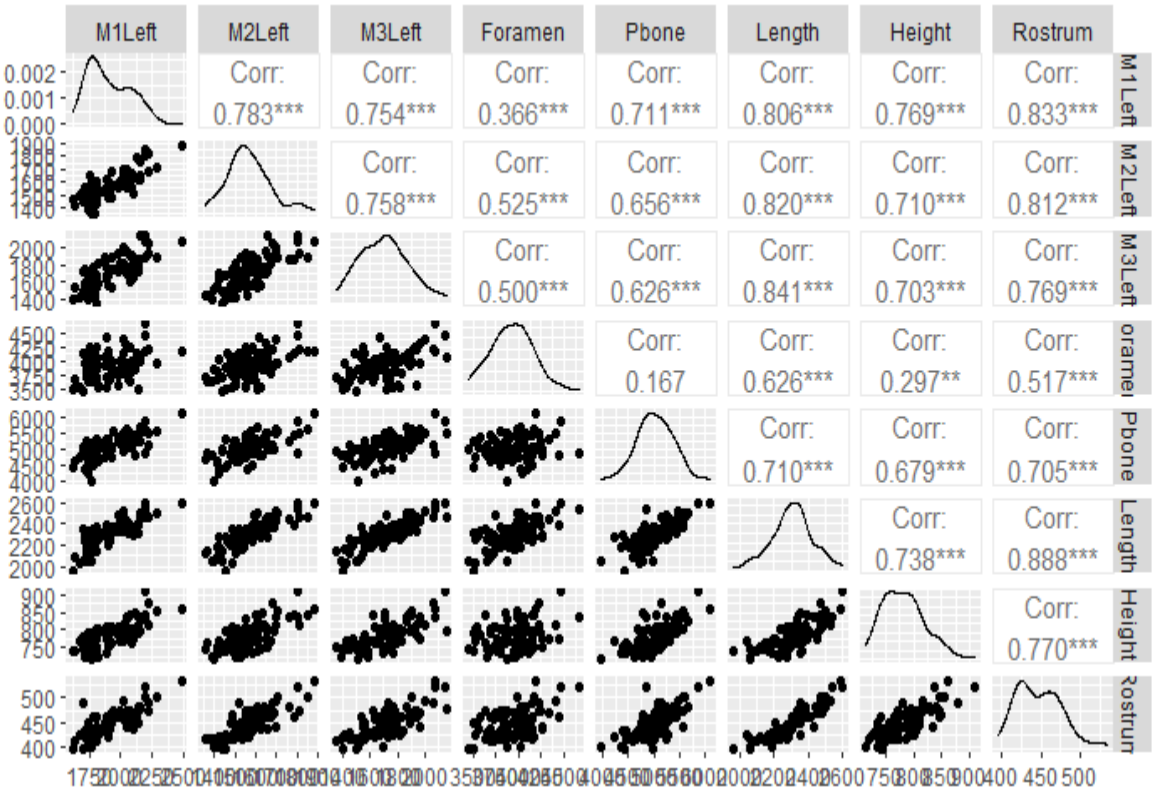

Table 3:Variable selection method

	M1Left	M2Left	M3Left	Foramen	Pbone	Length	Height	Rostrum
1 (1)	*							
2 (1)	*			*				
3 (1)	*			*				*
4 (1)	*			*		*		*
5 (1)	*		*	*		*		*
6 (1)	*		*	*	*	*		*
7 (1)	*		*	*	*	*	*	*
8 (1)	*	*	*	*	*	*	*	*

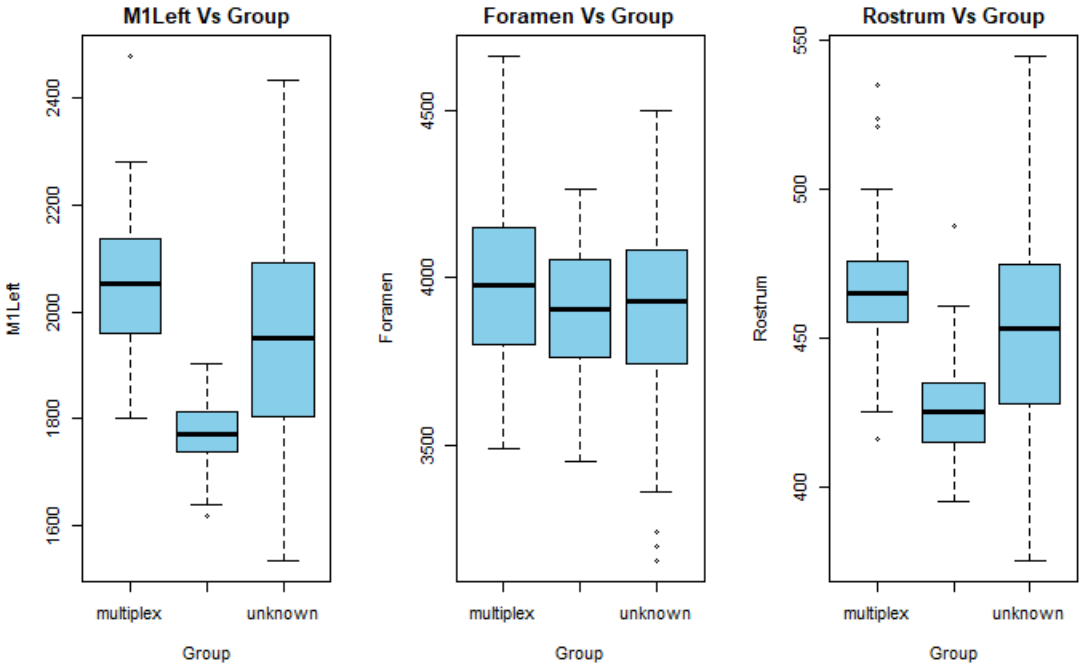

Table 4a:summary of model 1 (not the best model)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-62.4871316	20.0703580	-3.1134039	0.0018494
M1Left	0.0387737	0.0164437	2.3579754	0.0183749
Foramen	-0.0067068	0.0030955	-2.1666386	0.0302624
Rostrum	0.0357798	0.0518020	0.6907032	0.4897521

Table 4b: summary of the best model

	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 5:Model selection measures

model	AIC	CV	error.rate	accuracy
model1	29.55319	0.0530836	0.0561798	0.9438202
model2	28.04904	0.0449362	0.0449438	0.9550562

Table 6:anava table for the best model

term	df	Deviance	Resid..Df	Resid..Dev	p.value
NULL	NA	NA	88	123.27906	NA
M1Left	1	94.76201	87	28.51704	0.0000000
Foramen	1	6.46800	86	22.04904	0.0109834

Table 7: Number of obervations of the predicted chromosomes

	multiplex	subterraneus	unknown
multiplex	41	2	121
subterraneus	2	44	78

References

1.Niethammer, L. (1982). The Follower Factory: Denazification using Bavaria as an example . Dietz.

2.Krapp-Schickel, G. (1982). Family Amphilochidae. The Amphipoda of the Mediterranean, Part, 1, 70-83.

3.Ellerman, J. R., & Morrison-Scott, T. C. S. (1951). Checklist of Palaearctic and Indian mammals, 1758-1946 (Vol. 3). order of the Trustees of the British Museum.

4. Hothorn, T., Everitt, B. S., Data II, C. A. L., Scaling, C. M., & Partitioning, C. R. (2017). HSAUR3: A Handbook of Statistical Analyses Using R.